MANY: MERGE ANYTHING FOR MULTIMODAL CONTINUAL INSTRUCTION TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Continual Instruction Tuning (MCIT) is essential for sequential task adaptation of Multimodal Large Language Models (MLLMs) but is severely restricted by catastrophic forgetting. While existing literature focuses on the reasoning language backbone, in this work, we expose a critical yet neglected dual-forgetting phenomenon across both perception drift in Cross-modal Projection Space and reasoning collapse in Low-rank Parameter Space. To resolve this, we present **MAny** (**M**erge **Any**thing), a framework that merges task-specific knowledge through **C**ross-modal **P**rojection **M**erging (**CPM**) and **L**ow-rank **P**arameter **M**erging (**LPM**). Specifically, CPM recovers perceptual alignment by adaptively merging cross-modal visual representations via visual-prototype guidance, ensuring accurate feature recovery during inference. Simultaneously, LPM eliminates mutual interference among task-specific low-rank modules by recursively merging low-rank weight matrices. By leveraging recursive least squares, LPM provides a closed-form solution that mathematically guarantees an optimal fusion trajectory for reasoning stability. Notably, MAny operates as a training-free paradigm that achieves knowledge merging via efficient CPU-based algebraic operations, eliminating additional gradient-based optimization beyond initial tuning. Our extensive evaluations confirm the superior performance and robustness of MAny across multiple MLLMs and benchmarks. Specifically, on the UCIT benchmark, MAny achieves significant leads of up to 8.57% and 2.85% in final average accuracy over state-of-the-art methods across two different MLLMs, respectively. [1]

1 INTRODUCTION

Recent advancements in multimodal large language models (MLLMs) Yin et al. (2024); Bai et al. (2023); Zhu et al. (2023) have significantly extended traditional LLM Touvron et al. (2023) capabilities through sophisticated cross-modal alignment Liu et al. (2023) and autoregressive generation. To bridge the gap between general-purpose pretraining and domain-specific applications, multimodal instruction tuning Zhang et al. (2026); Tong et al. (2025) aligns model behavior with user intent. However, real-world scenarios typically require models to continuously adapt to diverse, sequentially arriving datasets to meet evolving demands. In Multimodal Continual Instruction Tuning (MCIT) scenario Chen et al. (2024a) , MLLMs must efficiently absorb new knowledge while preserving previously acquired skills. Unfortunately, severe catastrophic forgetting Liu et al. (2025); Gao et al. (2025a;b;c; 2024a;b) hinders this process, remaining a critical and unresolved challenge in the field.

To combat forgetting, current methods face difficult trade-offs: Mixture of Experts (MoE) based architectures Guo et al. (2025a); Yu et al. (2025) require heavy computation, regularization Chen et al. (2025); Zeng et al. (2025) limits model flexibility, and replay approaches Li et al. (2025); Lee et al. (2025) raise privacy and storage concerns. Crucially, these methods mainly share a fundamental oversight: they focus almost entirely on the anti-forgetting ability of the language backbone, assuming the multimodal projector is immune to forgetting.

To pinpoint the root causes of forgetting, we conduct a decoupling analysis on the UCIT Guo et al. (2025a) benchmark, as visualized in Figure 1. As shown, the naive fine-tuning baseline (green line) exhibits a severe performance collapse compared to the oracle (black dashed line). Crucially, even

[1] Our code and data are included in the supplementary material for examination.

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.

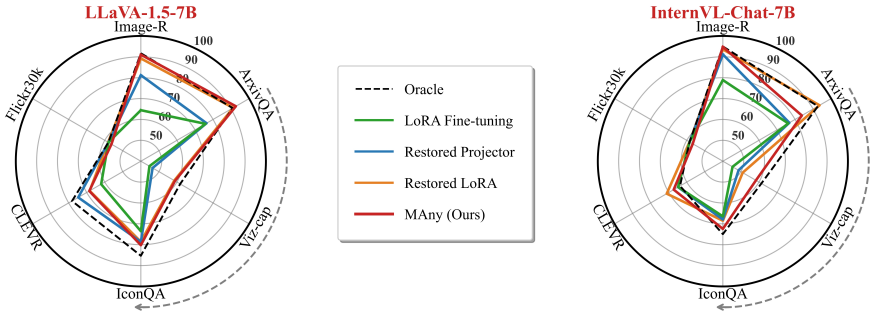


Figure 1: **Performance diagnosis on UCIT benchmark for LLaVA-1.5-7B and InternVL-Chat-7B using sequential LoRA fine-tuning. Tasks follow a clockwise sequence from Image-R.**

when replacing the drifted final-task LoRA with task-specific parameters (orange line), a substantial gap persists if the drifted final-task projector is maintained. For instance, on the Viz-cap Gurari et al. (2018) and IconQA Lu et al. (2021) datasets using the LLaVA Liu et al. (2024a) and InternVL Chen et al. (2024b) models, the performance fails to reach the oracle level despite having lossless reasoning weights. This gap proves that the projector itself is also a primary source of forgetting. Surprisingly, on the CLEVR Lindström & Abraham (2022) dataset using the LLaVA model, restoring only the projector (blue line) yields higher accuracy than restoring the LoRA (orange line). This confirms that for specific visual-reasoning tasks, perceptual misalignment can be more damaging than reasoning collapse. Finally, our proposed MAny (red lines) successfully resolves both failure modes. These findings confirm a dual-forgetting problem in MLLMs: perception drift in the cross-modal projection space and reasoning collapse in the low-rank parameter space.

Driven by these insights, we introduce **MAny** (**M**erge **Any**thing), a simple yet effective framework designed to fix both the perception and reasoning spaces. To reverse the perception drift seen in our experiments, we propose Cross-modal Projection Merging (CPM). Instead of using a single projector, CPM maintains lightweight, task-specific projection layers. These layers are highly parameter-efficient and ensure fast inference by adaptively merging visual features. Guided by visual prototypes, CPM can automatically recognize which task an image belongs to and recover the most accurate visual-language alignment. Simultaneously, to stop the reasoning collapse within the LLM, we develop Low-rank Parameter Merging (LPM). Unlike simple weight averaging that often causes severe parameter interference, LPM uses the proposed Recursive Least Squares algorithm to find a conflict-minimizing way to merge task-specific low-rank modules. This provides a closed-form solution that mathematically ensures the new model remembers all tasks while minimizing interference within the low-rank parameter space. Crucially, by performing dual-merging via fast CPU-based algebraic operations, MAny eliminates the need for additional GPU training or historical data, offering a lightweight solution for real-world multimodal learning.

In summary, our key contributions can be summarized as follows: (1) We show a neglected dual-forgetting phenomenon, proving that projector drift is a major cause of failure alongside LLM forgetting. (2) We introduce CPM, a plug-and-play module that adaptively merges task-specific visual features to restore perceptual alignment. (3) We develop LPM, which utilizes a recursive solution based on Recursive Least Squares to achieve conflict-minimizing parameter merging. (4) The proposed MAny framework achieves state-of-the-art performance on the UCIT and MLLM-DCL benchmarks, demonstrating strong knowledge retention without any additional GPU-based training.

2 Related Work

Multimodal Large Language Models. Driven by the success of LLMs Touvron et al. (2023); Zhang et al. (2023) in natural language processing Nirenburg (1993); Min et al. (2023), MLLMs Bai et al. (2023); Zhu et al. (2023); Dai et al. (2023) have emerged to bridge the gap between vision and language. Architecturally, MLLMs rely on a frozen vision encoder and a cross-modal alignment module (such as a projector Liu et al. (2023) or a cross-attention mechanism Dai et al. (2023)) to align visual features with the language space. While achieving impressive performance Liu et al. (2024b); Zhang et al. (2024); Zhao et al. (2025b), most existing MLLMs are trained in static multi-task settings. They struggle to adapt to continually arriving data streams Shi et al. (2021), where learning new knowledge often leads to catastrophic forgetting of previously acquired capabilities.

Continual Instruction Tuning for MLLMs. Existing solutions for overcoming forgetting in MCIT Liu et al. (2023); Longpre et al. (2023) generally fall into three categories: replay-based, regularization-based, and architecture-based methods. Replay-based methods Li et al. (2025); Lee et al. (2025) rehearse historical or synthetic data, introducing severe storage/computation overheads and privacy risks. Regularization-based approaches Zeng et al. (2025); Chen et al. (2025) penalize drastic parameter or cross-modal shifts to bypass memory burdens Wang et al. (2023a). Architecture-based strategies Guo et al. (2025a); Wang et al. (2023a) isolate new knowledge via task-specific modules or dynamic routing. While inherently preserving prior parameter spaces, they face critical challenges in precise task-agnostic routing. Crucially, these paradigms focus almost exclusively on the language backbone, assuming the multimodal projector is immune to forgetting. Our work challenges this assumption by revealing and addressing the drift in the projection space.

Model Merging. Model merging consolidates knowledge from multiple task-specific models into a unified architecture without costly retraining. Early works relied on weight averaging Matena & Raffel (2022) or feature distance minimization Jin et al. (2023), while recent methods Yadav et al. (2023); Du et al. (2024) utilize task vectors Ilharco et al. (2022) to integrate modules like LoRA Hu et al. (2022). To mitigate negative transfer, advanced training-free paradigms Sun et al. (2025c) have begun to explicitly minimize layer-wise feature drift. However, these methods are predominantly tailored for static, one-shot merging within the NLP domain, lacking a mechanism to handle continual learning scenarios. In contrast, our work is the first to bridge model merging with the MCIT paradigm. Moving beyond one-time integration, we develop a recursive merging paradigm that allows MLLMs to incrementally absorb new tasks while mathematically ensuring an optimal fusion of task-specific parameters across the entire task sequence.

