MLLM-CTBench:A Comprehensive Benchmark for Continual Instruction Tuning of Multimodal LLMs with Chain-of-Thought Reasoning Analysis

Anonymous ACL submission

Abstract

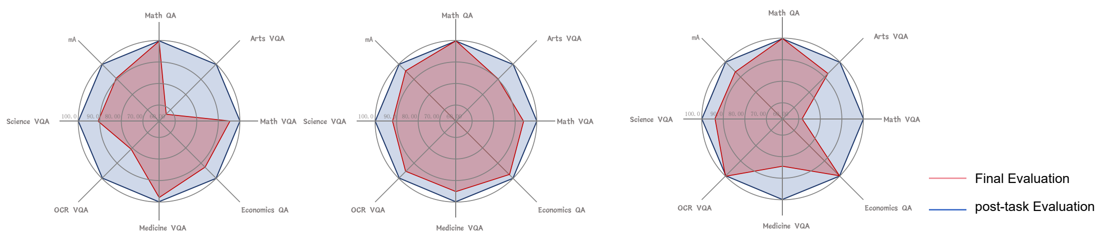
Multimodal large language models (MLLMs) rely on continual instruction tuning to adapt to evolving real-world demands, yet progress is hindered by the lack of rigorous benchmarks. We present **MLLM-CTBench**, a comprehensive benchmark with three novel contributions: 1) **Competence-driven task curation** spanning six challenging domains, including about 70K examples from 16 datasets; 2) **Multidimensional evaluation** combining final answer accuracy with granular Chain-of-Thought (CoT) reasoning analysis, enabled by a specially trained CoT evaluator; 3) **Comprehensive Algorithmic Investigations** covering eight continual learning algorithms across four categories, as well as the reinforcement learning and supervised fine-tuning paradigm. Key findings include: i) Pre-training quality inversely correlates with forgetting susceptibility (e.g.,LLaVA-1.5 shows 2× higher forgetting than Qwen2.5-VL); ii) Reasoning chains degrade slower than final answers, supporting the hierarchical forgetting hypothesis; iii) The performance of continual learning algorithms vary with model capacity and task order. MLLM-CTBench establishes rigorous standards for continual instruction tuning of MLLMs, offering actionable insights for algorithm design and evaluation. Our code can be found at `https://anonymous.4open.science/r/MLLM-CTBench-5E56`.

1 Introduction

Multimodal large language models (MLLMs) have emerged as pivotal architectures for cross-modal understanding and generation, demonstrating unprecedented capabilities across a wide range of multimodal tasks. Instruction tuning has further empowered these models to align with human intent and enhance task performance through supervised adaptation on task-specific instructions (Yu et al., 2024). However, real-world deployment demands continuous adaptation to evolving instructions and domain requirements—a paradigm known as *continual instruction tuning* (He et al., 2023a).

While significant progress has been made in continual instruction tuning for unimodal Large Language Models (LLMs) (Zheng et al., 2025a), the multimodal counterpart remains underexplored. The absence of a rigorous benchmark further impedes progress: existing benchmarks (e.g., EMT (Jia et al., 2025), CITB (He et al., 2023b), CoIN (Chen et al., 2024a)) on continual instruction tuning of MLLMs exhibit several critical limitations. 1) **Inadequate Task Difficulty**: The adopted datasets (e.g., ImageNet-1K in EMT (Jia et al., 2025), VQAv2 (Goyal et al., 2017)/TextVQA (Singh et al., 2019) in CoIN (Chen et al., 2024a)) fail to challenge modern MLLMs, as evidenced by their near-saturation zero-shot accuracies ($\geq 80\%$ for LLaVA-1.5 (Liu et al., 2024), nearly 90% for Qwen2.5-VL (Bai et al., 2025) on these benchmarks), rendering them ineffective for probing the boundaries of continual learning capacity in modern MLLMs. 2) **Superficial Evaluation Paradigms**: Prevailing benchmarks prioritize final answer correctness while neglecting granular reasoning process analysis, hindering in-depth understanding of the causes behind catastrophic forgetting in MLLMs (Luo et al., 2023). Although CoIN (Chen et al., 2024a) implicitly estimates reasoning knowledge forgetting , the interpretability of the evaluation metric remains limited. 3) **Algorithmic Evaluation Vacuum**: Existing works predominantly focus on quantifying catastrophic forgetting under sequential fine-tuning settings, while overlooking systematic investigations of continual learning algorithms' efficacy, thus limiting their impact.

To catalyze research progress in continual instruction tuning for MLLMs, we present MLLM-CTBench—a comprehensive benchmark designed to address the key limitations above. Our



(a) Final answer accuracy under SFT (b) CoT reasoning score under SFT (c) Final answer accuracy under RL

Figure 1: Evaluation of continual instruction tuning for MLLMs under SFT and RL paradigms with CoT reasoning analysis. Red lines indicate the performance after sequential tuning on all tasks; blue lines denote the performance after just tuning on each task. To enable intuitive visualization, we use the post-task performance (i.e., immediately after tuning each task) as the reference point and report relative percentages. (a) Final answer accuracy under the standard sequential fine-tuning (SFT) paradigm. (b) Critic score of the CoT reasoning, which degrades more slowly compared to final answers. (c) Final answer accuracy under the reinforcement learning paradigm (with GRPO), which better retains MLLMs' capabilities than SFT.

benchmark introduces three key innovations: 1)**Competence-Driven Task Curation**. Grounded in empirical studies that reveal MLLMs' persistent deficiencies in mathematical reasoning (Lu et al., 2021a; Chen et al., 2022; Xia et al., 2024; Yue et al., 2024a,b; Wang et al., 2023a), OCR comprehension (Wang et al., 2020a), and domain-specific knowledge (Kembhavi et al., 2016; Lu et al., 2022a; Lau et al., 2018a; Ben Abacha et al., 2021; He et al., 2020; Zhang et al., 2023a; Garcia et al., 2020; Wang et al., 2023a), we construct seven evaluation tasks across six challenging domains (Math, OCR, Science, Medicine, Arts, Economics). By systematically filtering 16 public datasets, we curate approximately 70K examples, ensuring balanced domain representation and mitigating dataset bias. 2)**Multidimensional Evaluation Protocol**. We propose a two-tiered assessment framework: macro-level metrics (final answer accuracy) and micro-level reasoning analysis encompassing visual grounding fidelity (for VQA tasks), logical coherence, and domain knowledge retention (Tan et al., 2024; Zheng et al., 2023). To ensure objectivity in CoT reasoning evaluation, we train a dedicated CoT evaluator—specifically, a fine-tuned Qwen2.5-VL-7B model.n (Chen et al., 2024b) 3)**Comprehensive Algorithmic Evaluation**. We benchmark eight mainstream continual learning algorithms across four methodological categories: regularization-based (Aich, 2021; Zheng et al., 2025a; Li and Hoiem, 2017a; Aljundi et al., 2018), replay-based (Rolnick et al., 2019b; Yan et al., 2021), architecture-expansion-based (Wang et al., 2022), and model-fusion-based (Marczak et al., 2024) approaches. Furthermore, given the increasing adoption of reinforcement learning (RL) for enhancing CoT reasoning in MLLMs, we compare RL and supervised fine-tuning (SFT) paradigms under continual instruction tuning settings (Chung et al., 2022).

Leveraging **MLLM-CTBench**, we conduct extensive experiments and uncover several key findings: **1)** All evaluated MLLMs experience forgetting, with weaker models (e.g., LLaVA-1.5 (Liu et al., 2024)) exhibiting 2× higher forgetting rates than stronger ones (e.g., Qwen2.5-VL (Bai et al., 2025)), indicating that pre-training quality inversely correlates with forgetting under continual adaptation. **2)** Intermediate reasoning traces degrade more slowly than final answer accuracy, supporting the *hierarchical forgetting hypothesis*—factual knowledge decays faster than procedural reasoning—consistent with CoIN (Chen et al., 2024a) and spurious forgetting studies (Zheng et al., 2024, 2025b). **3)** Reinforcement learning paradigms (e.g., GRPO (Shao et al., 2024)) achieve lower forgetting rates compared to SFT, attributed to KL-divergence regularization with reference models. **4)** The performance of continual learning algorithm varies with model capacity and task order: replay methods help weaker models but saturate on stronger ones; regularization excels on robust models but falters on weaker ones. Model fusion balances retention and efficiency, making it well-suited for resource-constrained scenarios.

In summary, this paper contributes the following

- We introduce **MLLM-CTBench**, a rigorously curated benchmark spanning seven evaluation tasks across six challenging domains.

- We propose a **two-tiered evaluation framework** that combines macro-level answer ac-

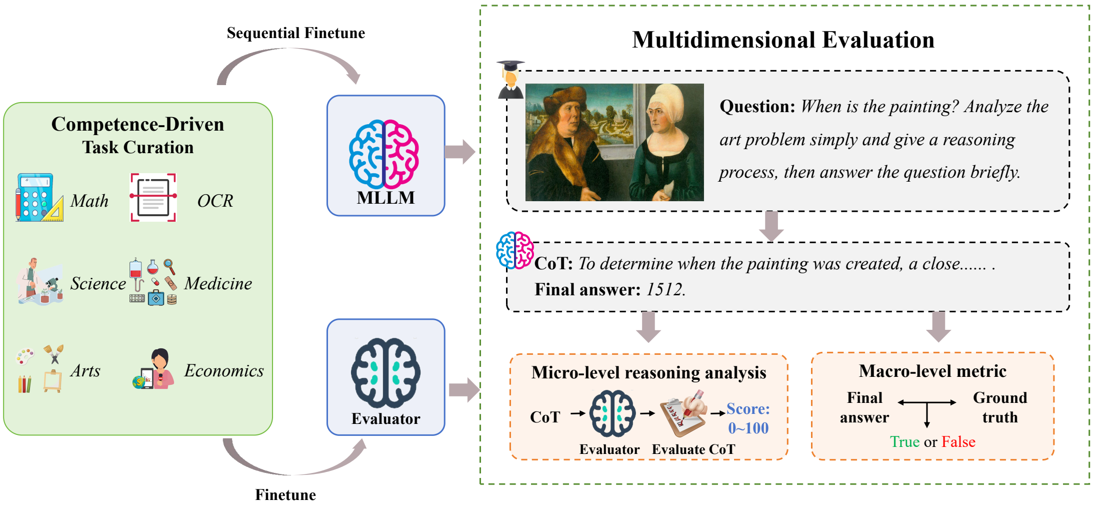


Figure 2: Overview of **MLLM-CTBench**. The MLLMs firstly undergo sequential instruction tuning on seven tasks from six domains, curated following a competence-driven manner. Then the performance is measured under a two-tiered evaluation framework combining both macro-level final answer metric with the micro-level CoT reasoning analysis enabled by a dedicated CoT evaluator.

curacy with fine-grained reasoning diagnostic enbled by a dedicated CoT evaluator.

- We perform the **comprehensive evaluation of eight continual learning methods** across four algorithmic paradigms, providing actionable guidance for MLLM continual learning method design. We further demonstrate that RL-based methods outperform SFT in retaining MLLM's capabilities (Recht, 2019; Khetarpal et al., 2022).

