RL FORGETS! TOWARDS CONTINUAL POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Continual post-training is becoming a central paradigm for adapting vision-language models to evolving tasks. Recent studies report that reinforcement learning is less prone to forgetting than supervised fine-tuning, motivating the view that RL is inherently resistant to forgetting. However, this view remains insufficiently validated, as existing evidence is largely drawn from outdated or homogeneous benchmarks. We revisit this assumption by introducing MRCL, a Multimodal Reasoning Continual Learning benchmark built from recent and diverse multimodal reasoning tasks. Experiments on MRCL show that standard reinforcement learning still suffers from catastrophic forgetting during continual post-training. We trace the failure to an objective mismatch. The KL regularization used in common policy optimization methods is evaluated on current-task data, whereas forgetting is caused by behavioral drift on prior-task distributions. To address this problem, we propose Continual Policy Optimization (CPO), a replay-free method grounded in a prior-task behavioral KL objective. CPO derives a local Fisher surrogate from the historical KL objective and uses parameter movement as a gradient-free proxy for Fisher sensitivity, enabling sparse regularization with negligible additional overhead. Experiments on three model scales and comparisons with multiple RL baselines show that CPO consistently reduces forgetting while maintaining effective adaptation and preserving broader pretrained capabilities.

1 INTRODUCTION

Vision-language models (VLMs) have demonstrated strong performance across a wide range of real-world tasks (Comanici et al., 2025; Achiam et al., 2023; Radford et al., 2021). As VLMs are increasingly deployed in dynamic environments, they require continually adapting to new data and tasks. However, such a process is hindered by the fundamental challenge of catastrophic forgetting (French, 1999), where learning new tasks degrades previously acquired capabilities. Most prior work addresses the problem under supervised fine-tuning (SFT) (Gu et al., 2026a; Chen et al., 2025c; Kurniawan et al., 2024). Recently, reinforcement learning (RL) has emerged as a promising alternative for continual post-training (Lai et al., 2025), with several studies reporting substantially less forgetting than SFT (Shenfeld et al., 2026; Zhang et al., 2026; Jin et al., 2025; Chen et al., 2025b). These findings suggest an appealing possibility: *is RL intrinsically resistant to forgetting in continual post-training?*

We show that the current evidence is incomplete. Existing evaluations have two blind spots that can make forgetting appear milder than it is. First, outdated benchmarks may obscure forgetting, as datasets that substantially predate the adapted model carry a higher risk of pretraining exposure. For example, Lai et al. (2025) evaluate Qwen2.5-VL (Bai et al., 2025b) on datasets released two to seven years before the model, potentially making the corresponding capabilities easier to retain. Second, some studies evaluate RL on narrow task distributions. For instance, Zhang et al. (2026) draw conclusions from a single jigsaw task. To provide a more reliable evaluation, we introduce MRCL, a Multimodal Reasoning Continual Learning benchmark built from diverse and recent multimodal reasoning datasets. Its construction jointly considers release recency, teacher-forced answer-token NLL, and zero-shot performance, selecting tasks that are recent and less familiar to the base model while leaving substantial scope for adaptation. Fig. 1 compares MRCL with RFT-SFT, the benchmark introduced by Lai et al. (2025), and shows that MRCL generally exhibits higher answer-token negative log likelihood (NLL). Together with its low zero-shot performance, providing complemen-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

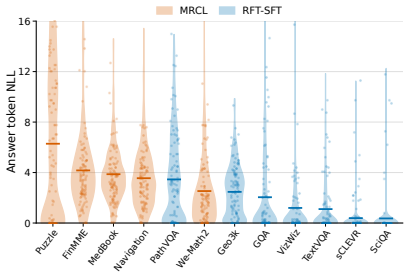
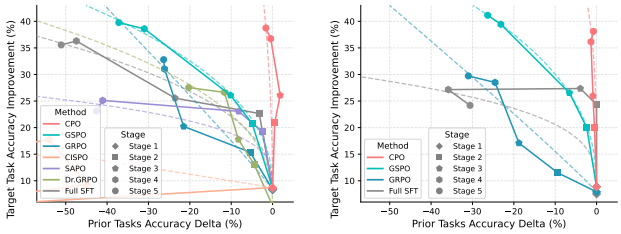


Figure 1: Teacher-forced answer-token NLL across continual learning benchmarks. Results are obtained with a frozen Qwen3-VL-4B. Higher NLL indicates greater surprisal of the canonical answers to the model. Horizontal bars denote task means. Details are shown in Appendix. E.2.



(a) Qwen3-VL-4B (b) Qwen3-VL-8B

Figure 2: Plasticity–stability Pareto frontier during continual post-training. The y-axis reports the cumulative mean accuracy improvement (CAI) on current tasks, while the x-axis reports the cumulative mean accuracy delta (CAD) on previously learned tasks. Definitions of CAI and CAD are in Appendix E.1. Both RL and SFT exhibit catastrophic forgetting, whereas our method preserves prior knowledge more effectively while maintaining plasticity.

tary evidence that MRCL remains challenging for the base model. Experiments on MRCL reveal that continual RL still suffers from catastrophic forgetting, as shown in Fig. 2, challenging the view that RL naturally solves continual learning.

We next examine why existing RL objectives fail to prevent forgetting. The key issue is an objective mismatch. Shenfeld et al. (2026) argue that the KL regularization in GRPO (Shao et al., 2024) can directly reflect forgetting, whereas Lai et al. (2025) and Chen et al. (2025b) suggest that KL regularization is not the main factor behind reduced forgetting. Our experiments further show that the GRPO KL constraint fails to mitigate forgetting and instead reduces plasticity on new tasks. GRPO evaluates KL divergence on current-task data, $\mathbb{E}_{x\sim\mathcal{T}_{new}}\left[\mathrm{KL}\left(\pi_\theta\|\pi_{old}\right)\right]$, which is mainly designed to prevent over-optimization (Shao et al., 2024). Recent methods such as DAPO (Yu et al., 2025) and GSPO (Zheng et al., 2025) even remove the KL term to improve learning capacity. Catastrophic forgetting is caused by behavioral drift on previously learned tasks. Our theoretical analysis shows that the relevant constraint should be evaluated on prior-task data, $\mathbb{E}_{x\sim\mathcal{T}_{old}}\left[\mathrm{KL}\left(\pi_{old}\|\pi_\theta\right)\right]$. Current-task KL can serve as a forgetting proxy only when the new and prior task distributions are sufficiently close. In continual learning, prior-task data are often unavailable due to efficiency constraints and data privacy concerns (Wang et al., 2024b; Zhou et al., 2024; Rolnick et al., 2019), motivating a replay-free mechanism for constraining prior-task behavioral drift.

