Overcoming Catastrophic Forgetting in Visual Continual Learning with Reinforcement Fine-Tuning

Meng Lou, Hanzhong Guo, Linwei Chen, Yizhou Yu*
School of Computing and Data Science, The University of Hong Kong
{ loumeng@connect.hku.hk, hanzhong@connect.hku.hk,
chenlinwei.ai@gmail.com, yizhouy@acm.org }

Abstract

Reinforcement Fine-Tuning (RFT) with verifiable rewards has emerged as a powerful learning paradigm for eliciting reasoning capabilities in Multi-modal Large Language Models (MLLMs). Recent studies suggest that Reinforcement Fine-Tuning (RFT) is inherently more resilient to catastrophic forgetting than Supervised Fine-Tuning (SFT). However, whether RFT (e.g., GRPO) can effectively overcome forgetting in challenging visual continual learning settings, such as class-incremental learning (CIL) and domain-incremental learning (DIL), remains an open problem. Through a pilot study, we confirm that while RFT consistently outperforms SFT, it still suffers from non-negligible forgetting. We empirically trace this bottleneck to Trajectory-level Drift Agnosticism: among candidate rollouts achieving identical task rewards, the KL divergence from the preceding-task policy varies substantially, which strongly correlates with catastrophic forgetting across sequential tasks. Motivated by this insight, we propose Retention-aware Policy Optimization (RaPO), a simple yet effective RFT method that explicitly mitigates forgetting through trajectory-level reward shaping. Specifically, RaPO comprises two core components: (1) Retention Reward that converts trajectory-level distribution drift into a continuous reward signal, preferentially reinforcing knowledge-preserving rollouts within each group; (2) Cross-Task Advantage Normalization (CTAN), which maintains a persistent exponential moving average of reward statistics across task boundaries to stabilize the optimization progress during continual learning. Leveraging the free-form textual generalization of MLLMs, we comprehensively evaluate RaPO across five visual continual learning settings. Extensive experiments demonstrate that RaPO achieves leading performance, substantially reducing catastrophic forgetting while preserving strong plasticity. To the best of our knowledge, this work represents the first systematic exploration of RFT in visual continual learning, offering insights that we hope will inspire future research. Code is publicly accessible at: https://github.com/LMMMEng/RaPO.

1 Introduction

Reinforcement Fine-Tuning (RFT) with verifiable rewards [1, 2, 3, 4, 5, 6] has demonstrated remarkable progress in eliciting reasoning capabilities in Multi-modal Large Language Models (MLLMs), outperforming classical Supervised Fine-Tuning (SFT). GRPO [7] stands as a representative work that leverages verifiable RFT to train powerful large reasoning models [8, 9]. Motivated by this success, several subsequent works [10, 11, 12, 13] have demonstrated that RFT can effectively improve vision tasks even under few-shot training regimes.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

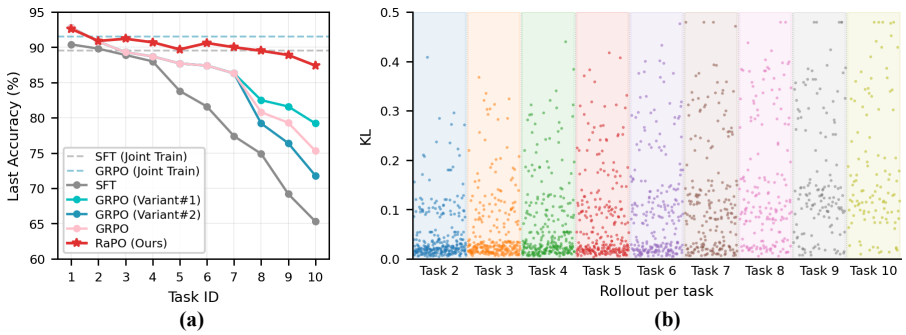


Figure 1: **(a)** RFT clearly outperforms SFT in rehearsal-free CIL, but still suffers from significant forgetting. **(b)** Among equally-rewarded rollouts, KL divergence from the policy in the preceding task varies substantially, and this difference enlarges as tasks progress.

However, real-world applications frequently encounter streaming data within continuously evolving environments [14], requiring large models to continually adapt to newly arriving data without suffering from catastrophic forgetting [15]. Although numerous SFT-centered approaches [16, 17, 18] have been developed for continual learning, recent studies [19, 20] have revealed that RFT is naturally more resilient to catastrophic forgetting than SFT. This property stems from the fact that on-policy learning in RFT implicitly biases the optimization toward solutions residing in low-drift distribution spaces, whereas SFT is prone to converge on solutions within arbitrary distribution drift [19]. Nevertheless, the efficacy of RFT in challenging visual continual learning, such as class-incremental learning (CIL) [21] and domain-incremental learning (DIL) [22], remains an open problem.

To investigate this, we conduct a pilot study on a challenging rehearsal-free few-shot CIL setting on the widely adopted ImageNet-R dataset [23]. Specifically, 200 image classes are randomly split over 10 non-overlapping tasks, with 20 classes and only 5 labeled examples per class at each task. We compare GRPO and SFT based on the Qwen2-VL-2B model [24], together with a joint-training upper bound. As shown in Figure 1 (a), GRPO consistently outperforms SFT, confirming that the forgetting resilience of RFT [19, 20] successfully transfers to visual tasks. Nevertheless, **GRPO still suffers from non-negligible forgetting, demonstrating that it remains insufficient to mitigate stability-plasticity tension in challenging visual continual learning**.

To further explore the rationale of this phenomenon, we analyze the trajectory-level learning patterns of GRPO during CIL. Specifically, we measure the distribution drift of each rollout generated by the current policy π_t as its token-level KL divergence from the frozen preceding-task policy π_{t-1}. We focus only on rollout groups with maximal task reward so that the comparison isolates drift differences among equal-reward trajectories. As illustrated in Figure 1 (b), **candidate rollouts achieving the same task reward exhibit vastly different KL divergence values**, a phenomenon we term **trajectory-level drift agnosticism**. For instance, two distinct trajectories generated from the same input can solve the current task equally well, but exhibit entirely different magnitudes of distributional drift. This discrepancy is increasingly pronounced as the task sequence progresses, which clearly correlates with the accuracy degradation trend shown in Figure 1 (a). This suggests that purely task-reward-driven behavior leads to drift-agnostic credit assignment, which may contribute to severe forgetting.

To validate this hypothesis, we design two simple GRPO variants that intervene on all-correct rollout groups in the later tasks, when forgetting becomes pronounced. Specifically, Variant#1 assigns zero reward to any rollout whose KL exceeds the group mean, thereby retaining positive reinforcement only for low-drift trajectories. Conversely, Variant#2 mirrors this operation, preserving positive rewards solely for heavily drifted rollouts. As demonstrated in Figure 1 (a), these variants exhibit opposed behaviors: Variant#1 effectively mitigates forgetting, whereas Variant#2 significantly exacerbates it. Collectively, these results validate trajectory-level drift agnosticism as a key empirical phenomenon in vanilla GRPO, i.e., **trajectory-level drift with respect to the preceding-task policy is an actionable signal closely tied to forgetting**. However, these hard-thresholding variants are impractical. First, binary gating collapses fine-grained credit assignment required for complex tasks such as dense predictions. Second, it is unstable under small rollout numbers, where even marginal KL differences may trigger winner-take-all updates.

Driven by the above observations, we propose Retention-aware Policy Optimization (RaPO), a simple yet effective RFT method that mitigates catastrophic forgetting through trajectory-level reward shaping. RaPO consists of two complementary components. First, a Retention Reward converts the trajectory-level drift from the preceding-task policy into a dense reward signal: rollouts that stay closer to the preceding policy receive proportionally higher rewards. This design differs fundamentally from standard GRPO, i.e., when two rollouts achieve comparable task rewards but exhibit different degrees of drift, RaPO explicitly reinforces the knowledge-preserving one, steering the policy toward regions that adapt to new data while remaining anchored to previously acquired knowledge. Second, Cross-Task Advantage Normalization (CTAN) maintains a persistent smoother of the reward scale, preventing the abrupt advantage fluctuations that arise from sharp reward-distribution shifts at task boundaries. Together, the Retention Reward directs credit assignment toward low-drift trajectories, while CTAN smoothly stabilizes its scale across the continual learning stream.

Leveraging the free-form textual generalization capabilities of MLLMs, our method has been comprehensively evaluated across a diverse suite of visual continual learning tasks, including class-incremental image classification, domain-incremental image classification, class-incremental object detection, domain-incremental object detection, and class-incremental video classification. Extensive experimental results in Section 4 demonstrate the promising performance and generalization capacity of RaPO. Overall, our goal is not to chase state-of-the-art performance on different benchmarks, but to systematically explore the potential of verifier-based RFT for visual continual learning. We hope this work will stimulate further research on RFT-based continual learning.

2 Related Work

Reinforcement Fine-Tuning (RFT) has demonstrated a superior capacity to incentivize reasoning capabilities in LLMs [25]. A foundational paradigm is reinforcement learning with human feedback, which aligns model outputs with human preferences [26, 27, 28, 29]. Recently, the research has increasingly shifted toward reinforcement learning with verifiable rewards. In particular, the remarkable success of Group Relative Policy Optimization (GRPO) [7, 8] has motivated a surge of research exploring this paradigm further, such as Dr.GRPO [30], DAPO [31], and GSPO [32]. Subsequently, many works [10, 12, 13, 33] have also demonstrated that RFT can significantly improve visual tasks over SFT by activating reasoning capabilities. Additionally, recent studies [19, 20] suggest that RFT is naturally more resistant to catastrophic forgetting than SFT when adapting to new data. Along this direction, a data-perspective study [34] reveals that RFT preserves prior knowledge better than SFT because it reinforces correct samples naturally aligned with the base model's probability landscape. CGL [35] further adapts RFT to continual GUI learning, employing policy entropy and gradient surgery strategies to prevent the overwriting of previously acquired interaction skills.

Visual Continual Learning. Recent years have witnessed rapid progress in diverse visual tasks [36, 37, 38, 39, 40, 41, 42]. However, visual continual learning remains a long-standing problem that aims to adapt models to non-stationary visual streams without catastrophic forgetting [22]. Among its various paradigms, rehearsal-free class-incremental learning (CIL) [43, 44] stands out as one of the most representative and challenging settings, which requires a model to continuously adapt to incrementally arriving classes without accessing any historical training data, while simultaneously preserving its recognition capabilities on all observed classes. During CIL, the label spaces across different tasks are strictly disjoint. Another prominent setting is rehearsal-free domain-incremental learning (DIL) [45, 46], which aims to enable a model to sequentially adapt to new domains while ensuring its previously acquired knowledge is not catastrophically degraded by domain shifts. Unlike CIL, different tasks in DIL share the same label space but exhibit distinct domain distributions. Existing progress [47, 48, 49, 50] has been driven primarily by SFT-based adaptation of vision-centric models. One of the most prevalent paradigms involves incrementally appending parameter-efficient modules [51, 52, 53, 54, 55, 56] into pre-trained models such as ViT [57] and CLIP [58], demonstrating promising results. Nevertheless, these methods are typically centered on a single scenario, such as class-incremental image classification, rather than a unified model that is able to handle diverse settings simultaneously, including class-incremental image/video classification and dense predictions. On the other hand, since real-world scenarios frequently lack abundant, high-quality annotations for each incremental task, vision models with SFT may be prone to overfitting under data-scarce conditions.

In this work, we explore the untapped potential of RFT for challenging visual continual learning paradigms. Without bells and whistles, our proposed RaPO achieves leading performance and strong generalizability compared with different baselines. To the best of our knowledge, this work is the first to systematically explore the potential of RFT in visual continual learning.

3 Method

3.1 Preliminaries

GRPO [7] optimizes a policy π_θ using group-relative advantages instead of a learned value function. For a given textual prompt x, GRPO samples a rollout group $\mathcal{G}(x) = \{y_1, \ldots, y_n\}$ of n trajectories from the current policy π_θ. Each rollout y_i receives a task-specific verifiable reward $R_{\text{task}}(y_i)$. These rewards are centered and rescaled within the group to produce the group-relative advantage A_i:

$$A_i \;=\; \frac{R_{\text{task}}(y_i) - \mu_{\text{group}}}{\sigma_{\text{group}} + \epsilon} \tag{1}$$

where μ_{group} and σ_{group} are the mean and standard deviation of the rewards within $\mathcal{G}(x)$, respectively. Then, the policy is updated using the clipped surrogate objective [26], where the standard advantage based on the value-function is replaced by the group-relative advantage A_i.

