REVISITING CONTINUAL LEARNING FOR MLLMS: A SIMPLE APPROACH REVEALS HIDDEN LIMITATIONS IN EXISTING BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) have emerged as a promising foundation for AI systems, and continual learning has attracted increasing research attention to enable MLLMs to adapt to evolving environments. Despite the proliferation of related work, a critical question remains: do current benchmarks adequately capture the task relationships needed for effective knowledge sharing in MLLM continual learning? In this paper, we take an initial step toward answering this question. First, we show that a small modification to one existing MoE-LoRA method—removing cross terms from its fused LoRA readout—is sufficient to substantially outperform state-of-the-art methods. To understand this unexpected improvement, we analyze the routing behavior and find that task identity can be readily inferred from the input, allowing models to exploit task-level shortcuts and revealing a potential benchmark-hacking issue. To systematically examine this issue, we introduce an entropy-parameterized framework that recasts existing benchmarks into variants with controllable task overlap. As task overlap increases, performance can improve even as task identity becomes harder to infer, while cross-task knowledge sharing becomes increasingly beneficial. These findings suggest that existing benchmarks may overstate the benefit of task specialization while underestimating cross-task knowledge transfer.

1 INTRODUCTION

Multimodal large language models (MLLMs) pair large language models with vision encoders and follow instructions about images. After deployment, adaptation typically arrives as a stream: new vision–language tasks come one after another, and the model is fine-tuned on each in turn. This setting, known as continual instruction tuning, requires models to acquire new capabilities while preserving previous knowledge, a challenge that neural networks notoriously struggle to overcome; the failure mode is catastrophic forgetting (McCloskey & Cohen, 1989; Kirkpatrick et al., 2017). A representative approach is MoE-LoRA (Chen et al., 2024), which freezes the backbone and allocates a dedicated LoRA expert to each task. Each LoRA expert is implemented as a low-rank update composed of two trainable matrices (Hu et al., 2021), optimized for the corresponding task. The expert pool grows as tasks arrive, and at inference time the model must decide which experts to use, although the identity of the current task is unavailable. However, existing MLLM-CL benchmarks are typically constructed from distinct sequential tasks, which may expose task-specific cues in the input. This motivates us to examine how readily task identity can be inferred under existing benchmarks, and whether their current evaluation settings adequately reflect the challenges of continual learning.

We revisit the expert composition mechanism in MoE-LoRA-based MLLM continual instruction tuning and find that expert composition does not always lead to effective knowledge sharing. Removing the cross-expert interactions introduced by LoRA parameter merging substantially improves performance on existing benchmarks, suggesting that current evaluations may favor expert specialization over knowledge integration. Further analysis reveals that these benchmarks operate in a highly separable regime, where task identity can be inferred from input samples and used as a shortcut for expert selection. Consequently, existing benchmarks mainly evaluate isolated task adaptation rather than the broader trade-off between expert specialization and knowledge sharing.

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.

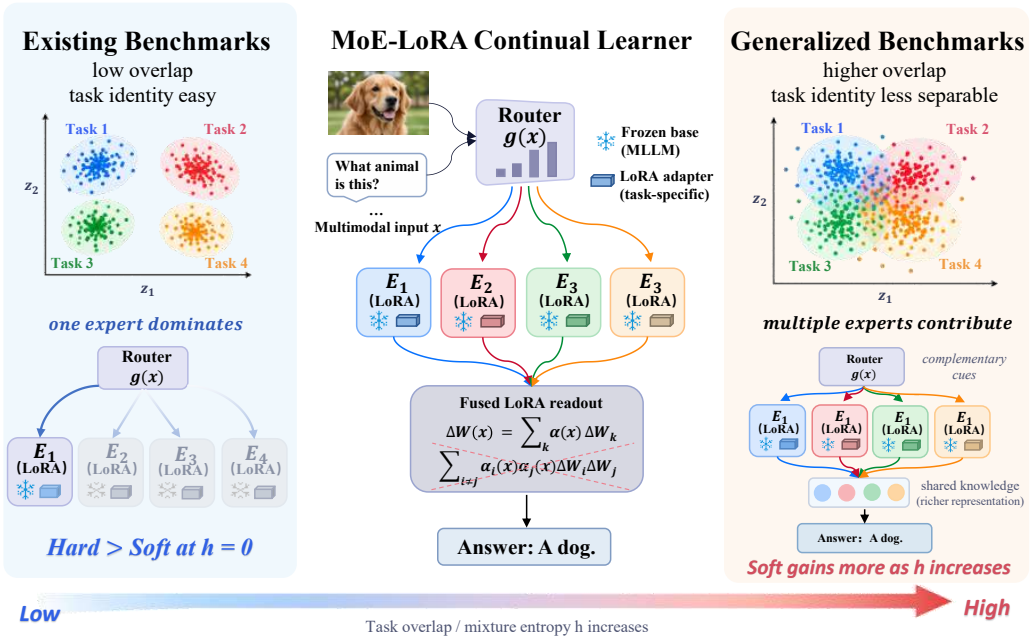


Figure 1: Overview of our analysis on task overlap and expert utilization. Existing MLLM-CL benchmarks exhibit highly separable tasks, which can favor expert isolation. By progressively increasing task overlap, our generalized setting better reflects the need for cross-task knowledge transfer and sharing.