To constrain prior-task behavior without storing historical data, a classical approach is to regularize parameters according to their estimated importance. EWC (Kirkpatrick et al., 2017) and related methods (Aljundi et al., 2018; Liu & Chang, 2026), developed primarily for supervised continual learning, estimate importance from Fisher information, parameter gradients, or optimization trajectories. Obtaining and maintaining these statistics requires additional backward computation, parameter-wise states, or training-time bookkeeping, making them difficult to scale in distributed post-training (Rajbhandari et al., 2020; Narayanan et al., 2021). In our RL setting, we instead derive parameter sensitivity directly from the prior-task behavioral KL and show that its local Fisher structure can be approximated by empirical parameter movement accumulated during training. Based on the insight, CPO maintains a cumulative set of large-movement parameters and applies masked sparse movement regularization during sequential policy optimization.

In summary, our key contributions are as follows:

- We introduce MRCL, a continual learning benchmark combining release recency, model familiarity, and zero-shot performance to reduce prior exposure risk while retaining scope for adaptation. MRCL provides a diverse and challenging testbed for continual post-training.
- We propose CPO, a continual RL method grounded in the prior-task behavioral KL. CPO derives a local Fisher surrogate from this objective and uses parameter movement as a gradient-free proxy for Fisher sensitivity, with negligible extra computation.
- Extensive experiments across multiple VLM scales demonstrate that CPO substantially reduces forgetting while maintaining effective adaptation to new tasks.

2 MULTIMODAL REASONING CONTINUAL LEARNING BENCHMARK

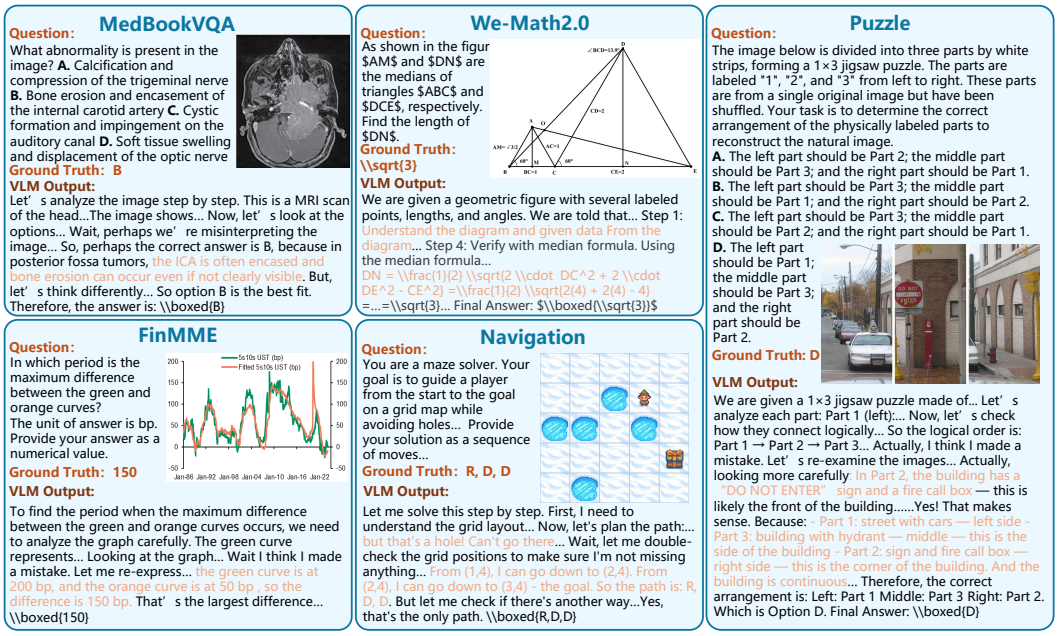


Figure 3: Overview of the MRCL benchmark. MRCL combines recently released multimodal reasoning tasks to evaluate both the plasticity and stability of VLMs during continual post-training.

To rigorously evaluate continual reinforcement learning while reducing the interference of pretraining data contamination, we construct MRCL by jointly considering benchmark recency, model familiarity, and scope for adaptation. We prioritize recently released datasets to reduce the risk that the base model has already encountered closely related data during pretraining. We further examine teacher-forced answer-token NLL to assess conditional familiarity with the canonical answer distribution, and zero-shot performance to ensure that the selected tasks leave sufficient scope for adaptation. As shown in Fig. 1, MRCL generally exhibits higher answer-token NLL than RFT-SFT, suggesting lower conditional-answer familiarity to the base model. MRCL covers diverse forms of multimodal reasoning, including medical understanding, navigation, geometry, visual-spatial reasoning, and financial chart analysis. Fig. 3 presents representative query and answer formats, and Tab. 1 summarizes the basic statistics of each task.

Table 1: Basic statistics and answer formats of the evaluated tasks.

Task	Release Date	# Train	# Test	Answer Format
MedBookVQA	2025-06	1.4k	0.6k	Multiple Choice
Navigation	2025-10	1.5k	0.5k	Direction Sequence
We-Math2.0	2025-08	5.7k	3.7k	Numerical & Multiple Choice
Puzzle	2025-10	6.0k	0.4k	Multiple Choice
FinMME	2025-05	8.0k	1.0k	Numerical & Multi-Answer Choice

3 METHOD

This section formalizes continual RL for VLM post-training and shows that prior-task performance degradation is bounded by behavioral KL on historical queries and old-policy prefixes. Directly evaluating the behavioral KL requires access to historical task data and old-policy rollouts. We therefore derive a parameter-space relaxation and develop a simple and effective approach named Continual Policy Optimization.

3.1 Continual RL Objective and Forgetting Penalty

We formulate continual reinforcement learning on a sequence of tasks denoted by $\{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$. Each task $\mathcal{T}_i$ is associated with an input distribution $\mathcal{D}_i$, where a sample $x \sim \mathcal{D}_i$ represents a multimodal query, typically comprising an image and a textual instruction. Conditioned on x, a policy π autoregressively generates a response sequence $y_{1:H}$:

$$P_\pi(y_{1:H} \mid x) = \prod_{t=1}^{H} \pi(y_t \mid x, y_{<t}), \tag{1}$$

where H denotes the maximum generation horizon, and t indexes the decoding step. Upon completion of the sequence, the environment evaluates the response quality using a terminal scalar reward $R_i(x, y_{1:H})$. The expected objective for task $\mathcal{T}_i$ under policy π is defined as $\eta_i(\pi) = \mathbb{E}_{x\sim\mathcal{D}_i,\, y\sim P_\pi(\cdot|x)}[R_i(x, y)]$.