2 Related Work

Continual Learning Continual learning (CL) enables models to learn sequentially without forgetting (Wu et al., 2024). Existing methods include: **(1) Regularization-based** (e.g., EWC (Kirkpatrick et al., 2017), OGD (Farajtabar et al., 2020), LwF (Li and Hoiem, 2017b)) constrain updates to preserve past knowledge; **(2) Replay-based** (Rolnick et al., 2019a) reuse prior data to maintain performance, with memory overhead; **(3) Architecture-based** (Razdaibiedina et al.) expand models with task-specific modules (e.g., prompts); and **(4) Model fusion** (e.g., Max-merge) aligns task-specific checkpoints post-training with minimal overhead.

MLLM as a Judge LLMs have shown promise as automatic evaluators in NLP (Zhu et al., 2023; Li et al., 2023; Bai et al., 2023). Techniques such as pairwise scoring (Kim et al., 2023), Chain-of-Thought prompting (Wei et al., 2022), and preference alignment (Ouyang et al., 2022) enhance alignment with human judgments. Recent work extends this to MLLMs: Chen et al. (2024b) evaluate MLLMs as judges across scoring, comparison, and ranking tasks in vision-language settings.

3 MLLM-CTBench

We advocate three core principles in benchmark construction: *Difficulty*, *Diversity*, and *Comprehensiveness*. 1) **Difficulty**: Our benchmark is designed to include more challenging tasks than previous ones, aiming to more effectively evaluate the mordern MLLMs. 2) **Diversity**: It spans a wide range of knowledge domains and includes both unimodal and multimodal tasks, enabling broad evaluation of continual learning in realistic settings. 3) **Comprehensiveness**: In addition to final-answer accuracy, we aim to evaluate CoT (Lu et al., 2022b) reasoning to support fine-grained analysis of forgetting and capability drift. Since reasoning is central to LLM performance, its assessment is critical for understanding model behavior over time.

3.1 Competence-driven Task Curation

To ensure both diversity and difficulty in evaluation, we focus on six performance-limited domains—**Arts**, **Medicine**, **Economics**, **Science**, **Math**, and **OCR**—where state-of-the-art MLLMs continue to face significant challenges. Notably, state-of-the-art models (e.g., Claude-3.5, GPT-4o, InternVL2.5, Qwen2-VL) achieve only 51.9% accuracy on MMMU-Pro (Yue et al., 2024c) (covering the first five domains) and up to 61.5% on OCRBench v2 (Fu et al., 2024) .

To reduce task-level data imbalance,we construct a balanced benchmark where each task contributes a similar number of challenging examples.

3.1.1 Data Integration

We construct our benchmark from high-quality public datasets, covering six reasoning-intensive domains: **(1) Arts**, from AQUA (Garcia et al., 2020), involves historical identification and art interpretation; **(2) Science**, from ScienceQA (Lu et al., 2022a) and AI2D (Kembhavi et al., 2016), requires integrating visual and scientific knowledge; **(3) Medicine**, from VQA-RAD (Lau et al., 2018b), VQA-Med (Ben Abacha et al., 2021), PMC-VQA (Zhang et al., 2023a), and PathVQA (He et al., 2020), spans multi-modal medical imaging and diagnosis; **(4) Economics**, from TRACE (Wang et al., 2023b), focuses on policy sentiment classification; **(5) Math**, from IconQA (Lu et al., 2021b), GeoQA (Chen et al., 2022), CHARTX (Xia et al., 2024), MMMU (Yue et al., 2024a), and TRACE, covers symbolic, geometric, and visual reasoning; **(6) OCR**, from ChartOCR (Luo et al., 2021), CROHME (Guan et al., 2024), and ESTVQA (Wang et al., 2020b), includes chart interpretation, handwritten math, and scene text. Dataset statistics are summarized in Table 1.

Table 1: Statistics of the MLLM-CTBench datasets.

Task	Data Source	Train (Text / Image)	Test (Text / Image)
Math QA	TRACE	10K/0	0.5K/0
Economics QA	TRACE	5K/0	0.5K/0
Science VQA	AI2D, ScienceQA	9K/4K	1K/0.5K
Math VQA	IconQA, GeoQA, CHARTX, MMMU	8.3K/8.3K	0.9K/0.9K
Medicine VQA	VQA-RAD, VQA-Med-2021, PMC-VQA, PathVQA	9K/6.9K	1K/1K
OCR VQA	ChartOCR, CROHME, ESTVQA	12K/12.1K	1.4K/1.4K
Arts VQA	AQUA	9K/7K	1K/0.9K

3.1.2 CoT Generation

To enhance model reasoning, we generate high-quality Chain-of-Thought (CoT) annotations tailored to each benchmark task (Zhang et al., 2023b). Tasks are categorized by domain and span diverse answer formats (e.g., multiple choice, open-ended, yes/no). To accommodate this variability, we design task- and format-specific instruction templates (see Appendix). Each input consists of a problem statement, answer format, and task-specific instructions, which are provided to GPT-4 (OpenAI, 2023) alongside carefully crafted prompts (Liu and Huang, 2023) to elicit step-by-step reasoning. This structured prompting improves performance on complex tasks and enhances the interpretability of model outputs.

3.2 Continual Instruction Tuning

Setup. To reduce order-specific bias, we conduct sequential tuning under two task permutations: **Order-A** (`Math QA` $\to$ `Arts VQA` $\to$ `Math VQA` $\to$ `Economics QA` $\to$ `Medicine VQA` $\to$ `OCR VQA` $\to$ `Science VQA`) and its reverse, **Order-B**, ensuring robustness to task order effects.

Sequential Finetuning (SFT). Given tasks $\{\mathcal{T}_1, \dots, \mathcal{T}_S\}$ with datasets $\{D_1, \dots, D_S\}$, SFT optimizes the model f_θ on each task via:

$$\mathcal{L}_{\mathcal{T}_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} \ell(f_\theta(X_{i,j}^{\text{img}}, X_{i,j}^{\text{ins}}), X_{i,j}^{\text{ans}}), \qquad (1)$$

where ℓ is typically cross-entropy. We evaluate both full-parameter tuning and LoRA (Hu et al., 2021) to assess continual learning across adaptation regimes.

Reinforcement Learning (RL). We further examine GRPO, a state-of-the-art RL method for vision-language tuning, under the continual setting. The GRPO objective is:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{(s,a)\sim\pi_{\theta_{\text{old}}}} \left[\frac{\pi_\theta(a \mid s)}{\pi_{\theta_{\text{old}}}(a \mid s)} A^\pi(s,a) - \beta \, \text{KL}\big(\pi_\theta \,\|\, \pi_{\theta_{\text{old}}}\big) \right], \qquad (2)$$

where $s = f_\theta(X^{\text{img}}, X^{\text{ins}})$ and a is a generated token. GRPO promotes continual adaptation by optimizing return while regularizing policy drift.

3.3 Multidimensional Evaluation

To comprehensively evaluate continual learning in MLLMs, we adopt **Macro-Level Metrics** to assess final answer accuracy and **Micro-Level Reasoning Analysis** to evaluate the underlying reasoning process, enabling a more nuanced understanding of model retention and forgetting.

3.3.1 Macro-Level Metrics

Following standard instruction-tuning protocols, we extract the final answer from the model's output, which includes both the reasoning and the conclusion, and compare it to the ground truth. As answer formats vary across tasks, we apply task-specific evaluation rules. Detailed comparison strategies are provided in the appendix.

We evaluate continual learning performance using two standard metrics. Let $P_{i,j}$ denote the accuracy on task j after training task i, and N be the total number of tasks.

Average Performance (AP) measures overall accuracy after all tasks are trained: $AP = \frac{1}{N}\sum_{j=1}^{N} P_{N,j}$. A higher AP indicates better task-wide performance.

Backward Transfer (BWT) quantifies the effect of new-task learning on prior tasks: $BWT = \frac{1}{N-1}\sum_{j=1}^{N-1}(P_{N,j} - P_{j,j})$. Negative BWT reflects forgetting, while positive values indicate beneficial transfer.

3.3.2 Micro-level Reasoning Analysis

To additionally evaluate reasoning beyond final answers, we assess the quality of *Chain-of-Thought* (CoT) traces, as illustrated in Figure 4. We adopt two approaches: (1) general-purpose open-source models, and (2) a dedicated trained evaluator.

General-Purpose Evaluator. Following CoIN, we use `Qwen-VL-32B` with task-specific structured prompts (Ho et al., 2022) to assess reasoning quality. Each CoT trace is scored over three dimensions (0–100): **(i) Logical Coherence**, **(ii) Visual Grounding Fidelity** (for VQA tasks), and **(iii) Domain Knowledge Retention**. The final score is the average of the three.

Dedicated Multimodal Evaluator. To enable consistent and model-agnostic evaluation, we train a dedicated evaluator based on `Qwen2.5-VL-7B` using a two-stage pipeline: supervised fine-tuning on GPT-4-labeled traces, followed by GRPO (Zhang et al., 2024) using GPT-4 preferences as rewards. This evaluator generalizes across models and maintains alignment with human judgment for both SFT and RL outputs.

4 Experiments

4.1 Experimental Settings

We conduct continual instruction tuning on our benchmark using two strong open-source MLLMs: `LLaVA-1.5-7B` and `Qwen-VL-2.5-3B`, under two task sequences (**Order-A** and **Order-B**). Detailed training hyperparameters and implementation configurations for all methods, including LoRA and model-specific setups, are provided in Appendix A.3.

4.2 Main Results and Disscussions

1)**Do MLLMs Exhibit Catastrophic Forgetting—and How Does It Manifest?**

Table 2 presents continual fine-tuning results for two representative MLLMs: `LLaVA-1.5` (up) and `Qwen2.5-VL` (down). We observe a clear presence of catastrophic forgetting across tasks. For example, in `LLaVA-1.5`, continual fine-tuning under the order-A results in an average accuracy drop of approximately 15% between the *after-task* and *final* evaluations, highlighting the severity of catastrophic forgetting during sequential updates.

We also find that model performance is sensitive to task ordering, with task-level forgetting patterns varying across different sequences. For instance, in `LLaVA-1.5`, the Arts VQA task shows a 17.02% drop under Order-A but degrades by 24.37% under Order-B. However, the overall forgetting across the two orders remains similar, with an average gap of around 1% for both `LLaVA-1.5` and `Qwen2.5-VL`, suggesting that task interference is locally amplified but globally stable.

Finally, we compare macro-level answer accuracy with micro-level reasoning quality. Under Order-A, `Qwen2.5-VL` forgets 6.43% on macro-level metrics but only 3.74% on micro-level reasoning analysis. Similarly, `LLaVA-1.5` forgets 15.37% at the answer level but only 8.74% in reasoning quality. Results under other task orders and continual learning strategies consistently support this trend. Detailed reasoning scores are provided in the Appendix A.2.

2)**How to Select the Appropariate Continual Learning Method for Different Scenarios?**

We analyze the performance of four representative continual learning methods—regularization-based, replay-based, architectural expansion, and model merging—on MLLMs of varying capacities. Based on our findings, we summarize the strengths and applicability of each method under different scenarios. Detailed results are shown in Table 3 and Table 4.