To move beyond this extreme regime, we introduce an entropy-parameterized framework that generalizes existing benchmarks by continuously increasing task overlap. Across two benchmarks and multiple overlap levels, we observe a consistent transition: as task overlap increases, the advantage of strict expert isolation gradually diminishes, while combining knowledge from multiple experts becomes increasingly effective and eventually surpasses single-expert selection. Notably, performance can even improve as task identity becomes harder to infer, suggesting that accurate task recognition is not a prerequisite for effective continual learning when knowledge can be shared across tasks. These results suggest that existing MLLM-CL benchmarks represent an overly separable and restrictive regime that favors expert isolation. More generalized task distributions are needed to properly evaluate knowledge transfer and sharing across tasks.

2 Related Work

Methods for MLLM continual learning. Sequential fine-tuning of an MLLM erodes both old-task knowledge and the vision–language alignment it depends on, and existing methods draw on the classical anti-forgetting strategies: regularization constrains updates to parameters important for old tasks (Kirkpatrick et al., 2017); replay rehearses on stored exemplars of past data (Rebuffi et al., 2017); and structural isolation gives each task dedicated capacity, e.g., masks or added modules (Rusu et al., 2016; Mallya & Lazebnik, 2018). In the multimodal setting, prompt-based methods freeze the backbone and store task knowledge in a learnable prompt pool retrieved at input time: L2P selects prompts by input similarity (Wang et al., 2022b), DualPrompt separates general from expert prompts (Wang et al., 2022a), CODA-Prompt composes prompts with attention (Smith et al., 2023), and ModalPrompt extends the design to vision–language inputs (Zeng et al., 2025); these methods are cheap at inference but limited in representational capacity. He et al. bring the classical strategies to multimodal instruction tuning and find data replay effective, at the cost of storing cross-modal instruction data (He et al., 2026). Cross-modal regularization constrains the relative geometry of visual and textual embeddings so that learning one modality does not corrupt the other (Liu et al., 2025).

Finally, parameter-efficient adaptation freezes the pretrained backbone and updates only a small set of newly introduced or designated parameters, thereby reducing interference with previously learned knowledge. Representative methods include LoRA (Hu et al., 2021), O-LoRA (Wang et al., 2023), which keeps the LoRA subspaces of different tasks orthogonal in the text-only setting. HiDe-LLaVA (Guo et al., 2025a) extends this direction to MLLM continual learning with a growing pool of task-specific LoRA experts. At inference time, it combines the trained experts through a fused LoRA readout conditioned on task-related similarity. Overall, existing methods largely inherit their anti-forgetting mechanisms from classical continual learning (Huo & Tang, 2025; Liu et al., 2025). Comparatively less attention has been paid to the evaluation settings of current MLLM-CL benchmarks. In this work, we shift the focus from method design to the evaluation setting itself and examine whether existing benchmarks provide a sufficiently general testbed for continual learning.

Benchmarks for MLLM continual learning. CLiMB (Srinivasan et al., 2022) studies continual learning over an early sequence of vision–language tasks, including visual entailment, scene recognition, question answering, and captioning, and evaluates both forgetting and forward transfer. For continual instruction tuning of MLLMs, He et al. (He et al., 2026) establish benchmark sequences of captioning and question-answering tasks on BLIP2 and InstructBLIP. CoIN (Chen et al., 2024) further introduces eight tasks spanning question answering, classification, and grounding to increase task diversity. Guo et al. (Guo et al., 2025a) later identify information leakage in CoIN, caused by overlap between several datasets and LLaVA pretraining data, and propose UCIT with six tasks screened to reduce such correlations.

Despite these advances, existing benchmarks still organize continual learning as sequences of distinct and largely separated tasks. We show that this structure makes task identity highly recoverable from the input and places current benchmarks in an extreme low-overlap regime. Such a regime can disproportionately favor task-specific expert isolation, while providing limited pressure for knowledge transfer and sharing across tasks. To study a more generalized setting, we retain the original tasks, data, and evaluation pipelines, but transform each benchmark into a spectrum of task-overlap conditions controlled by mixture entropy. This formulation allows us to systematically examine continual learning beyond sharply separated task boundaries.

3 Rethinking Knowledge Transfer in MLLM Continual Instruction Tuning

3.1 Preliminaries: MoE-LoRA continual instruction tuning

Given a sequence of instruction-tuning tasks $\{\mathcal{D}_1, \ldots, \mathcal{D}_T\}$, continual instruction tuning aims to adapt a pretrained multimodal large language model to new tasks while preserving previously learned knowledge. Parameter-efficient approaches address this challenge by freezing the pretrained backbone and introducing lightweight task-specific parameters. Instead of maintaining a single adaptation module, MoE-LoRA-based methods allocate different LoRA experts to different tasks and determine how these experts contribute during inference.

Following LoRA (Hu et al., 2021) and MoE-LoRA-based continual instruction tuning methods (Chen et al., 2024), a linear projection layer with pretrained weight $\mathbf{W}$ can be abstracted as

$$y = \mathbf{W}\mathbf{x} + \frac{\alpha}{r} \sum_{i=1}^{T} w_i(\mathbf{x}) \mathbf{B}_i \mathbf{A}_i \mathbf{x}, \tag{1}$$

where $(\mathbf{A}_i, \mathbf{B}_i)$ denotes the low-rank factors of the i-th LoRA expert, $w_i(\mathbf{x})$ represents the routing weight assigned to the corresponding expert, and α/r is the LoRA scaling factor.

The expert weighting strategy directly influences the balance between expert specialization and knowledge sharing. Selecting a single expert emphasizes task-specific adaptation, while combining multiple experts allows information from different tasks to be jointly utilized. In MLLM continual instruction tuning, the expert weighting strategy determines how knowledge from different tasks is utilized during inference. Understanding the behavior of different expert composition strategies is therefore essential for evaluating MoE-LoRA-based approaches.