RFT in Vision. To address visual continual learning, the input is defined as a multi-modal signal $x = (v, l)$, where v is a visual input (image or video) and l is a textual instruction template specifying the task. The output y_i is a textual response generated by MLLM. Following the recent paradigm [10], we employ diverse verifiable reward functions R_{task} depending on the task type, such as accuracy reward for image/video classification and IoU reward for object detection. More details concerning prompt templates and reward formulations are provided in Appendix B.

3.2 Retention-aware Policy Optimization

3.2.1 Overview

We study rehearsal-free visual continual learning (i.e., CIL and DIL) across a sequential stream of tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_N\}$. For an arriving task $\mathcal{T}_t$, the learner is optimized using only the training data from $\mathcal{T}_t$, with no access to historical training data from $\mathcal{T}_1$ to $\mathcal{T}_{t-1}$. The overall learning pipeline is simple. Specifically, at the onset of task $\mathcal{T}_t$, we initialize the actor policy π_t using the weights saved at the end of $\mathcal{T}_{t-1}$, while simultaneously maintaining a frozen copy of it to serve as the anchor policy π_{t-1}. During each optimization iteration, the actor π_t generates a group of candidate rollouts for a given multi-modal input $x = (v, l)$. Each rollout is then evaluated across two aspects: 1) The primary task reward is computed by a task-specific verifier. 2) A new retention reward is estimated with a trajectory-level drift metric against the anchor π_{t-1}. These two signals are aggregated into a unified objective to update π_t, explicitly steering the policy to explore parameter spaces that jointly maximize proficiency in new tasks and historical knowledge preservation. Concurrently, to counter the optimization instability caused by abrupt reward-distribution shifts when transitioning across task boundaries, a simple Cross-Task Advantage Normalization (CTAN) mechanism is introduced to regulate the scale of the credit assignment.

3.2.2 Retention Reward

As empirically validated in the Section 1, vanilla GRPO suffers from trajectory-level drift agnosticism. When the actor policy π_t adapts to a new task, the trajectory-level distribution drift from the anchor policy π_{t-1} is strongly correlated with the catastrophic forgetting of previously acquired knowledge. This inspires us to explicitly formulate this trajectory-level drift into a continuous reward signal. Specifically, let $y_i = (y_{i,1}, \ldots, y_{i,m_i})$ denote a rollout sample of length m_i sampled from π_t, where $i \in \{1, \ldots, n\}$ indexes the rollouts within the group and $s \in \{1, \ldots, m_i\}$ indexes the token positions. Suppose $y_{i,<s} = (y_{i,1}, \ldots, y_{i,s-1})$ represent the generated prefix up to step s, the trajectory-level distribution drift is calculated as:

$$\bar{D}_{\text{drift}}(y_i) \;=\; \max\Big(\frac{1}{m_i}\sum_{s=1}^{m_i}\left[\log \pi_t(y_{i,s} \mid x, y_{i,<s}) - \log \pi_{t-1}(y_{i,s} \mid x, y_{i,<s})\right], 0\Big) \tag{2}$$

There are two remarks on Equation (2) that warrant highlighting. First, the bracketed term computes the per-token log-probability ratio between the actor π_t and the anchor π_{t-1}. Averaging this ratio over the m_i generated tokens provides a length-normalized Monte Carlo estimate of the trajectory-level distribution drift evaluated along the sampled trajectory. This length normalization is crucial, as it ensures that $\bar{D}_{\text{drift}}(y_i)$ remains strictly comparable across candidate rollouts of varying lengths. Second, the outer $\max(\cdot, 0)$ applies a one-sided truncation. Specifically, a negative pre-truncation value indicates that the actor π_t has become less confident in the generated trajectory than the anchor π_{t-1}, signifying that π_t has not specialized toward this trajectory relative to π_{t-1}. Therefore, clamping these negative values to zero explicitly prevents reward hacking in the subsequent reward formulation (Equation (3)), as the actor π_t is possible to intentionally generate low-confidence outputs to inflate its reward.

To seamlessly incorporate this forgetting measurement into the RFT objective, we stop the gradient of $\bar{D}_{\text{drift}}(y_i)$ and convert it into a bounded and positive reward score using an exponentially decaying mapping:

$$R_{\text{ret}}(y_i) \;=\; \exp\big(-\alpha\, \bar{D}_{\text{drift}}(y_i)\big) \in (0, 1] \tag{3}$$

where $\alpha > 0$ is a scaling hyperparameter that controls the sensitivity to distribution drift. Rollouts that remain closer to the anchor π_{t-1} yield a lower $\bar{D}_{\text{drift}}$, thereby translating into a higher R_{ret} score approaching 1.0. Since this reward directly measures the retention of previously learned knowledge, it is termed the retention reward. Afterwards, the retention reward is seamlessly integrated with the task-specific reward via an additive formulation:

$$R_{\text{total}}(y_i) \;=\; R_{\text{task}}(y_i) + \lambda\, R_{\text{ret}}(y_i) \tag{4}$$

where $\lambda > 0$ balances task adaptation and retention. Crucially, R_{ret} enters the composite reward before the group-relative advantage computation. This explicitly changes the rollout ranking inside $\mathcal{G}(x)$: among candidate trajectories that achieve comparable task rewards on $\mathcal{T}_t$, those remaining closer to the anchor π_{t-1} are assigned larger advantages.

Despite its simplicity, this formulation provides three properties: *Continuous:* within a similar-R_{task} sub-group, the assigned advantage is an increasing function of the R_{ret}, ensuring that the anchor-closer rollout is always reinforced more strongly. *Disentangled:* The weighted-additive form limits the influence of retention relative to task reward. Since R_{ret} is strictly bounded, the retention term can change the total reward by at most λ, so its effect remains controlled and is most relevant when candidate trajectories have similar task rewards. *General:* The retention reward can be easily combined with any task reward, such as a binary accuracy reward for image classification and a continuous IoU reward for object detection. Note that retention reward differs fundamentally from standard loss-level KL regularization, as discussed in Section 3.3.

3.2.3 Cross-Task Advantage Normalization

During continual learning, abrupt changes in reward statistics may occur at task transitions. Specifically, near the end of $\mathcal{T}_{t-1}$, many rollouts are similarly correct and the within-batch reward spread σ_{batch} becomes small, which inflates normalized advantages. At the start of $\mathcal{T}_t$, rewards typically become lower and more variable, compressing advantages precisely when fast adaptation is required. This oscillation destabilizes GRPO training across task boundaries.

To stabilize the optimization scale, we replace the per-batch reward standard deviation with a running exponential moving average (EMA) $\hat{\sigma}$ that is persistent across optimization steps and across task boundaries. At every optimization step, after computing the batch-level reward standard deviation σ_{batch} from the current batch of rollout rewards, we perform the update:

$$\hat{\sigma} \;\leftarrow\; \beta\, \hat{\sigma} \;+\; (1 - \beta)\, \sigma_{\text{batch}} \tag{5}$$

where β is the smoothing coefficient (e.g., $\beta = 0.99$). We then compute the advantage as:

$$A_i \;=\; \frac{R_{\text{total}}(y_i) - \mu_{\text{group}}}{\hat{\sigma} + \epsilon} \tag{6}$$

where μ_{group} is still the mean total reward within the rollout group for the same prompt. The numerator preserves within-group ranking, while the denominator is stabilized across batches and across tasks. CTAN is persistent at task boundaries: the state $\hat{\sigma}$ is saved at the end of $\mathcal{T}_{t-1}$ and loaded as the initial EMA value at the beginning of $\mathcal{T}_t$, so the normalization scale does not reset

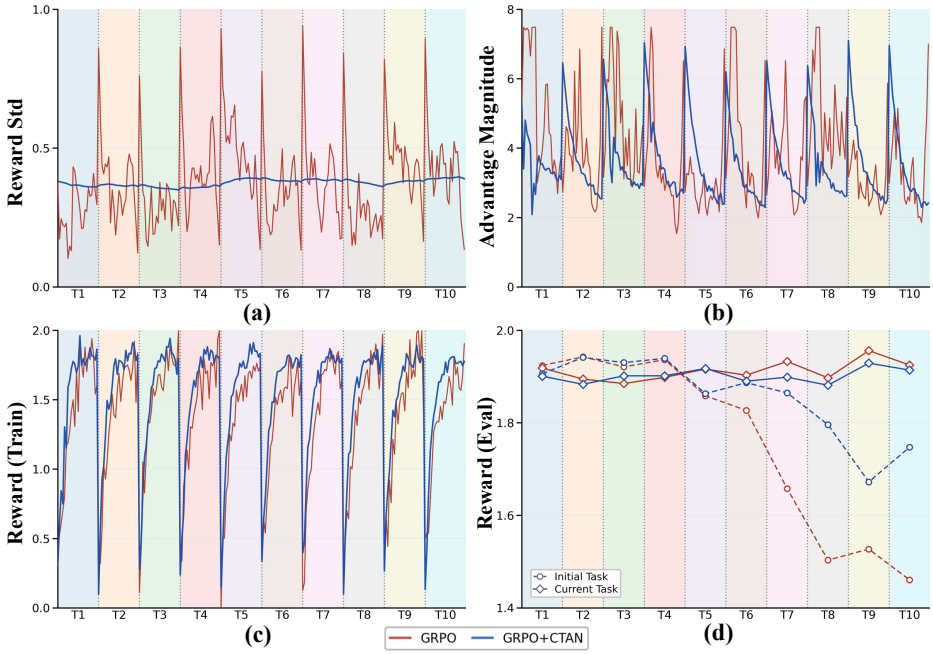


Figure 2: **(a)** GRPO relies on the instantaneous reward standard deviation, which fluctuates sharply across task boundaries, whereas CTAN maintains a persistent EMA normalizer (β=0.99). **(b)** CTAN produces a smoother advantage magnitude (sum of the absolute values of advantages) across the continual learning stream. **(c)** The stabilized advantage scale is accompanied by smoother acquisition of training reward. **(d)** Evaluation reward trends further reflect more robust continual learning behavior, with Initial Task denoting evaluation on task#1 and Current Task denoting evaluation on the task currently being learned. *Overall, CTAN stabilizes cross-task reward normalization, leading to smoother credit assignment, steadier reward acquisition, and stronger continual learning capability.*

abruptly when a new task arrives. As shown in Figure 2, CTAN exhibits a smoother advantage scale and steadier reward acquisition during continual learning by stabilizing the reward scale across task boundaries.

We have provided more in-depth analysis regarding optimization properties of RaPO in Appendix A. Briefly, the retention reward is compatible with the standard score-function policy-gradient update when treated as detached scalar feedback. The resulting reward and CTAN-normalized advantage remain bounded. Under the usual smoothness, bounded-variance, and score-function moment assumptions, the idealized detached surrogate inherits the standard stationary-point convergence behavior of stochastic policy-gradient methods.

3.3 Discussions of RaPO

Retention Reward vs. Standard KL Loss. Although both our retention reward and the KL loss exploit a divergence signal, they intervene at different stages of optimization. The standard loss-level KL term in GRPO, $\mathrm{KL}(\pi_t \,\|\, \pi_{\mathrm{ref}})$, is a regularizer added to the training objective $\mathcal{L}$ (with π_{ref} a fixed reference model) that globally constrains the magnitude of the policy update. However, its gradient contribution is independent of the task reward received by each rollout, and it cannot distinguish the degree of forgetting within a group of rollouts that share comparable task rewards. In contrast, our retention reward is injected before the group-relative advantage computation, i.e., it lives inside R_{total} rather than inside $\mathcal{L}$. Consequently, RaPO can change the ranking of rollouts within a group, whereas a standard KL loss term primarily controls the scale of policy movement and cannot differentiate rollouts that receive the same reward. Experimental results in Appendix C.4 confirm that the standard loss-level KL has only a modest effect on continual learning performance.

Relation to On-policy Distillation. On-policy Distillation (OPD) [59, 60] trains a student policy on its own sampled trajectories using dense token-level KL supervision from a superior external teacher.

Table 1: A comparison of different methods on class-incremental image classification.