During sequential adaptation to the i-th task, the learning process can be formulated as a dual-objective optimization problem that improves plasticity on new data while maintaining stability across prior tasks. Let $\pi^{(i)}$ denote the policy optimized on $\mathcal{T}_i$. The oracle continual RL objective is

$$\max_{\pi^{(i)}} \left[\eta_i(\pi^{(i)}) - \lambda \sum_{k=1}^{i-1} \Big(\eta_k(\pi^{(k)}) - \eta_k(\pi^{(i)}) \Big) \right], \tag{2}$$

where $\lambda > 0$ indicates the critical stability-plasticity trade-off. To optimize the joint objective, we explicitly constrain the performance degradation on previous tasks, represented by the penalty term $\eta_k(\pi^{(k)}) - \eta_k(\pi^{(i)})$. In the following, we establish a theoretical surrogate for the behavioral drift.

3.2 Prior-Task Behavioral KL Bounds Forgetting

To quantify the behavioral drift that causes catastrophic forgetting on a prior task $\mathcal{T}_{i-1}$, we analyze token-level KL divergence under the prefix visitation distribution of the old policy. We first formulate the oracle objective assuming previous-task data is accessible. Formally, let $\tilde{\pi}$ denote the old policy preserved after training on $\mathcal{T}_{i-1}$ and let π denote the current policy optimized on $\mathcal{T}_i$, the autoregressive factorization induces $P_{\tilde{\pi}}(\cdot \mid x)$ and $P_\pi(\cdot \mid x)$, and the old-task performance is given by $\eta_{i-1}(\tilde{\pi})$. Assuming the terminal reward function is bounded, its maximum variation range is defined as $B_{i-1} = \sup_{x,y,y'} |R_{i-1}(x, y) - R_{i-1}(x, y')|$, where y and y' represent any two possible response sequences generated for the same input query x.

To capture the distribution of generated prefixes at each step during the autoregressive generation, we define the normalized prefix visitation distribution under the old policy:

$$d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}(t, x, y_{<t}) = \frac{1}{H} \mathcal{D}_{i-1}(x) P_{\tilde{\pi}}(y_{<t} \mid x), \qquad t \in \{1, \ldots, H\}. \tag{3}$$

The mean token-level KL divergence over this visitation distribution characterizes the behavioral discrepancy between the two policies:

$$\bar{D}_{\mathrm{KL}}^{d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}}(\tilde{\pi}, \pi) = \mathbb{E}_{(t,x,y_{<t})\sim d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}} \left[D_{\mathrm{KL}}\big(\tilde{\pi}(\cdot \mid x, y_{<t}) \,\|\, \pi(\cdot \mid x, y_{<t})\big) \right]. \tag{4}$$

The above formulation differs from the KL regularization typically used in standard reinforcement learning algorithms such as GRPO, which is computed only over inputs from the current task $\mathcal{T}_i$. Current-task regularization can prevent forgetting only when the new data distribution is sufficiently close to the old data distribution. The KL divergence in Eq. 4 instead provides a theoretical guarantee by bounding performance degradation on prior tasks, as shown in the following proposition.

Proposition 3.1 (Finite horizon KL forgetting bound)**.** *Given a bounded reward range $B_{i-1} < \infty$, the expected return degradation on task $\mathcal{T}_{i-1}$ is bounded by the token level KL divergence:*

$$|\eta_{i-1}(\tilde{\pi}) - \eta_{i-1}(\pi)| \le B_{i-1} \sqrt{\frac{H}{2} \bar{D}_{\mathrm{KL}}^{d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}}(\tilde{\pi}, \pi)}. \tag{5}$$

The bound follows by first relating expected return degradation to total variation distance, then applying Pinsker's inequality and Jensen's inequality (Cover & Thomas, 2006) to control it by the expected sequence-level KL. The autoregressive KL chain rule further decomposes the sequence-level divergence into the mean token-level KL in Eq. 5. The detailed proof is provided in Appendix B.1.

Eq. 5 shows that forgetting on $\mathcal{T}_{i-1}$ can be controlled by keeping the current policy π close to the old policy $\tilde{\pi}$ under the historical prefix distribution. However, directly optimizing the joint objective in Eq. 2 requires access to historical distributions $\mathcal{D}_{i-1}$, while optimizing the KL constraint in Eq. 4 requires sampling old-policy prefixes and forwarding $\tilde{\pi}$. For large-scale post-training, these requirements are impractical due to computational overhead and data privacy constraints.

3.3 Fisher Relaxation of Prior-Task Behavioral KL

To eliminate the dependence of the oracle KL divergence on historical data, we derive a replay-free surrogate. We denote the old and active policies as $\pi_{\tilde{\theta}} \equiv \tilde{\pi}$ and $\pi_\theta \equiv \pi$, where $\tilde{\theta}$ represents the frozen weights from task $\mathcal{T}_{i-1}$ and θ denotes the active weights optimized on $\mathcal{T}_i$. Since the KL divergence reaches its minimum when the two distributions are identical, we have $K_{i-1}(\tilde{\theta}) = 0$. The score identity implies that the first-order derivative vanishes $\nabla_\theta K_{i-1}(\theta)|_{\theta=\tilde{\theta}} = 0$. Thus, defining the oracle old-task KL and expanding it around $\tilde{\theta}$ gives:

$$\begin{aligned} K_{i-1}(\theta) &\coloneqq \bar{D}_{\mathrm{KL}}^{d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}}(\pi_{\tilde{\theta}}, \pi_\theta) \\ &= \frac{1}{2}(\theta - \tilde{\theta})^\top H_{\tilde{\theta}}^{\mathcal{T}_{i-1}}(\theta - \tilde{\theta}) + R_{i-1}^{(3)}(\theta), \end{aligned} \tag{6}$$

where $H_{\tilde{\theta}}^{\mathcal{T}_{i-1}} = \nabla_\theta^2 K_{i-1}(\theta)|_{\theta=\tilde{\theta}}$ denotes the Hessian matrix, and $R_{i-1}^{(3)}(\theta)$ represents the third-order Taylor remainder. For the KL divergence, the Hessian is mathematically equivalent to the Fisher Information Matrix (FIM) of the old task:

$$F_{\tilde{\theta}}^{\mathcal{T}_{i-1}} = \mathbb{E}_{(t,x,y_{<t})\sim d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}} \mathbb{E}_{a\sim\tilde{\pi}(\cdot|x,y_{<t})} \left[\nabla_\theta \log \pi_\theta(a \mid x, y_{<t}) \nabla_\theta \log \pi_\theta(a \mid x, y_{<t})^\top\right]_{\theta=\tilde{\theta}}. \tag{7}$$