3.2 Revisiting Expert Composition in MoE-LoRA

HiDe-LLaVA (Guo et al., 2025a) instantiates this paradigm by maintaining task-specific LoRA experts and determining expert contributions based on the similarity between the input representation and multimodal task anchors. It adopts different expert combination strategies across layers. Specifically, the top layer aggregates expert outputs using softmax routing weights:

$$w_i = \frac{\exp(s_i/\tau)}{\sum_j \exp(s_j/\tau)}, \tag{2}$$

where $\mathbf{s}_i$ denotes the similarity between the input representation and the i-th task anchor, and τ is the temperature parameter controlling the sharpness of the routing distribution.

For the remaining layers, multiple LoRA experts are combined by directly merging their parameter updates. Despite this design, we observe that removing the cross-expert interactions introduced by direct parameter merging substantially improves performance on existing benchmarks (Table 1). We decompose the fused LoRA update into individual expert contributions and cross-expert interactions:

$$\Delta\mathbf{W} = \left(\sum_i \mathbf{B}_i\right)\left(\sum_j \mathbf{A}_j\right) = \underbrace{\sum_i \mathbf{B}_i\mathbf{A}_i}_{\text{individual expert contributions}} + \underbrace{\sum_{i\neq j} \mathbf{B}_i\mathbf{A}_j}_{\text{cross-expert interactions}}. \tag{3}$$

The first term preserves the original contribution of each expert, whereas the second term emerges from factor-level composition and introduces interactions between different experts.

3.3 Diagnosing the Factors of Improvement

The effectiveness of removing cross-expert interactions raises a fundamental question: why does expert isolation become more effective than expert composition on existing benchmarks?

We conjecture that existing benchmarks operate in an overly separable regime, where each task corresponds to a distinct capability and data distribution. Under such settings, task identity itself can serve as a shortcut for expert selection, allowing task-specific experts to be accurately activated while reducing the need for knowledge sharing.

To validate this hypothesis, we first measure whether task identity can be inferred directly from input samples using the learned router. On UCIT, the router achieves 100% top-1 routing accuracy across all six tasks, evaluated on 100 sampled test inputs per task, demonstrating that existing task partitions contain strong distributional cues. Consequently, models can identify the corresponding expert through task recognition alone, reducing the need for cross-task knowledge sharing.

We further examine whether this task separability explains the preference for expert specialization over knowledge sharing. Specifically, we compare single-expert selection and multi-expert composition during inference. Surprisingly, hard routing consistently outperforms soft routing on both existing benchmarks (Tables 1 and 2), indicating that selecting a single expert is more effective than composing multiple experts under the current task settings.

4 Benchmark Reformulation: An Entropy-Parameterized Task-Overlap Spectrum

4.1 Motivation

Existing continual instruction tuning benchmarks typically organize learning as sequences of isolated tasks, where each stage focuses on a single capability with its corresponding data distribution. As shown in Section 3.3, such a design can result in an overly separable task structure, where task identity becomes a strong cue for expert selection and favors task-specific specialization over knowledge sharing.

However, highly separable task arrangements represents only one extreme of the broader spectrum of possible task relationships. In realistic continual learning scenarios, tasks may exhibit different degrees of overlap, leading to different trade-offs between expert specialization and knowledge sharing. Consequently, a fixed benchmark construction cannot fully characterize the behavior of continual learning methods across varying task relationships.

To address this limitation, we seek to construct a controllable task-overlap spectrum from existing benchmarks, where continual instruction tuning can be evaluated under different similarity regimes. Such a spectrum should satisfy the following desiderata:

(i) a single, interpretable scalar parameter, allowing model behavior to be analyzed along a single axis;

(ii) comparability across benchmarks with different numbers of tasks and task configurations;

(iii) meaningful endpoints, with at least one endpoint exactly recovering the original benchmark;

(iv) no change to the task datasets, the model, or the training pipeline, so that observed behavioral changes can be attributed to the controlled variation in task similarity.

4.2 Entropy-parameterized Similarity Framework

To enable such controllable similarity variation, we propose an **entropy-parameterized similarity control framework**. The entropy parameter controls the degree of task mixing in each training stage. Specifically, for stage t, we construct a training distribution by interpolating between the original task-specific distribution and a uniform mixture over all C tasks:

$$p_t = 1 - \lambda + \frac{\lambda}{C}, \qquad p_{i \neq t} = \frac{\lambda}{C}, \tag{4}$$

where $\lambda \in [0, 1]$ controls the amount of cross-task mixing. Rather than using λ directly, we measure the resulting task mixture using its normalized entropy:

$$h = \frac{-\sum_i p_i \ln p_i}{\ln C} \in [0, 1]. \tag{5}$$

For a desired entropy level h, λ is obtained by solving Eq. 5. At $h = 0$, each stage contains only its original task, recovering the original benchmark exactly. At $h = 1$, all stages adopt the same uniform mixture over the C tasks, representing the maximum task overlap.

The framework converts a single benchmark into a *spectrum* of settings indexed by h, ranging from disjoint task distributions to fully mixed ones. The entropy normalization places benchmarks with different numbers of tasks (e.g., UCIT with C=6 and CoIN with C=8) on a shared $[0, 1]$ scale, allowing their behaviors under different similarity regimes to be compared directly. Given an existing continual-instruction-tuning benchmark, the framework requires only three steps: specify the task set, invert Eq. 5 to obtain the corresponding λ and generate stage-wise mixtures, then train and evaluate the model using the original pipeline.