Method	ImageNet-R				ImageNet-A			
	10 Tasks		20 Tasks		10 Tasks		20 Tasks	
	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$
Joint Training								
SFT	89.55	-	89.55	-	67.61	-	67.61	-
GRPO	91.50	-	91.50	-	69.81	-	69.81	-
Continual Learning								
SFT	$64.49_{\pm 2.44}$	$39.83_{\pm 4.99}$	$57.84_{\pm 4.87}$	$41.91_{\pm 4.96}$	$28.18_{\pm 4.97}$	$66.92_{\pm 2.80}$	$14.20_{\pm 5.99}$	$79.71_{\pm 6.78}$
L2	$60.99_{\pm 3.50}$	$39.39_{\pm 4.59}$	$54.09_{\pm 4.05}$	$45.74_{\pm 7.38}$	$28.22_{\pm 2.57}$	$66.97_{\pm 3.65}$	$12.16_{\pm 4.86}$	$82.30_{\pm 5.43}$
EWC	$62.47_{\pm 2.69}$	$37.75_{\pm 4.65}$	$56.97_{\pm 3.49}$	$42.74_{\pm 3.67}$	$30.37_{\pm 6.56}$	$64.24_{\pm 2.96}$	$12.64_{\pm 3.43}$	$81.75_{\pm 4.21}$
LwF	$69.81_{\pm 1.63}$	$26.82_{\pm 2.36}$	$57.83_{\pm 2.97}$	$40.40_{\pm 6.81}$	$23.00_{\pm 3.18}$	$68.53_{\pm 3.25}$	$11.39_{\pm 3.56}$	$81.85_{\pm 3.69}$
GRPO	$74.67_{\pm 1.27}$	$20.02_{\pm 4.91}$	$73.18_{\pm 3.32}$	$17.40_{\pm 5.04}$	$37.37_{\pm 2.68}$	$28.88_{\pm 4.43}$	$22.80_{\pm 3.21}$	$48.14_{\pm 5.37}$
RaPO (Ours)	$\mathbf{85.92}_{\pm 1.82}$	$\mathbf{4.69}_{\pm 1.71}$	$\mathbf{81.97}_{\pm 2.88}$	$\mathbf{9.54}_{\pm 3.44}$	$\mathbf{44.61}_{\pm 2.36}$	$\mathbf{20.16}_{\pm 4.62}$	$\mathbf{32.15}_{\pm 0.92}$	$\mathbf{35.19}_{\pm 2.09}$

Method	TinyImageNet				CUB-200			
	10 Tasks		20 Tasks		10 Tasks		20 Tasks	
	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$
Joint Training								
SFT	34.71	-	34.71	-	48.79	-	48.79	-
GRPO	71.51	-	71.51	-	52.16	-	52.16	-
Continual Learning								
SFT	$15.94_{\pm 5.93}$	$43.29_{\pm 6.74}$	$8.71_{\pm 1.82}$	$53.68_{\pm 5.15}$	$19.09_{\pm 2.72}$	$71.29_{\pm 2.07}$	$15.78_{\pm 3.55}$	$72.46_{\pm 3.19}$
L2	$15.18_{\pm 3.83}$	$44.10_{\pm 4.22}$	$8.60_{\pm 3.26}$	$53.98_{\pm 4.51}$	$18.39_{\pm 3.56}$	$71.56_{\pm 2.63}$	$16.70_{\pm 2.66}$	$71.72_{\pm 3.48}$
EWC	$17.17_{\pm 6.40}$	$41.55_{\pm 6.77}$	$11.41_{\pm 0.29}$	$51.02_{\pm 0.95}$	$18.40_{\pm 2.80}$	$71.73_{\pm 1.69}$	$14.15_{\pm 4.08}$	$74.38_{\pm 3.29}$
LwF	$15.07_{\pm 3.13}$	$36.41_{\pm 4.10}$	$12.51_{\pm 3.40}$	$45.82_{\pm 3.26}$	$23.05_{\pm 2.22}$	$54.49_{\pm 2.70}$	$12.28_{\pm 2.10}$	$71.80_{\pm 1.12}$
GRPO	$46.84_{\pm 2.80}$	$36.24_{\pm 3.74}$	$42.08_{\pm 4.49}$	$40.85_{\pm 5.25}$	$28.67_{\pm 1.54}$	$45.50_{\pm 2.15}$	$24.67_{\pm 4.59}$	$47.04_{\pm 1.79}$
RaPO (Ours)	$\mathbf{62.36}_{\pm 2.26}$	$\mathbf{16.94}_{\pm 3.94}$	$\mathbf{62.20}_{\pm 3.63}$	$\mathbf{14.68}_{\pm 4.56}$	$\mathbf{45.15}_{\pm 1.20}$	$\mathbf{19.10}_{\pm 1.79}$	$\mathbf{37.33}_{\pm 2.05}$	$\mathbf{25.40}_{\pm 1.13}$

This mechanism is inherently tailored to a single-task knowledge transfer paradigm: every generated token contributes an independent gradient that uniformly pulls the student's per-token distribution toward that of the teacher. In visual continual learning, however, none of these premises hold. At task $\mathcal{T}_t$, there is no stronger teacher for the incoming data, while the only available reference is the policy in the preceding task (i.e., π_{t-1}). Consequently, enforcing a dense, token-level KL penalty against π_{t-1} on the data of $\mathcal{T}_t$ can inject an independent gradient term that actively pushes π_t's per-token distribution back toward π_{t-1} at every position. This suppresses the plasticity required to learn $\mathcal{T}_t$, transforming a retention mechanism into an obstacle for adaptation to new tasks.

4 Experiments

We conduct a comprehensive experimental evaluation on challenging visual continual learning problems across different visual modalities and task formulations. Specifically, we focus on class-incremental image classification, domain-incremental image classification, class-incremental video classification, class-incremental object detection, and domain-incremental object detection. The default hyperparameters of RaPO are $n = 8$ (Section 3.1), $\alpha = 20$ (Section 3.2.2), $\lambda = 0.5$ (Section 3.2.2), and $\beta = 0.999$ (Section 3.2.3). The retention reward is activated starting from task 2, since the pre-trained base model lacks domain-specific knowledge, while anchor π_1 is obtained only after the model has been adapted to task 1. All experiments are conducted on 8 NVIDIA H100 GPUs. Due to page limits, more experimental results are provided in Appendix C.

4.1 Class-Incremental Image Classification

We first evaluate RaPO under class-incremental image classification, a representative and challenging continual learning setting. The model receives a sequence of tasks, each introducing a disjoint subset of novel classes without access to previous training data. At test time, the model is evaluated on all classes observed so far.

Datasets. Experiments are presented on four commonly used image classification datasets: ImageNet-R [23], ImageNet-A [61], TinyImageNet [62], and CUB-200 [63]. For each dataset, we employ two task settings: 10 tasks with 20 classes per task and 20 tasks with 10 classes per task. All results are averaged over three random class orders to reduce the influence of any particular ordering. In

Table 2: A comparison of different methods on class-incremental object detection using the COCO 2017 dataset.

Method	5 Tasks		10 Tasks	
	$\mathcal{A}_b \uparrow$	$\mathcal{F}_b \downarrow$	$\mathcal{A}_b \uparrow$	$\mathcal{F}_b \downarrow$
Joint Training				
SFT	20.39	-	20.39	-
GRPO	23.48	-	23.48	-
Continual Learning				
SFT	$7.09_{\pm 0.28}$	$13.37_{\pm 1.04}$	$7.41_{\pm 3.37}$	$12.51_{\pm 4.36}$
L2	$7.64_{\pm 0.56}$	$14.24_{\pm 1.50}$	$7.82_{\pm 3.95}$	$13.72_{\pm 4.67}$
EWC	$7.14_{\pm 0.52}$	$14.51_{\pm 0.43}$	$7.87_{\pm 4.01}$	$13.91_{\pm 4.79}$
LwF	$13.93_{\pm 0.50}$	$7.31_{\pm 0.07}$	$13.41_{\pm 2.25}$	$7.34_{\pm 1.91}$
GRPO	$14.64_{\pm 2.18}$	$6.67_{\pm 2.71}$	$14.30_{\pm 3.02}$	$6.73_{\pm 3.51}$
RaPO (Ours)	$\mathbf{19.31}_{\pm 1.08}$	$\mathbf{1.39}_{\pm 1.20}$	$\mathbf{19.12}_{\pm 0.09}$	$\mathbf{1.37}_{\pm 0.56}$

practice, we adopt a 5-shot training protocol: only five labeled samples per class are available at each task. This scarce-annotation regime mirrors the constraints of real-world deployment, which lacks large and high-quality per-task annotation budgets, and poses a technical challenge that demands efficient few-shot adaptation while retaining prior knowledge [64]. This is also the standard setup in recent vision-centric RFT works [10, 12, 13], where learning under limited annotations is central to the problem formulation.

Evaluation Metrics. Two standard metrics for continual learning [21, 22, 65] are reported. Specifically, Last Accuracy ($\mathcal{A}$) is defined as the accuracy over the test set of all observed classes after the model finishes learning the final task. This metric evaluates the overall recognition ability retained by the final model. Forgetting ($\mathcal{F}$) measures the average performance drop from the historical best accuracy on previous tasks to their final accuracy after learning all tasks, which reflects the degree of forgetting throughout the continual learning process. **Baselines.** We employ Qwen2-VL-2B [24] as the base MLLM. This choice is motivated by two considerations. First, stronger frontier MLLMs (e.g., Qwen3 series [4, 66]) may already possess sufficient recognition ability on these benchmarks, making the gains from continual learning difficult to observe. Second, image classification is a relatively simple visual task. A relatively small-scale model is sufficiently expressive for continual image classification, while larger models risk overfitting. We have conducted experiments on a diverse set of baselines. As an upper bound, we jointly train the model on the entire dataset using both SFT and GRPO. For standard continual training, we sequentially apply SFT and GRPO, i.e., each method is used to train the model task by task. We also compare three representative SFT-based baselines tailored for continual learning: 1) A simple L2 regularizer pulls the current policy parameters toward the previous-task policy during training on each new task; 2) EWC [67] slows down updates on parameters that are important for previous tasks with Fisher information matrices; 3) LwF [68] uses the previous-task model as a teacher and distills its output distributions on the current task data into the current model. In practice, we train all models for 2 epochs, since more epochs do not improve performance and may lead to overfitting. Note that our goal is not to exhaustively benchmark existing continual learning algorithms, many of which rely on architecture- or task-specific mechanisms, but to isolate whether RFT can serve as a visual continual learning paradigm and whether our design further improves it. We therefore compare representative SFT-based continual-learning baselines and a GRPO-based RFT baseline.

Results. Table 1 summarizes the results on four class-incremental image classification benchmarks. RaPO consistently achieves the best performance across all datasets and task splits. For example, on ImageNet-R with 10 tasks, RaPO improves accuracy from 74.67% to 85.92% and reduces forgetting from 20.02% to 4.69% over the strongest baseline, GRPO. Similar trends hold on ImageNet-A, where accuracy rises from 37.37% to 44.61% and forgetting drops from 28.88% to 20.16%. RaPO also delivers clear gains on the more challenging TinyImageNet and CUB-200 benchmarks. In contrast, classical continual learning methods such as L2, EWC, and LwF provide only marginal or even detrimental benefits, while GRPO already serves as a strong RFT baseline that RaPO further improves by a large margin.

Table 3: A comparison of different methods on class-incremental video classification.