Thus, $H_{\tilde{\theta}}^{\mathcal{T}_{i-1}} = F_{\tilde{\theta}}^{\mathcal{T}_{i-1}}$. Omitting the higher-order remainder term defines a second-order local surrogate objective:

$$Q_{i-1}(\theta; \tilde{\theta}) = \frac{1}{2}(\theta - \tilde{\theta})^\top F_{\tilde{\theta}}^{\mathcal{T}_{i-1}}(\theta - \tilde{\theta}). \tag{8}$$

To reduce the prohibitive computational and memory costs of the dense FIM in large scale models, we relax the dense matrix $F_{\tilde{\theta}}^{\mathcal{T}_{i-1}}$ with its diagonal components yields:

$$Q_{i-1}^{\mathrm{diag}}(\theta; \tilde{\theta}) = \frac{1}{2}\sum_{k=1}^{d} F_{kk}^{\mathcal{T}_{i-1}}(\theta_k - \tilde{\theta}_k)^2, \tag{9}$$

where the k-th diagonal entry is computed as:

$$F_{kk}^{\mathcal{T}_{i-1}} = \mathbb{E}_{(t,x,y_{<t})\sim d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}} \mathbb{E}_{a\sim\tilde{\pi}(\cdot|x,y_{<t})} \left[\left(\frac{\partial \log \pi_\theta(a \mid x, y_{<t})}{\partial \theta_k}\right)^2\right]_{\theta=\tilde{\theta}}. \tag{10}$$

The Fisher in Eq. 10 differs from EWC supervised empirical Fisher computed from fixed input-label pairs (Kirkpatrick et al., 2017). Eq. 10 is an old-policy token-level Fisher averaged over autoregressive prefix visitation, and thus arises from the local curvature of the prior-task behavioral KL.

Eq. 9 and Eq. 10 show that the oracle KL constraint admits a local parameter-space surrogate that penalizes parameter movement according to the sensitivity of the old policy distribution. However, explicitly estimating $F_{kk}^{\mathcal{T}_{i-1}}$ still requires sampling from the old-policy prefix distribution and additional gradient computation. This additional backward computation undermines the efficiency required for scalable continual reinforcement learning, motivating a gradient-free proxy.

3.4 Parameter Movement as a Gradient-Free Fisher Proxy

To avoid the computational cost of explicit Fisher estimation, we use empirical parameter movement during task training as a proxy for parameter importance. Let $\theta^{(i-2)}$ and $\tilde{\theta}$ denote the model

Algorithm 1 Continual Policy Optimization

Require: Task sequence $\{\mathcal{T}_i\}_{i=1}^{T}$, top ratio p, regularization strength λ
1: Initialize protected set $\mathcal{S}^{(0)} \leftarrow \varnothing$ and checkpoint $\theta^{(0)}$
2: **for** $i = 1$ to T **do**
3: Initialize $\tilde{\theta} \leftarrow \theta^{(i-1)}$
4: Train on $\mathcal{T}_i$ with Eq. 15 using protected set $\mathcal{S}^{(i-1)}$
5: Save the task-end checkpoint $\theta^{(i)}$
6: Add the top $p\%$ moving coordinates to obtain $S^{(i)}$ with Eq. 16
7: **end for**

checkpoints after tasks $\mathcal{T}_{i-2}$ and $\mathcal{T}_{i-1}$, respectively. The movement score for the k-th coordinate on task $\mathcal{T}_{i-1}$ is

$$s_k^{(i-1)} = \left|\tilde{\theta}_k - \theta_k^{(i-2)}\right|. \tag{11}$$

The proxy is theoretically motivated by the relationship between policy gradient updates and Fisher sensitivity. Let $h_t = (x, y_{<t})$ denote the generation history and let $\phi_k(t, h_t, a)$ be the token-level score evaluated on the old policy:

$$\phi_k(t, h_t, a) = \frac{\partial}{\partial \theta_k} \log \pi_\theta(a \mid h_t)\Big|_{\theta=\tilde{\theta}}. \tag{12}$$

Specifically, the diagonal elements of the old task FIM correspond to the expectation of the squared score function:

$$F_{kk}^{\mathcal{T}_{i-1}} = \mathbb{E}_{(t,h_t)\sim d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}} \left[\mathbb{E}_{a\sim\tilde{\pi}(\cdot|h_t)}[\phi_k(t, h_t, a)^2]\right]. \tag{13}$$

Since policy gradient updates are proportional to reward-weighted scores, parameters with larger Fisher sensitivity tend to exhibit larger accumulated movement during training.

Proposition 3.2 (Parameter movement as a noisy Fisher proxy)**.** *Under locally stationary approximation with learning rate η for stochastic updates, the expected squared movement satisfies:*

$$\mathbb{E}\left[\left|\tilde{\theta}_k - \theta_k^{(i-2)}\right|^2\right] = \eta^2 \sigma_A^2 C_k F_{kk}^{\mathcal{T}_{i-1}}. \tag{14}$$

where σ_A^2 is the second moment of the advantage and C_k is a coordinate coefficient reflecting both reward-score alignment and stochastic-gradient variance. Consequently, when C_k is approximately uniform across coordinates, ranking by the expected squared movement $\mathbb{E}[(s_k^{(i-1)})^2]$ is consistent with ranking by $F_{kk}^{\mathcal{T}_{i-1}}$.

We defer the rigorous proof and the explicit formulation of C_k to Appendix B.2. Proposition 3.2 shows that empirical parameter movement approximately preserves Fisher importance ordering when C_k varies mildly. Appendix C further shows that approximate uniformity is not necessary. It is sufficient that a substantial fraction of the Fisher-important coordinates have sufficiently large C_k, while C_k may vary arbitrarily elsewhere.