The resulting spectrum provides a controlled way to study how continual instruction tuning changes as task similarity varies. As h increases, task boundaries become less distinct, enabling analysis of the transition between expert specialization and knowledge sharing.

The spectrum is an analysis instrument rather than a benchmark proposal: any continual-learning method can be evaluated along h to reveal the similarity regime under which its behavior changes.

5 Experiments

5.1 Experimental Setup

Benchmarks. We evaluate on two standard multimodal continual instruction tuning benchmarks: UCIT (T=6: ImageNet-R, ArxivQA, VizWiz, IconQA, CLEVR-Math, Flickr30k) and CoIN (T=8: ScienceQA, TextVQA, ImageNet, GQA, VizWiz, REC-COCO, VQAv2, OCRVQA). Following

Table 1: UCIT per-task accuracy and Last Average (%). Baseline and Multi-task results are taken from Guo et al. (2025a). † denotes evaluation under the MCITlib pipeline (Guo et al., 2025c), and $*$ denotes our reproduction of the corresponding method.

Method	Image-R	ArxivQA	VizWiz	IconQA	CLEVR	Flickr30k	Last Avg
Fine-tuning	37.63	72.33	43.47	41.70	35.63	57.95	48.12
LwF (Li & Hoiem, 2017)	40.27	75.93	42.76	44.38	37.43	56.34	49.52
EWC (Kirkpatrick et al., 2017)	39.05	77.88	43.24	45.33	39.72	55.94	50.20
L2P (Wang et al., 2022b)	32.73	80.41	43.72	42.16	39.25	52.77	48.51
O-LoRA (Wang et al., 2023)	69.36	82.42	48.64	53.66	42.53	53.52	58.36
MoELoRA (Chen et al., 2024)	49.87	77.63	43.65	46.40	36.47	**58.34**	52.06
HiDe-LLaVA (Guo et al., 2025a)	80.50	89.83	48.78	62.90	47.97	55.15	64.19
SEFE† (Chen et al., 2025)	80.83	78.00	47.01	69.63	65.83	57.92	66.54
DISCO† (Guo et al., 2025b)	**87.43**	**93.07**	46.96	68.13	65.70	56.69	69.66
HiDe-LLaVA* (Guo et al., 2025a)	83.93	90.63	48.51	66.60	61.63	55.32	67.77
Soft	86.23	93.03	47.87	73.47	69.87	57.75	71.37
Hard	87.27	92.00	**59.06**	**75.00**	**69.90**	57.86	**73.51**
Multi-task	90.63	91.30	61.81	73.90	73.60	57.45	74.78

prior work, we report *Last Average*, the mean accuracy over all tasks after the final continual learning stage.

Entropy sweep construction. For each benchmark, we generate task-mixture variants using the entropy-parameterized framework in Section 4.2. For UCIT, we sweep $h \in \{0.25, 0.5, 0.75, 1.0\}$, while $h{=}0$ corresponds to the original benchmark. The corresponding mixing coefficients are $\lambda \in \{0.108, 0.270, 0.496, 1.0\}$, yielding main-task proportions of $\{91.0\%, 77.5\%, 58.6\%, 16.7\%\}$. For CoIN, the same entropy grid corresponds to $\lambda \in \{0.114, 0.282, 0.512, 1.0\}$ and main-task proportions of $\{90.0\%, 75.3\%, 55.2\%, 12.5\%\}$. When a task data pool cannot provide its target proportion, the remaining samples are assigned back to the main task. The resulting stages have entropy values within 0.01 of the target. All training hyperparameters remain identical to the $h{=}0$ baseline; only the mixture composition changes.

Model and training. We use LLaVA-1.5-7B with the MoE-LoRA framework described in Section 3.2. The LoRA rank is set to 48 for UCIT and 64 for CoIN. We use a learning rate of 2×10^{-4}, one training epoch for each stage, an effective batch size of 48, and fp16 precision. Entropy-sweep training follows the same optimization settings, differing only in the stage-wise mixture composition.

Inference modes. We compare three inference-time expert readout strategies. HiDe-LLaVA* reproduces the original HiDe-LLaVA formulation, where LoRA experts are combined through parameter-level fusion in the Remain layers and routed aggregation is applied in the Top layer. *Soft* applies routing-based readout to all layers using softmax routing weights with temperature $\tau = 0.1$, whereas *Hard* selects the expert with the highest routing score. Soft and Hard use the same expert pool and routing scores, differing only in the aggregation strategy: Soft combines multiple experts using the routing weights, while Hard selects a single expert. Both avoid the cross-expert interactions introduced by parameter-level fusion.

5.2 Cross-term-free readout improves performance

Tables 1 and 2 compare different expert composition strategies with existing continual-instruction-tuning methods. Baseline and Multi-task results are those reported by Guo et al. (2025a); SEFE and DISCO, marked †, were evaluated under the MCITlib pipeline (Guo et al., 2025c). The HiDe-LLaVA*, Soft, and Hard variants use the same trained checkpoint and differ only in their inference-time expert readout strategy, allowing us to analyze how cross-expert interactions affect expert composition.

On UCIT (Table 1), Hard achieves a Last Average accuracy of 73.51%, improving over HiDe-LLaVA* by 5.74 percentage points. This gain is obtained solely by modifying the inference-time

Table 2: CoIN per-task accuracy and Last Average (%). Per-task cells follow our task sequence. Baseline and Multi-task results are taken from Guo et al. (2025a), and $*$ denotes our reproduction of HiDe-LLaVA.