Method	UCF-101				Kinetics-200			
	5 Tasks		10 Tasks		5 Tasks		10 Tasks	
	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$
Joint Training								
SFT	79.54	-	79.54	-	75.41	-	75.41	-
GRPO	81.39	-	81.39	-	72.37	-	72.37	-
Continual Learning								
SFT	$72.57_{\pm 2.57}$	$26.00_{\pm 8.07}$	$62.96_{\pm 2.65}$	$27.35_{\pm 4.38}$	$66.29_{\pm 1.05}$	$53.28_{\pm 2.61}$	$33.89_{\pm 4.10}$	$61.53_{\pm 3.59}$
L2	$72.27_{\pm 2.50}$	$26.65_{\pm 8.40}$	$63.80_{\pm 1.73}$	$25.80_{\pm 3.19}$	$66.32_{\pm 1.21}$	$53.95_{\pm 3.20}$	$33.81_{\pm 3.85}$	$61.21_{\pm 3.88}$
EWC	$72.43_{\pm 2.98}$	$27.01_{\pm 9.37}$	$63.76_{\pm 1.96}$	$25.92_{\pm 3.23}$	$66.62_{\pm 0.91}$	$51.59_{\pm 3.41}$	$34.67_{\pm 3.62}$	$60.67_{\pm 2.90}$
LwF	$74.07_{\pm 0.70}$	$17.28_{\pm 2.57}$	$65.00_{\pm 0.30}$	$21.63_{\pm 0.92}$	$68.56_{\pm 0.14}$	$30.89_{\pm 1.20}$	$42.14_{\pm 3.76}$	$49.27_{\pm 3.24}$
GRPO	$77.62_{\pm 0.52}$	$13.81_{\pm 2.15}$	$68.12_{\pm 1.72}$	$16.94_{\pm 4.48}$	$70.33_{\pm 1.92}$	$30.37_{\pm 5.35}$	$50.67_{\pm 2.09}$	$34.65_{\pm 3.21}$
RaPO (Ours)	$\mathbf{79.57}_{\pm 0.77}$	$\mathbf{11.19}_{\pm 2.63}$	$\mathbf{71.79}_{\pm 1.84}$	$\mathbf{10.92}_{\pm 0.15}$	$\mathbf{74.18}_{\pm 2.36}$	$\mathbf{16.76}_{\pm 3.73}$	$\mathbf{54.85}_{\pm 2.81}$	$\mathbf{27.28}_{\pm 1.38}$

4.2 Class-Incremental Object Detection

Setup. We further evaluate RaPO on class-incremental object detection on the COCO 2017 dataset [69], which contains 80 object categories. We design two task splits: a 5-task setting with 16 classes per task and a 10-task setting with 8 classes per task. Since an image may contain instances from multiple categories, we enforce a strict task partition: each training image is assigned to only one task, and there is no training sample overlap across tasks. For each class, at most five training images are selected, resulting in a few-shot detection setting. On the other hand, different from image classification, object detection requires fine-grained spatial perception and structured output generation. Hence, we use Qwen2-VL-7B [24] as the base MLLM. We train each task for 5 epochs so that the model can sufficiently adapt to the detection objective while avoiding overfitting. For evaluation, we follow the metric definitions in Section 4.1, while replacing classification accuracy with box average precision (AP) under the standard COCO protocol [69, 70]. Accordingly, we report last box AP $\mathcal{A}_b$ and box forgetting $\mathcal{F}_b$.

Results. As listed in Table 2, RaPO achieves the best $\mathcal{A}_b$ and the lowest $\mathcal{F}_b$ in both task settings. Compared with GRPO, it improves $\mathcal{A}_b$ from 14.64% to 19.31% and reduces $\mathcal{F}_b$ from 6.67% to 1.39% in the 5-task setting, and further improves $\mathcal{A}_b$ from 14.30% to 19.12% while reducing $\mathcal{F}_b$ from 6.73% to 1.37% in the 10-task setting. Unlike in image classification, LwF becomes competitive with GRPO in object detection, suggesting that distilling predictions from the preceding-task model is particularly useful for preserving structured localization behavior. Nevertheless, RaPO remains clearly superior to all baselines.

4.3 Class-Incremental Video Classification

Setup. We extend the evaluation to class-incremental video classification, where the model sequentially learns new classes from a stream of video clips. The primary difference from image classification is the modality shift from images to video, which introduces temporal cues across tasks. We conduct experiments on two commonly used video recognition datasets: UCF-101 [71] and Kinetics-200 [72, 73]. Due to the substantial computational cost of video training and limited hardware resources, we adopt 5-task and 10-task settings for both datasets. All other experimental settings follow the configuration described in Section 4.1.

Results. As listed in Table 3, our method achieves leading performance on both datasets. On UCF-101 under the 10-task setting, RaPO attains an $\mathcal{A}$ of 71.79% with only 10.92% $\mathcal{F}$, compared to 68.12% and 16.94% for GRPO. On Kinetics-200 under the 5-task setting, RaPO improves $\mathcal{A}$ from 70.33% to 74.18% and reduces $\mathcal{F}$ from 30.37% to 16.76%. These results demonstrate that RaPO generalizes effectively to the video domain, preserving both spatial and temporal knowledge throughout the incremental learning process.

4.4 Domain-Incremental Image Classification and Object Detection

Setup. We evaluate RaPO under domain-incremental learning (DIL), where the model sequentially adapts to new visual domains while the label space remains fixed and no data replay from previous domains is allowed. For image classification, we use DomainNet [74] with 6 domains and Office-

Table 4: A comparison of Domain-Incremental Learning (DIL) across image classification and object detection.

Method	Image Classification				Object Detection	
	DomainNet (6 Domains)		OfficeHome (4 Domains)		Pascal Series (4 Domains)	
	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\bar{\mathcal{A}}_b\uparrow$	$\mathcal{F}_b\downarrow$
Joint Training						
SFT	67.76	-	93.84	-	35.40	-
GRPO	68.46	-	94.65	-	39.13	-
Continual Learning						
SFT	$63.09_{\pm 0.21}$	$1.34_{\pm 0.33}$	$91.07_{\pm 1.13}$	$1.02_{\pm 1.02}$	$34.55_{\pm 2.16}$	$\mathbf{0.55}_{\pm 0.73}$
L2	$63.90_{\pm 1.55}$	$2.36_{\pm 1.49}$	$91.62_{\pm 0.63}$	$2.29_{\pm 0.67}$	$33.45_{\pm 2.98}$	$1.07_{\pm 1.14}$
EWC	$63.08_{\pm 1.60}$	$1.28_{\pm 1.33}$	$91.29_{\pm 1.19}$	$2.52_{\pm 0.86}$	$33.52_{\pm 2.21}$	$0.61_{\pm 0.70}$
LwF	$62.63_{\pm 4.15}$	$1.75_{\pm 1.90}$	$88.33_{\pm 1.40}$	$2.82_{\pm 1.55}$	$33.86_{\pm 1.41}$	$0.63_{\pm 0.81}$
GRPO	$64.54_{\pm 1.03}$	$0.31_{\pm 0.26}$	$92.80_{\pm 0.28}$	$0.55_{\pm 0.09}$	$34.58_{\pm 1.75}$	$1.18_{\pm 1.44}$
RaPO (Ours)	$\mathbf{66.27}_{\pm 0.62}$	$\mathbf{0.17}_{\pm 0.27}$	$\mathbf{93.63}_{\pm 0.14}$	$\mathbf{0.45}_{\pm 0.49}$	$\mathbf{37.18}_{\pm 1.44}$	$0.84_{\pm 0.62}$

Home [75] with 4 domains. All remaining settings follow Section 4.1. For domain-incremental object detection, we use the four Pascal Series domains [76], including Pascal VOC [77], Clipart, Watercolor, and Comic. All four domains share 6 common object classes, ensuring a consistent label space throughout the task sequence. Since the four domains have test sets of vastly different sizes, a direct all-sample AP is dominated by the largest domain. Therefore, we compute the AP of each domain separately and then average these four values after the model finishes the last task, termed $\bar{\mathcal{A}}_b$. All other training and evaluation protocols remain identical to those in Section 4.2.

Results. Table 4 shows that RaPO consistently improves performance across domain-incremental classification and detection. For example, RaPO improves $\mathcal{A}$ from 64.54% to 66.27% and reduces $\mathcal{F}$ from 0.31% to 0.17% on DomainNet compared with GRPO. For domain-incremental object detection, RaPO achieves the best $\bar{\mathcal{A}}_b$ among all continual learners at 37.18%, improving over GRPO by 2.60%. This result is also higher than joint-training SFT at 35.40% and remains close to joint-training GRPO at 39.13%. Although RaPO exhibits a slightly higher $\mathcal{F}_b$ than plain SFT, the accompanying boost in absolute detection performance is substantial, confirming that RaPO learns to preserve both location and category knowledge more effectively rather than merely resisting change. Compared with CIL, we also observe that DIL exhibits intrinsically lower forgetting. This is because DIL maintains a fixed label space across distribution shifts, which inherently avoids the severe inter-task confusion that drives catastrophic forgetting when learning disjoint novel classes in CIL [78].

5 Conclusion

In this work, we investigate the potential of RFT for visual continual learning. Through a pilot study, we empirically confirm that RFT is more resilient to catastrophic forgetting than SFT, yet still suffers from non-negligible forgetting, due to trajectory-level drift agnosticism. To this end, we propose **RaPO**, which addresses catastrophic forgetting through two complementary components: a Retention Reward that preferentially reinforces low-drift, knowledge-preserving rollouts within each group, and CTAN that stabilizes the optimization scale across task-wise boundaries. Extensive results across diverse visual continual learning settings demonstrate that RaPO consistently achieves leading performance. As the first systematic exploration of RFT in visual continual learning, we hope this work will stimulate further research in this area.

Acknowledgements and Disclosures

This work was supported in part by Hong Kong Research Grants Council under NSFC/RGC Collaborative Research Scheme (Grant CRS HKU703/24). The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1.5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

[2] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[3] Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, et al. Seed1.5-vl technical report. *arXiv preprint arXiv:2505.07062*, 2025.

[4] Qwen Team. Qwen3.5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

[5] Kaiyan Zhang, Yuxin Zuo, Bingxiang He, Youbang Sun, Runze Liu, Che Jiang, Yuchen Fan, Kai Tian, Guoli Jia, Pengfei Li, et al. A survey of reinforcement learning for large reasoning models. *arXiv preprint arXiv:2509.08827*, 2025.

[6] Hanzhong Guo, Jie Wu, Jie Liu, Yu Gao, Zilyu Ye, Linxiao Yuan, Xionghui Wang, Yizhou Yu, and Weilin Huang. Leveraging verifier-based reinforcement learning in image editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[7] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[8] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[9] Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, et al. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*, 2025.

[10] Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-rft: Visual reinforcement fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2034–2044, 2025.

[11] Ming Li, Jike Zhong, Shitian Zhao, Yuxiang Lai, Haoquan Zhang, Wang Bill Zhu, and Kaipeng Zhang. To think or not to think: A study of thinking in rule-based visual reinforcement fine-tuning. In *Advances in Neural Information Processing Systems*, 2025.

[12] Huajie Tan, Yuheng Ji, Xiaoshuai Hao, Xiansheng Chen, Pengwei Wang, Zhongyuan Wang, and Shanghang Zhang. Reason-rft: Reinforcement fine-tuning for visual reasoning of vision language models. In *Advances in Neural Information Processing Systems*, 2025.

[13] Hulingxiao He, Zijun Geng, and Yuxin Peng. Fine-r1: Make multi-modal llms excel in fine-grained visual recognition by chain-of-thought reasoning. In *International Conference on Learning Representations*, 2026.

[14] Heitor Murilo Gomes, Jean Paul Barddal, Fabrício Enembreck, and Albert Bifet. A survey on ensemble learning for data stream classification. *ACM Computing Surveys*, 50(2):1–36, 2017.

[15] Junhao Zheng, Shengjie Qiu, Chengming Shi, and Qianli Ma. Towards lifelong learning of large language models: A survey. *ACM Computing Surveys*, 57(8):1–35, 2025.

[16] Haizhou Shi, Zihao Xu, Hengyi Wang, Weiyi Qin, Wenyuan Wang, Yibin Wang, Zifeng Wang, Sayna Ebrahimi, and Hao Wang. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025.

[17] Yutao Yang, Jie Zhou, Xuanwen Ding, Tianyu Huai, Shunyu Liu, Qin Chen, Yuan Xie, and Liang He. Recent advances of foundation language models-based continual learning: A survey. *ACM Computing Surveys*, 57(5):1–38, 2025.

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

[19] Song Lai, Haohan Zhao, Rong Feng, Changyi Ma, Wenzhuo Liu, Hongbo Zhao, Xi Lin, Dong Yi, Qingfu Zhang, Hongbin Liu, et al. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training. *arXiv preprint arXiv:2507.05386*, 2025.

[20] Idan Shenfeld, Jyothish Pari, and Pulkit Agrawal. RL's razor: Why online reinforcement learning forgets less. In *International Conference on Learning Representations*, 2026.

[21] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9851–9873, 2024.

[22] Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[24] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[25] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[27] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[28] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[29] Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. In *International Conference on Learning Representations*, 2024.

[30] Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective. In *Second Conference on Language Modeling*, 2025.

[31] Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025.

[32] Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, et al. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025.

[33] Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[34] Zhihao Zhang, Qiaole Dong, Qi Zhang, Enyu Zhou, Jun Zhao, Zhiheng Xi, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Mingqi Wu, Yanwei Fu, Tao Ji, Tao Gui, Xuanjing Huang, and Kai Chen. Why reinforcement fine-tuning enables mllms preserve prior knowledge better: A data perspective. In *International Conference on Learning Representations*, 2026.