3 PRELIMINARY

3.1 MULTIMODAL CONTINUAL INSTRUCTION TUNING

We consider an MLLM Liu et al. (2023) composed of three primary components: a vision encoder f_v parameterized by Φ_v, a cross-modal projector $\mathcal{P}$ parameterized by $\Phi_{\mathcal{P}}$, and a large language model (LLM). In the context of MCIT, the model is required to learn a sequence of n tasks, denoted as $\mathcal{T} = \{1, 2, \ldots, n\}$, which arrive sequentially. For each task $t \in \mathcal{T}$, the corresponding dataset is defined as $\mathcal{D}_t = \{(x_v^{t,j}, x_{ins}^{t,j}, x_{ans}^{t,j})\}_{j=1}^{N_t}$, where $x_v^{t,j}$, $x_{ins}^{t,j}$, and $x_{ans}^{t,j}$ represent the input raw image, instruction tokens, and answer tokens, respectively, and N_t is the total number of samples in $\mathcal{D}_t$.

For a given sample $(x_v, x_{ins}, x_{ans}) \in \mathcal{D}_t$, the vision encoder first extracts visual features from the raw image. Specifically, we follow the prevailing paradigm Liu et al. (2023) to extract dense spatial feature $\mathbf{Z}_v^{\mathrm{s}}$ from the penultimate layer of f_v, which are then mapped into the LLM's embedding space via the projector, yielding projected visual tokens $\mathbf{P}_v = \mathcal{P}(\mathbf{Z}_v^{\mathrm{s}}; \Phi_{\mathcal{P}})$. The LLM subsequently processes the concatenated sequence of visual tokens $\mathbf{P}_v$ and the embedded instruction tokens x_{ins} to generate the response. Following standard paradigms, we keep the vision encoder parameters Φ_v frozen to preserve foundational visual representations. The trainable parameters $\Theta = \{\Phi_{\mathcal{P}}, \Delta\Phi_l\}$ typically include the projector weights $\Phi_{\mathcal{P}}$ and a set of Parameter-Efficient Fine-Tuning (PEFT) weights $\Delta\Phi_l$ (e.g., LoRA Hu et al. (2022)) added to the LLM. The model is optimized via an autoregressive next-token prediction loss:

$$\mathcal{L}_{MLLM} = -\sum_{k=1}^{N_{ans}} \log p(x_{ans,k} \mid \mathbf{P}_v, x_{ins}, x_{ans,<k}; \Theta), \tag{1}$$

where N_{ans} is the sequence length of the answer. Upon encountering task t, the objective is to minimize the empirical risk over $\mathcal{D}_t$ while ensuring that Θ retains the knowledge acquired from preceding tasks $\{1, \ldots, t-1\}$.

3.2 PERCEPTION DRIFT AND REASONING COLLAPSE

To empirically motivate our approach, we investigate the internal dynamics of forgetting in MLLMs. As visualized in Figure 2, our analysis reveals a distinct dual-forgetting pattern that provides a mechanistic explanation for the performance drops diagnosed in Figure 1. Specifically, we examine the representation drift in both the projector and the LLM layers to identify the root causes of failure

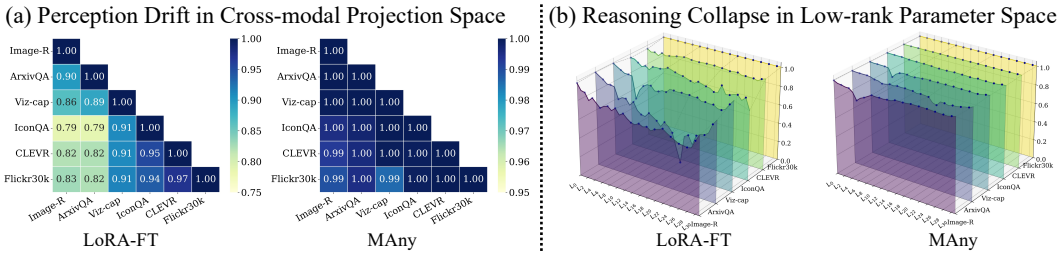


Figure 2: **Visualizing Perception Drift and Reasoning Collapse.** (a) Cosine similarity of projector features across tasks. (b) Layer-wise feature similarity between oracle states and the final model. In both aspects, MAny effectively mitigates internal representation drift.

in sequential instruction tuning. Figure 2(a) evaluates the stability of the cross-modal projection space. It measures the cosine similarity between the projector's output features in their oracle state (immediately after training on task i) and the features extracted after learning subsequent tasks $j > i$, using the exact samples from task i. In parallel, Figure 2(b) focuses on the internal layers of the LLM hierarchy, visualizing the similarity between the features of task i at its oracle state and those extracted by the final model (after n tasks). This analysis adds a depth dimension to observe how reasoning structures evolve across different model layers.

Our observations confirm two distinct failure modes that hinder successful continual learning. As shown in Figure 2(a), naive LoRA fine-tuning (LoRA-FT) suffers from a noticeable perception drift, where the similarity for initial tasks steadily drops to approximately 0.83. This drift disrupts the vision-language alignment, causing the projector to misinterpret visual inputs from earlier learning stages. In stark contrast, Figure 2(b) reveals a severe reasoning collapse within the LLM layers. For LoRA-FT, we observe a consistent and significant decline in feature similarity across all model layers. This degradation is particularly pronounced for the earliest trained tasks, reflecting the cumulative interference that compromises the model's reasoning stability over the task sequence. Conversely, our proposed MAny demonstrates superior stability across both the cross-modal projection space and the hierarchical LLM layers. By maintaining feature consistency, our framework effectively overcomes both perception drift and reasoning collapse throughout the entire task sequence, ensuring the model's internal representations remain robust.

4 Method

4.1 Decoupling Perception from Reasoning

To resolve the structural conflicts and divergent representation dynamics identified in Sec. 3.2, we propose a dual-track architecture that explicitly decouples perception from reasoning. Our framework independently stabilizes these functional spaces through the Perceptual Track (Cross-modal Projection Merging, detailed in Sec. 4.2) and the Reasoning Track (Low-rank Parameter Merging, detailed in Sec. 4.3). This component-aware design ensures that both perceptual alignment and reasoning stability are preserved throughout the task sequence, alleviating the dual-forgetting inherent in standard MLLM continual learning paradigms. Figure 3 shows the overall design of our MAny.

4.2 Cross-modal Projection Merging

To alleviate the perception drift, we introduce Cross-modal Projection Merging (CPM) and move away from the shared-projector bottleneck and instead treat the projection space as a collection of task-specific alignment layers. For each task $i \in \{1, \ldots, t\}$, we maintain a task-specific lightweight projector $\mathcal{P}_i : \mathbb{R}^{d_v} \to \mathbb{R}^{d_l}$. Since these projectors are typically implemented as Multi-Layer Perceptrons (MLPs), the additional parameter overhead remains minimal compared to the LLM backbone, ensuring memory efficiency throughout the task sequence. To enable task-agnostic inference, CPM employs a dynamic merging strategy that adaptively activates these projectors based on the visual context. During the task t, we compute a task-representative prototype μ_t by aggregating visual features through the frozen vision encoder. Specifically, we utilize the 0-th index feature (i.e., the [CLS] token) from the last layer of f_v as the global representation $\mathbf{Z}_v^{\mathrm{g}}$ and compute prototype

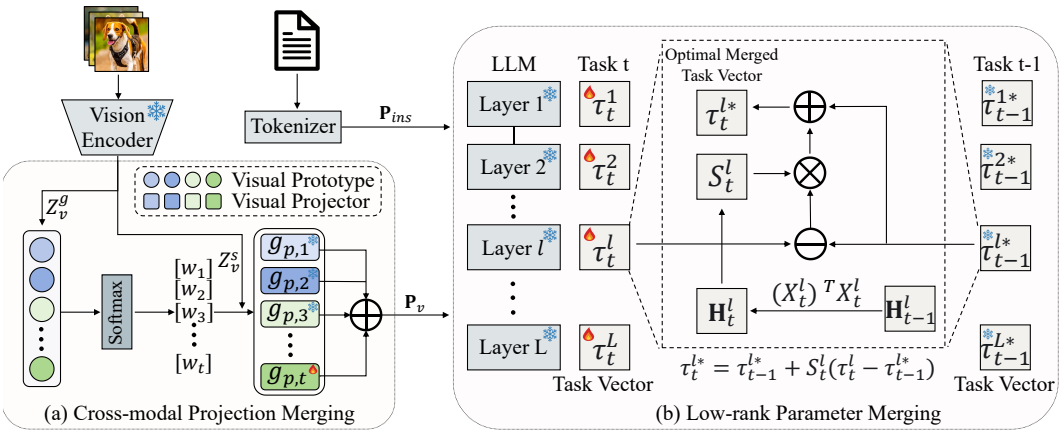


Figure 3: Overview of MAny. The dual-track design explicitly decouples perception from reasoning.

$\mu_t = \frac{1}{N_t}\sum_{j=1}^{N_t} \mathbf{Z}_v^{g,j}$. This prototype μ_t serves as a stable anchor in the visual domain, capturing the specific distribution of task t without requiring any textual metadata.