Method	SciQA	TextVQA	ImageNet	GQA	VizWiz	REC	VQAv2	OCRVQA	Last Avg
Fine-tuning	57.43	51.39	28.90	50.76	41.88	30.05	53.28	64.78	47.31
LwF	60.71	52.80	30.58	47.07	41.49	36.01	53.43	65.12	48.40
EWC	59.75	51.06	31.88	51.84	42.26	34.96	55.30	64.55	48.95
L2P	70.21	56.25	23.31	58.46	44.21	43.76	62.32	64.11	52.83
O-LoRA	72.56	57.66	62.84	59.14	48.43	58.97	63.21	63.31	60.77
MoELoRA	62.02	52.05	37.21	53.12	43.32	33.22	57.92	65.75	50.58
HiDe-LLaVA	73.20	56.92	69.28	61.33	50.76	59.18	67.12	64.76	63.95
HiDe-LLaVA*	76.30	52.53	74.17	47.03	55.83	58.71	55.23	38.96	57.35
Soft	82.26	57.33	48.23	60.37	61.33	66.87	**68.47**	67.68	64.07
Hard	**84.20**	**58.30**	**77.37**	**61.83**	**61.87**	**69.73**	66.40	**68.18**	**68.48**
Multi-task	82.36	61.27	89.63	59.93	52.51	65.83	65.67	62.03	67.40

readout, without additional training. Compared with published baselines, Hard also improves over HiDe-LLaVA (64.19%) and DISCO (69.66%) by 9.32 and 3.85 percentage points, respectively.

On CoIN (Table 2), Hard reaches a Last Average accuracy of 68.48%, improving over HiDe-LLaVA* by 11.14 percentage points. Notably, removing cross-expert interactions alone raises the Last Average from 57.35% to 64.07%, accounting for 6.72 percentage points of the overall gain.

5.3 Existing benchmarks favor expert isolation

The improvement over HiDe-LLaVA* can be decomposed into two factors. HiDe-LLaVA* and Soft share the same routing weights but differ in whether the Remain layers introduce cross-expert interactions, while Soft and Hard share the same cross-term-free readout and differ only in expert selection. Therefore, the two gaps isolate the effects of cross-term removal and routing, respectively. On UCIT, the $+5.74\%$ percentage-point gain from HiDe-LLaVA* to Hard consists of $+3.60\%$ points from removing cross-expert interactions and $+2.14\%$ points from hard selection. On CoIN, the corresponding gains are $+6.72\%$ and $+4.42\%$ points.

The improvement mainly comes from removing the cross-expert interactions introduced by parameter-level fusion rather than from changing the routing strategy itself. Interestingly, hard routing consistently outperforms soft routing on the original benchmarks, although soft routing explicitly enables knowledge combination across experts. This preference for single-expert selection persists even after removing cross-expert interactions, suggesting that current benchmarks favor expert specialization over expert collaboration: when tasks are highly separable, selecting a task-specific expert is sufficient and cross-expert knowledge sharing becomes less critical.

This observation motivates the entropy-parameterized similarity spectrum, which allows us to study how the balance between expert specialization and knowledge sharing changes as task overlap increases.

5.4 Entropy-Parameterized Similarity Framework

The previous analysis suggests that existing continual instruction tuning benchmarks mainly evaluate models in a task-separable regime, where task-specific expert selection is highly effective. Such a regime makes it difficult to assess when knowledge sharing across experts becomes necessary. To investigate how the balance between expert specialization and knowledge sharing changes with task similarity, our entropy-parameterized framework enables controlled variation of task similarity along a continuous spectrum.

We analyze soft and hard routing across the similarity spectrum induced by our entropy-parameterized framework. Figure 2 shows a systematic transition between the two strategies. At $h = 0$, corresponding to existing benchmarks, hard routing consistently outperforms soft routing on both datasets. As task overlap increases, this advantage gradually disappears and reverses in favor

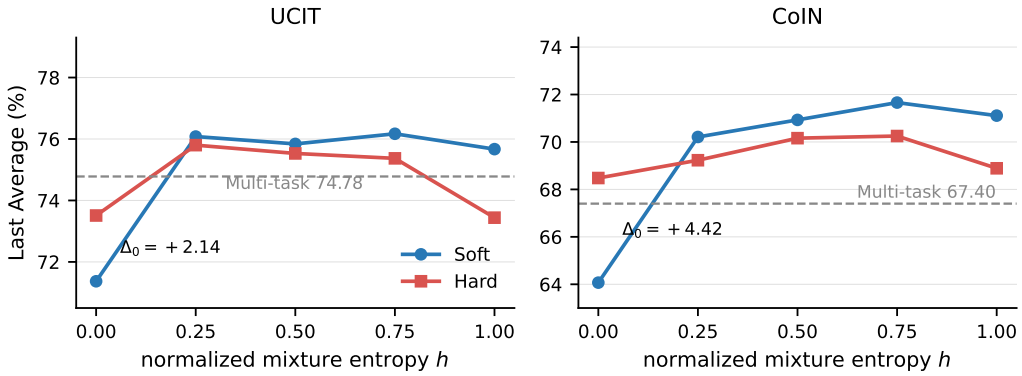


Figure 2: Last Average along the entropy spectrum for the two inference modes of the same checkpoints, on UCIT (left) and CoIN (right); h=0 is the original benchmark, and the dashed lines mark the multi-task references reported by Guo et al. (2025a) for UCIT and CoIN. Hard has an advantage only at h=0. With increasing task overlap, Soft surpasses Hard at h=0.25 and remains superior at higher entropy levels, with the largest gap at h=1.

of soft routing. On UCIT, the performance gap changes from a $+2.14\%$ percentage-point advantage for hard routing at $h = 0$ to a -2.23% point gap at $h = 1$. CoIN exhibits the same transition, where the initial $+4.42\%$ point advantage of hard routing is replaced by a consistent gain from soft routing in higher-similarity regimes.