[35] Zhenquan Yao, Zitong Huang, Yihan Zeng, Jianhua Han, Hang Xu, Chun-Mei Feng, Jianwei Ma, and Wangmeng Zuo. Cgl: Advancing continual gui learning via reinforcement fine-tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15212–15221, 2026.

[36] Meng Lou and Yizhou Yu. Overlock: An overview-first-look-closely-next convnet with context-mixing dynamic kernels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 128–138. IEEE, 2025.

[37] Meng Lou, Yunxiang Fu, and Yizhou Yu. Sparx: A sparse cross-layer connection mechanism for hierarchical vision mamba and transformer networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 19104–19114, 2025.

[38] Meng Lou, Shu Zhang, Hong-Yu Zhou, Sibei Yang, Chuan Wu, and Yizhou Yu. Transxnet: learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6):11534–11547, 2025.

[39] Yunxiang Fu, Meng Lou, and Yizhou Yu. Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19077–19087. IEEE, 2025.

[40] Chaoqi Chen, Jiongcheng Li, Xiaoguang Han, Xiaoqing Liu, and Yizhou Yu. Compound domain generalization via meta-knowledge encoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7109–7119. IEEE, 2022.

[41] Boyang Dai, Zeng Fan, Zihao Qi, Meng Lou, and Yizhou Yu. Cgsa: Class-guided slot-aware adaptation for source-free object detection. In *International Conference on Learning Representations*, 2026.

[42] Cheng Shi, Yizhou Yu, and Sibei Yang. Vision transformers need more than registers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022.

[44] Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Revisiting class-incremental learning with pre-trained models: Generalizability and adaptivity are all you need. *International Journal of Computer Vision*, 133(3):1012–1032, 2025.

[45] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022.

[46] Qiang Wang, Yuhang He, Songlin Dong, Xinyuan Gao, Shaokun Wang, and Yihong Gong. Non-exemplar domain incremental learning via cross-domain concept integration. In *European Conference on Computer Vision*, pages 144–162. Springer, 2024.

[47] Da-Wei Zhou, Hai-Long Sun, Jingyi Ning, Han-Jia Ye, and De-Chuan Zhan. Continual learning with pre-trained models: a survey. In *International Joint Conference on Artificial Intelligence*, pages 8363–8371, 2024.

[48] Meng Lou, Yunxiang Fu, and Yizhou Yu. Scaling continual learning to 300+ tasks with bi-level routing mixture-of-experts. In *International Conference on Machine Learning*, pages 80632–80649. PMLR, 2026.

[49] Yunxiang Fu, Meng Lou, Zicheng Liao, and Yizhou Yu. Clare: Scalable class-incremental continual learning via a sparsity-based framework. In *British Machine Vision Conference*, 2026.

[50] Yunxiang Fu, Meng Lou, and Yizhou Yu. One adapter, many tasks: Task-conditioned feature transformations for continual learning. *arXiv preprint arXiv:2608.31096*, 2026.

[51] Meng Lou, Stanley Yu, and Yizhou Yu. Parameters as experts: Adapting vision models with dynamic parameter routing. In *International Conference on Machine Learning*, pages 80746–80763. PMLR, 2026.

[52] Yan-Shuo Liang and Wu-Jun Li. Inflora: Interference-free low-rank adaptation for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23638–23647, 2024.

[53] Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Ping Hu, Dong Wang, Huchuan Lu, and You He. Boosting continual learning of vision-language models via mixture-of-experts adapters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23219–23230, 2024.

[54] Yan Wang, Da-Wei Zhou, and Han-Jia Ye. Integrating task-specific and universal adapters for pre-trained model-based class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 806–816, 2025.

[55] Hai-Long Sun, Da-Wei Zhou, Hanbin Zhao, Le Gan, De-Chuan Zhan, and Han-Jia Ye. Mos: Model surgery for pre-trained model-based class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 20699–20707, 2025.

[56] Da-Wei Zhou, Zi-Wen Cai, Han-Jia Ye, Lijun Zhang, and De-Chuan Zhan. Dual consolidation for pre-trained model-based domain-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20547–20557, 2025.

[57] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

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

[59] Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024.

[60] Wenkai Yang, Weijie Liu, Ruobing Xie, Kai Yang, Saiyong Yang, and Yankai Lin. Learning beyond teacher: Generalized on-policy distillation with reward extrapolation. *arXiv preprint arXiv:2602.12125*, 2026.

[61] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[62] Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

[63] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[64] Jinghua Zhang, Li Liu, Olli Silvén, Matti Pietikäinen, and Dewen Hu. Few-shot class-incremental learning for classification and object detection: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(4):2924–2945, 2025.

[65] Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European Conference on Computer Vision*, pages 556–572, 2018.

[66] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[67] 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.

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

[69] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[70] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[71] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[72] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[73] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 318–335, 2018.

[74] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[75] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5385–5394, 2017.

[76] Naoto Inoue, Ryosuke Furuta, Toshihiko Yamasaki, and Kiyoharu Aizawa. Cross-domain weakly-supervised object detection through progressive domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018.

[77] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[78] Gido M Van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. *Nature Machine Intelligence*, 4(12):1185–1197, 2022.

[79] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020.

[80] Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization. *arXiv preprint arXiv:2511.20347*, 2025.

A Policy-Gradient Compatibility and Optimization Stability of RaPO

We provide a standard reinforcement-learning view of RaPO. This analysis does not claim global optimality or a formal no-forgetting guarantee. Instead, it focuses on three optimization properties of the proposed retention reward: compatibility with the score-function policy-gradient estimator under a detached-reward surrogate view, bounded reward and advantage magnitudes, and the standard stationary-point guarantee inherited under idealized non-convex stochastic policy-gradient assumptions.

Policy-gradient compatibility. At optimization step k, let θ_k denote the policy used to sample rollouts and compute the retention reward. For a sampled trajectory $y = (y_1, \ldots, y_m)$, define the length-normalized pre-truncation log-ratio drift as

$$d_{\theta_k}(y) = \frac{1}{m} \sum_{s=1}^{m} \left[\log \pi_{\theta_k}(y_s \mid x, y_{<s}) - \log \pi_{t-1}(y_s \mid x, y_{<s}) \right]. \tag{7}$$

This quantity is the trajectory-level counterpart of the drift proxy used in Equation (2). Under on-policy sampling from π_{θ_k}, its expectation corresponds to a token-averaged forward log-ratio contribution of π_{θ_k} relative to the anchor π_{t-1}, whereas a single sampled value can be negative and should not be interpreted as a non-negative KL value. RaPO therefore uses the one-sided truncation $[d_{\theta_k}(y)]_+$ inside the reward, but does not optimize this signal as an explicit differentiable regularizer. As in standard PPO/GRPO implementations, the reward value assigned to each sampled rollout is treated as stop-gradient scalar feedback when forming the policy-gradient update. Concretely, after rollouts are collected at θ_k, RaPO constructs the detached shaped reward

$$\widetilde{R}_k(y) = R_{\text{task}}(y) + \lambda \exp\left(-\alpha \left[d_{\theta_k}(y)\right]_+\right), \tag{8}$$

where the subscript k emphasizes that the scalar reward is fixed during the subsequent gradient computation. This distinction is important: if the log-ratio inside R_{ret} were differentiated directly, it would become a loss-level KL-style regularizer. RaPO instead uses the drift only to shape the sampled rollout's reward, so the policy-gradient term remains a likelihood-ratio update weighted by a scalar return.

To make this precise, consider the prompt-conditioned local detached objective

$$\mathcal{J}_k(\theta; x) = \mathbb{E}_{y \sim \pi_\theta(\cdot|x)} \left[\widetilde{R}_k(y) \right]. \tag{9}$$

This objective is an analysis device for a single update: $\widetilde{R}_k$ is fixed while differentiating with respect to θ, although the detached reward rule is recomputed after new rollouts are collected at later optimization steps. Refer to [30], for any baseline $b(x)$ that does not depend on the sampled trajectory y, the score-function identity gives

$$\mathbb{E}_{y\sim\pi_{\theta_k}(\cdot|x)}\left[\left(\widetilde{R}_k(y)-b(x)\right)\nabla_\theta\log\pi_\theta(y\mid x)\Big|_{\theta=\theta_k}\right]=\nabla_\theta\mathcal{J}_k(\theta;x)\Big|_{\theta=\theta_k}. \tag{10}$$

The equality follows because $\widetilde{R}_k(y)$ is detached: the gradient acts only on $\log\pi_\theta(y\mid x)$, while the reward is treated exactly like a verifier-provided return. Therefore, in the ideal score-function form, adding the retention reward does not introduce an additional pathwise gradient term; it simply changes the scalar credit assigned to each sampled trajectory. In practical GRPO, clipping and finite-group normalization make the estimator an approximation to this ideal identity, and the group mean is a sample-dependent baseline rather than a fixed $b(x)$. The key point is narrower: the retention term itself does not add a separate token-level distillation gradient that directly pulls every token distribution toward the anchor.

Bounded reward and advantage. Assume the task verifier is bounded, i.e.,

$$0\le R_{\text{task}}(y)\le R_{\max}. \tag{11}$$

Since the retention reward is defined as

$$R_{\text{ret}}(y)=\exp(-\alpha\bar{D}_{\text{drift}}(y)), \tag{12}$$

and $\bar{D}_{\text{drift}}(y)\ge 0$, we have

$$0<R_{\text{ret}}(y)\le 1. \tag{13}$$

Therefore, the total reward used by RaPO is bounded as

$$0\le R_{\text{total}}(y)=R_{\text{task}}(y)+\lambda R_{\text{ret}}(y)\le R_{\max}+\lambda. \tag{14}$$

Let $B=R_{\max}+\lambda$. For a rollout group, the group mean reward also lies in $[0,B]$, and CTAN computes the advantage with denominator $\hat{\sigma}+\epsilon$. Since $\hat{\sigma}$ is an EMA of non-negative batch standard deviations, $\hat{\sigma}\ge 0$ whenever it is initialized non-negatively. Thus,

$$|A_i^{\text{RaPO}}|=\left|\frac{R_{\text{total}}(y_i)-\mu_{\text{group}}}{\hat{\sigma}+\epsilon}\right|\le\frac{B}{\epsilon}. \tag{15}$$

This shows that the retention reward can enlarge the reward range by at most λ, while CTAN keeps the normalization denominator lower-bounded by ϵ. Consequently, RaPO does not introduce unbounded advantages. To connect this bound to the usual bounded-gradient condition, we additionally require the standard score-function moment assumption: the trajectory score $\nabla_\theta\log\pi_\theta(y\mid x)=\sum_s\nabla_\theta\log\pi_\theta(y_s\mid x,y_{<s})$ has bounded second moment. Under this assumption, the score-function estimator weighted by A_i^{RaPO} also has a bounded second moment.

Idealized convergence to a stationary point and proof. The practical RaPO update uses clipped GRPO, finite rollout groups, and a detached reward that is recomputed from the current sampling policy. A complete convergence theorem for this exact time-varying procedure would need to track the clipping bias and sample-dependent group normalization. Here we state the standard idealized stochastic policy-gradient result that RaPO inherits once its shaped reward is viewed as a bounded detached scalar feedback signal.

Let $\mathcal{J}_{\text{RaPO}}(\theta)$ denote a fixed detached shaped surrogate objective over the analysis window, where verifier rewards and retention rewards are treated as stop-gradient scalar feedback when forming the score-function estimator. Suppose that:

(i) $\mathcal{J}_{\text{RaPO}}$ is L-smooth and upper bounded by $\mathcal{J}^\star$;

(ii) the stochastic policy-gradient estimator g_k is unbiased for this surrogate, i.e., $\mathbb{E}[g_k\mid\theta_k]=\nabla\mathcal{J}_{\text{RaPO}}(\theta_k)$;

(iii) its variance is bounded, i.e., $\mathbb{E}[\|g_k-\nabla\mathcal{J}_{\text{RaPO}}(\theta_k)\|^2\mid\theta_k]\le\sigma_g^2$. These assumptions are standard in non-convex stochastic policy-gradient analyses. Assumption (i) is a local smoothness and bounded-objective condition on the detached surrogate. Assumption (ii) follows from the score-function identity in Equation (10) before practical approximations such as clipping and finite-group

normalization are applied. Assumption (iii) is supported by the bounded reward-and-advantage result above: if the trajectory score has a bounded second moment, i.e.,

$$\mathbb{E}\left[\|\nabla_\theta \log \pi_\theta(y \mid x)\|^2 \mid \theta\right] \le C_{\text{score}}, \tag{16}$$

Then, Equation (15) implies that the idealized score-function estimator has a bounded second moment,

$$\mathbb{E}\left[\left\|A^{\text{RaPO}}\nabla_\theta \log \pi_\theta(y \mid x)\right\|^2 \mid \theta\right] \le \left(\frac{B}{\epsilon}\right)^2 C_{\text{score}}, \tag{17}$$

which in turn yields a bounded variance term σ_g^2.