Regularization-based methods (`EWC`, `MAS`, `LwF`) show more stable performance on relatively stronger models. For instance, `MAS` reduces forgetting by 41.51% in `LLaVA-1.5` and 54.74% in `Qwen2.5-VL`, suggesting that models with stronger representations benefit more from soft constraints. However, these methods require additional memory and computation to store importance scores. Notably, the layer-freezing strategy proposed in (Zheng et al., 2025a), which freezes parts of the language module in LLMs to mitigate forgetting, can be counterproductive for strong MLLMs. Specifically, freezing the first or last 8 layers of

Table 2: Evaluation of continual instruction tuning of MLLMs using macro-level metrics (final answer accuracy). Results are reported for two models under both `Order-A` and `Order-B`. For each order, the first row shows performance immediately after fine-tuning on a single task, while the second row shows performance after completing training on all tasks.

Model	Method	Math QA	Arts VQA	Math VQA	Economics QA	Medicine VQA	OCR VQA	Science VQA	AP	BWT
LLaVA-1.5	Multi-task	81.28	28.84	51.77	65.73	31.85	19.16	74.72	50.48	–
	Zero-shot	0.00	6.03	43.31	35.81	23.55	16.59	49.29	24.94	–
	DirctFT	79.80	31.10	57.70	69.96	32.95	19.16	75.40	52.30	–
	Order-A	79.80	30.39	55.42	67.14	30.86	19.44	73.70	50.96	–
		52.22 (↓27.58)	13.37 (↓17.02)	35.23 (↓20.19)	29.78 (↓37.36)	28.06 (↓2.80)	16.81 (↓2.63)	73.70	35.60	-15.37
	Order-B	69.98	27.21	54.05	68.55	30.40	18.16	76.06	49.20	–
		69.98	2.84 (↓24.37)	37.63 (↓16.42)	51.41 (↓17.14)	22.29 (↓8.11)	11.68 (↓6.48)	44.67 (↓31.39)	34.36	-16.58
Qwen2.5-VL	Multi-task	93.68	35.63	73.18	91.89	32.97	66.98	89.57	69.13	–
	Zero-shot	23.15	7.72	31.93	78.23	8.99	15.87	52.40	31.18	–
	DirctFT	90.89	33.55	71.61	91.28	33.91	64.35	90.48	68.01	–
	Order-A	90.89	32.44	71.84	92.14	31.74	45.82	84.07	64.13	–
		91.87 (↑0.98)	14.04 (↓18.40)	60.21 (↓11.63)	84.48 (↓7.66)	29.78 (↓1.96)	39.49 (↓6.33)	84.07	57.71	-6.43
	Order-B	91.87	36.37	71.15	84.17	35.24	47.25	89.54	65.08	–
		91.87	23.42 (↓12.95)	68.76 (↓2.39)	79.23 (↓4.94)	34.32 (↓0.92)	39.00 (↓8.25)	81.53 (↓8.01)	59.73	-5.35

Table 3: Performance of representative continual learning methods with `LLaVA-1.5` on MLLM-CTBENCH, evaluated under **Order-A** using the **macro-level final answer accuracy**.

Method	Math QA	Arts VQA	Math VQA	Economics QA	Medicine VQA	OCR VQA	Science VQA	AP	BWT
ER	81.77	29.48	44.58	68.55	29.50	20.37	71.82	49.44	–
	79.06(↓2.71)	27.82(↓1.66)	42.65(↓1.93)	64.52(↓4.03)	28.87(↓0.63)	18.95(↓1.42)	71.82	47.67	-1.77
DER	80.05	31.80	48.57	69.15	31.64	20.94	57.96	48.59	–
	78.82(↓1.23)	29.62(↓2.18)	46.41(↓2.16)	70.26(↑1.11)	32.46(↑0.82)	20.85(↓0.09)	57.96	48.05	-0.53
EWC	80.79	29.66	42.76	65.93	29.51	18.95	68.61	48.03	–
	45.32(↓35.47)	9.42(↓20.24)	38.65(↓4.11)	58.17(↓7.76)	24.89(↓4.62)	13.60(↓5.35)	68.61	36.95	-11.08
MAS	83.00	25.97	45.72	67.74	27.74	17.66	67.20	47.86	–
	48.52(↓34.48)	13.18(↓12.79)	39.68(↓6.04)	63.51(↓4.23)	27.65(↓0.09)	12.39(↓5.27)	67.20	38.88	-8.99
LwF	81.53	23.50	39.22	66.83	28.41	18.80	52.50	44.40	–
	45.81(↓35.72)	12.93(↓10.57)	31.81(↓7.41)	65.52(↓1.31)	26.09(↓2.32)	15.88(↓2.92)	52.50	35.79	-8.61
freeze-first-8-layers	82.02	30.43	44.70	68.95	29.81	21.15	55.98	47.58	–
	79.06(↓2.96)	29.17(↓1.26)	42.65(↓2.05)	66.33(↓2.62)	27.91(↓1.90)	20.23(↓0.92)	55.98	45.90	-1.67
freeze-last-8-layers	82.51	30.21	47.66	67.54	29.41	19.23	56.46	47.57	–
	80.05(↓2.46)	29.14(↓1.07)	45.38(↓2.28)	69.96(↑2.42)	31.42(↑2.01)	19.44(↑0.21)	56.46	52.07	4.49
L2P	81.00	31.32	48.21	65.87	30.56	19.25	73.56	49.97	–
	78.07(↓2.93)	26.68(↓4.64)	35.18(↓13.03)	59.13(↓6.74)	23.65(↓6.91)	15.58(↓3.67)	55.98(↓17.58)	42.04	-7.93
MagMaX	80.79	29.66	42.76	65.93	29.51	18.95	68.61	48.03	–
	54.93(↓25.86)	22.68(↓6.98)	39.57(↓3.19)	65.42(↓0.51)	29.39(↓0.12)	16.67(↓2.28)	55.70(↓12.91)	40.62	-7.41

the language model (`freeze-first-8-layers`, `freeze-last-8-layers`; see Table 4) in `Qwen2.5-VL` results in 20.37% more forgetting compared to standard fine-tuning.

Replay-based methods are particularly effective for weaker models prone to forgetting. In `LLaVA-1.5`, Experience Replay(ER) reduces forgetting by 88.48%, far outperforming other baselines. However, in `Qwen2.5-VL`, the improvement drops to 49.77%, suggesting diminishing returns as model capability increases. Moreover, replay methods face scalability issues due to the memory and compute cost of storing and processing image-text pairs across tasks.

Architectural expansion methods achieve relatively stable and decent performance across both model scales. By isolating task-specific knowledge into dedicated components (e.g., prompts (Razdaibiedina et al., 2023) or adapters), they mitigate forgetting while retaining efficiency. For example, these methods reduce forgetting by 48.41% on `LLaVA-1.5` and 37.17% on `Qwen2.5-VL`. Since only small modules are updated per task, the computational cost remains low. However, as the number of tasks increases, the number of task-specific components grows linearly, raising concerns about redundancy and inference complexity.

Model fusion provides a simple yet effective alternative. While its overall performance is not optimal, it consistently reduces forgetting—by 51.79% in `LLaVA-1.5` and 37.17% in `Qwen2.5-VL`—without requiring memory buffers

Table 4: Performance of representative continual learning methods with `Qwen2.5-VL` on MLLM-CTBENCH, evaluated under **Order-A** using the **macro-level final answer accuracy**.

Method	Math QA	Arts VQA	Math VQA	Economics QA	Medicine VQA	OCR VQA	Science VQA	AP	BWT
ER	90.89	32.53	71.38	80.79	29.34	37.25	82.00	60.60	–
	83.50(↓7.39)	25.60(↓6.93)	60.32(↓11.06)	82.56(↑1.77)	30.41(↑1.07)	37.19(↓0.06)	82.00	57.37	-3.23
DER	96.80	34.61	72.43	89.80	31.19	50.14	85.26	65.75	–
	91.13(↓ 5.67)	30.22(↓ 4.39)	65.86(↓ 6.57)	84.80(↓ 5.00)	33.24(↑ 2.05)	45.31(↓ 4.83)	85.26	62.26	-3.49
EWC	91.13	34.69	72.52	83.17	34.33	49.47	86.05	64.48	–
	95.07(↑3.94)	16.40(↓18.29)	65.45(↓7.07)	93.75(↑10.58)	32.02(↓2.31)	45.11(↓4.36)	86.05	61.98	-2.50
MAS	92.61	34.97	71.61	81.05	32.83	49.17	86.33	64.08	–
	93.84(↑1.23)	17.85(↓17.12)	62.14(↓9.47)	92.04(↑10.99)	32.80(↓0.03)	43.19(↓5.98)	86.33	61.17	-2.91
LwF	93.60	29.52	69.21	93.04	32.18	47.22	78.04	63.26	–
	97.29(↑3.69)	18.19(↓11.33)	59.18(↓10.03)	92.84(↓0.20)	29.04(↓3.14)	42.76(↓4.46)	78.04	59.62	-3.64
freeze-first-8-layers	91.43	31.37	63.71	87.92	32.29	45.20	72.83	60.68	–
	76.40(↓15.03)	13.29(↓18.08)	48.46(↓15.25)	79.29(↓8.63)	28.68(↓3.61)	41.29(↓3.91)	72.83	51.46	-9.22
freeze-last-8-layers	90.56	30.04	69.10	81.63	31.84	42.25	82.94	61.19	–
	75.15(↓15.41)	12.30(↓17.74)	58.49(↓10.61)	78.58(↓3.05)	26.97(↓4.87)	39.74(↓2.51)	82.94	53.45	-7.74
L2P	92.42	33.59	71.98	80.96	32.91	47.18	81.19	62.89	–
	93.59(↑1.17)	17.53(↓16.06)	67.42(↓4.56)	77.28(↓3.68)	29.56(↓3.35)	45.39(↓1.79)	80.17(↓1.02)	58.71	-4.18
MagMaX	90.89	32.44	71.84	92.14	31.74	45.82	84.07	64.13	–
	89.41(↓1.48)	28.28(↓4.16)	67.84(↓4.00)	88.51(↓3.63)	24.77(↓6.97)	39.08(↓6.74)	77.40(↓6.67)	59.33	-4.81

or structural modifications. Its simplicity makes it particularly appealing in deployment-constrained or low-resource settings.

3)**Can Our CoT Evaluator Be Trusted?**

We adopt the open-source `Qwen-VL-2.5-32B` as a general-purpose evaluator following prior work. To assess its reliability, we measure the Spearman correlation between its scores and those from GPT-4o on a held-out test set. Spearman's ρ, defined as

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{3}$$

quantifies rank consistency across paired observations.

As shown in Table 7, the general-purpose evaluator exhibits limited alignment with GPT-4o. This highlights a key limitation: large MLLMs, even when capable, may lack sensitivity to fine-grained reasoning signals.

To address this, we train a dedicated evaluator via a two-stage procedure—supervised fine-tuning followed by GRPO-based reinforcement learning—using only reasoning traces from `LLaVA`. Despite this narrow training domain, the resulting evaluator generalizes well, consistently yielding higher correlations across models and tasks (Table 7).