During inference, the model encounters a query image x_v with an unknown task identity. We first extract its global semantic feature $\mathbf{Z}_v^g$ from the last layer and dense spatial feature $\mathbf{Z}_v^s$ from the penultimate layer of the vision encoder. We then measure the cosine similarity s_i between the query's global feature and each stored prototype μ_i for $i \in \{1, \ldots, t\}$:

$$s_i = \frac{(\mathbf{Z}_v^g)^\top \mu_i}{\|\mathbf{Z}_v^g\| \|\mu_i\|}. \tag{2}$$

These similarities are then normalized into merging weights w_i using a temperature-scaled softmax operation. The final projected visual tokens $\mathbf{P}_v$ are computed as an adaptive merging of task-specific projectors acting on the spatial feature:

$$w_i = \frac{\exp(s_i/\eta)}{\sum_{j=1}^{t} \exp(s_j/\eta)}, \quad \mathbf{P}_v = \sum_{i=1}^{t} w_i \mathcal{P}_i(\mathbf{Z}_v^s), \tag{3}$$

where η is a temperature hyperparameter, which is empirically set to 0.1. This soft merging mechanism ensures robustness: for samples with ambiguous boundaries, the model can adaptively blend alignment knowledge from multiple relevant projectors. By isolating task-specific mappings and merging them only at test time, CPM effectively prevents the overwriting of prior alignments, ensuring the stability of the vision-language interface throughout the task sequence.

4.3 Low-rank Parameter Merging

While CPM preserves perceptual alignments, the internal reasoning layers of the LLM still suffer from severe reasoning collapse. Here, we propose Low-Rank Parameter Merging (LPM), which recursively consolidates task-specific LoRA weights into a consolidated global task vector.

4.3.1 Parameter-Efficient Task Vectors

To continually incorporate the new LoRA weights in each layer of the LLM, we formulate our solution through the lens of model merging Sun et al. (2025c); Ilharco et al. (2022). Consider the L-layer pre-trained LLM h_l with frozen weights $W_{\text{pre}} = \{W_{\text{pre}}^l\}_{l=1}^L$, where $W_{\text{pre}}^l \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$. Following the parameterization $\Theta = \{\Phi_{\mathcal{P}}, \Delta\Phi_l\}$ defined in Sec. 3.1, we implement the reasoning track by embedding LoRA modules into the linear layers of the LLM. For a specific layer l, the task-specific update for task t is represented as a low-rank task vector $\tau_t^l = B_t^l A_t^l$, where $B_t^l \in \mathbb{R}^{d_{\text{out}} \times r}$ and $A_t^l \in \mathbb{R}^{r \times d_{\text{in}}}$. Critically, following the LoRA Hu et al. (2022) fine-tuning strategy used in most MLLMs Liu et al. (2024a), we also employ a strictly bias-free LoRA architecture to ensure linear additivity. Under our PEFT regime, since the LLM backbone W_{pre} remains frozen, the task vector for the l-th layer is equivalent to the learned low-rank increment:

$$\tau_t^l = W_t^l - W_{\text{pre}}^l = \Delta W_t^l = B_t^l A_t^l. \tag{4}$$

This design allows us to reframe the challenge of continual learning as a layer-wise model merging problem in the parameter subspace: our objective is to continuously integrate newly acquired task vectors τ_t^l into a consolidated global task vector τ^{l*}, thereby achieving exemplar-free knowledge consolidation while minimizing interference with preceding tasks.

4.3.2 Objective: Resolving Layer-wise Task Vector Conflicts

While the task vector formulation allows for layer-wise merging, a straightforward arithmetic summation (e.g., $\tau^l = \sum_{i=1}^{t} \tau_i^l$ Ilharco et al. (2022)) ignores the fact that different task vectors often occupy interfering directions in the weight space. This conflict manifests as negative transfer Yadav et al. (2023); Sun et al. (2025b), where the performance degradation on task i is formally defined as the loss gap between the merged model $W_{\text{merge}} = W_{\text{pre}} + \tau$ and the task-specific fine-tuned model $W_i = W_{\text{pre}} + \tau_i$:

$$\Delta\mathcal{L}_i = \mathbb{E}_{(x_v,x_{ins},x_{ans})\sim\mathcal{D}_i}[\mathcal{L}(W_{\text{merge}}; x_v, x_{ins}, x_{ans})] - \mathbb{E}_{(x_v,x_{ins},x_{ans})\sim\mathcal{D}_i}[\mathcal{L}(W_i; x_v, x_{ins}, x_{ans})]. \quad (5)$$

Our goal is to find the optimal task vectors τ^* that jointly minimize this gap $\Delta\mathcal{L}_i$ across all encountered tasks. Directly analyzing this global degradation is intractable due to the complex hierarchical structure of MLLMs. However, recent studies Sun et al. (2025a) demonstrate that this global conflict is strictly bounded by the accumulation of layer-wise feature drift:

$$|\Delta\mathcal{L}_i| \leq \beta \sum_{l=1}^{L} \left(\prod_{m=l+1}^{L} \gamma_m \right) \|\Delta f_i^l\|_F, \quad (6)$$

where β and γ_m are the Lipschitz constants of the loss function and the m-th layer, respectively, and $\Delta f_i^l = f_i^l(W_{\text{pre}}^l + \tau^l) - f_i^l(W_{\text{pre}}^l + \tau_i^l)$ denotes the feature drift induced at layer l. This theorem establishes a critical guideline: mitigating global reasoning collapse inherently requires minimizing $\|\Delta f_i^l\|_F$ layer by layer. Leveraging the strictly bias-free LoRA architecture established in Sec. 4.3.1, the feature transformation at any linear layer l simplifies to a pure matrix multiplication: $f_i^l(W^l) = \mathbf{X}_i^l W^l$, where $\mathbf{X}_i^l$ denotes the input feature matrix for task i. Substituting this into the drift definition, the feature drift at layer l is expressed as:

$$\Delta f_i^l = \mathbf{X}_i^l(W_{\text{pre}}^l + \tau^l) - \mathbf{X}_i^l(W_{\text{pre}}^l + \tau_i^l) = \mathbf{X}_i^l(\tau^l - \tau_i^l). \quad (7)$$

To maximize knowledge retention up to task t, the goal of the knowledge consolidation mechanism is to find an optimal merged vector τ_t^{l*} that incorporate the historical task vectors $\{\tau_1^l, \ldots, \tau_{t-1}^l\}$ with the current vector τ_t^l while minimizing the cumulative feature drift. Assuming that all historical features $\{\mathbf{X}_i^l\}_{i=1}^t$ and task vectors $\{\tau_i^l\}_{i=1}^t$ are available simultaneously, we define the following equivalent global objective:

$$\tau_t^{l*} = \arg\min_{\tau^l} \sum_{i=1}^{t} \|\mathbf{X}_i^l(\tau^l - \tau_i^l)\|_F^2. \quad (8)$$

By setting the gradient of Eq. 8 with respect to τ_t^{l*} to zero, we obtain the batch least squares solution:

$$\tau_t^{l*} = \left(\sum_{i=1}^{t} \mathbf{X}_i^{l\top} \mathbf{X}_i^l \right)^{-1} \sum_{i=1}^{t} \mathbf{X}_i^{l\top} \mathbf{X}_i^l \tau_i^l. \quad (9)$$

Directly computing Eq. 9 necessitates joint access to all historical features $\{\mathbf{X}_i^l\}_{i=1}^t$ and task vectors $\{\tau_i^l\}_{i=1}^t$, which is structurally intractable in exemplar-free continual learning.

4.3.3 Exemplar-free Recursive Parameter Merging

To satisfy the exemplar-free constraint, we reformulate the batch solution in Eq. 9 into a recursive update rule that allows for incremental parameter merging. We first represent the historical data distribution through a sufficient statistic, namely the cumulative feature covariance:

Definition 1 (Cumulative Feature Covariance)**.** *For each layer l, the cumulative feature covariance matrix $\mathbf{H}_t^l \in \mathbb{R}^{d\times d}$ is defined as the unnormalized second-order moment of all input features encountered up to task t:*

$$\mathbf{H}_t^l = \sum_{i=1}^{t} \mathbf{X}_i^{l\top} \mathbf{X}_i^l = \mathbf{H}_{t-1}^l + \mathbf{X}_t^{l\top} \mathbf{X}_t^l. \quad (10)$$

The matrix $\mathbf{H}_t^l$ captures the relative importance of different parameter directions, enabling the model to remember historical feature structures without storing raw samples.

By maintaining the matrix $\mathbf{H}_t^l$, the optimal solution can be updated recursively as summarized in the following theorem:

Theorem 1 (Recursive Least Squares (RLS) Solution)**.** *The optimal merged task vector τ_t^{l*} that minimizes the cumulative feature drift up to task t can be obtained recursively:*

$$\tau_t^{l*} = \tau_{t-1}^{l*} + \mathbf{S}_t^l(\tau_t^l - \tau_{t-1}^{l*}), \quad \textit{where } \mathbf{S}_t^l = (\mathbf{H}_t^l)^{-1}(\mathbf{X}_t^{l\top}\mathbf{X}_t^l). \tag{11}$$

Here, $\mathbf{S}_t^l$ denotes the gain matrix that adaptively modulates the task residual $(\tau_t^l - \tau_{t-1}^{l*})$. This update rule achieves exact mathematical equivalence to the batch least squares solution in Eq. 9 without requiring raw historical data.