Then stochastic gradient ascent with $\theta_{k+1} = \theta_k + \eta g_k$ and $\eta \le 1/L$ satisfies

$$\min_{0\le k<K} \mathbb{E}\left[\|\nabla \mathcal{J}_{\text{RaPO}}(\theta_k)\|^2\right] \le \frac{2(\mathcal{J}^\star - \mathcal{J}_{\text{RaPO}}(\theta_0))}{\eta K} + L\eta\sigma_g^2. \tag{18}$$

Choosing $\eta = \mathcal{O}(K^{-1/2})$ yields

$$\min_{0\le k<K} \mathbb{E}\left[\|\nabla \mathcal{J}_{\text{RaPO}}(\theta_k)\|^2\right] = \mathcal{O}(K^{-1/2}). \tag{19}$$

To prove this statement, use the smoothness of $\mathcal{J}_{\text{RaPO}}$. One step of stochastic gradient ascent gives

$$\mathcal{J}_{\text{RaPO}}(\theta_{k+1}) \ge \mathcal{J}_{\text{RaPO}}(\theta_k) + \eta\,\langle \nabla \mathcal{J}_{\text{RaPO}}(\theta_k), g_k\rangle - \frac{L\eta^2}{2}\|g_k\|^2. \tag{20}$$

Taking the conditional expectation and using the unbiasedness assumption,

$$\mathbb{E}\left[\mathcal{J}_{\text{RaPO}}(\theta_{k+1}) - \mathcal{J}_{\text{RaPO}}(\theta_k) \mid \theta_k\right] \ge \eta\,\|\nabla \mathcal{J}_{\text{RaPO}}(\theta_k)\|^2 - \frac{L\eta^2}{2}\left(\|\nabla \mathcal{J}_{\text{RaPO}}(\theta_k)\|^2 + \sigma_g^2\right). \tag{21}$$

When $\eta \le 1/L$, this implies

$$\mathbb{E}\left[\mathcal{J}_{\text{RaPO}}(\theta_{k+1}) - \mathcal{J}_{\text{RaPO}}(\theta_k)\right] \ge \frac{\eta}{2}\mathbb{E}\left[\|\nabla \mathcal{J}_{\text{RaPO}}(\theta_k)\|^2\right] - \frac{L\eta^2}{2}\sigma_g^2. \tag{22}$$

Summing the above inequality from $k = 0$ to $K - 1$ and using the upper bound $\mathcal{J}^\star$ gives Equation (18).

B More Detailed Implementations

B.1 Prompt Template and Output Format

Image and Video Classification. As shown in Figure 3, at task $\mathcal{T}_t$, let $\Delta\mathcal{Y}_t$ denote the set of novel class names introduced by the current classification task, and let $\mathcal{Y}_{\le t} = \mathcal{Y}_{\le t-1} \cup \Delta\mathcal{Y}_t$ be the cumulative vocabulary of all class names observed so far. The prompt explicitly lists every class in $\mathcal{Y}_{\le t}$ as the candidate answer space and never exposes future-task class names, so the model is evaluated strictly on its ability to recognize and retain classes that have already been introduced. The final answer is constrained to a closed-set prediction over this predefined vocabulary, while the reasoning span remains free-form and unconstrained. Video classification follows the same template, with the image placeholder replaced by a video placeholder. Note that in domain-incremental learning, the class vocabulary remains unchanged across all tasks since only the visual domain varies. Although we acknowledge the practical importance of open-set recognition in real-world applications, our primary goal is to investigate the potential of RFT in visual continual learning, where the central challenge lies in balancing plasticity to newly introduced classes against the stability of previously acquired ones. Allowing fully open-set generation may introduce confounding factors such as synonyms, hypernyms, or explanatory paraphrases that fall outside the evaluation taxonomy, making it difficult to isolate whether performance changes stem from genuine recognition gains or merely from naming variability.

Object Detection. For object detection, let $\Delta\mathcal{C}_t$ denote the novel class names introduced by the current task, and let $\mathcal{C}_{\le t} = \mathcal{C}_{\le t-1} \cup \Delta\mathcal{C}_t$ be the cumulative seen class vocabulary up to task $\mathcal{T}_t$. As shown in Figure 4, the model is required to output a JSON list of detected instances, where the

Input. Visual input v and cumulative class-name vocabulary $\mathcal{Y}_{\leq t}$.

Prompt Template.

<image> or <video>
Perform image/video classification on the given visual input.
You MUST choose exactly one class name from this list:
{class names in $\mathcal{Y}_{\leq t}$}
Output the reasoning process in <think> </think>
and final answer in <answer> </answer>.
Your answer MUST be one class name from the list,
character-for-character. No extra words.

Required Output.

<think> reasoning process </think>
<answer> one class name from $\mathcal{Y}_{\leq t}$ </answer>

Figure 3: Prompt template and required output format for image and video classification.

"category" field of each entry is restricted to a closed-set prediction over $\mathcal{C}_{\leq t}$. This constraint ensures that the evaluation faithfully reflects the model's ability to recognize and retain previously learned object classes, consistent with the closed-set protocol adopted in classification. At the same time, the model freely predicts the number of instances and their spatial locations. Bounding box coordinates are expressed as integer values normalized to the range $[0, 1000]$, making the verifier independent of the original image resolutions.

B.2 Task Reward Function Design

For the reward definitions below, y_i denotes the i-th response sampled for visual input v within a rollout group.

Classification Reward. For image and video classification, the task reward combines answer correctness and format compliance:

$$R_{\text{task}}(y_i) = R_{\text{acc}}(y_i) + R_{\text{fmt}}(y_i). \tag{23}$$

Here, $R_{\text{acc}} \in \{0, 1\}$ is an exact-match reward for the final class name extracted from the <answer> span. The prediction and the ground-truth class name are normalized by lower-casing and by mapping underscores, hyphens, and periods to spaces before matching, making the verifier robust to superficial formatting variation while preserving class identity. The format term $R_{\text{fmt}} \in \{0, 1\}$ checks the complete <think></think> <answer></answer> structure.

Detection Reward. For object detection, the task reward combines localization quality, class-name accuracy, and format compliance:

$$R_{\text{task}}(y_i) = R_{\text{iou}}(y_i) + R_{\text{cls}}(y_i) + R_{\text{fmt}}(y_i). \tag{24}$$

The verifier first parses the JSON output from the <answer> span into a set of predicted boxes. Class names are lower-cased, coordinates are rounded to the nearest integer and clipped to $[0, 1000]$, and degenerate boxes (e.g., those with zero area) are discarded. Following DETR [79], we construct a bipartite matching score matrix where same-class pairs are scored by their IoU and mismatched pairs receive zero score. The Hungarian algorithm then finds an optimal one-to-one assignment between predictions and ground truth. This matching is critical for meaningful reward computation: it ensures each ground-truth instance is associated with at most one prediction, preventing the policy from inflating its reward by outputting multiple near-identical boxes for the same object. Based on the matching, R_{iou} is the sum of IoU values from all assigned same-class pairs divided by $\max(N_{\text{gt}}, N_{\text{pred}})$, where N_{gt} and N_{pred} denote the number of ground-truth and predicted instances, respectively. This denominator normalizes the reward by the larger of the two counts, penalizing both missed detections and excessive false positives in a single term. R_{cls} is the fraction of assigned same-class pairs whose IoU meets or exceeds 0.5, computed under the same denominator. The format term R_{fmt} verifies both the output tag structure and JSON schema. Outputs with valid tags but unparsable JSON receive partial format credit, while R_{iou} and R_{cls} are set to zero.

Input. Image v and cumulative class-name vocabulary $\mathcal{C}_{\leq t}$.

Prompt Template.

```
<image>
Detect all objects in the image that belong to:
{class names in C_{≤t}}
For each detected object, provide its class name
in the "category" field and its bounding box
as [x1, y1, x2, y2],
with coordinates normalized to 0-1000.
Output reasoning in <think> </think>
and detection results in <answer> </answer>.
```

Required Output.

```
<think> reasoning process </think>
<answer>[
  {"category": "cat_name", "bbox": [x1, y1, x2, y2]}, ...
]</answer>
Use <answer>[]</answer> if no listed category is detected.
```

Figure 4: Prompt template and required output format for object detection.

C Ablation Studies

We provide a comprehensive ablation analysis of different components and hyperparameters in RaPO. Experiments are conducted on class-incremental image classification using the 10-task setting with ImageNet-R, and on class-incremental object detection using the 5-task setting with COCO 2017. All other experimental settings follow the configurations described in Sections 4.1 and 4.2. Unless otherwise specified, RaPO uses the default hyperparameters: rollout group size $n = 8$ (Section 3.1), retention reward scale $\alpha = 20$ (Section 3.2.2), retention reward weight $\lambda = 0.5$ (Section 3.2.2), and smoothing coefficient $\beta = 0.999$ (Section 3.2.3).

C.1 Analysis of Core Components in RaPO

We first verify the contribution of the two core components, Retention Reward (Section 3.2.2) and CTAN (Section 3.2.3), by gradually adding them to the vanilla GRPO baseline. As shown in Table 5, both components individually improve performance on the two benchmarks. Adding CTAN alone raises $\mathcal{A}$ from 74.67 to 79.43 and $\mathcal{A}_b$ from 14.64 to 16.27, while reducing $\mathcal{F}$ from 20.02 to 13.17 and $\mathcal{F}_b$ from 6.67 to 4.73. This demonstrates that stabilizing the advantage scale across task boundaries already benefits both task adaptation and knowledge retention. Incorporating the Retention Reward provides a substantially larger gain, pushing $\mathcal{A}$ to 82.57 and $\mathcal{A}_b$ to 18.11, with forgetting rates dropping to 8.38 and 2.61, respectively. Combining both components yields the best results across all metrics, achieving $\mathcal{A}$ of 85.92 and $\mathcal{F}$ of 4.69 on ImageNet-R, and $\mathcal{A}_b$ of 19.31 and $\mathcal{F}_b$ of 1.39 on COCO. These results confirm that CTAN and the Retention Reward play complementary roles: CTAN stabilizes the update magnitude across tasks while the Retention Reward steers credit assignment toward low-drift trajectories that preserve previously acquired knowledge.

On the other hand, we further analyze the retention-reward dynamics. For RaPO, R_{ret} is optimized with the default coefficient $\lambda = 0.5$. For a controlled diagnostic comparison, we also attach the same retention-reward estimator to vanilla GRPO while fixing its coefficient to $\lambda = 0$. Therefore, the computed R_{ret} for GRPO is used only for measurement: it is not used in the advantage computation, and never contributes to policy updates. As shown in Figure 5, RaPO consistently maintains higher retention rewards than GRPO after task transitions on both datasets. Specifically, the GRPO curves reveal that task-reward-only optimization can move toward high-drift trajectories, even though it adapts to the current task. In contrast, RaPO preserves higher R_{ret} throughout the continual sequence, showing that the retention reward continuously reshapes credit assignment toward low-drift, knowledge-preserving trajectories.