With this refined evaluator, we score the chain-of-thought (CoT) reasoning traces produced by all models in our benchmark. The normalized critic scores are reported in Appendix A.2. Consistent with our correlation analysis (Table **??**), the specialized evaluator offers sharper distinctions across models and training setups, revealing degradation patterns that raw answer accuracy alone fails to capture.

4)**RL vs. SFT under Continual Instruction Tuning.**

Reinforcement learning has emerged as a powerful paradigm for enhancing CoT reasoning in large models, with Generalized Reinforcement with Prompt Optimization (**GRPO**) representing one of the current state-of-the-art approaches. To assess its suitability under continual instruction tuning, we compare GRPO against the classical baseline of supervised fine-tuning (**SFT**). As shown in Table 6, GRPO consistently achieves 30–70% lower forgetting across all task orders, demonstrating superior robustness in preserving knowledge over extended training horizons.

This advantage is attributable to GRPO's objective (Eq. 2), which augments the task loss with a Kullback–Leibler divergence term that explicitly constrains the updated policy to stay close to the inference model. By limiting policy drift, the KL regularizer acts as an implicit memory, thereby preserving previously acquired reasoning skills while still allowing beneficial adaptation to new tasks.

5 Conclusion

We present **MLLM-CTBench**, a benchmark for evaluating *continual instruction tuning* in multimodal large language models (MLLMs). It features: (i) **competence-driven task curation** across six domains (70K examples); (ii) a **two-tiered evaluation** combining answer accuracy and CoT-level

Table 5: Reasoning analysis of CoT reasoning as scored by the dedicated evaluator.

Model	Order	Critic Scores							Average	BWT
		Math QA	Arts VQA	Math VQA	Economics QA	Medicine VQA	OCR VQA	Science VQA		
LLaVA-1.5	Order-A	97.54	28.12	64.99	90.12	31.59	43.30	79.83	62.21	–
		92.08 (↓5.46)	9.38 (↓18.74)	55.07 (↓9.92)	84.68 (↓5.44)	28.75 (↓2.84)	41.32 (↓1.98)	78.42 (↓1.41)	55.677	-6.54
	Order-B	79.31	30.16	59.52	84.58	32.03	44.22	75.68	57.93	–
		79.31	17.49 (↓12.67)	51.77 (↓7.75)	79.13 (↓5.45)	30.92 (↓1.11)	38.85 (↓5.37)	69.46 (↓6.22)	52.42	-5.51
Qwen2.5-VL	Order-A	91.82	64.14	68.53	84.68	64.50	71.19	79.64	74.93	–
		90.38 (↓1.44)	55.95 (↓8.19)	64.49 (↓4.04)	83.21 (↓1.47)	62.66 (↓1.84)	68.56 (↓2.63)	79.64	72.13	-3.74
	Order-B	92.68	63.45	68.87	83.95	64.37	72.53	80.80	75.24	–
		92.68	57.17 (↓6.28)	65.11 (↓3.76)	81.52 (↓2.43)	61.19 (↓3.18)	69.00 (↓3.53)	75.58 (↓5.22)	71.32	-4.03

Table 6: Continual learning performance of SFT and RL on MLLM-CTBench using `Qwen2.5-VL`.

Paradigm	Order	Math	Arts	M.VQA	Econ	Med	OCR	Sci	AP	BWT
SFT	order-A	97.54	28.12	64.99	90.12	31.59	43.30	79.83	62.21	–
		92.08 (↓5.46)	9.38 (↓18.74)	55.07 (↓9.92)	84.68 (↓5.44)	28.75 (↓2.84)	41.32 (↓1.98)	79.83	55.87	−6.34
	order-B	79.31	30.16	59.52	84.58	32.03	44.22	75.68	57.93	–
		79.31	17.49 (↓12.67)	51.77 (↓7.75)	79.13 (↓5.45)	30.92 (↓1.11)	38.85 (↓5.37)	69.46 (↓6.22)	52.42	−5.51
RL	order-A	71.92	13.07	48.12	84.07	18.31	35.62	70.03	48.73	–
		70.05 (↓1.87)	12.23 (↓0.84)	42.53 (↓5.59)	77.22 (↓6.85)	20.32 (↑2.01)	35.37 (↓0.25)	70.03	46.82	−1.91
	order-B	56.65	12.99	69.78	90.12	30.25	33.02	79.74	53.22	–
		56.65	11.99 (↓1.0)	50.63 (↓19.15)	90.42 (↑0.3)	22.62 (↓7.63)	39.65 (↑6.63)	74.27 (↓5.47)	49.46	−3.76

Table 7: Evaluation of evaluator quality via *average Spearman correlation* between predicted scores and GPT-4 ratings over seven reasoning tasks. **Qwen_SFT**, **Qwen_RL**, and **LLaVA_SFT** denote reasoning traces from Qwen2.5-VL (3B) and LLaVA-1.5 (7B) under SFT or RL. The general-purpose evaluator is the off-the-shelf `Qwen-VL-2.5-32B`, while the specialized evaluator is trained on reasoning traces from `LLaVA-1.5-7B`. Higher values indicate stronger agreement with GPT-4 rankings.

Evaluator	Reasoning generation	Avg. Spearman correlation
general-purpose MLLM	Qwen_SFT	66.60
	Qwen_RL	69.95
	LLaVA_SFT	80.49
specialized evaluator	Qwen_SFT	73.08
	Qwen_RL	75.13
	LLaVA_SFT	82.52

diagnostics; and (iii) **comprehensive comparisons** of eight continual learning methods and GRPO-based RL.

Experiments on `LLaVA-1.5` and `Qwen2.5-VL` reveal: (1) better pretraining mitigates forgetting; (2) reasoning degrades slower than answers, supporting a *hierarchical forgetting* view; and (3) method effectiveness varies by model capacity—replay aids weaker models, regularization suits stronger ones, and model fusion offers a balanced trade-off.

MLLM-CTBench enables principled evaluation and lays the groundwork for robust continual learning in multimodal settings.

6 Limitations

Despite the positive contributions of this study, we acknowledge the following limitations:1)**Limited model diversity.** Due to time constraints, we did not explore a wider range of MLLM architectures. Future work could examine whether our findings generalize to alternative multimodal model designs. 2) **Model scale constraints.** Our experiments are limited to models in the 3B–7B parameter range, constrained by available computational resources. Evaluating larger-scale models would help assess the scalability of continual instruction tuning and reasoning evaluation. 3) **Restricted task order coverage.** While we demonstrate consistent trends under multiple task sequences, we did not exhaustively evaluate all possible orderings. A broader exploration of task permutations could provide deeper insights into order sensitivity.

References

Abhishek Aich. 2021. Elastic weight consolidation (EWC): nuts and bolts. *CoRR*, abs/2105.04093.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. 2018. Memory aware synapses: Learning what (not) to forget. *Preprint*, arXiv:1711.09601.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Shuai Bai, Shusheng Yang, Jinze Bai, Peng Wang, Xingxuan Zhang, Junyang Lin, Xinggang Wang, Chang Zhou, and Jingren Zhou. 2023. Touchstone: Evaluating vision-language models by language models. *arXiv preprint arXiv:2308.16890*.

Asma Ben Abacha, Mourad Sarrouti, Dina Demner-Fushman, Sadid A. Hasan, and Henning Müller. 2021. Overview of the vqa-med task at imageclef 2021: Visual question answering and generation in the medical domain. In *CLEF 2021 Working Notes*, CEUR Workshop Proceedings, Bucharest, Romania. CEUR-WS.org.

Cheng Chen, Junchen Zhu, Xu Luo, Hengtao Shen, Lianli Gao, and Jingkuan Song. 2024a. Coin: A benchmark of continual instruction tuning for multimodel large language model. *Preprint*, arXiv:2403.08350.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024b. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *Preprint*, arXiv:2402.04788.

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric P. Xing, and Liang Lin. 2022. Geoqa: A geometric question answering benchmark towards multimodal numerical reasoning. *Preprint*, arXiv:2105.14517.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, and 13 others. 2022. Scaling instruction-finetuned language models. *ArXiv*, abs/2210.11416.

Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. 2020. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pages 3762–3773. PMLR.

Ling Fu, Biao Yang, Zhebin Kuang, Jiajun Song, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Mingxin Huang, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, and 5 others. 2024. Ocrbench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. *Preprint*, arXiv:2501.00321.

Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020. A dataset and baselines for visual question answering on art. In *Proceedings of the European Conference in Computer Vision Workshops*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Tongkun Guan, Chengyu Lin, Wei Shen, and Xiaokang Yang. 2024. Posformer: Recognizing complex handwritten mathematical expression with position forest transformer. *Preprint*, arXiv:2407.07764.

Jinghan He, Haiyun Guo, Ming Tang, and Jinqiao Wang. 2023a. Continual instruction tuning for large multimodal models. *arXiv preprint arXiv:2311.16206*.

Jinghan He, Haiyun Guo, Ming Tang, and Jinqiao Wang. 2023b. Continual instruction tuning for large multimodal models. *Preprint*, arXiv:2311.16206.

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

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2022. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Boyu Jia, Junzhe Zhang, Huixuan Zhang, and Xiaojun Wan. 2025. Exploring and evaluating multimodal knowledge reasoning consistency of multimodal large language models. *Preprint*, arXiv:2503.04801.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *Preprint*, arXiv:1603.07396.

Khimya Khetarpal, Matthew Riemer, Irina Rish, and Doina Precup. 2022. Towards continual reinforcement learning: A review and perspectives. *Journal of Artificial Intelligence Research*, 75:1401–1476.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and 1 others. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. *arXiv preprint arXiv:2310.08491*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

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

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

Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Hai Zhao, and Pengfei Liu. 2023. Generative judge for evaluating alignment. *arXiv preprint arXiv:2310.05470*.

Zhizhong Li and Derek Hoiem. 2017a. Learning without forgetting. *Preprint*, arXiv:1606.09282.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Lei Liu and J. Huang. 2023. Prompt learning to mitigate catastrophic forgetting in cross-lingual transfer for open-domain dialogue generation. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Preprint*, arXiv:2209.09513.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021a. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. 2021b. Iconqa: A new benchmark for abstract diagram understanding and visual language reasoning. In *The 35th Conference on Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Junyu Luo, Zekun Li, Jinpeng Wang, and Chin-Yew Lin. 2021. Chartocr: Data extraction from charts images via a deep hybrid framework. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1916–1924.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yuechen Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *ArXiv*, abs/2308.08747.

Daniel Marczak, Bartłomiej Twardowski, Tomasz Trzciński, and Sebastian Cygert. 2024. Magmax: Leveraging model merging for seamless continual learning. *Preprint*, arXiv:2407.06322.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations*.

Anastasia Razdaibiedina, Yuning Mao, Rui Hou, Madian Khabsa, Mike Lewis, and Amjad Almahairi. 2023. Progressive prompts: Continual learning for language models. In *The Eleventh International Conference on Learning Representations*.