This crossover reflects a fundamental trade-off between expert specialization and knowledge sharing. In low-similarity regimes, each expert primarily captures task-specific knowledge, and task identity provides sufficient cues for selecting the relevant expert. Hard routing therefore benefits from specialization while avoiding interference from unrelated experts. As task similarity increases, however, experts encounter increasingly overlapping distributions, making complementary knowledge across experts more valuable. Soft routing can exploit such shared information through expert combination, leading to improved performance.

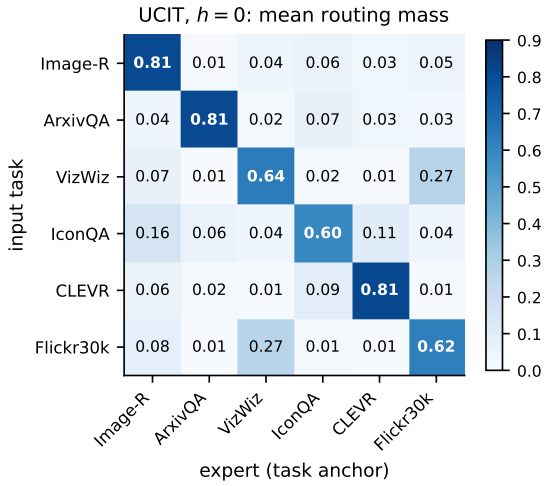
These results reveal a limitation of existing continual instruction tuning benchmarks. Their highly separable task construction favors task-specific expert selection while underestimating the role of knowledge sharing. By making task similarity controllable, our entropy-parameterized framework exposes this evaluation bias and enables systematic analysis across diverse task relationships.

Finally, we examine whether the entropy parameter induces the intended change in task similarity. As shown in Fig. 3, increasing h increases the mean pairwise anchor similarity while reducing top-1 routing accuracy, indicating a transition from task-separable to task-overlapping regimes. The routing mass distribution further illustrates the original task separability: at $h = 0$, UCIT shows nearly perfect expert alignment, with each task assigning dominant mass to its corresponding expert. These results verify that the proposed framework provides a controllable spectrum of task relationships for evaluating continual instruction tuning under diverse similarity regimes.

5.5 Discussion

For **method design**, our findings suggest that expert specialization and knowledge sharing should be viewed as complementary strategies whose effectiveness depends on task relationships. Existing MoE-based continual learning approaches often assume that composing multiple experts naturally facilitates knowledge transfer. However, chances are when task distributions are highly separable, models can rely on task-specific cues to retrieve specialized experts, leaving limited opportunities for cross-task knowledge sharing. Under such conditions, expert interactions may provide little benefit and can even introduce unnecessary interference.

Our analysis further highlights that **benchmark design** plays an important role in evaluating continual instruction tuning methods. Strong performance on existing task sequences may reflect effective expert specialization and retrieval rather than genuine knowledge transfer across tasks. Therefore,



	UCIT		CoIN	
h	sim	ACC	sim	ACC
0	0.57	100.00	0.66	78.38
0.25	0.61	100.00	0.71	74.63
0.5	0.71	100.00	0.78	64.50
0.75	0.82	100.00	0.87	50.13
1.0	1.00	16.67	0.95	12.88

Figure 3: Validation of the entropy-parameterized framework. **Left:** Mean routing mass on UCIT at h=0 with τ=0.1, where rows denote input tasks and columns denote experts. Each row is averaged over 100 sampled test inputs and sums to one. **Right:** Mean pairwise anchor similarity (sim) and top-1 anchor-routing accuracy (ACC, %) along the entropy spectrum. Chance accuracy is $1/T$ (16.67% for UCIT and 12.50% for CoIN).

future evaluations should consider diverse task relationships and examine whether models can adaptively balance specialization and knowledge sharing. The current form is similar to Task Incremental Learning (TIL) (Kirkpatrick et al., 2017), and our work moves it towards Generalized TIL. This is also in line with the trend of related learning paradigms, including CIL (Rebuffi et al., 2017) to GCIL (Mi et al., 2020). While our entropy-parameterized framework provides a controlled way to study this trade-off, task overlap represents only one dimension of continual learning scenarios, leaving other factors such as task order and data imbalance for future exploration. Additionally, the current formulation of organizing existing VQA benchmarks into an MLLM-CL benchmark needs to be adjusted. While existing benchmarks provide a useful starting point, real-world needs may be better addressed by data collected from real environments.

6 Conclusion

Continual instruction tuning enables multimodal large language models to adapt to evolving tasks, but whether existing benchmarks properly evaluate knowledge transfer across tasks remains unclear. In this work, we reveal that current benchmarks are largely dominated by a task-separable regime, where task identity can be inferred from the input and task-specific expert selection is sufficient. This explains why removing cross-expert interactions from MoE-LoRA based models can substantially improve performance: under such settings, expert collaboration provides limited benefit and may introduce unnecessary interactions.