Proof. See Appendix A.1 for the detailed algebraic derivation based on the matrix inversion lemma and quadratic optimization. □

Unlike simple averaging, this recursive solution offers a clear advantage by using feature strength to prioritize important updates. As explained via Singular Value Decomposition (SVD) Wall et al. (2003) in Appendix A.2, this method naturally puts updates where they are most useful and removes parts that cause conflicts to prevent critical information from being lost even when different tasks update the same parts of the model. During testing, the merged update τ_t^{l*} is added into the frozen weights W_{pre}^l using a scaling factor λ:

$$W_{\text{final}}^l = W_{\text{pre}}^l + \lambda\tau_t^{l*} \tag{12}$$

Adjusting λ follows common practice in model merging Yadav et al. (2023); Du et al. (2024); Sun et al. (2025c) to balance new task knowledge with the pre-trained model for better performance across downstream continual learning tasks.

Additionally, to alleviate the storage burden, we avoid storing the full task vector τ_t^{l*} by merging it directly into the pre-trained weights. We recover the task vector only when it is needed by calculating the difference between the model and the pre-trained weights W_{pre}^l. Our method then focuses on managing the cumulative feature covariance $\mathbf{H}_t^l$ through two versions: **MAny** and **MAny***. In **MAny**, we maintain the full $d \times d$ matrix $\mathbf{H}_t^l$. To handle the matrix memory cost, **MAny*** uses SVD to compress $\mathbf{H}_t^l$. We keep only the top-r components where the sum of their squared singular values represents a sufficiently large portion of the total feature strength. Specifically, we choose the smallest r such that the ratio meets an energy threshold γ ($\frac{\sum_{i=1}^r \sigma_i^2}{\sum_{i=1}^d \sigma_i^2} \geq \gamma$), yielding the low-rank approximation $\mathbf{H}_t^l \approx \mathbf{U}_r^l\mathbf{\Sigma}_r^l\mathbf{V}_r^{l\top}$. This approach allows the model to remember historical feature structures using much less storage space while keeping the most useful information.

In summary, CPM maintains stable visual alignments through adaptive projector merging, while LPM recursively consolidates task vectors to resolve parameter conflicts. This dual-track design preserves both perceptual and structural knowledge across tasks and eliminates the need for expensive additional GPU training. The pseudo code is included in Algorithm 1.

5 Experiment

5.1 Experiment Setup

Datasets. We evaluate our framework on two MCIT benchmarks: UCIT Guo et al. (2025a) and MLLM-DCL Zhao et al. (2025a). UCIT benchmark rigorously assesses true learning capability without pre-training data contamination. Based on LLaVA's Liu et al. (2023) zero-shot performance, it comprises six sequential tasks strictly uncorrelated with standard MLLM corpora: ArxivQA Li et al. (2024), CLEVR-Math Lindström & Abraham (2022), IconQA Lu et al. (2021), ImageNet-R Hendrycks et al. (2021), VizWiz-caption (Viz-cap) Gurari et al. (2018), and Flickr30k Plummer et al. (2015). Conversely, MLLM-DCL benchmark evaluates the sequential acquisition of specialized knowledge across five distinct professional domains: remote sensing (RSVQA Lobry et al. (2020)), medicine (PathVQA He et al. (2020)), autonomous driving (DriveLM Sima et al. (2024)), science (AI2D Kembhavi et al. (2016), SciVerse Guo et al. (2025c), MapQA Chang et al. (2022), TQA Kembhavi et al. (2017)), and finance (StockQA Wang et al. (2023b)).

Table 1: Results on the UCIT benchmark using LLaVA-1.5-7B and InternVL-Chat-7B models.

	Method	Venue	Image-R	ArxivQA	Viz-cap	IconQA	CLEVR	Flickr30k	FAA (↑)	CAA (↑)	FFM (↓)
LLaVA-1.5-7B	Zero-shot	–	16.27	57.73	38.39	19.20	20.63	41.88	–	–	–
	LoRA-FT Hu et al. (2022)	ICLR'22	65.50	76.23	44.41	66.37	54.87	**57.80**	60.86	67.61	17.44
	O-LoRA Wang et al. (2023a)	EMNLP'23	83.40	**94.20**	41.51	58.50	56.80	53.37	64.63	75.98	8.03
	MoELoRA Chen et al. (2024a)	NeurIPS'24	68.57	77.73	44.26	65.43	45.20	57.31	59.75	75.57	17.54
	ModalPrompt Zeng et al. (2025)	EMNLP'25	51.07	87.27	48.11	39.23	46.57	42.93	52.53	57.67	**0.15**
	CL-MoE Huai et al. (2025)	CVPR'25	67.80	73.47	44.58	69.37	47.80	57.40	60.07	74.89	18.25
	HiDe-LLaVA Guo et al. (2025a)	ACL'25	86.93	91.37	48.68	67.80	47.87	52.95	65.93	79.03	5.58
	SEFE Chen et al. (2025)	ICML'25	83.53	65.77	45.72	75.23	63.03	57.74	65.17	76.64	12.80
	MAny*	**Ours**	**91.30**	92.63	59.54	79.63	**69.37**	56.96	**74.91** (+8.98)	**83.89** (+4.86)	0.92
	MAny	**Ours**	90.87	91.97	**60.48**	**80.20**	67.37	56.09	74.50 (+8.57)	83.42 (+4.39)	0.37
InternVL-Chat-7B	Zero-shot	–	21.10	63.20	40.59	24.70	21.20	44.67	–	–	–
	LoRA-FT Hu et al. (2022)	ICLR'22	78.87	76.23	45.36	66.43	70.13	**58.45**	65.91	77.89	10.12
	O-LoRA Wang et al. (2023a)	EMNLP'23	86.00	**94.23**	39.68	63.83	46.97	45.16	62.65	79.06	6.97
	MoELoRA Chen et al. (2024a)	NeurIPS'24	70.50	77.53	44.98	68.17	69.57	57.80	64.76	77.59	10.40
	ModalPrompt Zeng et al. (2025)	EMNLP'25	57.10	63.90	42.11	37.40	47.50	45.76	48.96	53.52	**0.02**
	CL-MoE Huai et al. (2025)	CVPR'25	76.80	78.43	44.72	69.07	**73.07**	58.43	66.75	79.68	9.24
	HiDe-LLaVA Guo et al. (2025a)	ACL'25	89.33	91.07	49.55	67.80	63.63	55.62	69.50	81.33	2.03
	SEFE Chen et al. (2025)	ICML'25	86.77	78.13	45.90	70.67	69.80	58.18	68.24	79.87	6.18
	MAny*	**Ours**	**94.60**	81.80	57.86	70.97	69.20	56.62	71.84 (+2.34)	83.10 (+1.77)	3.00
	MAny	**Ours**	**94.60**	83.60	**59.86**	**71.93**	67.63	56.48	**72.35** (+2.85)	**83.35** (+2.02)	1.77

Evaluation Metrics: We utilize three widely-recognized evaluation metrics for continual learning - Final Average Accuracy (FAA), Cumulative Average Accuracy (CAA), and Final Forgetting Measure (FFM) as detailed in Wang et al. (2024). We define the accuracy on the task $\mathcal{T}^j$ after learning the task $\mathcal{T}^i$ as A_{ij}. The average accuracy after learning task $\mathcal{T}^i$ is denoted as $AA_i = \frac{1}{i}\sum_{j=1}^{i} A_{ij}$. Upon completing all n tasks, we report $\text{FAA} = AA_n$, $\text{CAA} = \frac{1}{n}\sum_{i=1}^{n} AA_i$, and $\text{FFM} = \frac{1}{n-1}\sum_{j=1}^{n-1}(A_{jj} - A_{nj})$. The FAA is a critical metric highlighting performance discrepancies between CL methods and joint learning. The CAA provides a comprehensive view of overall historical performance, and the FFM quantifies the model's capability to mitigate forgetting.

Implementation Details. All algorithms are implemented using PyTorch Paszke et al. (2019) and the MCITlib toolbox Guo et al. (2025b), and are executed on an NVIDIA RTX A100 GPU cluster. We employ LLaVA-1.5-7B Liu et al. (2023) and InternVL-Chat-7B Chen et al. (2024b) as our foundational backbones, utilizing CLIP-L/14-336 Radford et al. (2021) for visual feature extraction. For parameter-efficient fine-tuning, we inject LoRA modules with a rank of 16. We set the scaling factor λ to 3 and the energy threshold γ in MAny* to 0.999. See Appendix A.3 for more details.

Baseline. We compare our MAny with recent MCIT methods, including LoRA-FT Hu et al. (2022), O-LoRA Wang et al. (2023a), MoELoRA Chen et al. (2024a), ModalPrompt Zeng et al. (2025), CL-MoE Huai et al. (2025), HiDe-LLaVA Guo et al. (2025a), SEFE Chen et al. (2025). The training process follows the replay-free continual learning setting Chen et al. (2025).

5.2 Main Results

Table 1 summarizes the core performance of MAny on the UCIT Guo et al. (2025a) benchmark, with full results and analysis on the more challenging MLLM-DCL benchmark deferred to the Table 4 in Appendix A.4. Across both LLaVA-1.5-7B and InternVL-Chat-7B, MAny consistently sets a new state-of-the-art (SOTA), resolving the core dilemma by achieving an unprecedented balance between plasticity and stability. Concretely, MAny outperforms the prior strongest baseline HiDe-LLaVA Guo et al. (2025a) by significant margins: +8.57% in FAA on LLaVA-1.5-7B, with FFM reduced to an ultra-low 0.37. This superiority generalizes seamlessly to InternVL-Chat-7B, with a +2.85% FAA gain over HiDe-LLaVA Guo et al. (2025a) and lower forgetting, validating robust cross-architecture generalization. Our memory-efficient variant MAny*, which leverages SVD to drastically reduce storage overhead, fully retains the competitive performance of MAny without degradation.