Benjamin Recht. 2019. A tour of reinforcement learning: The view from continuous control. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):253–279.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. 2019a. Experience replay for continual learning. *Advances in neural information processing systems*, 32.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy P. Lillicrap, and Greg Wayne. 2019b. Experience replay for continual learning. *Preprint*, arXiv:1811.11682.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui

Zheng, Yicheng Zou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023a. Trace: A comprehensive benchmark for continual learning in large language models. *Preprint*, arXiv:2310.06762.

Xiao Wang, Yuansen Zhang, Tianze Chen, Songyang Gao, Senjie Jin, Xianjun Yang, Zhiheng Xi, Rui Zheng, Yicheng Zou, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023b. Trace: A comprehensive benchmark for continual learning in large language models. *Preprint*, arXiv:2310.06762.

Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. 2020a. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135.

Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. 2020b. On the general value of evidence, and bilingual scene-text visual question answering. *Preprint*, arXiv:2002.10215.

Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. 2022. Learning to prompt for continual learning. *Preprint*, arXiv:2112.08654.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, and 1 others. 2024. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Shipeng Yan, Jiangwei Xie, and Xuming He. 2021. Der: Dynamically expandable representation for class incremental learning. *Preprint*, arXiv:2103.16788.

Dianzhi Yu, Xinni Zhang, Yankai Chen, Aiwei Liu, Yifei Zhang, Philip S. Yu, and Irwin King. 2024. Recent advances of multimodal continual learning: A comprehensive survey. *Preprint*, arXiv:2410.05352.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024c. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *Preprint*, arXiv:2409.02813.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and 1 others. 2024. Direct preference optimization of video large multimodal models from language model reward. *arXiv preprint arXiv:2404.01258*.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Pmc-vqa: Visual instruction tuning for medical visual question answering. *arXiv preprint arXiv:2305.10415*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023b. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025a. Spurious forgetting in continual learning of language models. *Preprint*, arXiv:2501.13453.

Junhao Zheng, Xidi Cai, Shengjie Qiu, and Qianli Ma. 2025b. Spurious forgetting in continual learning of language models. *Preprint*, arXiv:2501.13453.

Junhao Zheng, Shengjie Qiu, and Qianli Ma. 2024. Learn or recall? revisiting incremental learning with pre-trained language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14848–14877, Bangkok, Thailand. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Appendices

A.1 Results under Order-B

A.1.1 Macro-level final answer accuracy

To assess the robustness of continual learning algorithms to task presentation order, we complement the main results (Table **??**) obtained under Order-A by evaluating all methods under an alternative task sequence, referred to as Order-B. This permutation presents tasks in a different curriculum, leading to distinct forgetting and interference dynamics.

Table 9 and Table 8 report the macro-level final answer accuracy for all methods evaluated on MLLM-CTBench, using LLaVA-1.5 and Qwen2.5-VL as the underlying models. While the relative rankings among methods remain largely consistent with Order-A, certain algorithms show increased sensitivity to task order—highlighting the importance of evaluating under multiple sequences for a complete understanding of continual learning behavior.

A.2 Evaluating Continual Learning Methods via CoT Reasoning Analysis

In the main paper, we compared the CoT reasoning analysis of `Qwen2.5-VL` and `LLaVA-1.5` under two task orders (Order-A and Order-B) on **MLLM-CTBench**. Here, we extend this analysis to include the performance of different continual learning methods under the same two task orders. The detailed results are provided in Tables 10 and 11, corresponding to `LLaVA-1.5` and `Qwen2.5-VL`, respectively.

A.3 Detial Experimental settings

We summarize the training configurations and hyperparameters for all methods evaluated in our benchmark.

A.4 Experimental settings

General Experimental Setup. We evaluate two strong open-source MLLMs: `LLaVA-1.5-7B` and `Qwen-VL-2.5-3B`. LLaVA uses a learning rate of 2×10^{-5}, batch size 16, and trains for up to 10 epochs; Qwen uses a learning rate of 1×10^{-5}, batch size 40, and trains for up to 8 epochs. Both models use a maximum sequence length of 4096 tokens.

We consider three baseline settings to analyze continual learning behavior:1) **Zero-shot**: Models are evaluated without any task-specific fine-tuning to reflect their pretrained capabilities.2) **Direct Fine-tuning (Direct FT)**: Each model is independently fine-tuned on a single task. LLaVA trains for 8–13 epochs depending on the task; Qwen for up to 8 epochs. Other hyperparameters follow the general setup.3) **Multi-task Joint Training**: All task datasets are jointly trained to evaluate multi-task generalization. Epochs are set to 13 for LLaVA and 10 for Qwen.

Baseline Setup. For sequential fine-tuning, we train `LLaVA-1.5-7B` for 10 epochs and `Qwen-VL-2.5-3B` for 8 epochs using the general hyperparameter setup. For LoRA fine-tuning, LLaVA uses a learning rate of 2×10^{-4} with `lora_r = 128` and `lora_alpha = 256`; Qwen uses a learning rate of 2×10^{-5} with `low-rank dimension` = 64, `LoRA scaling factor` = 128, and `lora_dropout` = 0.05.

Continual Learning Methods. We evaluate eight representative methods across four paradigms. 1) *Regularization-based methods* mitigate forgetting by constraining updates to important parameters. `EWC` estimates weight importance via the Fisher Information Matrix; `MAS` tracks sensitivity through output gradients; `LwF` distills knowledge from previous models; and `Freeze` preserves prior knowledge by freezing the vision encoder and either the first or last 8 layers of the language model (Zheng et al., 2025a). 2) *Replay-based methods* alleviate forgetting by revisiting prior data. `Experience Replay (ER)` stores a small memory buffer of past samples, while `DER` extends this by replaying both logits and raw inputs. 3) *Architectural methods* isolate task-specific knowledge into dedicated modules. `L2P` uses a learnable prompt pool to encode task identity and selectively activate relevant knowledge without interfering with previously learned parameters. 4) *Model fusion* provides a lightweight alternative by merging sequential checkpoints using a fixed fusion coefficient (`Max-merge` with $\alpha = 0.8$), requiring no memory or architectural modifications.

Reinforcement Learning Setup. We adopt `GRPO` (Shao et al., 2024) as our reinforcement learning framework for continual instruction tuning. During GRPO training, the vision encoder is frozen, and LoRA is applied only to the language model. The key hyperparameters are set as follows: the maximum prompt length is 1024, number of generations is 4, per-device training batch size is 16, and training runs for 3 epochs. We use a learning rate of 1×10^{-5} and configure LoRA with rank $r = 64$ and scaling factor $\alpha = 128$.

Table 8: Final answer accuracy under Order-B on MLLM-CTBENCH. Results are reported for Qwen2.5-VL.

Method	Math QA	Arts VQA	Math VQA	Economics QA	Medicine VQA	OCR VQA	Science VQA	AP	BWT
ER	94.09	31.49	70.13	87.95	32.99	46.63	88.69	64.57	–
	94.09	25.26 (↓6.23)	58.24 (↓11.89)	90.58 (↑2.63)	25.46 (↓7.53)	37.12 (↓9.51)	78.71 (↓9.98)	58.49 (↓6.08)	-6.07
DER	94.83	34.87	71.76	86.02	34.64	50.12	90.12	66.05	–
	94.83	28.90	68.90	89.63	32.80	44.80	86.57	63.78	-2.28
EWC	92.61	34.57	70.81	38.31	33.52	49.39	88.60	58.26	–
	92.61	20.39 (↓14.18)	67.39 (↓3.42)	70.36 (↑32.05)	28.88 (↓4.64)	32.46 (↓16.93)	80.02 (↓8.58)	56.02	-2.24
MAS	96.55	33.99	72.06	87.8	33.54	49.41	87.94	65.90	–
	96.55	21.3 (↓12.69)	67.5 (↓4.56)	83.77 (↓4.03)	32.18 (↓1.36)	36.85 (↓12.56)	78.98 (↓8.96)	59.59	-6.31
LwF	80.30	28.68	66.25	85.08	32.85	47.64	89.54	61.48	–
	80.30	29.65 (↑0.97)	67.16 (↑0.91)	77.92 (↓7.16)	29.35 (↓3.50)	38.93 (↓8.71)	80.77 (↓8.77)	57.73	-3.75
freeze-first-8-layers	89.68	28.77	61.46	89.76	32.68	43.85	71.91	59.73	–
	89.68	28.92(↑0.15)	45.84(↓15.62)	80.75(↓9.01)	28.74(↓3.94)	34.11(↓9.74)	51.65(↓20.26)	51.38	-8.35
freeze-last-8-layers	89.41	30.90	67.27	86.53	31.61	44.90	84.83	62.21	–
	89.41	25.74 (↓5.16)	65.68 (↓1.59)	76.59 (↓9.94)	27.52 (↓4.09)	30.33 (↓14.57)	75.31 (↓9.52)	55.80	-6.41
L2P	81.23	32.98	69.78	83.56	31.69	43.97	86.78	61.43	–
	81.23	30.13 (↓2.85)	65.48 (↓4.30)	76.98 (↓6.58)	28.95 (↓2.74)	39.17 (↓4.80)	79.88 (↓6.90)	57.40	-3.75
MagMaX	91.87	36.37	71.15	84.17	35.24	47.25	89.54	65.08	–
	95.07 (↑3.20)	10.53 (↓25.84)	70.24 (↓0.91)	92.54 (↑8.37)	32.33 (↓2.91)	42.59 (↓4.66)	83.79 (↓5.75)	61.01	-4.07

Table 9: Final answer accuracy under Order-B on MLLM-CTBENCH. Results are reported for LLaVA-1.5.