3.5 Cumulative Masked Movement Regularization

CPO maintains a cumulative protected set $\mathcal{S}^{(i-1)}$ before training task $\mathcal{T}_i$, initialized as $\mathcal{S}^{(0)} = \varnothing$. Let $\tilde{\theta} \equiv \theta^{(i-1)}$ be the previous checkpoint and let $m^{(i-1)}$ be the binary mask of $\mathcal{S}^{(i-1)}$. We optimize the current policy with RL loss $\mathcal{L}_i^{\text{RL}}(\theta)$ and masked L_1 movement regularization:

$$\mathcal{L}_i(\theta) = \mathcal{L}_i^{\text{RL}}(\theta) + \frac{\lambda}{\|m^{(i-1)}\|_0} \left\| m^{(i-1)} \odot (\theta - \tilde{\theta}) \right\|_1. \tag{15}$$

When $\mathcal{S}^{(i-1)}$ is empty, the regularization term is omitted. We employ L_1 regularization because the movement proxy provides ranking information rather than calibrated Fisher magnitudes. While the Fisher regularizer in Eq. 9 assigns continuous parameter weights, top-p movement selection yields a binary support. An unweighted masked L_2 regularizer treats all protected parameters equally and can still allow dense parameter drift. Masked L_1 regularization promotes sparse movement,

encouraging most selected parameters to remain close to the previous checkpoint. After task T_i, we rank parameter movement independently within each trainable parameter and add the top p moving coordinates from each tensor to the protected set. Let $\mathcal{G} = \{G_1, \ldots, G_M\}$ denote the partition of trainable coordinates induced by the M parameter tensors. We update the protected set as

$$S^{(i)} = S^{(i-1)} \cup \bigcup_{G_m \in \mathcal{G}} \mathrm{TopP}_p \left(\left\{ \left| \theta_k^{(i)} - \tilde{\theta}_k \right| \right\}_{k \in G_m} \right). \tag{16}$$

We summarize the overall training procedure in Algorithm 1.

4 Experiments

4.1 Experimental Setup

Evaluation Metrics. We report three aggregated accuracy metrics and final accuracy $a_{T,j}$ of each task. Let T denote the total number of tasks and $a_{i,j}$ represent the accuracy on task j evaluated after the model completes training on task i. Mean Finetune Accuracy (MFT) measures the average performance on each task when evaluated immediately after training, aiming to quantify the ability to learn new tasks. Mean Final Accuracy (MFN) measures the average accuracy at the end of the training sequence, reflecting the severity of forgetting. Mean Task Accuracy (MTA) provides a holistic measure of performance stability by averaging the accuracy throughout learning process.

$$\text{MFT} = \frac{1}{T}\sum_{i=1}^{T} a_{i,i}, \quad \text{MFN} = \frac{1}{T}\sum_{j=1}^{T} a_{T,j}, \quad \text{MTA} = \frac{1}{T}\sum_{j=1}^{T}\left(\frac{1}{T-j+1}\sum_{i=j}^{T} a_{i,j}\right). \tag{17}$$

Reward Design. The reward consists of an accuracy reward and a length penalty. The accuracy reward aligns exactly with the aforementioned evaluation scoring rules for most tasks, except for *Navigation*, where we provide dense stepwise feedback by awarding $1/k$ points for each correctly predicted step as k is the total number of steps in the ground truth path. To prevent the policy from collapsing into degenerate short responses, we apply a length penalty of -0.5 if the generated output contains fewer than 20 tokens.

4.2 Continual Learning Performance on MRCL

We evaluate all methods over a five-stage continual learning sequence, with each dataset treated as one task. As shown in Tab. 2, 5, and 6, CPO achieves the highest MFN and MTA across all three model scales while maintaining competitive MFT. On Qwen3-VL-4B, CPO reaches 71.56% MFN, exceeding the highest competing RL baseline by 14.73%. The same trend holds on the 2B and 8B models, indicating that the retention gains persist across model scales. GRPO*, which uses current-task KL regularization, consistently lowers MFT without yielding a consistent MFN gain, suggesting that such regularization limits adaptation without reliably improving retention.

Table 2: Comparison of our proposed method with baseline approaches on Qwen3-VL-4B.

Method	Final Accuracy on Each Task (%)					Aggregated Acc. (%)		
	MedBook	Naviga.	We-Math2	Puzzle	FinMME	MFT↑	MFN↑	MTA↑
Zero-shot[1]	56.88	3.31	34.60	36.00	41.60	–	–	34.48
Full SFT	56.71	3.12	37.84	**84.50**	37.90	70.04	44.01	55.85
Zero-shot[2]	49.75	20.77	38.27	24.75	38.95	–	–	34.50
GRPO* (Shao et al., 2024)	63.16	<u>40.26</u>	<u>62.84</u>	65.50	<u>48.50</u>	61.77	56.05	58.51
GRPO (Shao et al., 2024)	70.63	39.34	58.18	68.25	47.75	67.29	<u>56.83</u>	60.04
Dr.GRPO (Liu et al., 2025)	68.76	30.51	56.60	54.00	43.50	62.04	50.67	56.61
SAPO (Gao et al., 2025)	<u>78.27</u>	0.00	28.37	34.00	29.00	57.62	33.93	44.53
GSPO (Zheng et al., 2025)	65.03	30.70	45.17	<u>83.00</u>	44.80	**74.26**	53.74	<u>63.94</u>
CISPO (Chen et al., 2025a)	57.39	0.18	23.37	48.50	27.55	49.97	31.40	38.77
CPO (Ours)	**90.83**	**76.84**	**63.64**	77.50	**49.00**	<u>73.24</u>	**71.56**	**72.81**

4.3 Pretrained Capability Retention on External Benchmarks

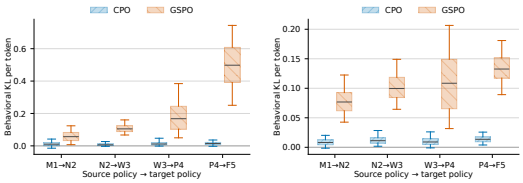
We further evaluate pretrained capability retention after MRCL training on diverse multimodal benchmarks. As shown in Tab. 3, 7, and 8, CPO consistently preserves pretrained capabilities better than competing post-training methods across model scales. On Qwen3-VL-4B, CPO achieves an average score of 64.14%, compared with 49.55% for GSPO and 59.50% for the pretrained model. The advantage extends across the 2B and 8B models and multiple capability dimensions, indicating broad retention of the pretrained capability profile with further gains on several benchmarks. We hypothesize that these gains reflect cumulative positive transfer across continual stages, as transferable capabilities acquired on each task can accumulate only when they survive subsequent updates. By suppressing cross-task behavioral drift, CPO better preserves and accumulates such knowledge across the training stream, an interpretation further supported by the substantially lower prior-task behavioral KL in Fig. 5.

Table 3: Comparison of Qwen3-VL-4B multimodal performance under different fine-tuning settings across diverse benchmarks.