To systematically study this issue, we introduce an entropy-parameterized similarity framework that enables controlled variation of task similarity over a continuous spectrum. Our analysis shows that the preferred expert composition strategy depends on the similarity regime: hard routing is effective when tasks are highly separable, whereas soft routing becomes increasingly beneficial as cross-task knowledge sharing becomes necessary. These findings suggest that existing continual instruction tuning benchmarks may overemphasize expert specialization while underestimating knowledge transfer across tasks. We hope this work encourages the development of more general and comprehensive continual learning benchmarks that better reflect the diverse challenges of real-world learning, where models must not only acquire new knowledge but also effectively reuse and integrate previously learned knowledge.

AI USE STATEMENT

In this work, we used generative AI tools for limited purposes, including editing the manuscript to improve readability, grammar, and overall presentation. Generative AI tools were not used for developing the proposed method, conducting experiments, analyzing results, or generating scientific claims. All AI-assisted edits were carefully reviewed and revised by the authors to ensure that the final manuscript accurately reflects our original ideas, contributions, and findings. The authors take full responsibility for the content of this work, including all text, claims, and experimental results.

REPRODUCIBILITY STATEMENT

All experiments in this work are conducted on publicly available benchmarks (UCIT and CoIN) using the open-source LLaVA-1.5-7B model. We build our experiments upon the MoE-LoRA implementation of HiDe-LLaVA and provide all necessary configurations, including model settings, LoRA parameters, optimization details, and evaluation procedures. For the proposed entropy-parameterized framework, we provide the complete task mixture generation procedure and corresponding configurations. The reported expert composition results are obtained from the same trained checkpoints with different inference-time readout strategies, without additional training. Source code, training and evaluation scripts, and configuration files will be released to facilitate reproduction of all reported results.

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. *Advances in neural information processing systems*, 37:57817–57840, 2024.

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ho Shing Ip, and Sam Kwong. Sefe: Superficial and essential forgetting eliminator for multimodal continual instruction tuning. *arXiv preprint arXiv:2505.02486*, 2025.

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, Fanhu Zeng, Fei Zhu, Wenzhuo Liu, Da-Han Wang, Jian Xu, Xu-Yao Zhang, and Cheng-Lin Liu. Federated continual instruction tuning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1325–1335. IEEE, 2025b.

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*, 2025c.

Jinghan He, Haiyun Guo, Kuan Zhu, Ming Tang, and Jinqiao Wang. Continual instruction tuning for large multimodal models. *IEEE Transactions on Image Processing*, 2026.

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. *arXiv preprint arXiv:2106.09685*, 2021.

Yukang Huo and Hao Tang. When continue learning meets multimodal large language model: A survey. *arXiv preprint arXiv:2503.01887*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

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

Yuyang Liu, Qiuhe Hong, Linlan Huang, Alexandra Gomez-Villa, Dipam Goswami, Tiantian Peng, 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.

Arun Mallya and Svetlana Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7765–7773. IEEE, 2018.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Fei Mi, Lingjing Kong, Tao Lin, Kaicheng Yu, and Boi Faltings. Generalized class incremental learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 970–974. IEEE, 2020.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5533–5542. IEEE, 2017.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11909–11919. IEEE, 2023.

Tejas Srinivasan, Ting-Yun Chang, Leticia Pinto Alva, Georgios Chochlakis, Mohammad Rostami, and Jesse Thomason. Climb: A continual learning benchmark for vision-and-language tasks. *Advances in Neural Information Processing Systems*, 35:29440–29453, 2022.

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, 2023.

Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pp. 631–648. Springer, 2022a.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 139–149. IEEE, 2022b.

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.

A Per-task accuracy along the entropy spectrum

Tables 3 and 4 report per-task and average accuracy along the entropy spectrum of Section 5.4. The $h{=}0$ column repeats the per-task entries of Tables 1 and 2.

Table 3: UCIT per-task accuracy (%) along the entropy spectrum (h=0 is the original benchmark).

Task	h=0	0.25	0.5	0.75	1.0
Soft					
ImageNet-R	86.23	87.40	88.60	90.03	90.27
ArxivQA	93.03	94.13	93.83	94.40	93.40
VizWiz	47.87	65.92	66.06	64.57	65.10
IconQA	73.47	79.67	79.43	79.47	81.03
CLEVR-Math	69.87	71.27	70.10	71.67	68.43
Flickr30k	57.75	58.10	57.04	56.91	55.76
Hard					
ImageNet-R	87.27	86.90	87.63	89.13	88.87
ArxivQA	92.00	93.87	93.23	93.83	91.20
VizWiz	59.06	65.90	65.67	60.64	63.09
IconQA	75.00	79.63	80.47	79.70	79.43
CLEVR-Math	69.90	70.17	68.40	70.97	63.63
Flickr30k	57.86	58.35	57.77	57.94	54.41

Table 4: CoIN per-task accuracy (%) along the entropy spectrum (h=0 is the original benchmark).