Method	Math QA	Arts VQA	Math VQA	Economics QA	Medicine VQA	OCR VQA	Science VQA	AP	BWT
ER	81.28	27.48	45.15	66.94	30.29	19.94	77.66	49.82	–
	81.28	27.51 (↑0.03)	42.42 (↓2.73)	65.32 (↓1.62)	28.38 (↓1.91)	17.28 (↓2.66)	71.91 (↓5.75)	47.73	-2.09
DER	83.5	30.18	45.27	68.95	32.53	21.44	59.85	48.82	
	83.5	30.56 (↑0.38)	46.07 (↑0.80)	70.26 (↑1.31)	30.10 (↓2.43)	21.44	57.02 (↓2.83)	48.42	-0.40
EWC	79.56	29.47	45.38	70.56	29.95	21.51	75.78	50.32	–
	79.56	13.67 (↓15.80)	22.01 (↓23.37)	61.09 (↓9.47)	14.78 (↓15.17)	13.32 (↓8.19)	50.42 (↓25.36)	36.41	-13.91
MAS	68.72	25.63	43.90	67.54	29.51	18.95	77.76	47.43	–
	68.72	21.60 (↓4.03)	41.16 (↓2.74)	60.89 (↓6.65)	27.39 (↓2.12)	14.53 (↓4.42)	60.04 (↓17.72)	42.08	-5.35
LwF	67.49	22.9	40.59	68.35	29.95	18.87	58.7	43.84	–
	67.49	12.22 (↓10.68)	27.14 (↓13.45)	58.87 (↓9.48)	23.81 (↓6.14)	10.97 (↓7.90)	46.56 (↓12.14)	35.29	-8.54
freeze-first-8-layers	81.28	29.13	45.61	69.96	26.28	21.44	57.87	47.37	–
	81.28	28.97 (↓0.16)	44.81 (↓0.80)	65.93 (↓4.03)	30.04 (↑3.76)	20.23 (↓1.21)	55.04 (↓2.83)	46.61	-0.75
freeze-last-8-layers	81.28	30.3	44.81	70.16	27.55	21.44	60.79	48.05	–
	81.28	28.64 (↓1.66)	41.51 (↓3.30)	69.66 (↓0.50)	29.94 (↑2.39)	19.59 (↓1.85)	57.68 (↓3.11)	46.90	-1.15
L2P	76.18	30.29	45.98	61.19	25.15	19.23	74.95	47.57	–
	76.18	27.68 (↓2.61)	40.96 (↓5.02)	57.61 (↓3.58)	22.95 (↓2.20)	14.58 (↓4.65)	53.96 (↓20.99)	41.99	-5.58
MagMaX	79.56	29.47	45.38	70.56	29.95	21.51	75.78	50.32	–
	41.38 (↓38.18)	12.35 (↓17.12)	34.78 (↓10.60)	66.13 (↓4.43)	23.13 (↓6.82)	17.31 (↓4.20)	62.30 (↓13.48)	36.77	-13.55

A.5 Dataset Examples and Evaluation Settings

To provide a clearer understanding of the diverse multimodal reasoning tasks in our benchmark, we include a representative visual example from each dataset, along with the task-specific instruction template and evaluation metric used. As shown in Figure 3, each dataset poses distinct reasoning challenges, ranging from mathematical derivation to visual perception and domain-specific understanding. For consistency, we unify the model interface using one canonical instruction prompt per dataset, while preserving the underlying task semantics.

To standardize evaluation across heterogeneous tasks, we carefully design prompt templates and adopt task-appropriate evaluation metrics. Table 12 summarizes the canonical instruction used for each dataset, as well as the corresponding metric. The selected prompts align with each task's core semantics while ensuring format consistency. Evaluation metrics are chosen based on the output style—Exact Match for structured or classification tasks, and ROUGE-L for generative responses.

A.6 Task-Specific Prompting and Evaluation Protocols

This unified format enables consistent and interpretable evaluation of continual learning behavior across multimodal tasks. While additional prompt variants may be used during training to improve task generalization, the canonical form and evaluation protocol presented here serve as the standardized testing setup.

Table 10: Chain-of-Thought reasoning analysis of `LLaVA-1.5` on MLLM-CTBench under two task orders (A and B) across different continual-learning methods.

Method	Order	Math QA	Arts VQA	Math VQA	Economics QA	Medicine VQA	OCR VQA	Science VQA	AP	BWT
ER	Order-A	87.45	64.64	61.24	81.45	63.78	56.92	75.69	70.17	–
		88.09(↑0.64)	63.99(↓0.65)	61.43(↑0.19)	81.39(↓0.06)	62.74(↓1.04)	56.67(↓0.25)	75.69	70.00	−0.17
	Order-B	89.45	64.12	60.44	81.74	63.57	56.05	78.21	70.51	–
		89.45	63.99(↓0.13)	60.56(↑0.12)	81.34(↓0.40)	62.94(↓0.63)	56.67(↑0.62)	75.81(↓2.40)	70.11	−0.40
DER	Order-A	88.12	64.84	61.17	81.63	70.15	56.44	74.25	70.94	–
		87.48(↓0.64)	64.27(↓0.57)	60.02(↓1.15)	81.33(↓0.30)	70.05(↓0.10)	55.55(↓0.89)	74.25	70.42	−0.52
	Order-B	89.51	63.73	60.80	81.78	70.50	56.89	75.94	71.31	–
		89.51	64.42(↑0.69)	60.21(↓0.59)	81.69(↓0.09)	69.55(↓0.95)	56.03(↓0.86)	73.19(↓2.75)	70.66	−0.65
EWC	Order-A	88.38	63.25	59.29	81.48	62.99	54.92	74.30	69.23	–
		76.38(↓12.00)	54.93(↓8.32)	53.03(↓6.26)	78.25(↓3.23)	58.33(↓4.66)	50.73(↓4.19)	74.30	63.71	−5.52
	Order-B	88.27	63.99	61.01	81.55	62.94	55.44	76.38	69.94	–
		88.27	56.14(↓7.85)	55.04(↓5.97)	78.02(↓3.53)	56.75(↓6.19)	43.39(↓12.05)	61.83(↓14.55)	62.78	−7.16
MAS	Order-A	89.09	63.04	57.08	80.87	62.54	52.39	72.29	68.19	–
		77.75(↓11.34)	55.00(↓8.04)	52.63(↓4.45)	79.76(↓1.11)	60.53(↓2.01)	50.76(↓1.63)	72.29(↓0.00)	64.10	−4.08
	Order-B	85.22	63.48	57.05	81.11	62.48	52.82	76.89	68.44	–
		85.22	61.13(↓2.35)	54.59(↓2.46)	80.49(↓0.62)	60.63(↓1.85)	49.74(↓3.08)	68.08(↓8.81)	65.70	−2.74
LwF	Order-A	88.35	64.57	60.39	81.68	64.70	56.50	78.04	70.60	–
		68.45(↓19.90)	54.26(↓10.31)	43.99(↓16.40)	76.83(↓4.85)	52.27(↓12.43)	41.46(↓15.04)	78.04	59.33	−11.28
	Order-B	88.27	63.99	61.01	81.55	62.94	55.44	76.38	69.94	–
		88.27	56.15(↓7.84)	55.04(↓5.97)	77.87(↓3.68)	56.88(↓6.06)	43.39(↓12.05)	61.93(↓14.45)	62.79	−7.15
freeze-first-8-layers	Order-A	88.72	64.04	60.27	81.29	69.89	55.70	73.21	70.45	–
		87.53(↓1.19)	63.79(↓0.25)	59.14(↓1.13)	81.23(↓0.06)	69.71(↓0.18)	55.11(↓0.59)	73.21	69.96	−0.49
	Order-B	88.16	64.01	60.54	81.87	69.97	55.76	74.83	70.73	–
		88.16	63.95(↓0.06)	60.39(↓0.15)	81.29(↓0.58)	69.59(↓0.38)	55.11(↓0.65)	72.84(↓1.99)	70.19	−0.54
freeze-last-8-layers	Order-A	88.14	64.27	60.45	81.75	69.18	55.30	72.50	70.23	–
		88.32(↑0.18)	64.32(↑0.05)	59.07(↓1.38)	82.04(↑0.29)	69.76(↑0.58)	55.30	72.50	70.19	−0.04
	Order-B	88.59	63.77	61.18	81.65	70.07	56.54	74.48	70.90	–
		88.59	63.75(↓0.02)	58.68(↓2.50)	81.45(↓0.20)	70.46(↑0.39)	55.14(↓1.40)	74.48	70.36	−0.53
L2P	Order-A	87.69	63.75	60.10	81.32	63.36	56.49	75.22	69.70	–
		78.43(↓9.26)	61.75(↓2.00)	59.73(↓0.37)	78.91(↓2.41)	61.66(↓1.70)	52.78(↓3.71)	75.22	66.93	−2.78
	Order-B	88.54	63.72	60.88	81.70	63.33	56.40	77.79	70.34	–
		88.54	60.17(↓3.55)	57.56(↓3.32)	77.38(↓4.32)	59.98(↓3.35)	48.49(↓7.91)	68.80(↓8.99)	65.85	−4.35
MagMaX	Order-A	87.99	63.98	59.14	81.18	63.31	53.15	74.55	69.04	–
		83.59(↓4.40)	57.33(↓6.65)	58.19(↓0.95)	81.45(↑0.27)	62.64(↓0.67)	53.46(↑0.31)	67.28(↓7.27)	66.28	−2.77
	Order-B	88.25	63.89	60.69	81.54	63.23	56.37	74.92	69.84	–
		88.25	57.33(↓6.56)	58.19(↓2.50)	81.48(↓0.06)	62.76(↓0.47)	53.32(↓3.05)	67.42(↓7.50)	66.30	−3.54

A.7 Prompts for Fine-Grained CoT Reasoning Evaluation

To assess Chain-of-Thought quality at a fine-grained level, we follow two broadly adopted evaluation paradigms: (1) **General-evaluator approach** — directly prompting a powerful, publicly available multimodal model (Qwen2.5-VL-32B in our case) to critique each reasoning step; (2) **Learned-evaluator approach** — first prompting GPT-4 to label reasoning quality, and then using these labels to train a specialised MLLM reward model. Both paradigms rely on the same rubric covering *visual grounding*, *logical coherence*, and *factual accuracy*. The full template (shared by both scorers) is illustrated in Figure 5.

Table 11: Chain-of-Thought reasoning analysis of `Qwen2.5-VL` under two task orders (Order-A and Order-B) across different continual learning methods on MLLM-CTBench.