5.3 Ablation Study

Table 2 ablates MAny to validate our dual-track architecture. The naive baseline suffers from a severe 17.44% FFM. Introducing CPM alone reduces FFM to 10.45% and improves FAA to 68.71%, mitigating perception drift. Conversely, applying LPM alone drops FFM to 5.85%, resolving internal reasoning conflicts. Combining both modules in MAny yields the optimal FAA (74.50%) and virtually eliminates forgetting, with highly consistent results on InternVL-Chat-7B. Further analyses

Table 2: Ablation study of different components in MAny on the UCIT benchmark.

CPM	LPM	LLaVA-1.5-7B			InternVL-Chat-7B		
		FAA (↑)	CAA (↑)	FFM (↓)	FAA (↑)	CAA (↑)	FFM (↓)
		60.86	67.61	17.44	65.91	77.89	10.12
✓		68.71	75.65	10.45	68.20	80.14	8.22
	✓	70.87	81.93	5.85	69.99	81.85	3.71
✓	✓	**74.50**	**83.42**	**0.37**	**72.35**	**83.35**	**1.77**

on the synergy between perceptual and reasoning merging, the performance gap against theoretical bounds, and CPM's modality selection are detailed in Appendix A.5 and A.6, respectively.

5.4 Plug-and-Play Performance

Table 3: Comparison of different methods w/ and w/o the CPM module on the UCIT benchmark.

Method	w/o CPM			w/ CPM		
	FAA (↑)	CAA (↑)	FFM (↓)	FAA (↑)	CAA (↑)	FFM (↓)
Task-specific LoRA	73.99	82.56	4.40	**76.98**	**84.01**	**0.23**
LoRA-FT	60.86	67.61	17.44	68.71	75.65	10.45
O-LoRA	64.63	75.98	8.03	68.53	80.58	2.78
HiDe-LLaVA	65.93	79.03	5.58	68.93	80.23	1.96
SEFE	65.17	76.64	12.80	69.04	79.37	5.76
MAny (Ours)	70.87	81.93	5.85	74.50	83.42	0.37

Table 3 underscores CPM's versatility. In the "w/o CPM" setting, baselines are forced to utilize the projector from the final task, which triggers severe "perception drift" (e.g., LoRA-FT's 18.64% FFM). Notably, Task-specific LoRA with CPM achieves an FAA of 76.98%, outstripping the ideal task-specific oracle (75.40% detailed in Table 5). This super-oracle performance indicates that CPM's adaptive soft-merging provides more robust cross-modal alignment than isolated task projectors. By mitigating the shared-projector bottleneck, CPM consistently reduces forgetting across various baselines. For example, it reduces HiDe-LLaVA's FFM from 5.58% to 1.96%.

5.5 Further Analysis

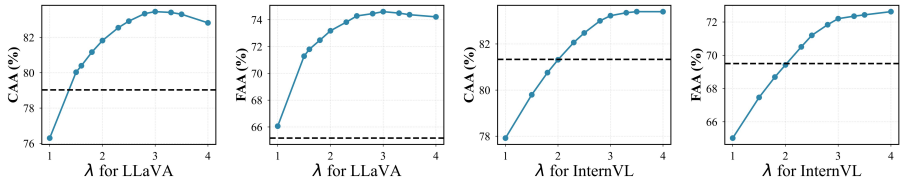


Figure 4: Sensitivity of λ. Dashed lines denote the SOTA performance of all baseline.

Sensitivity of the Scaling Factor λ. Figure 4 evaluates MAny's performance across various λ values, where dashed lines indicate the current SOTA performance. Our recursive merging neutralizes interfering parameter directions, which naturally attenuates the magnitude of the merged task vector τ_t^{l*}. Consequently, a larger scaling factor is required to amplify the purified knowledge against the frozen backbone. For both LLaVA and InternVL, MAny outperforms the SOTA over a broad range of λ values. Notably, the FAA metric demonstrates exceptional robustness and maintains superior accuracy once λ exceeds 2.0. This saturation around $\lambda \approx 3.0$ confirms that an appropriate scaling factor effectively compensates for RLS-induced attenuation without compromising stability. We set $\lambda = 3$ as the default for all experiments. See Appendix A.7 for the hyperparameter γ analysis.

6 Conclusion

In this work, we identified a neglected dual-forgetting phenomenon in MCIT involving perception drift in the projection space and reasoning collapse in the parameter space. To resolve these conflicts, we presented MAny, a training-free framework that decouples perception from reasoning through a dual-track merging design. CPM restores perceptual alignment via adaptive prototype guidance, while LPM ensures optimal reasoning consolidation using a recursive least squares mechanism. Extensive evaluations on the UCIT and MLLM-DCL benchmarks demonstrate that MAny consistently sets a new state-of-the-art across various MLLM architectures. By leveraging a training-free merging paradigm, MAny eliminates additional gradient-based optimization beyond naive task tuning, providing a robust and lightweight solution for multimodal continual instruction tuning.

AI USE STATEMENT

In this work, we used generative AI tools for writing polish, grammar correction, and improving text readability across the manuscript. We have not used generative AI tools for research ideation, study design, theoretical proofs, algorithm development, code implementation, data collection, or experimental analysis, and all other research components with required disclosure are not applicable to this work. Additionally, we used generative AI tools for refining text clarity and styling. We have reviewed all AI-assisted work. All LLM-polished text was manually checked and edited by the authors to ensure that the original technical meaning, mathematical statements, and nuances remained precise and untranslated in intent. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Shuaichen Chang, David Palzer, Jialin Li, Eric Fosler-Lussier, and Ningchuan Xiao. Mapqa: A dataset for question answering on choropleth maps. *arXiv preprint arXiv:2211.08545*, 2022.

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

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 *Forty-second International Conference on Machine Learning*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024b.

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. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Guodong Du, Junlin Lee, Jing Li, Runhua Jiang, Yifei Guo, Shuyang Yu, Hanting Liu, Sim K Goh, Ho-Kin Tang, Daojing He, et al. Parameter competition balancing for model merging. *Advances in Neural Information Processing Systems*, 37:84746–84776, 2024.

Zijian Gao, Kele Xu, Huiping Zhuang, Li Liu, Xinjun Mao, Bo Ding, Dawei Feng, and Huaimin Wang. Less confidence, less forgetting: Learning with a humbler teacher in exemplar-free class-incremental learning. *Neural Networks*, 179:106513, 2024a. ISSN 0893-6080.

Zijian Gao, Xingxing Zhang, Kele Xu, Xinjun Mao, and Huaimin Wang. Stabilizing zero-shot prediction: A novel antidote to forgetting in continual vision-language tasks. *Advances in Neural Information Processing Systems*, 37:128462–128488, 2024b.

Zijian Gao, Shanhao Han, Xingxing Zhang, Kele Xu, Dulan Zhou, Xinjun Mao, Yong Dou, and Huaimin Wang. Maintaining fairness in logit-based knowledge distillation for class-incremental learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 16763–16771, 2025a.

Zijian Gao, Wangwang Jia, Xingxing Zhang, Dulan Zhou, Kele Xu, Feng Dawei, Yong Dou, Xinjun Mao, and Huaimin Wang. Knowledge memorization and rumination for pre-trained model-based class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20523–20533, 2025b.

Zijian Gao, Kele Xu, Xingxing Zhang, Huiping Zhuang, Tianjiao Wan, Bo Ding, Xinjun Mao, and Huaimin Wang. Rethinking obscured sub-optimality in analytic learning for exemplar-free class-incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025c.

Haiyang Guo, Fanhu Zeng, Ziwei Xiang, Fei Zhu, Da-Han Wang, Xu-Yao Zhang, and Cheng-Lin Liu. Hide-llava: Hierarchical decoupling for continual instruction tuning of multimodal large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13572–13586, 2025a.

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

Ziyu Guo, Renrui Zhang, Hao Chen, Jialin Gao, Dongzhi Jiang, Jiaze Wang, and Pheng-Ann Heng. Sciverse: Unveiling the knowledge comprehension and visual reasoning of lmms on multi-modal scientific problems. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19683–19704, 2025c.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

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

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021.

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

Tianyu Huai, Jie Zhou, Xingjiao Wu, Qin Chen, Qingchun Bai, Ze Zhou, and Liang He. Cl-moe: Enhancing multimodal large language model with dual momentum mixture-of-experts for continual visual question answering. In *Proceedings of the computer vision and pattern recognition conference*, pp. 19608–19617, 2025.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Suchin Gururangan, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. Editing models with task arithmetic. *arXiv preprint arXiv:2212.04089*, 2022.

Xisen Jin, Xiang Ren, Daniel Preotiuc-Pietro, and Pengxiang Cheng. Dataless knowledge fusion by merging weights of language models. In *International Conference on Learning Representations*, 2023.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern recognition*, pp. 4999–5007, 2017.

Minjae Lee, Minhyuk Seo, Tingyu Qu, Tinne Tuytelaars, and Jonghyun Choi. Oasis: Online sample selection for continual visual instruction tuning. *arXiv preprint arXiv:2506.02011*, 2025.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14369–14387, 2024.