Dimension	Benchmark	Qwen3-VL-4B							
		Base	**SFT**	**GRPO**	**Dr.GR**	**SAPO**	**GSPO**	**CISPO**	**CPO**
STEM	$\text{MMMU}_{\text{pro_full}}$	51.20	25.70	<u>40.56</u>	38.88	17.03	33.12	7.71	**49.59**
	$\text{MathVista}_{\text{mini}}$	56.90	51.70	<u>65.20</u>	62.90	33.30	57.40	28.10	**73.90**
	$\text{MathVision}_{\text{TEST}}$	33.78	18.39	29.81	<u>30.66</u>	20.92	22.43	15.79	**45.00**
	$\text{MathVerse}_{\text{mini}}$	45.61	29.49	45.61	<u>45.63</u>	26.57	37.84	19.16	**57.69**
General VQA	RealworldQA	68.24	51.63	43.40	<u>59.22</u>	39.48	58.17	37.25	**64.58**
	MMStar	58.93	55.27	<u>63.20</u>	60.20	28.40	51.20	39.00	**67.73**
Hallucination	POPE	87.16	<u>85.49</u>	83.17	84.33	1.96	84.99	75.71	**86.93**
Chart/Doc Understanding	$\text{DocVQA}_{\text{VAL}}$	65.04	53.51	<u>63.47</u>	62.67	3.29	53.88	12.17	**77.08**
	$\text{CharXiv}_{\text{(DQ)}}$	71.63	41.98	48.93	<u>52.60</u>	11.50	42.15	9.15	**75.15**
	$\text{CharXiv}_{\text{(RQ)}}$	33.10	23.70	<u>31.60</u>	30.40	8.00	27.70	7.30	**39.20**
Grounding	CountBenchQA	82.89	61.30	<u>70.26</u>	66.40	0.00	**76.17**	9.78	68.64
Average		59.50	45.29	53.20	<u>53.99</u>	17.31	49.55	23.74	**64.14**

4.4 Prior-Task Behavioral Drift

Proposition 3.1 identifies behavioral drift on prior-task distributions as the quantity directly associated with forgetting. For each transition T_i to T_{i+1}, the task-end checkpoints after T_i and T_{i+1} serve as the source and target policies, respectively. We sample queries from T_i, generate autoregressive prefixes with the source policy, and compute the mean token-level KL between the two policies at these prefixes. As shown in Fig. 5, CPO yields substantially lower prior-task behavioral KL than GSPO across transitions. The difference becomes more pronounced in later stages, while CPO maintains consistently low behavioral drift.



(a) Qwen3-VL-4B (b) Qwen3-VL-8B

Figure 5: Prior-task behavioral KL between consecutive task-end policies across continual transitions on Qwen3-VL-4B and Qwen3-VL-8B. Lower values indicate smaller behavioral drift on the preceding task.

Table 4: Ablation and sensitivity analysis of protected-parameter selection in CPO on Qwen3-VL-4B. Random replaces movement-based top-p selection with a random mask of the same size, while the remaining rows vary the protected-parameter ratio p.

Method	**MFT↑**	**MFN↑**	**MTA↑**
Random ($p = 10\%$)	73.26	57.05	66.64
CPO ($p = 2.5\%$)	**74.58**	69.10	72.28
CPO ($p = 5\%$)	73.75	69.46	71.74
CPO ($p = 10\%$)	73.24	**71.56**	**72.81**

4.5 Protected-Parameter Selection and Sensitivity

Tab. 4 examines the importance of movement-based parameter selection and sensitivity to the protected ratio p. CPO consistently outperforms random selection, confirming that parameter movement provides a meaningful signal for identifying parameters important to previously learned behavior. A smaller p better preserves plasticity and yields higher MFT, whereas a larger p strengthens stability, with the highest MTA. CPO is insensitive to the exact choice of p and remains stable across a wide range of ratios. The masked L_1 regularizer permits limited adaptation of protected parameters, while strong performance at small p suggests that effective retention does not require rapid expansion of the protected set, reducing the risk of exhausting available parameter space.

5 Related Work

Continual post-training seeks to adapt VLMs to evolving tasks without erasing previously acquired capabilities (Yu et al., 2026; Wang et al., 2024b). Research has largely focused on forgetting under SFT, while recent work has increasingly examined the performance of RL in continual learning.

Reinforcement learning optimizes models through reward signals and has become central to eliciting deliberative reasoning capabilities (Oh et al., 2025; Wu et al., 2026). Policy optimization has evolved from PPO (Schulman et al., 2017) to GRPO (Shao et al., 2024) and following methods such as DAPO (Yu et al., 2025) and GSPO (Zheng et al., 2025), motivating interest in continual RL (Shenfeld et al., 2026; Abel et al., 2023). A series of recent studies (Shenfeld et al., 2026; Zhang et al., 2026; Chen et al., 2025b; Lai et al., 2025) report that RL is more resistant to catastrophic forgetting than SFT, and Jin et al. (2025) further show that RL fine-tuning can recover out-of-distribution capabilities degraded by SFT. The mechanism behind this robustness remains debated. Shenfeld et al. (2026) attribute the effect partly to the KL regularization used in standard policy optimization, whereas Chen et al. (2025b) and Lai et al. (2025) find that KL regularization plays a limited role in retention. Other studies instead emphasize on-policy data generation (Zhang et al., 2026; Chen et al., 2025b), while Lou et al. (2026) shows that RL can still suffer from non-negligible forgetting on classification. Despite these observations, catastrophic forgetting in multi-modal reasoning remains underexplored, and the role of policy regularization in continual RL is still not well understood.

Replay-free continual learning mitigates forgetting by estimating parameter importance and constraining subsequent updates. EWC (Kirkpatrick et al., 2017) uses Fisher information to identify important parameters, requiring an additional gradient-based estimation over task data. MAS (Aljundi et al., 2018) similarly estimates importance from output sensitivity via parameter gradients. SI (Zenke et al., 2017) and RWalk (Chaudhry et al., 2018) avoid separate importance estimation by accumulating gradient-dependent statistics during optimization. More recently, Luo et al. (2026); Liang & Li (2024); Saha et al. (2021) constrain adaptation through protected subspaces constructed from task gradients. Although these methods avoid storing historical data, importance estimation still requires extra backward computation, parameter-wise statistics, or training-time state tracking beyond standard task optimization. These overheads increase the difficulty of integration into distributed training, where memory, computation, and communication are major scaling constraints (Rajbhandari et al., 2020; Narayanan et al., 2021), limiting the practicality for large model training.