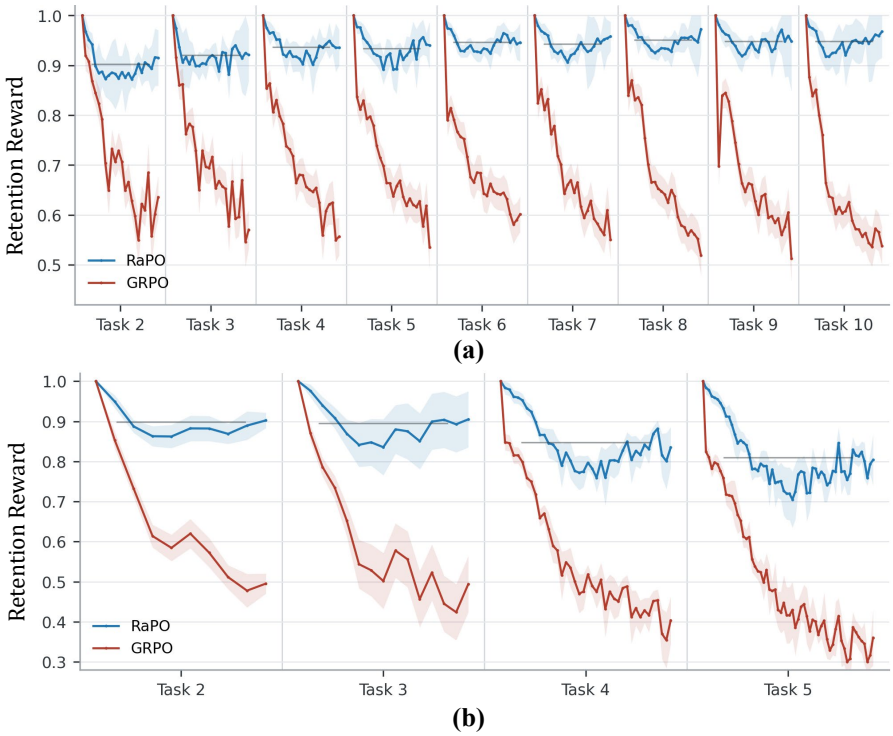


Figure 5: Retention reward dynamics on (a) ImageNet-R 10-task class-incremental classification and (b) COCO 5-task class-incremental detection.

Table 5: Ablation study of core components in RaPO.

Method	**ImageNet-R (10 Tasks)**		**COCO (5 Tasks)**	
	$\mathcal{A} \uparrow$	$\mathcal{F} \downarrow$	$\mathcal{A}_b \uparrow$	$\mathcal{F}_b \downarrow$
GRPO	$74.67_{\pm 1.27}$	$20.02_{\pm 4.91}$	$14.64_{\pm 2.18}$	$6.67_{\pm 2.71}$
GRPO + CTAN	$79.43_{\pm 1.68}$	$13.17_{\pm 3.26}$	$16.27_{\pm 1.63}$	$4.73_{\pm 1.92}$
GRPO + Retention Reward	$82.57_{\pm 1.74}$	$8.38_{\pm 2.42}$	$18.11_{\pm 1.31}$	$2.61_{\pm 1.37}$
RaPO (Ours)	$\mathbf{85.92}_{\pm 1.82}$	$\mathbf{4.69}_{\pm 1.71}$	$\mathbf{19.31}_{\pm 1.08}$	$\mathbf{1.39}_{\pm 1.20}$

C.2 Effect of Anchor Policy

We further evaluate the effect of different anchor models used for computing the retention reward, as described in Section 3.2.2. By default, RaPO uses the immediate preceding-task policy π_{t-1} as the anchor. We compare this against three alternatives: an EMA of historical policies π_{ema} (with a decay factor of 0.1), the fixed policy from the first task π_1, and the pre-trained base model π_p without any fine-tuning. As reported in Table 6, the preceding-task anchor achieves the best results on both classification and detection. Replacing it with the EMA anchor π_{ema} slightly reduces $\mathcal{A}$ from 85.92 to 84.86 on ImageNet-R and $\mathcal{A}_b$ from 19.31 to 18.61 on COCO. Further reverting to the task#1 anchor π_1 drops $\mathcal{A}$ to 83.88 and $\mathcal{A}_b$ to 17.34, with corresponding increases in forgetting. Using the pre-trained model π_p as an anchor leads to the most degradation, decreasing $\mathcal{A}$ to 82.68 and $\mathcal{A}_b$ to 16.88. This trend is expected: the preceding-task policy provides a tighter and more recent reference for measuring distribution drift. While the EMA incorporates historical information, it responds slowly to abrupt distribution shifts at task boundaries, making it less effective at constraining the large distribution changes that cause forgetting. Meanwhile, static anchors like π_1 or π_p are progressively less informative about the accumulated knowledge that should be preserved as the task sequence progresses. Nevertheless, even with the weakest anchor, RaPO still outperforms vanilla GRPO by a clear margin on both benchmarks.

Table 6: Effect of retention anchor strategy.

Anchor Policy	ImageNet-R (10 Tasks)		COCO (5 Tasks)	
	$\mathcal{A} \uparrow$	$\mathcal{F} \downarrow$	$\mathcal{A}_b \uparrow$	$\mathcal{F}_b \downarrow$
π_{t-1}	**85.92**$_{\pm 1.82}$	**4.69**$_{\pm 1.71}$	**19.31**$_{\pm 1.08}$	**1.39**$_{\pm 1.20}$
π_{ema}	84.86$_{\pm 2.01}$	5.65$_{\pm 2.78}$	18.61$_{\pm 1.27}$	2.11$_{\pm 1.32}$
π_1	83.88$_{\pm 1.93}$	7.24$_{\pm 2.34}$	17.34$_{\pm 1.48}$	2.82$_{\pm 1.57}$
π_p	82.68$_{\pm 2.17}$	9.43$_{\pm 2.66}$	16.88$_{\pm 1.51}$	3.58$_{\pm 1.93}$

Table 7: Hyperparameter sensitivity in RaPO.

Hyperparameter	ImageNet-R (10 Tasks)		COCO (5 Tasks)	
	$\mathcal{A} \uparrow$	$\mathcal{F} \downarrow$	$\mathcal{A}_b \uparrow$	$\mathcal{F}_b \downarrow$
$\lambda = 0.2$	84.23$_{\pm 1.64}$	7.28$_{\pm 2.47}$	18.13$_{\pm 1.39}$	2.48$_{\pm 1.52}$
$\lambda = 0.5$	85.92$_{\pm 1.82}$	4.69$_{\pm 1.71}$	19.31$_{\pm 1.08}$	1.39$_{\pm 1.20}$
$\lambda = 0.8$	85.37$_{\pm 1.84}$	5.14$_{\pm 1.83}$	18.78$_{\pm 1.21}$	1.76$_{\pm 1.29}$
$\lambda = 1.0$	84.43$_{\pm 2.12}$	6.47$_{\pm 2.16}$	18.09$_{\pm 1.44}$	2.34$_{\pm 1.58}$
$\alpha = 10$	84.73$_{\pm 1.79}$	5.86$_{\pm 2.11}$	18.29$_{\pm 1.32}$	2.39$_{\pm 1.47}$
$\alpha = 20$	85.92$_{\pm 1.82}$	4.69$_{\pm 1.71}$	19.31$_{\pm 1.08}$	1.39$_{\pm 1.20}$
$\alpha = 40$	85.18$_{\pm 1.97}$	5.31$_{\pm 1.96}$	18.68$_{\pm 1.27}$	1.92$_{\pm 1.33}$
$\beta = 0.9$	84.58$_{\pm 2.03}$	6.27$_{\pm 2.36}$	18.14$_{\pm 1.41}$	2.53$_{\pm 1.62}$
$\beta = 0.99$	85.33$_{\pm 1.81}$	5.27$_{\pm 1.91}$	18.82$_{\pm 1.22}$	1.89$_{\pm 1.31}$
$\beta = 0.999$	85.92$_{\pm 1.82}$	4.69$_{\pm 1.71}$	19.31$_{\pm 1.08}$	1.39$_{\pm 1.20}$
$\beta = 0.9999$	85.63$_{\pm 1.87}$	4.91$_{\pm 1.78}$	19.07$_{\pm 1.12}$	1.56$_{\pm 1.24}$

C.3 Hyperparameter Sensitivity

We further examine the sensitivity of RaPO to its three key hyperparameters: the retention weight λ, the retention sensitivity α, and the CTAN smoothing coefficient β. In each experiment, only one hyperparameter is varied while the other two are held fixed. As reported in Table 7, varying λ from 0.2 to 1.0, α from 10 to 40, and β from 0.9 to 0.9999 all produce comparable results on both benchmarks. Across all tested configurations, the last accuracy on ImageNet-R stays within a ~1.7% range and the forgetting rate within a ~2.6% range. Similarly, on COCO, $\mathcal{A}_b$ varies by at most ~1.2% and $\mathcal{F}_b$ by at most ~1.1%. These results indicate that RaPO is robust to hyperparameter choice and does not rely on delicate tuning.

C.4 Impact of Standard Loss-level KL Regularization

We evaluate the effect of the standard KL regularization term by removing it from both GRPO and RaPO during training. As shown in Table 8, removing the KL term does not lead to a large drop in performance for both methods. This suggests that the loss-level KL penalty does not play the primary role of stronger resistance to catastrophic forgetting in RFT, which aligns with the findings of [19, 20].

C.5 Effect of Reasoning and Number of Rollouts

We further examine two practical factors: the effect of removing the reasoning step before answering, and the number of rollouts per prompt. For the reasoning ablation, the default prompt forces the model to reason before outputting the final answer, while the ablated version directly requests the answer without intermediate reasoning. As listed in Table 9, removing reasoning degrades both accuracy and forgetting for GRPO and RaPO. On class-incremental image classification, GRPO drops from 74.67 to 72.83 in $\mathcal{A}$ and RaPO from 85.92 to 83.87 in $\mathcal{A}$, a similar absolute decline of roughly 2% for both methods. On class-incremental object detection, however, the gap is much larger for GRPO: $\mathcal{A}_b$ falls from 14.64 to 9.82, and forgetting nearly doubles. RaPO also declines on detection, but to a far lesser extent ($\mathcal{A}_b$ from 19.31 to 17.64, $\mathcal{F}_b$ from 1.39 to 2.46). We attribute this asymmetry to the retention reward mechanism. During training, the model naturally learns to

Table 8: Effect of loss-level KL regularization.

Method	ImageNet-R (10 Tasks)		COCO (5 Tasks)	
	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}_b\uparrow$	$\mathcal{F}_b\downarrow$
GRPO (w/o KL Loss)	$73.63_{\pm 1.73}$	$22.87_{\pm 4.37}$	$13.08_{\pm 1.66}$	$8.17_{\pm 2.52}$
GRPO (w/ KL Loss)	$74.67_{\pm 1.27}$	$20.02_{\pm 4.91}$	$14.64_{\pm 2.18}$	$6.67_{\pm 2.71}$
RaPO (w/o KL Loss)	$84.38_{\pm 1.86}$	$6.13_{\pm 1.94}$	$17.89_{\pm 1.28}$	$2.72_{\pm 1.36}$
RaPO (w/ KL Loss)	$85.92_{\pm 1.82}$	$4.69_{\pm 1.71}$	$19.31_{\pm 1.08}$	$1.39_{\pm 1.20}$

Table 9: Ablation study on reasoning capacity and number of rollouts (n).

Method	ImageNet-R (10 Tasks)		COCO (5 Tasks)	
	$\mathcal{A}\uparrow$	$\mathcal{F}\downarrow$	$\mathcal{A}_b\uparrow$	$\mathcal{F}_b\downarrow$
GRPO (w/ Reasoning)	$74.67_{\pm 1.27}$	$20.02_{\pm 4.91}$	$14.64_{\pm 2.18}$	$6.67_{\pm 2.71}$
GRPO (w/o Reasoning)	$72.83_{\pm 1.96}$	$22.14_{\pm 5.27}$	$9.82_{\pm 2.21}$	$13.05_{\pm 2.96}$
RaPO (w/ Reasoning)	$85.92_{\pm 1.82}$	$4.69_{\pm 1.71}$	$19.31_{\pm 1.08}$	$1.39_{\pm 1.20}$
RaPO (w/o Reasoning)	$83.87_{\pm 2.04}$	$6.84_{\pm 2.17}$	$17.64_{\pm 1.41}$	$2.46_{\pm 1.62}$
GRPO ($n=4$)	$72.93_{\pm 2.01}$	$23.24_{\pm 5.42}$	$13.73_{\pm 2.27}$	$7.48_{\pm 2.82}$
GRPO ($n=6$)	$74.12_{\pm 1.62}$	$21.07_{\pm 5.11}$	$14.28_{\pm 2.16}$	$6.92_{\pm 2.77}$
GRPO ($n=8$)	$74.67_{\pm 1.27}$	$20.02_{\pm 4.91}$	$14.64_{\pm 2.18}$	$6.67_{\pm 2.71}$
GRPO ($n=10$)	$74.88_{\pm 1.59}$	$19.73_{\pm 4.76}$	$14.78_{\pm 2.23}$	$6.48_{\pm 2.68}$
RaPO ($n=4$)	$83.82_{\pm 2.09}$	$7.16_{\pm 2.31}$	$17.83_{\pm 1.42}$	$2.78_{\pm 1.67}$
RaPO ($n=6$)	$85.13_{\pm 1.86}$	$5.36_{\pm 1.91}$	$18.83_{\pm 1.23}$	$1.82_{\pm 1.32}$
RaPO ($n=8$)	$85.92_{\pm 1.82}$	$4.69_{\pm 1.71}$	$19.31_{\pm 1.08}$	$1.39_{\pm 1.20}$
RaPO ($n=10$)	$86.07_{\pm 1.77}$	$4.53_{\pm 1.68}$	$19.36_{\pm 1.11}$	$1.34_{\pm 1.18}$

produce implicit reasoning traces to maximize the task reward without an explicit chain-of-thought. RaPO then preferentially reinforces rollouts that remain close to the preceding-task policy, which anchors the policy to these already learned implicit reasoning patterns. Consequently, even when the prompt omits the explicit reasoning instruction, the retention signal continues to guide the policy toward trajectories that exhibit a similar expressive pattern. In contrast, GRPO's behavior is fully prompt-dependent and changes sharply when reasoning is removed.