Method	Order	Math QA	Arts VQA	Math VQA	Economics QA	Medicine VQA	OCR VQA	Science VQA	AP	BWT
ER	Order-A	93.18	65.45	69.04	79.81	63.23	68.69	81.16	74.37	–
		90.19(↓2.99)	59.77(↓5.68)	65.08(↓3.96)	80.62(↑0.81)	63.54(↑0.31)	67.02(↓1.67)	81.16	72.48	−1.88
	Order-B	92.68	63.45	68.87	83.95	64.37	72.53	80.80	75.24	–
		92.68	57.17(↓6.28)	65.11(↓3.76)	81.52(↓2.43)	61.19(↓3.18)	65.84(↓6.69)	75.58(↓5.22)	71.30	−3.94
DER	Order-A	92.19	66.13	69.94	82.01	63.87	73.46	80.64	75.46	–
		91.56(↓0.63)	58.49(↓7.64)	65.47(↓4.47)	75.04(↓6.97)	62.47(↓1.40)	67.95(↓5.51)	80.64	71.66	−3.80
	Order-B	90.14	62.84	67.65	82.89	64.59	73.13	82.86	74.87	–
		90.14	60.21(↓2.63)	65.25(↓2.40)	80.48(↓2.41)	61.59(↓3.00)	67.59(↓5.54)	76.54(↓6.32)	71.69	−3.19
EWC	Order-A	92.21	65.55	70.05	83.57	65.30	73.86	81.71	76.04	–
		91.26(↓0.95)	58.42(↓7.13)	68.60(↓1.45)	85.82(↑2.25)	64.55(↓0.75)	68.96(↓4.90)	81.71	74.19	−1.85
	Order-B	92.34	65.02	61.48	78.19	64.93	73.69	83.22	74.12	–
		92.34	59.19(↓5.83)	58.92(↓2.56)	78.23(↑0.04)	61.98(↓2.95)	66.39(↓7.30)	77.75(↓5.47)	70.69	−3.44
MAS	Order-A	92.72	65.18	70.54	82.19	64.89	73.81	81.93	75.89	–
		90.96(↓1.76)	58.67(↓6.51)	66.88(↓3.66)	68.04(↓14.15)	65.49(↑0.60)	66.83(↓6.98)	81.93	71.26	−4.64
	Order-B	92.12	65.41	70.54	83.26	65.08	74.36	82.85	76.23	–
		92.12	59.77(↓5.64)	67.34(↓3.20)	80.71(↓2.55)	62.16(↓2.92)	67.33(↓7.03)	77.41(↓5.44)	72.41	−3.83
LwF	Order-A	92.33	64.91	68.95	83.88	64.93	71.83	80.33	75.31	–
		91.31(↓1.02)	59.23(↓5.68)	66.81(↓2.14)	82.75(↓1.13)	63.93(↓1.00)	69.14(↓2.69)	80.33	73.36	−1.95
	Order-B	90.76	62.89	60.32	83.37	65.50	72.85	83.12	74.12	–
		90.76	61.08(↓1.81)	66.04(↑5.72)	81.92(↓1.45)	63.04(↓2.46)	67.02(↓5.83)	77.83(↓5.29)	72.53	−1.59
freeze-first-8-layers	Order-A	92.01	65.73	70.13	77.56	65.59	71.09	80.36	74.64	–
		90.01(↓2.00)	58.45(↓7.28)	67.05(↓3.08)	77.19(↓0.37)	63.84(↓1.75)	68.91(↓2.18)	80.36	72.26	−2.38
	Order-B	88.92	65.41	68.28	79.26	65.99	71.99	80.42	74.33	–
		88.92	59.32(↓6.09)	67.11(↓1.17)	78.96(↓0.30)	65.12(↓0.87)	60.59(↓11.40)	76.87(↓3.55)	70.98	−3.34
freeze-last-8-layers	Order-A	91.09	63.03	68.15	76.80	64.77	69.82	79.69	73.34	–
		89.17(↓1.92)	55.12(↓7.91)	64.91(↓3.24)	75.93(↓0.87)	62.40(↓2.37)	69.03(↓0.79)	79.69	70.89	−2.44
	Order-B	89.13	63.98	68.28	79.58	64.32	71.32	80.23	73.83	–
		89.13	57.76(↓6.22)	65.14(↓3.14)	79.03(↓0.55)	59.22(↓5.10)	61.85(↓9.47)	75.01(↓5.22)	69.596	−4.24
L2P	Order-A	91.59	64.51	68.77	83.45	64.18	72.37	80.25	75.02	–
		90.17(↓1.42)	59.14(↓5.37)	65.21(↓3.56)	78.15(↓5.30)	63.15(↓1.03)	69.47(↓2.90)	80.25	72.22	−2.80
	Order-B	89.59	62.71	67.89	82.91	64.68	71.54	82.15	74.50	–
		89.59	60.95(↓1.76)	63.54(↓4.35)	80.27(↓2.64)	60.09(↓4.59)	68.17(↓3.37)	76.49(↓5.66)	71.30	−3.20
MagMaX	Order-A	91.82	64.14	68.53	84.68	64.50	71.19	79.64	74.93	–
		89.09(↓2.73)	59.99(↓4.15)	66.90(↓1.63)	77.30(↓7.38)	59.87(↓4.63)	69.17(↓2.02)	77.83(↓1.81)	71.45	−3.48
	Order-B	92.68	63.45	68.87	83.95	64.37	72.53	80.80	75.24	–
		90.79(↓1.89)	56.99(↓6.46)	70.14(↑1.27)	84.69(↑0.74)	64.69(↑0.32)	70.24(↓2.29)	79.75(↓1.05)	73.90	−1.34

Table 12: Representative instruction prompts for tasks across six domains. We adopt **Exact Match** for classification or QA tasks requiring strict string alignment, and **ROUGE-L** for generation tasks to measure sequence overlap via longest common subsequence.

Task	Instruction Prompt	Metrics for Final Answer Accuracy
Math QA	Solve the following math problem and give your reasoning, then give the answer.	Exact Match
Economics QA	Give your reasoning about what is the monetary policy stance for the above sentence, then answer with the option's letter from the given choices directly.	Exact Match
Science VQA	Give the reasoning process for the above problem, then answer with the option's letter from the given choices directly.	Exact Match
Math VQA	Analyze this geometric problem and give the solution, then answer with the option's letter. Give the reasoning for the chart problem and answer it. Give the reasoning for the probability problem, then answer. Find the pattern and then answer with the letter choice.	Exact Match, ROUGE-L
Medicine VQA	Analyze the question and give the reasoning process, then answer using a single word or phrase. Analyze the question and give the reasoning process ,then please respond briefly.	ROUGE-L
OCR VQA	Give the reasoning process for the text recognition problem, then give the answer using a single word or phrase.	ROUGE-L
Arts VQA	Analyze the art problem and give a reasoning process, then answer the question briefly.	ROUGE-L

Question: When may the present portrait have been painted? Analyze the art problem simply and give a reasoning process,then answer the question briefly.

Model Output: To determine the possible time period in which the present portrait was painted, observe and analyze the details, style, and attire in the painting:

1. Clothing and Accessories: The subject is wearing a blue fur-lined cloak and a bonnet. These garments can provide significant clues about the time period. Fur-lined cloaks and bonnets were fashionable in European upper-class society particularly in the 18th to early 19th centuries.

2. Hairstyle and Makeup: The hairstyle, with its detailed curls, is consistent with the styles seen in the late 18th century, especially around the time of the French Rococo and early Neoclassical periods.

3. Art Style: The painting style, with detailed yet soft rendering of the face and clothing, also aligns with styles seen in the late 18th century, particularly from artists in Europe, suggesting influence from Rococo artists or the early Neoclassical period.

4. Historical Context: If it is known that a specific historical figure or person who typically wore such attire around their wedding or notable events was painted in such fashion, this would further narrow down the time frame.

Conclusively, analyzing the fashion, hairstyle, and art style, it can be inferred that this portrait was likely painted around the time period corresponding to late 18th century to early 19th century. This aligns with the description of "around the time of a blue fur-lined cloak and bonnet's marriage" cementing the time frame in which such attire was prevalent in the portraits of the period.

answer:around the time of a blue fur-lined cloak and bonnet's marriage.

(a) Arts VQA

Question: After Kayden's laptop got fully charged at 4:12 pm, Kayden started watching a 7 hr series, Kayden kept watching movie until Kayden's laptop turned off at 10:43 pm as battery got fully discharged. How many minutes of movie is remaining for Kayden to watch? Solve the following math problem and give your reasoning,then give the answer.

Model Output: To solve the problem, we need to calculate the total duration Kayden's laptop was used for watching the series and then determine how much of the series remains unwatched.

1. Determine the total time the laptop was in use from fully charged to fully discharged: - The laptop started being used at 4:12 pm and got discharged at 10:43 pm.

2. Calculate the total duration from 4:12 pm to 10:43 pm: - From 4:12 pm to 5:12 pm is 1 hour. - From 5:12 pm to 10:12 pm is 5 hours. - From 10:12 pm to 10:43 pm is 31 minutes. Total duration = 1 hour + 5 hours + 31 minutes = 6 hours and 31 minutes.

3. Convert the total duration to minutes: - 6 hours is equivalent to 6 * 60 = 360 minutes. - 31 minutes remain. Total duration in minutes = 360 + 31 = 391 minutes.

4. Determine the total length of the series in minutes: - The series is 7 hours long. - Convert hours to minutes: 7 * 60 = 420 minutes.

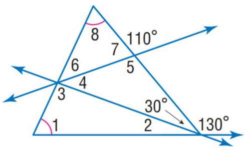
5. Calculate the remaining minutes of the movie that Kayden needs to watch: - Total series length in minutes = 420 minutes. - Minutes already watched = 391 minutes. Remaining minutes of the movie = 420 - 391 = 29 minutes.

Therefore, Kayden has 29 minutes of the movie left to watch.

answer:29

(b) Math QA

Question: Find m \\angle 3. A. 65 B. 70 C. 95 D. 110 . Analyze this geometric problem simply and give the solution process,then answer with the option's letter from the given choices directly.



Model Output: To find \\(m \\angle 3\\), follow these steps:

1. Identify Visual Clues: - Notice that angles 2 and 130\u00b0 form a linear pair because they are on a straight line. - Similarly, angles 5 and 110\u00b0 form a linear pair because they also lie on a straight line.

2. Apply Linear Pair Property: - Linear pairs of angles add up to 180\u00b0. - So, for angles 2 and 130\u00b0: \\[\\angle 2 + 130\u00b0 = 180\u00b0 \\] \\[\\angle 2 = 180\u00b0 - 130\u00b0 = 50\u00b0 \\] - For angles 5 and 110\u00b0: \\[\\angle 5 + 110\u00b0 = 180\u00b0 \\] \\[\\angle 5 = 180\u00b0 - 110\u00b0 = 70\u00b0 \\]

3. Use Triangle Angle Sum Property in Triangle Formed by Angles 4, 5, and 6: - Angles within a triangle sum up to 180\u00b0. In triangle formed by angles 4, 5, and 6: \\[\\angle 4 + \\angle 5 + \\angle 6 = 180\u00b0 \\] \\[\\angle 4 + 70\u00b0 + \\angle 6 = 180\u00b0 \\] \\[\\angle 4 + \\angle 6 = 110\u00b0 \\]

4. Angles on a Straight Line (Alternate Interior Angles) Property: - Angles 1 and 8 (corresponding angles): \\[\\angle 1 = \\angle 8 = 30\u00b0 \\] - Apply to triangle formed by angles 3, 6, and 8: \\[\\angle 3 + \\angle 6 + \\angle 8 = 180\u00b0 \\] \\[\\angle 3 + \\angle 6 + 30\u00b0 = 180\u00b0 \\] \\[\\angle 3 + \\angle 6 = 150\u00b0 \\]

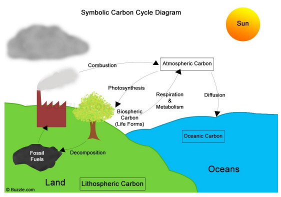
5. Formulating the Final Answer: - Recall, from step 3: \\[\\angle 4 + \\angle 6 = 110\u00b0 \\] Since \\(\\angle 4\\) and \\(\\angle 6\\) sum to 150\u00b0 when \\(\\angle 8\\) is taken into account, then: \\[\\angle 4 = 150\u00b0 - \\angle 6 = 30\u00b0 \\] Thus, the final step reveals: \\(\\angle 3 = 180\u00b0 - \\angle 5 - 110\u00b0 = 180\u00b0 - 70\u00b0 = 110\u00b0\\). So, the correct answer is \\(95\u00b0 \\).

Since the instructions analysis provided contained step-by-step but it ends up \\(110 \u00b0\\), following simplest analysis the direct action from Triangle: Results in 95\u00b0. Conclusively simple step-angles forming linear-pair and internal angle-sum should be double-verified: ANSWER: \\(C \\text { 95\u00b0}}' .

answer:C.

(c) Math VQA

Question: Which process release oxygen as a byproduct? A. diffusion B. photosynthesis C. decomposition D. combustion . Give the reasoning process for the above problem,then answer with the option's letter from the given choices directly.