6 Conclusion

In this work, we challenge the view that reinforcement learning is inherently resistant to catastrophic forgetting in continual VLM post-training. Designed around release recency, model familiarity, and scope for adaptation, MRCL reveals catastrophic forgetting across RL methods. We further identify an objective mismatch between current-task KL regularization and the prior-task behavioral KL that bounds performance degradation. Motivated by this mismatch, CPO derives a local Fisher surrogate for the prior-task behavioral KL and uses parameter movement as a gradient-free proxy for Fisher sensitivity in the RL setting with negligible additional overhead. MRCL and CPO move continual RL beyond presumed robustness by exposing forgetting under challenging evaluation and explicitly constraining prior-task policy drift, providing a simple and scalable path for continual VLM post-training. This perspective informs continual learning benchmark construction and supports an effective approach to continual RL, grounded in both theory and empirical evidence.

AI USE STATEMENT

Generative AI tools were used solely for language editing to improve the readability and clarity of text originally written by the authors. All AI-assisted edits were reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide detailed information to facilitate reproduction of our results. Complete theoretical proofs and additional analyses are provided in Appendices B and C. The construction and data-processing procedures of MRCL are described in Appendix D. Training configurations, hyperparameters, evaluation protocols, and scoring rules are documented in Appendix E. Complete stage-by-stage results and prompt templates are provided in Appendices G and H, respectively. Anonymous source code for reproducing the experiments is provided in the supplementary material.

REFERENCES

David Abel, André Barreto, Benjamin Van Roy, Doina Precup, Hado van Hasselt, and Satinder Singh. A definition of continual reinforcement learning. *Advances in Neural Information Processing Systems*, 36:50377–50407, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European conference on computer vision*, pp. 144–161. Springer, 2018.

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*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *European conference on computer vision*, pp. 556–572. Springer, 2018.

Aili Chen, Aonian Li, Bangwei Gong, Binyang Jiang, Bo Fei, Bo Yang, Boji Shan, Changqing Yu, Chao Wang, Cheng Zhu, et al. Minimax-m1: Scaling test-time compute efficiently with lightning attention. *arXiv preprint arXiv:2506.13585*, 2025a.

Howard Chen, Noam Razin, Karthik Narasimhan, and Danqi Chen. Retaining by doing: The role of on-policy data in mitigating forgetting. *arXiv preprint arXiv:2510.18874*, 2025b.

Jinpeng Chen, Runmin Cong, Yuzhi Zhao, Hongzheng Yang, Guangneng Hu, Horace Ip, and Sam Kwong. SEFE: Superficial and essential forgetting eliminator for multimodal continual instruction tuning. In *Forty-second International Conference on Machine Learning*, 2025c.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. A Wiley-Interscience publication. Wiley, 2006. ISBN 9780471748816. URL https://books.google.com/books?id=0QuawYmc2pIC.

Robert M French. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135, 1999.

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.

Hao Gu, Mao-Lin Luo, Zi-Hao Zhou, Han-Chen Zhang, Min-Ling Zhang, and Tong Wei. Spectral imbalance causes forgetting in low-rank continual adaptation. In *Forty-third International Conference on Machine Learning*, 2026a.

Jiawei Gu, Yunzhuo Hao, Huichen Will Wang, Linjie Li, Michael Qizhe Shieh, Yejin Choi, Ranjay Krishna, and Yu Cheng. Thinkmorph: Emergent properties in multimodal interleaved chain-of-thought reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Hangzhan Jin, Sitao Luan, Sicheng Lyu, Guillaume Rabusseau, Reihaneh Rabbany, Doina Precup, and Mohammad Hamdaqa. RL fine-tuning heals ood forgetting in sft. *arXiv preprint arXiv:2509.12235*, 2025.

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

Muhammad Rifki Kurniawan, Xiang Song, Zhiheng Ma, Yuhang He, Yihong Gong, Yang Qi, and Xing Wei. Evolving parameterized prompt memory for continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2024.

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.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

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*, pp. 23638–23647, 2024.

Xuan Liu and Xiaobin Chang. Elastic weight consolidation done right for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3930–3940, June 2026.

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.

Meng Lou, Hanzhong Guo, Linwei Chen, and Yizhou Yu. Overcoming catastrophic forgetting in visual continual learning with reinforcement fine-tuning. *arXiv preprint arXiv:2605.09640*, 2026.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, et al. Finmme: Benchmark dataset for financial multimodal reasoning evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29465–29489, 2025.

Mao-Lin Luo, Zi-Hao Zhou, Yi-Lin Zhang, Yuanyu Wan, Min-Ling Zhang, and Tong Wei. KeepLoRA: Continual learning with residual gradient adaptation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the international conference for high performance computing, networking, storage and analysis*, pp. 1–15, 2021.

Junhyuk Oh, Gregory Farquhar, Iurii Kemaev, Dan A Calian, Matteo Hessel, Luisa Zintgraf, Satinder Singh, Hado Van Hasselt, and David Silver. Discovering state-of-the-art reinforcement learning algorithms. *Nature*, 648(8093):312–319, 2025.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching clip to count to ten. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3170–3180, 2023.

Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Guanting Dong, Shiqiang Lang, Sitong Zhou, Yida Xu, Yuchen Zeng, Jie Wang, Chong Sun, Chen Li, and Honggang Zhang. We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. In *The Fourteenth International Conference on Learning Representations*, 2026.

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

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: international conference for high performance computing, networking, storage and analysis*, pp. 1–16. IEEE, 2020.

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

Gobinda Saha, Isha Garg, and Kaushik Roy. Gradient projection memory for continual learning. In *International Conference on Learning Representations*, 2021.

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

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

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

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024a.

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, 2024b.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024c.

Yongliang Wu, Yizhou Zhou, Ziheng Zhou, Yingzhe Peng, Xinyu Ye, Xinting Hu, Wenbo Zhu, Lu Qi, Ming-Hsuan Yang, et al. On the generalization of sft: A reinforcement learning perspective with reward rectification. In *International Conference on Learning Representations*, volume 2026, pp. 27550–27571, 2026.

xAI. Realworldqa: A benchmark for real-world spatial understanding. `https://huggingface.co/datasets/xai-org/RealworldQA`, 2024. Accessed: 2025-04-26.

Sau Lai Yip, Sunan He, Yuxiang Nie, Shu Pui Chan, Yilin Ye, Sum Ying Lam, and Hao Chen. Medbookvqa: A systematic and comprehensive medical benchmark derived from open-access book. *arXiv preprint arXiv:2506.00855*, 2025.