Task	h=0	0.25	0.5	0.75	1.0
Soft					
ScienceQA	82.26	84.34	86.56	87.80	89.88
TextVQA	57.33	57.53	57.97	59.00	58.30
ImageNet	48.23	94.07	95.07	95.80	95.70
GQA	60.37	62.33	61.87	61.23	61.73
VizWiz	61.33	61.90	62.10	64.57	63.60
REC	66.87	67.27	68.20	69.13	68.57
VQAv2	68.47	66.50	67.10	67.93	67.17
OCRVQA	67.68	67.75	68.58	67.81	63.94
Hard					
ScienceQA	84.20	85.52	86.83	87.04	88.01
TextVQA	58.30	58.33	58.13	59.27	57.80
ImageNet	77.37	87.13	91.87	93.37	94.70
GQA	61.83	62.43	62.17	60.90	59.93
VizWiz	61.87	63.20	63.07	62.63	57.53
REC	69.73	68.03	68.80	69.63	66.87
VQAv2	66.40	65.90	66.70	67.13	65.27
OCRVQA	68.18	63.27	63.74	62.04	61.00

B PER-TASK ROUTING MEASUREMENTS

Tables 5 and 6 report per-task top-1 routing accuracy and mean self-routing mass at h=0. These measurements quantify whether task identities can be inferred from input samples and whether the router assigns dominant weights to the corresponding task experts. All results are computed over 100 sampled test inputs per task with τ=0.1 and seed 42; the chance level is $1/T$.

On UCIT, the router achieves perfect task identification accuracy for all tasks, and each task retains at least 0.60 routing mass on its own expert. The routing-mass matrix in Figure 3 further shows that most probability mass concentrates on the diagonal, indicating strong task separability. The largest off-diagonal mass occurs between VizWiz and Flickr30k, reflecting their semantic relatedness.

Although CoIN exhibits more routing ambiguity than UCIT, strong task-specific cues remain evident in its routing behavior, as shown in Figure 4. These results support our observation that existing benchmarks still operate in relatively separable task regimes, where task identity can serve as an effective shortcut for expert selection.

Table 5: UCIT per-task top-1 routing accuracy (acc, %) and mean softmax mass on the own-task expert (self mass) at h=0. Results are computed over 100 sampled test inputs per task (τ=0.1, seed 42); random-guess accuracy is 16.67%.

Task	acc	self mass
ImageNet-R	100.00	0.81
ArxivQA	100.00	0.81
VizWiz	100.00	0.64
IconQA	100.00	0.60
CLEVR-Math	100.00	0.81
Flickr30k	100.00	0.62
Mean	100.00	0.72

Table 6: CoIN per-task top-1 routing accuracy (acc, %) and mean softmax mass on the own-task expert (self mass) at h=0. Results are computed over 100 sampled test inputs per task (τ=0.1, seed 42); random-guess accuracy is 12.5%.

Task	acc	self mass
ScienceQA	100.00	0.49
TextVQA	100.00	0.45
ImageNet	73.00	0.36
GQA	76.00	0.25
VizWiz	75.00	0.31
REC	100.00	0.58
VQAv2	39.00	0.23
OCRVQA	64.00	0.28
Mean	78.38	0.37

C Additional Experimental Details

C.1 Implementation Details

All experiments use LLaVA-1.5-7B and build upon the MoE-LoRA implementation of HiDe-LLaVA (Guo et al., 2025a). The LoRA modules use total ranks of 48 on UCIT and 64 on CoIN, with a scaling factor of α/r=2, corresponding to rank 8 for each task expert. We train each stage for one epoch with an effective batch size of 48 and a learning rate of 2×10^{-4}. We use fp16 precision. The same configuration is used across all entropy levels, with only the training-task composition being changed.

C.2 Evaluation Protocol

We use greedy decoding with a maximum of 128 newly generated tokens. The softmax routing weight used by the Soft and Hard readouts employ a temperature of τ=0.1. For classification and VQA tasks, predictions and references are normalized before matching. REC is evaluated using an IoU threshold of 0.5 between the predicted and ground-truth bounding boxes. For the UCIT captioning tasks, VizWiz and Flickr30k, we follow Guo et al. (2025a) and report the aggregate score over BLEU-1–4, METEOR, ROUGE-L, and CIDEr. Last Average is the unweighted mean over the final-stage task scores.

C.3 Entropy-Mixture Construction

For each entropy level, we vary the task composition of each training stage while keeping its total number of samples unchanged. Let N_t denote the original number of training samples for task t and C the number of tasks. Given the mixing coefficient λ, each of the other $C-1$ tasks contributes

$$\left\lfloor \frac{N_t \lambda}{C} \right\rfloor$$

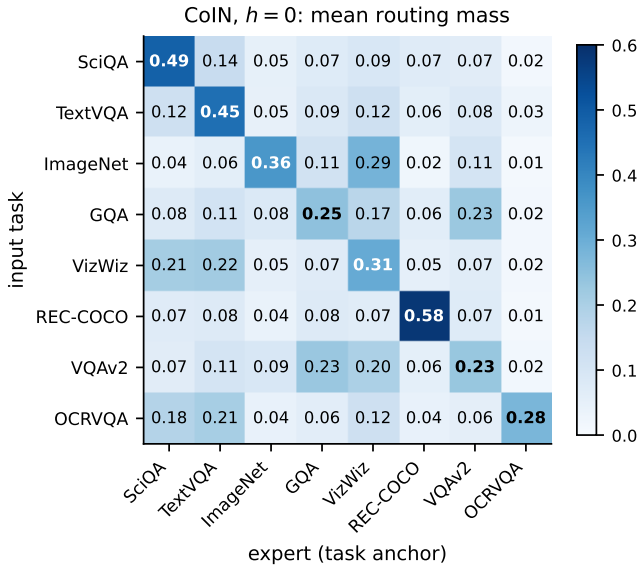


Figure 4: Mean routing mass on CoIN at h=0 with τ=0.1. Rows denote input tasks and columns denote experts. Each row is averaged over 100 sampled test inputs and sums to one.

samples to stage t, with the remaining samples drawn from task t itself. Samples are drawn without replacement and shuffled before training. All entropy levels use the same training and evaluation configuration.