Songze Li, Mingyu Gao, Tonghua Su, Xu-Yao Zhang, and Zhongjie Wang. Multimodal continual instruction tuning with dynamic gradient guidance. *arXiv preprint arXiv:2511.15164*, 2025.

Adam Dahlgren Lindström and Savitha Sam Abraham. Clevr-math: A dataset for compositional language, visual and mathematical reasoning. *arXiv preprint arXiv:2208.05358*, 2022.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024b.

Yuyang Liu, Qiuhe Hong, Linlan Huang, Alexandra Gomez-Villa, Dipam Goswami, Xialei Liu, Joost van de Weijer, and Yonghong Tian. Continual learning for vlms: A survey and taxonomy beyond forgetting. *arXiv preprint arXiv:2508.04227*, 2025.

Sylvain Lobry, Diego Marcos, Jesse Murray, and Devis Tuia. Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566, 2020.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V Le, Barret Zoph, Jason Wei, et al. The flan collection: Designing data and methods for effective instruction tuning. In *International conference on machine learning*, pp. 22631–22648. PMLR, 2023.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Michael S Matena and Colin A Raffel. Merging models with fisher-weighted averaging. *Advances in Neural Information Processing Systems*, 35:17703–17716, 2022.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2):1–40, 2023.

Sergei Nirenburg. *Progress in machine translation*. Ios Press, 1993.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

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

Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in neural information processing systems*, 34:6747–6761, 2021.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pp. 256–274. Springer, 2024.

Wenju Sun, Qingyong Li, Yangli-ao Geng, and Boyang Li. Cat merging: A training-free approach for resolving conflicts in model merging. *arXiv preprint arXiv:2505.06977*, 2025a.

Wenju Sun, Qingyong Li, Wen Wang, Yangliao Geng, and Boyang Li. Task arithmetic in trust region: A training-free model merging approach to navigate knowledge conflicts. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 5178–5187, 2025b.

Wenju Sun, Qingyong Li, Wen Wang, Yang Liu, Yangliao Geng, and Boyang Li. Towards minimizing feature drift in model merging: Layer-wise task vector fusion for adaptive knowledge integration. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025c.

Shengbang Tong, David Fan, Jiachen Li, Yunyang Xiong, Xinlei Chen, Koustuv Sinha, Michael Rabbat, Yann LeCun, Saining Xie, and Zhuang Liu. Metamorph: Multimodal understanding and generation via instruction tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17001–17012, 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pp. 91–109. Springer, 2003.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuan-Jing Huang. Orthogonal subspace learning for language model continual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10658–10671, 2023a.

Ziao Wang, Yuhang Li, Junda Wu, Jaehyeon Soon, and Xiaofeng Zhang. Finvis-gpt: A multimodal large language model for financial chart analysis. *arXiv preprint arXiv:2308.01430*, 2023b.

Prateek Yadav, Derek Tam, Leshem Choshen, Colin A Raffel, and Mohit Bansal. Ties-merging: Resolving interference when merging models. *Advances in neural information processing systems*, 36:7093–7115, 2023.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

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

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

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12401–12430, 2024.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Guoyin Wang, et al. Instruction tuning for large language models: A survey. *ACM Computing Surveys*, 58(7):1–36, 2026.

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

Xuanle Zhao, Xianzhen Luo, Qi Shi, Chi Chen, Shuo Wang, Zhiyuan Liu, and Maosong Sun. Chartcoder: Advancing multimodal large language model for chart-to-code generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7333–7348, 2025b.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

A Appendix

A.1 Derivation and Equivalence Proof of Theorem 1

This appendix provides the formal algebraic derivation of the Recursive Least Squares (RLS) merging rule introduced in Sec. 4.3.3. We prove that the incremental update in Theorem 1 is mathematically equivalent to the batch least squares solution.

For a specific layer l, let $\mathbf{H}_t^l \in \mathbb{R}^{d\times d}$ be the cumulative feature covariance and momentum $\mathbf{Q}_t^l \in \mathbb{R}^{d\times r}$ be the cumulative weighted task vector. These sufficient statistics are defined as:

$$\mathbf{H}_t^l = \sum_{i=1}^{t} \mathbf{X}_i^{l\top}\mathbf{X}_i^l, \quad \mathbf{Q}_t^l = \sum_{i=1}^{t} \mathbf{X}_i^{l\top}\mathbf{X}_i^l\tau_i^l. \tag{13}$$

As derived in Eq. 9, the optimal batch solution τ_t^{l*} that minimizes the cumulative feature drift is given by:

$$\tau_t^{l*} = (\mathbf{H}_t^l)^{-1}\mathbf{Q}_t^l. \tag{14}$$

Upon encountering the t-th task, the statistics $\mathbf{H}_t^l$ and $\mathbf{Q}_t^l$ can be updated from their previous states at task $t-1$:

$$\mathbf{H}_t^l = \mathbf{H}_{t-1}^l + \mathbf{X}_t^{l\top}\mathbf{X}_t^l, \tag{15}$$

$$\mathbf{Q}_t^l = \mathbf{Q}_{t-1}^l + \mathbf{X}_t^{l\top}\mathbf{X}_t^l\tau_t^l. \tag{16}$$

From Eq. 14, we have the relationship $\mathbf{Q}_{t-1}^l = \mathbf{H}_{t-1}^l\tau_{t-1}^{l*}$.

To derive the update rule for τ_t^{l*}, we substitute the recurrence equation 16 into the batch form equation 14:

$$\begin{aligned}\tau_t^{l*} &= (\mathbf{H}_t^l)^{-1}\mathbf{Q}_t^l = (\mathbf{H}_t^l)^{-1}\left(\mathbf{Q}_{t-1}^l + \mathbf{X}_t^{l\top}\mathbf{X}_t^l\tau_t^l\right)\\ &= (\mathbf{H}_t^l)^{-1}\left(\mathbf{H}_{t-1}^l\tau_{t-1}^{l*} + \mathbf{X}_t^{l\top}\mathbf{X}_t^l\tau_t^l\right).\end{aligned} \tag{17}$$

By rearranging Eq. 15 as $\mathbf{H}_{t-1}^l = \mathbf{H}_t^l - \mathbf{X}_t^{l\top}\mathbf{X}_t^l$, we can eliminate the historical covariance $\mathbf{H}_{t-1}^l$:

$$\begin{aligned}\tau_t^{l*} &= (\mathbf{H}_t^l)^{-1}\left[(\mathbf{H}_t^l - \mathbf{X}_t^{l\top}\mathbf{X}_t^l)\tau_{t-1}^{l*} + \mathbf{X}_t^{l\top}\mathbf{X}_t^l\tau_t^l\right]\\ &= (\mathbf{H}_t^l)^{-1}\mathbf{H}_t^l\tau_{t-1}^{l*} - (\mathbf{H}_t^l)^{-1}\mathbf{X}_t^{l\top}\mathbf{X}_t^l\tau_{t-1}^{l*} + (\mathbf{H}_t^l)^{-1}\mathbf{X}_t^{l\top}\mathbf{X}_t^l\tau_t^l\\ &= \tau_{t-1}^{l*} + (\mathbf{H}_t^l)^{-1}(\mathbf{X}_t^{l\top}\mathbf{X}_t^l)(\tau_t^l - \tau_{t-1}^{l*}).\end{aligned} \tag{18}$$

Algorithm 1 The Proposed MAny Method

Input: Pre-trained weights $\{W^l_{\text{pre}}, \Phi_v\}$, Task sequence $\{\mathcal{T}^t\}_{t=1}^n$, Scaling factor λ, Temperature η
Output: Merged task vectors $\{\tau_n^{l*}\}_{l=1}^L$, Task-specific Projectors $\{\mathcal{P}_i\}_{i=1}^n$, and Prototypes $\{\mu_i\}_{i=1}^n$

Training Phase:
Initialize cumulative covariance $\mathbf{H}_0^l = \mathbf{0}$ and global task vector $\tau_0^{l*} = \mathbf{0}$ for each layer $l \in \{1, \dots, L\}$
for $t = 1$ **to** n **do**
  1. Task-specific Training:
  Fine-tune LoRA weights τ_t^l and task-specific projector $\mathcal{P}_t$ on $\mathcal{T}^t$
  2. Perceptual Track (Cross-modal Projection Merging):
    Extract [CLS] features $\mathbf{Z}_v^{\text{g}}$ from the last layer of Φ_v
    Compute task-representative prototype: $\mu_t = \frac{1}{N_t}\sum_{j=1}^{N_t} \mathbf{Z}_v^{\text{g},j}$
    Store $\{\mathcal{P}_t, \mu_t\}$ for test-time adaptive merging
  3. Reasoning Track (Low-rank Parameter Merging):
  for each reasoning layer $l = 1$ **to** L **do**
      Update cumulative feature covariance: $\mathbf{H}_t^l = \mathbf{H}_{t-1}^l + \mathbf{X}_t^{l\top}\mathbf{X}_t^l$
      Compute gain matrix: $\mathbf{S}_t^l = (\mathbf{H}_t^l)^{-1}(\mathbf{X}_t^{l\top}\mathbf{X}_t^l)$
      Recursive parameter merging: $\tau_t^{l*} = \tau_{t-1}^{l*} + \mathbf{S}_t^l(\tau_t^l - \tau_{t-1}^{l*})$
  end for
end for
Inference Phase:
Extract global feature $\mathbf{Z}_v^{\text{g}}$ and spatial tokens $\mathbf{Z}_v^{\text{s}}$ from $f_v(x_v; \Phi_v)$
Compute similarity between $\mathbf{Z}_v^{\text{g}}$ and μ_i via Eq. 2
Calculate merged projected tokens $\mathbf{P}_v$ according to Eq. 3
Construct reasoning weights $\{W^l_{\text{final}}\}_{l=1}^L$ where $W^l_{\text{final}} = W^l_{\text{pre}} + \lambda\tau_n^{l*}$
return $\mathbf{P}_v$ and $\{W^l_{\text{final}}\}_{l=1}^L$ to generate answer

Defining the gain matrix as $\mathbf{S}_t^l = (\mathbf{H}_t^l)^{-1}(\mathbf{X}_t^{l\top}\mathbf{X}_t^l)$, we arrive at the final recursive fusion rule:

$$\tau_t^{l*} = \tau_{t-1}^{l*} + \mathbf{S}_t^l(\tau_t^l - \tau_{t-1}^{l*}). \tag{19}$$

This derivation rigorously confirms that the incremental update is mathematically identical to the joint batch optimum. Crucially, this equivalence is maintained without requiring the raw historical features $\mathbf{X}_{1:t-1}^l$ storage, as all necessary information is compressed within the sufficient statistic $\mathbf{H}_t^l$.