For the rollout experiments, we vary the number of sampled responses per prompt from 4 to 10. Both GRPO and RaPO improve steadily as the number of rollouts increases from 4 to 8. Beyond 8, the gains become marginal. For RaPO on ImageNet-R, moving from 8 to 10 rollouts improves $\mathcal{A}$ from 85.92 to 86.07 and $\mathcal{F}$ from 4.69 to 4.53. On COCO, $\mathcal{A}_b$ changes from 19.31 to 19.36 and $\mathcal{F}_b$ from 1.39 to 1.34. These differences are small, suggesting that 8 rollouts already provide sufficient diversity for stable advantage estimation, and further increasing the rollout count yields negligible improvement with additional computational overhead.

D Analytics Experiments

D.1 Impact of Different Policy Optimization Methods

To examine whether RaPO is tied to GRPO, we integrate its two core components, the retention reward and CTAN, into SAPO [80], a recent policy optimization algorithm. As shown in Table 10, SAPO alone yields an $\mathcal{A}$ of 76.83% and $\mathcal{F}$ of 18.07% on ImageNet-R, together with an $\mathcal{A}_b$ of 15.28% and $\mathcal{F}_b$ of 6.13% on COCO. After equipping SAPO with the retention reward and CTAN, accuracy rises to 84.46% on classification and 18.43% on detection, while forgetting drops to 6.52% and 2.07%, respectively. These improvements indicate that the retention mechanism and advantage normalization are not specific to a single optimizer. Rather, they serve as generic components that can strengthen different policy optimization algorithms for visual continual learning.

Table 10: Effect of RaPO components applied to SAPO.

Method	ImageNet-R (10 Tasks)		COCO (5 Tasks)	
	$\mathcal{A} \uparrow$	$\mathcal{F} \downarrow$	$\mathcal{A}_b \uparrow$	$\mathcal{F}_b \downarrow$
SAPO	$76.83_{\pm 1.94}$	$18.07_{\pm 4.24}$	$15.28_{\pm 1.84}$	$6.13_{\pm 2.38}$
SAPO+Ours	$84.46_{\pm 1.72}$	$6.52_{\pm 2.14}$	$18.43_{\pm 1.33}$	$2.07_{\pm 1.42}$

Table 11: Performance comparison across different MLLM backbones (Qwen2-VL vs. Qwen2.5-VL). RaPO consistently mitigates forgetting and improves overall accuracy regardless of the underlying model version.

Method	ImageNet-R (10 Tasks)		COCO (5 Tasks)	
	$\mathcal{A} \uparrow$	$\mathcal{F} \downarrow$	$\mathcal{A}_b \uparrow$	$\mathcal{F}_b \downarrow$
Qwen2-VL				
GRPO	$74.67_{\pm 1.27}$	$20.02_{\pm 4.91}$	$14.64_{\pm 2.18}$	$6.67_{\pm 2.71}$
RaPO (Ours)	$\mathbf{85.92}_{\pm 1.82}$	$\mathbf{4.69}_{\pm 1.71}$	$\mathbf{19.31}_{\pm 1.08}$	$\mathbf{1.39}_{\pm 1.20}$
Qwen2.5-VL				
GRPO	$76.41_{\pm 1.38}$	$20.73_{\pm 4.82}$	$16.89_{\pm 2.31}$	$7.52_{\pm 2.58}$
RaPO (Ours)	$\mathbf{88.27}_{\pm 1.72}$	$\mathbf{5.22}_{\pm 1.68}$	$\mathbf{21.68}_{\pm 1.19}$	$\mathbf{2.11}_{\pm 1.09}$

D.2 Impact of Different MLLMs

We assess both GRPO and RaPO on a more powerful Qwen2.5-VL model (3B for classification and 7B for detection). As shown in Table 11, both GRPO and RaPO gain accuracy from the stronger base models, while forgetting also increases slightly. This is because more capable MLLMs reach higher accuracy on early tasks, resulting in a higher performance ceiling from which any subsequent degradation is measured. Since forgetting captures the absolute drop from historical best accuracy, a stronger starting point naturally amplifies the measured forgetting for both methods.

D.3 Qualitative Results

We present qualitative examples of class-incremental image classification on ImageNet-R, ImageNet-A, CUB-200-2011, and TinyImageNet. For each input image, we present the model response after learning the first task and the response on the same validation image after the model finishes learning the final task. As shown in Figure 6, at the initial task, both GRPO and RaPO correctly recognize the target class. After the full continual learning sequence, however, GRPO often changes its final answer to a semantically related category, which reflects forgetting of previously learned classes. In contrast, our RaPO preserves the correct class more consistently. The reasoning process and final answer further make this difference clear: RaPO continues to rely on the visual cues that support the old class, whereas GRPO becomes more susceptible to partial cues that lead to inaccurate decisions after more tasks are learned. Meanwhile, Figure 7 follows the same protocol for class-incremental object detection on COCO. We show the model response on the same validation image after learning the initial task that first introduces the earlier category, and again after the model finishes learning the final task. At the initial task, both GRPO and RaPO correctly localize the bicycle. After the full continual learning sequence, GRPO detects later categories such as traffic light, person, and car, but misses the bicycle learned earlier. This behavior reflects forgetting in continual detection: the bicycle is small and visually coupled with the rider, so its response is more easily dominated by newly introduced categories after later-task training. Conversely, RaPO still locates the bicycle while also recognizing the later categories, showing stronger retention of earlier knowledge throughout continual learning.

E Limitations

Although RaPO demonstrates broad effectiveness across diverse visual continual learning settings, several limitations suggest promising directions for future work. First, RFT-based MLLM training is computationally expensive, and our hardware budget limits us from scaling to longer task sequences. Second, this study focuses on CIL and DIL with Qwen2-VL at the 2B and 7B scales, leaving open

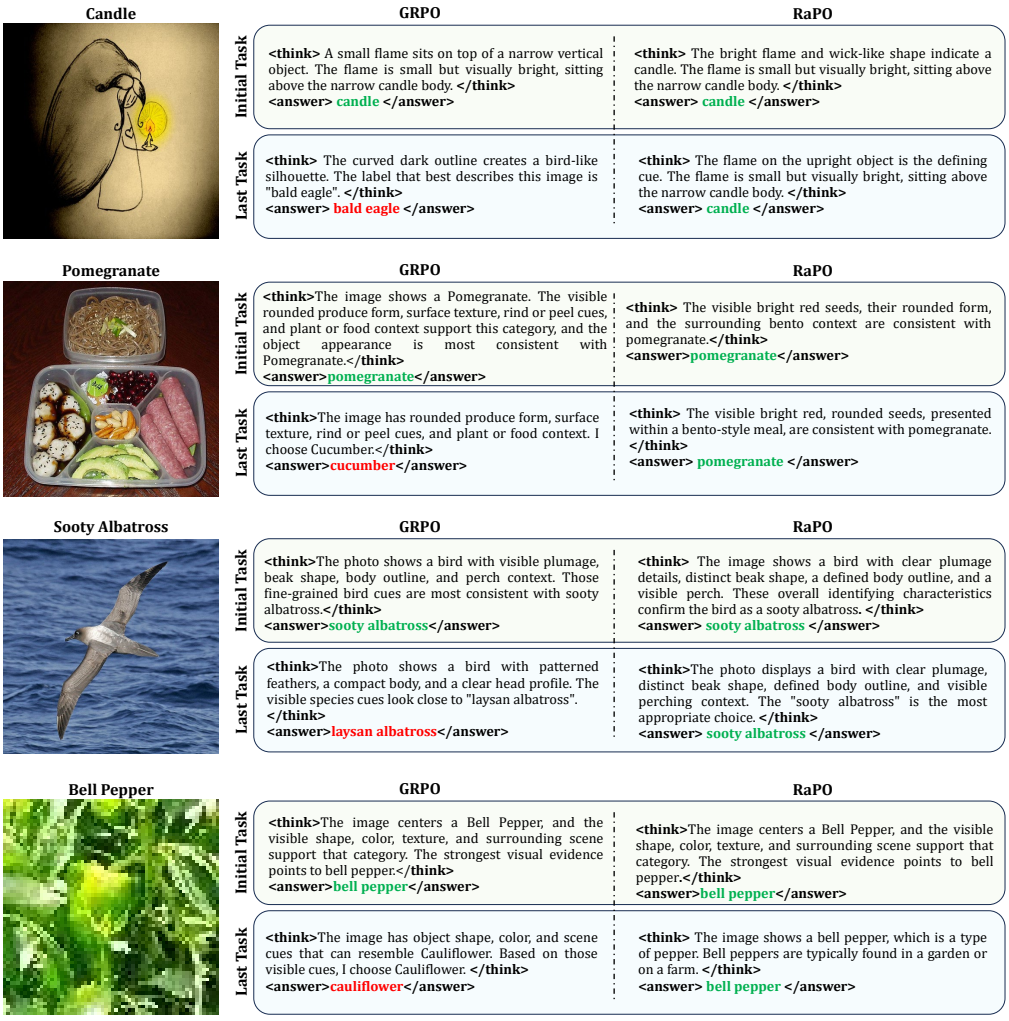


Figure 6: Qualitative class-incremental image classification examples from different datasets.

whether RaPO extends to larger MLLMs and more complex task formulations. Third, to isolate the properties of RFT in visual continual learning, we adopt closed-world settings with explicit task boundaries and verifiable rewards. Extending RaPO to open-world streams without such boundaries is an important direction for future work. However, these limitations do not weaken the central contribution: RaPO is the first systematic study of RFT for visual continual learning and offers a simple yet effective strategy that consistently suppresses catastrophic forgetting phenomenon across diverse settings.

F Impact Statement

This work proposes a reinforcement fine-tuning method for visual continual learning. All experiments rely on publicly available datasets and open-source pre-trained models, which ensures that no private or sensitive data is involved. Our research is domain-agnostic, not tailored towards potentially harmful applications such as surveillance or misinformation dissemination, and therefore poses no ethical concerns.



Figure 7: Qualitative class-incremental object detection examples on COCO 2017 dataset.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions and scope, which are supported by the method in Section 3 and the experiments in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Limitations and potential future works are discussed in Section E.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Full assumptions and theoretical proofs are provided in Appendix A.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided implementation details in Section 4 and Appendix B. Code is publicly available at https://github.com/LMMMEng/RaPO.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is publicly available at https://github.com/LMMMEng/RaPO, with implementation details in Section 4 and Appendix B. All datasets used are public and cited in the paper.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All training and evaluation details (e.g., data splits, hyperparameters, and prompt designs) are described in Section 4 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All experimental results report mean and standard deviation over multiple runs, as noted in the experimental setup.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided the type and number of GPU devices used in Section 4.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This paper involves no ethical issues, as it focuses on general machine learning topics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work focuses on general algorithmic frameworks for continual learning and does not involve sensitive data or real ethical risks. Its societal impacts are confined to positive technical progress with no adverse social consequences.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: This paper does not release any high-risk datasets, pre-trained models, or generative models. Therefore, no additional safeguards for responsible release are required.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All adopted datasets, baseline models, and open-source resources are publicly available. We properly cite their original papers and strictly comply with the corresponding licenses and terms of use.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: This paper focuses on technical contribution and does not introduce new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper involves no crowdsourcing experiments and does not conduct any research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve any human subjects or participant studies. No potential risks to human participants exist.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We leverage LLMs to revise the manuscript after initial drafting.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.