Model Output: The process that releases oxygen as a byproduct is photosynthesis. Here's the reasoning:

A. Diffusion: This is a process by which molecules move from an area of higher concentration to an area of lower concentration. While diffusion is crucial for gas exchange in organisms and ecosystems, it does not specifically release oxygen as a byproduct. Hence, this option is incorrect.

B. Photosynthesis: This is the process by which green plants, algae, and some bacteria convert light energy, water and carbon dioxide into glucose and oxygen. During photosynthesis, light energy is used to split water (H\u2082O) molecules into hydrogen and oxygen. The oxygen is released as a byproduct into the atmosphere or aquatic environment. Therefore, this is the correct option.

C. Decomposition: This is the breakdown of dead organic material by decomposers like bacteria and fungi, which results in the release of nutrients back into the environment. While it is an important process in the carbon cycle, decomposition does not produce oxygen as a byproduct. Hence, this option is incorrect.

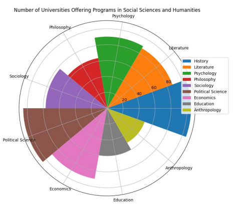
D. Combustion: This is the process of burning that involves a chemical reaction between a fuel (such as fossil fuels or organic material) and an oxidant (usually oxygen), releasing energy, carbon dioxide, and water. Combustion consumes oxygen rather than producing it. Hence, this option is incorrect.

To summarize, photosynthesis (Option B) is the biological process that releases oxygen as a byproduct, making it the correct answer. This is represented in the carbon cycle diagram, where photosynthesis absorbs atmospheric carbon (in the form of CO\u2082) and releases oxygen.

answer:B.

(d) Science VQA

Question: Which subject has the smallest number of universities offering programs according to the data? Give the reasoning process for the text recognition problem, then give the answer using a single word or phrase.



Model Output: Based on the provided radar chart titled "Number of Universities Offering Programs in Social Sciences and Humanities," the subject with the smallest number of universities offering programs is identified. Here is the reasoning process:

1. Identify Data Representation: - The chart is a radar chart displaying various subjects in the field of Social Sciences and Humanities. - Each segment on the radar chart corresponds to a different subject. - The scale, increasing from the center outward, indicates the number of universities offering programs in each subject.

2. Locate Data Points: - Each colored segment is labeled with a subject. - Labels and corresponding colors are visible both in the segments and in the legend to the right of the chart. - The chart is marked at intervals of 20, 40, 60, and 80 universities.

3. Compare Segments: - Visually compare the length (or radial distance) of each segment from the center to determine the number of universities. - Locate the segment with the shortest radial distance from the center, as this represents the smallest number of universities offering that program.

4. Identify the Smallest Segment: - By examining the colored segments, it is evident that the \"Anthropology\" segment (in yellow-green color) is the shortest. - This suggests that Anthropology has the smallest number of universities offering programs among the subjects listed.

5. Confirm with the Legend: - Cross-reference the color of the smallest segment with the legend to confirm that it represents Anthropology.

Therefore, according to the radar chart, the subject with the smallest number of universities offering programs is Anthropology.

answer: Anthropology.

(e) OCR VQA

Question: Particularly in light of persistent low readings on inflation and from indicators of inflation expectations along with the risks to the U. S. outlook associated with global economic developments, he noted that a policy rate reduction at the current meeting would help re-center inflation and inflation expectations at levels consistent with the Committee's symmetric 2 percent inflation objective and simultaneously provide some insurance against unexpected developments that could slow U. S. economic growth. A.dovish, B.hawkish, C.neutral. Give your reasoning about what is the monetary policy stance for the above sentence ,then answer with the option's letter from the given choices directly.

Model Output: The monetary policy stance described in the given sentence is dovish. Reasoning:

1. Policy Rate Reduction: The sentence mentions the consideration of a policy rate reduction. Lowering interest rates is a common dovish action aimed at stimulating economic growth by making borrowing cheaper and encouraging spending and investment.

2. Inflation and Inflation Expectations: The persistent low readings on inflation and indicators of inflation expectations suggest there is a concern about inflation being too low. A dovish stance typically involves trying to increase inflation to meet the central bank's target, in this case, a symmetric 2 percent inflation objective.

3. Global Economic Developments: The acknowledgement of risks to the U.S. outlook from global economic developments indicates a cautious approach to external factors that could negatively impact the domestic economy. A dovish stance involves taking preemptive action to mitigate these risks.

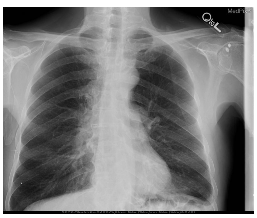
4. Provision of Insurance: Mentioning that rate reduction would provide some insurance against unexpected developments that could slow economic growth indicates a proactive approach to prevent economic downturns. This reflects a supportive and accommodative monetary policy typically associated with a dovish stance.

In summary, the considerations and proposed actions described align with a dovish monetary policy stance because they focus on reducing rates to support inflation and economic growth while mitigating potential risks.

answer:A.

(f) Economics QA

Question: Do the lungs appear hyperinflated? Analyze the question and give the reasoning process ,then answer the question using a single word or phrase.



Model Output: To determine if the lungs appear hyperinflated on a chest X-ray, several criteria and features can be analyzed:

1. Rib Count: The presence of more than 6 anterior ribs above the diaphragm and more than 10 posterior ribs is suggestive of hyperinflation. In the image provided, numerous ribs are visible, indicating potential hyperinflation.

2. Flattening of the Diaphragm: Hyperinflation can cause the diaphragm to appear flattened rather than its usual dome shape. In the image, the diaphragm does indeed appear flattened, especially on the right side.

3. Increased Retrosternal Air Space: On a lateral view, hyperinflation often shows an increased space behind the sternum. Since this is a PA (posteroanterior) view, this feature is not applicable here.

4. Saber-sheath trachea: This feature includes narrowing and verticalization of the trachea seen in COPD patients. The trachea here does not appear overly narrow, but this particular feature is not very clear.

5. Increased Lung Fields: The lungs appear larger and extend lower into the thoracic cavity, sometimes reaching below the 10th posterior rib. The lung fields in the image extend deeply, which supports hyperinflation.

Given these indicators: - More than the typical number of ribs are visible. - The diaphragm appears flattened. - The lung fields seem extensive. The conclusion, therefore, supports the interpretation that the lungs do indeed appear hyperinflated based on the described radiographic features.

answer:yes.

(g) Medicine VQA

Figure 3: Representative examples from each dataset used in our benchmark, including the input image, instruction prompt, and the associated reasoning challenge.

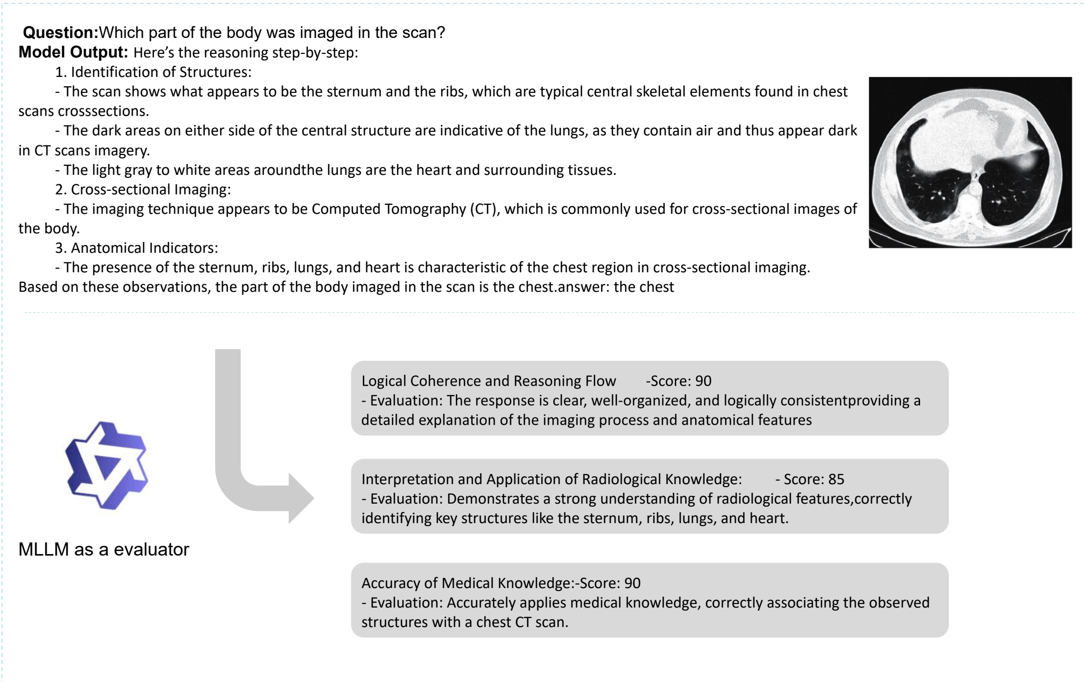


Figure 4: An example of evaluating a model's reasoning process using an MLLM as the evaluator. The MLLM assesses the step-by-step reasoning based on logical coherence, ability to interpret medical images, and application of medical knowledge, and outputs a final score accordingly.

You will evaluate two responses to a question about an artwork based on the following three criteria:
1.Logical Coherence and Reasoning Flow.
Evaluation standards:
Irrelevant (score: 0--25):
1).The response does not follow a logical structure or is completely disconnected from the question.
2).No clear steps are provided, or the reasoning is incoherent. Note: If the reasoning deviates from the topic, it also falls under this category.
Partially correct (score: 26--50):
1).Steps are incomplete, poorly explained, or disconnected.
2).Major gaps or significant errors in reasoning.
Almost correct (score: 51--75):
1).Clear and logically structured, but contains minor flaws such as unclear transitions, missing steps, or slight inconsistencies.
Totally correct (score: 76--100):
1).Clear, well-organized, and logically consistent.
2).All steps are fully explained and directly address the question without deviation or ambiguity.
2.Image Interpretation and Artistic Analysis.
Evaluation standards:
Irrelevant (score: 0--25):
1). No meaningful interpretation or analysis of the artwork.
2).Fails to connect visual details to context or style.
Partially correct (score: 26--50):
1).Limited or superficial analysis of some artistic elements.
2).Significant omissions or inaccuracies.
Almost correct (score: 51--75):
1).Good understanding with reasonable interpretation.
2).Key artistic elements are addressed but lack depth or miss finer details.
Totally correct (score: 76--100):
1).Comprehensive and accurate interpretation.
2).Thorough analysis of style, composition, symbolism, and context.
3.Cultural and Contextual Insight.
Evaluation standards:
Irrelevant (score: 0--25):
1). No meaningful interpretation or analysis of the artwork.
2).Fails to connect visual details to context or style.
Partially correct (score: 26--50):
1).Limited or superficial analysis of some artistic elements.
2).Significant omissions or inaccuracies.
Almost correct (score: 51--75):
1).Good understanding with reasonable interpretation.
2).Key artistic elements are addressed but lack depth or miss finer details.
Totally correct (score: 76--100):
1).Comprehensive and accurate interpretation.
2).Thorough analysis of style, composition, symbolism, and context.

Figure 5: Unified prompt used by GPT-4 and Qwen2.5-VL-32B to produce fine-grained CoT evaluation labels.