A.2 Theoretical Mechanism and Subspace Analysis of Low-Rank Parameter Merging

As established in Appendix A.1, our recursive task vector fusion rule yields an exact mathematical equivalent to the joint batch optimal solution $\tau_t^{l*} = (\sum_{i=1}^t \mathbf{X}_i^{l\top}\mathbf{X}_i^l)^{-1}\sum_{i=1}^t \mathbf{X}_i^{l\top}\mathbf{X}_i^l\tau_i^l$. To elucidate the underlying mechanics of this approach, we present a theoretical analysis focusing on the fusion behavior of matrix-multiplication parameters.Specifically, we apply SVD to the input feature matrix of a given task i: $\mathbf{X}_i^l = \mathbf{U}_i^l\mathbf{\Sigma}_i^l\mathbf{V}_i^{l\top}$. To rigorously handle potential rank deficiency in feature spaces, we employ the Moore-Penrose pseudoinverse (†) in the subsequent derivations. By analyzing two extreme cases of feature distribution, we uncover the fusion dynamics at play.

Ideal case (Mutually orthogonal features). Assume that for any pair of distinct tasks $i \neq j$, their right singular vectors satisfy the strict orthogonality condition ($\mathbf{V}_i^{l\top}\mathbf{V}_j^l = \mathbf{0}$). Under this condition, the batch optimal solution simplifies significantly to the summation of task vectors projected onto their respective feature subspaces:

$$\begin{aligned}\tau_{t,\text{ideal}}^{l*} &= \left(\sum_{i=1}^t \mathbf{V}_i^l(\mathbf{\Sigma}_i^l)^2\mathbf{V}_i^{l\top}\right)^\dagger \sum_{j=1}^t \mathbf{V}_j^l(\mathbf{\Sigma}_j^l)^2\mathbf{V}_j^{l\top}\tau_j^l \\ &= \sum_{i=1}^t \left(\mathbf{V}_i^l(\mathbf{\Sigma}_i^l)^2\mathbf{V}_i^{l\top}\right)^\dagger \left(\mathbf{V}_i^l(\mathbf{\Sigma}_i^l)^2\mathbf{V}_i^{l\top}\right)\tau_i^l = \sum_{i=1}^t \mathbf{V}_i^l\mathbf{V}_i^{l\top}\tau_i^l.\end{aligned} \tag{20}$$

Here, $\mathbf{V}_i^l \mathbf{V}_i^{l\top} \tau_i^l$ represents the orthogonal projection of the task vector τ_i^l onto the subspace spanned by the singular vectors $\mathbf{V}_i^l$. This projection mechanism exclusively retains the components of τ_i^l that align with the corresponding feature space $\mathbf{X}_i^l$, effectively filtering out irrelevant update directions.Consequently, in this ideal orthogonal setting, where $\mathbf{V}_i^{l\top} \mathbf{V}_j^l = \mathbf{0}$ for any $i \neq j$, the optimal fusion strategy achieves lossless knowledge integration. We prove that the cumulative layer-wise feature drift remains strictly zero by substituting the ideal solution $\tau_{t,\text{ideal}}^{l*} = \sum_{j=1}^t \mathbf{V}_j^l \mathbf{V}_j^{l\top} \tau_j^l$ into the objective:

$$\begin{aligned}
\sum_{i=1}^t \|\mathbf{X}_i^l(\tau_{t,\text{ideal}}^{l*} - \tau_i^l)\|_F^2 &= \sum_{i=1}^t \left\| \mathbf{U}_i^l \mathbf{\Sigma}_i^l \mathbf{V}_i^{l\top} \left(\sum_{j=1}^t \mathbf{V}_j^l \mathbf{V}_j^{l\top} \tau_j^l - \tau_i^l \right) \right\|_F^2 \\
&= \sum_{i=1}^t \|\mathbf{U}_i^l \mathbf{\Sigma}_i^l \sum_{j=1}^t (\mathbf{V}_i^{l\top} \mathbf{V}_j^l) \mathbf{V}_j^{l\top} \tau_j^l - \mathbf{U}_i^l \mathbf{\Sigma}_i^l \mathbf{V}_i^{l\top} \tau_i^l\|_F^2 \\
&= \sum_{i=1}^t \|\mathbf{U}_i^l \mathbf{\Sigma}_i^l \mathbf{V}_i^{l\top} \tau_i^l - \mathbf{U}_i^l \mathbf{\Sigma}_i^l \mathbf{V}_i^{l\top} \tau_i^l\|_F^2 = 0.
\end{aligned} \tag{21}$$

Worst case (Fully collinear features). Conversely, we consider the opposite extreme where all task features share an identical set of right singular vectors, i.e., $\mathbf{V}_i^l = \mathbf{V}^l$ for all $i \in \{1, \ldots, t\}$. This scenario represents a state of maximum interference, where tasks compete for the same parameter directions. Under this highly overlapping condition, the batch optimal solution degenerates into an adaptive weighted average within the shared subspace:

$$\begin{aligned}
\tau_{t,\text{worst}}^{l*} &= \left(\sum_{i=1}^t \mathbf{V}^l (\mathbf{\Sigma}_i^l)^2 \mathbf{V}^{l\top} \right)^\dagger \sum_{i=1}^t \mathbf{V}^l (\mathbf{\Sigma}_i^l)^2 \mathbf{V}^{l\top} \tau_i^l \\
&= \mathbf{V}^l \left(\sum_{j=1}^t (\mathbf{\Sigma}_j^l)^2 \right)^\dagger \mathbf{V}^{l\top} \sum_{i=1}^t \mathbf{V}^l (\mathbf{\Sigma}_i^l)^2 \mathbf{V}^{l\top} \tau_i^l \\
&= \sum_{i=1}^t \mathbf{V}^l \left[\left(\sum_{j=1}^t (\mathbf{\Sigma}_j^l)^2 \right)^\dagger (\mathbf{\Sigma}_i^l)^2 \right] \mathbf{V}^{l\top} \tau_i^l.
\end{aligned} \tag{22}$$

Here, the term $\mathbf{V}^l [(\sum_{j=1}^t (\mathbf{\Sigma}_j^l)^2)^\dagger (\mathbf{\Sigma}_i^l)^2] \mathbf{V}^{l\top}$ acts as a normalized importance weight in the spectral domain. Unlike the ideal case, feature drift becomes inevitable due to the structural overlap of task features. However, our recursive formulation ensures that the merged vector τ_t^{l*} is not a naive arithmetic average, but an optimal synthesis that prioritizes task components with higher spectral energy. By adaptively weighting updates according to their respective singular values, LPM ensures that the parameter directions most critical to historical knowledge are preserved with high fidelity, even under severe collinearity. This mechanism effectively minimizes the cumulative reconstruction error of task features within the shared parameter subspace, thereby mitigating catastrophic interference and safeguarding the structural integrity of the model's reasoning capabilities.

These cases theoretically show that the cumulatively optimized τ_t^{l*} strikes an intrinsic balance between task-specific specialization and shared structural patterns. By locally projecting independent transformations within orthogonal subspaces while robustly synthesizing overlapping features, LPM yields a theoretically-grounded consolidated vector that encapsulates the essence of the task stream.

A.3 Implementation Details

On the UCIT benchmark, all methods are trained for 1 epoch with a batch size of 16. The learning rates for LLaVA and InternVL are set to 2e-4 and 1e-4, respectively. On the MLLM-DCL benchmark, the batch size is reduced to 8, with learning rates of 2e-5 for LLaVA and 1e-4 for InternVL. Training epochs for this benchmark are task-specific: 3 epochs for the Medicine task, 2 epochs for the Science task, and 1 epoch for all remaining tasks.

Table 4: Results on the MLLM-DCL benchmark using LLaVA-1.5-7B and InternVL-Chat-7B models.

	Method	Venue	RS	Med	AD	Sci	Fin	FAA (↑)	CAA (↑)	FFM (↓)
LLaVA-1.5-7B	Zero-shot	–	32.29	28.28	15.59	35.55	62.56	–	–	–
	LoRA-FT Hu et al. (2022)	ICLR'22	74.13	47.63	33.53	41.42	89.66	57.27	63.23	11.43
	O-LoRA Wang et al. (2023a)	EMNLP'23	74.79	39.78	30.68	39.11	83.34	53.54	59.34	7.03
	MoELoRA Chen et al. (2024a)	NeurIPS'24	76.96	47.09	31.19	41.96	89.74	57.39	63.11	11.65
	ModalPrompt Zeng et al. (2025)	EMNLP'25	47.92	29.10	13.35	32.77	89.03	42.44	49.75	30.52
	CL-MoE Huai et al. (2025)	CVPR'25	73.05	48.00	30.27	41.48	**90.11**	56.58	62.81	12.74
	HiDe-LLaVA Guo et al. (2025a)	ACL'25	70.38	38.58	27.60	35.11	60.90	46.51	58.39	11.81
	SEFE Chen et al. (2025)	ICML'25	76.38	47.78	40.16	45.64	89.50	59.89	65.13	7.98
	MAny*	**Ours**	79.58	**59.02**	**49.56**	**49.53**	89.62	**65.46** (+5.64)	**66.82** (+1.69)	-1.43
	MAny	**Ours**	**79.72**	58.85	47.63	49.39	89.43	65.00 (+5.18)	66.43 (+1.30)	**-1.61**
InternVL-Chat-7B	Zero-shot	–	31.16	29.81	14.06	33.93	64.32	–	–	–
	LoRA-FT Hu et al. (2022)	ICLR'22	77.39	53.32	35.12	43.50	91.05	60.08	67.35	11.58
	O-LoRA Wang et al. (2023a)	EMNLP'23	76.93	40.32	29.55	34.61	72.39	50.76	63.14	12.46
	MoELoRA Chen et al. (2024a)	NeurIPS'24	74.41	53.89	39.67	43.46	**91.27**	60.54	67.35	10.24
	ModalPrompt Zeng et al. (2025)	EMNLP'25	58.19	28.76	13.77	34.41	65.32	40.09	41.91	0.07
	CL-MoE Huai et al. (2025)	CVPR'25	68.10	47.26	30.92	40.55	89.40	55.25	63.20	15.01
	HiDe-LLaVA Guo et al. (2025a)	ACL'25	77.73	58.10	35.87	**45.55**	90.14	61.48	66.77	5.81
	SEFE Chen et al. (2025)	ICML'25	78.57	55.09	45.27	44.88	90.29	62.82	67.85	8.33
	MAny*	**Ours**	**81.25**	**65.44**	**50.84**	45.43	90.55	**66.70** (+3.88)	**69.18** (+1.33)	-0.26
	MAny	**Ours**	**81.25**	65.13	50.28	44.99	90.55	66.44 (+3.62)	68.90 (+1.05)	**-0.56**

A.4 Detailed Results on the MLLM-DCL Benchmark

The MLLM-DCL benchmark focuses on continual instruction tuning for MLLMs across highly specialized vertical domains, where significant cross-task distribution shifts pose stringent challenges for mitigating catastrophic forgetting. As summarized in Table 4, our MAny framework consistently establishes a new SOTA across both tested backbones, achieving unprecedented anti-forgetting capabilities that substantially outperform all existing baselines. On LLaVA-1.5-7B, MAny outperforms the prior strongest baseline, SEFE Chen et al. (2025), by an absolute margin of +5.18% in FAA. More notably, it achieves a negative FFM of -1.61%—the only method across all competitors to do so—indicating that our framework not only effectively eliminates catastrophic forgetting, but successfully enables positive backward transfer to enhance old task performance during sequential learning. This consistent superiority extends to InternVL-Chat-7B, where MAny surpasses SEFE by +3.62% in FAA and delivers an ultra-low FFM of -0.56%, in stark contrast to the prominent forgetting exhibited by SEFE (8.33%) and HiDe-LLaVA (5.81%). These results validate that our decoupled dual-merging design effectively resolves the pervasive plasticity-stability dilemma: whereas prior works either suffer from severe forgetting when pursuing plasticity (e.g., LoRA-FT and CL-MoE with FFM > 10%) or drastically sacrifice new task performance to maintain stability (e.g., ModalPrompt, whose FAA drops by over 20% compared to SOTA), MAny excels at both. Furthermore, our memory-efficient variant, MAny*, retains the full model's exceptional capabilities, achieving even slightly higher FAA scores (65.46% on LLaVA-1.5-7B and 66.70% on InternVL-Chat-7B).

A.5 The Performance Gap and Synergy between Perceptual and Reasoning Merging

Table 5: Under the two settings of the projector corresponding to the task and the merged projector of CPM, the impact of using the corresponding task's LoRA and the LPM's Merged LoRA and the last task LoRA using LLaVA-1.5-7B.

Projector Setting	**LoRA Setting**	**UCIT**						**Metrics**		
		Img-R	Arxiv	VizCap	Icon	CLEVR	F30k	FAA (↑)	CAA (↑)	FFM (↓)
Task Projector	Task LoRA (Oracle)	**91.10**	**91.77**	**60.45**	76.17	**75.10**	57.80	**75.40**	-	-
	Merged LoRA	88.83	89.43	56.29	75.43	58.57	56.77	70.89	**82.05**	**5.92**
	Final Task LoRA	81.20	76.40	46.49	**78.63**	74.70	**57.99**	69.24	77.80	10.08
Merged Projector	Task LoRA (Oracle)	**91.63**	91.10	**60.79**	**83.07**	**77.17**	58.13	**76.98**	**84.01**	**0.23**
	Merged LoRA	90.87	**91.97**	60.48	80.20	67.37	56.09	74.50	83.24	0.37
	Final Task LoRA	78.43	76.73	46.03	79.30	73.50	**58.28**	68.71	75.65	10.45

To deeply evaluate the knowledge consolidation capability of LPM, we compare it against two extreme settings in Table 5: (1) **Task LoRA**, which serves as an ideal Oracle upper bound requiring

ground-truth task IDs at test time, and (2) **Final Task LoRA**, representing the lower bound where only the LoRA weights from the last trained task are used for inference. Under the CPM setting, LPM (FAA 74.50%, FFM 0.37%) demonstrates exceptional performance retention, closely approximating the Oracle Task LoRA (FAA 76.98%, FFM 0.23%) while vastly outperforming the naive Final LoRA integration. This establishes that the recursive least squares parameter merging mechanism in LPM can almost perfectly parse and preserve the structural reasoning weights of sequential tasks, even under strict exemplar-free and task-agnostic constraints.

A.6 Modality Choices for CPM Prototype Construction

Table 6: Comparison of modality choices for CPM prototype construction.

Modality	**UCIT**						**Metrics**		
	Img-R	Arxiv	VizCap	Icon	CLEVR	F30k	FAA (↑)	CAA (↑)	FFM (↓)
Text	**91.20**	**92.67**	57.81	**80.70**	67.67	**56.32**	74.40	**83.48**	0.70
Text + Image	91.03	92.63	58.89	80.23	**68.57**	56.29	**74.61**	83.46	0.38
Image (Ours)	90.87	91.97	**60.48**	80.20	67.37	56.09	74.50	83.24	**0.37**

In Section 4.2, we opted to construct the task-representative prototypes μ_t using solely visual features. We ablate this design choice in Table 6 by comparing prototypes built from text-only instructions, fused text-image features, and image-only features. To extract these multimodal embeddings, we consistently adopt CLIP-L/14-336 Radford et al. (2021). The results indicate highly consistent overall performance across all three modality choices, with FAA ranging tightly between 74.4% and 74.6% alongside minimal forgetting. This implies that the tasks exhibit sufficient distinctiveness in both their visual distributions and textual instructions. However, the image-only approach presents a distinct computational advantage: during inference, it completely frees the model from forwarding textual instructions through an additional text encoder to execute routing decisions. Consequently, our vision-centric design maximizes computational efficiency without sacrificing routing accuracy, ultimately yielding the optimal FFM of 0.37%.

A.7 Sensitivity of the Energy Threshold γ

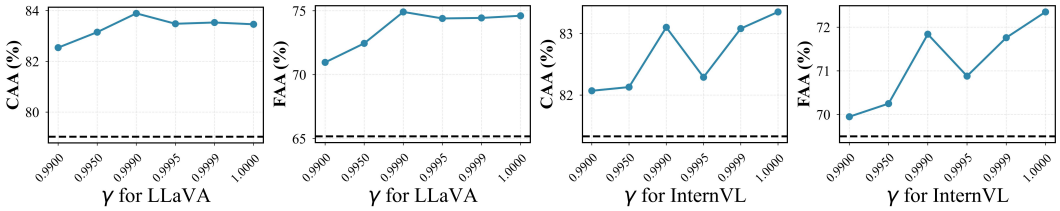


Figure 5: Impact of γ. Dashed lines denote the SOTA performance of all baselines.

Figure 5 evaluates MAny* across varying energy thresholds, where dashed lines represent the SOTA. MAny* consistently outperforms the SOTA across the entire range. Notably, a performance peak emerges at $\gamma = 0.999$, which surprisingly surpasses the uncompressed full-rank setting ($\gamma = 1.0$). This indicates that truncating long-tail singular values acts as an implicit regularizer by filtering noisy directions that exacerbate cross-task interference. This optimal configuration reduces storage overhead to merely 1/27 of the original requirement while achieving superior accuracy and stability. We therefore set $\gamma = 0.999$ as the default for MAny*.