Dianzhi Yu, Xinni Zhang, Yankai Chen, Aiwei Liu, Yifei Zhang, Philip S Yu, and Irwin King. Recent advances of multimodal continual learning: A comprehensive survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Ru Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Yonghui Wu, and Mingxuan Wang. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Friedemann Zenke, Ben Poole, and Surya Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pp. 3987–3995. Pmlr, 2017.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pp. 169–186. Springer, 2024.

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 *The Fourteenth International Conference on Learning Representations*, 2026.

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.

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.

A PERFORMANCE ON QWEN3-VL-2B AND QWEN3-VL-8B

This section reports the complete results on Qwen3-VL-2B and Qwen3-VL-8B, complementing the Qwen3-VL-4B results in the main text. As shown in Tab. 5 and Tab. 6, CPO achieves the highest MFN and MTA on both model scales while maintaining competitive MFT. Tab. 7 and Tab. 8 further show that CPO better preserves pretrained capabilities than competing post-training methods across external benchmarks. These results confirm that the effectiveness of CPO persists across model scales and evaluation settings.

Table 5: Comparison of our proposed method with baseline approaches on Qwen3-VL-2B.

Method	Final Accuracy on Each Task (%)					Aggregated Acc. (%)		
	MedBook	Naviga.	We-Math2	Puzzle	FinMME	MFT↑	MFN↑	MTA↑
Zero-shot[1]	52.63	0.00	17.23	35.75	32.75	–	–	27.67
Full SFT	44.65	0.55	33.41	83.25	33.60	66.32	39.09	52.02
Zero-shot[2]	47.20	1.65	40.40	22.00	31.45	–	–	28.54
GRPO* (Shao et al., 2024)	**76.23**	17.10	**51.24**	61.50	42.15	53.72	49.64	51.59
GRPO (Shao et al., 2024)	62.48	28.86	48.11	66.25	41.10	60.86	49.36	55.53
GSPO (Zheng et al., 2025)	69.27	11.95	45.20	75.75	41.05	68.23	48.64	57.96
CPO (Ours)	76.06	**43.38**	49.03	**78.00**	**42.65**	**68.28**	**57.82**	**62.28**

Table 6: Comparison of our proposed method with baseline approaches on Qwen3-VL-8B.

Method	Final Accuracy on Each Task (%)					Aggregated Acc. (%)		
	MedBook	Naviga.	We-Math2	Puzzle	FinMME	MFT↑	MFN↑	MTA↑
Zero-shot[1]	62.48	0.00	35.00	36.50	47.60	–	–	36.32
Full SFT	79.80	33.46	38.94	35.00	32.75	60.50	43.99	51.21
Zero-shot[2]	51.27	25.55	34.61	25.25	39.65	–	–	35.27
GRPO (Shao et al., 2024)	73.51	6.62	55.43	72.25	45.80	64.99	50.72	57.06
GSPO (Zheng et al., 2025)	77.08	44.67	52.68	**86.25**	48.15	**76.40**	61.77	68.85
CPO (Ours)	**96.94**	**81.25**	**64.75**	74.25	**49.40**	73.37	**73.32**	**73.04**

Table 7: Comparison of Qwen3-VL-2B multimodal performance under different fine-tuning settings across diverse benchmarks.

Dimension	Benchmark	Qwen3-VL-2B				
		Base	SFT	GRPO	GSPO	CPO
STEM	$\text{MMMU}_{\text{pro_full}}$	33.74	23.35	**30.41**	27.69	29.48
	$\text{MathVista}_{\text{mini}}$	48.00	41.60	52.40	51.60	**56.70**
	$\text{MathVision}_{\text{TEST}}$	14.87	15.99	17.43	18.19	**19.93**
	$\text{MathVerse}_{\text{mini}}$	31.12	22.84	32.51	32.16	**36.12**
General VQA	RealworldQA	57.78	51.11	54.51	46.01	**54.77**
	MMStar	52.80	48.60	**53.47**	50.53	53.13
Hallucination	POPE	85.61	82.98	83.39	81.41	**84.47**
Chart/Document Understanding	$\text{DocVQA}_{\text{VAL}}$	56.50	**55.21**	43.84	42.81	48.10
	$\text{CharXiv}_{\text{(DQ)}}$	60.98	38.70	41.60	42.38	**45.75**
	$\text{CharXiv}_{\text{(RQ)}}$	26.70	21.00	22.20	**23.60**	21.90
Grounding	CountBenchQA	83.30	62.32	37.88	60.08	**65.99**
Average		50.13	42.15	42.69	43.31	**46.94**

Table 8: Comparison of Qwen3-VL-8B multimodal performance under different fine-tuning settings across diverse benchmarks.

Dimension	Benchmark	Qwen3-VL-8B				
		Base	SFT	GRPO	GSPO	CPO
STEM	$\text{MMMU}_{\text{pro_full}}$	56.71	24.91	<u>40.56</u>	31.41	**53.70**
	$\text{MathVista}_{\text{mini}}$	50.80	47.40	61.10	<u>64.90</u>	**76.00**
	$\text{MathVision}_{\text{TEST}}$	36.78	18.82	<u>31.78</u>	26.35	**51.61**
	$\text{MathVerse}_{\text{mini}}$	43.17	28.22	<u>44.19</u>	42.84	**57.51**
General VQA	RealworldQA	66.93	55.29	27.45	<u>59.48</u>	**63.27**
	MMStar	55.87	51.93	59.47	<u>61.67</u>	**71.67**
Hallucination	POPE	76.73	82.02	82.51	<u>83.20</u>	**85.43**
Chart/Document Understanding	$\text{DocVQA}_{\text{VAL}}$	68.16	<u>54.16</u>	40.77	47.86	**66.83**
	$\text{CharXiv}_{\text{(DQ)}}$	82.00	36.30	<u>50.62</u>	50.25	**77.53**
	$\text{CharXiv}_{\text{(RQ)}}$	32.70	23.40	28.80	<u>30.00</u>	**44.20**
Grounding	CountBenchQA	86.97	63.14	64.15	<u>79.63</u>	**85.74**
Average		59.71	44.14	48.31	<u>52.51</u>	**66.68**

B Theoretical Proofs

B.1 Finite-Horizon Terminal-Reward KL Forgetting Bound

Proposition B.1 (Finite horizon KL forgetting bound)**.** *Consider an old task $\mathcal{T}_{i-1}$ with prompts $x \sim \mathcal{D}_{i-1}$, finite-horizon autoregressive generations $y_{1:H}$, and terminal reward $R_{i-1}(x, y_{1:H})$. Let $\tilde{\pi}$ be the policy obtained after training on $\mathcal{T}_{i-1}$ and let π be the policy after subsequent adaptation. Their sequence distributions are*

$$P_{\tilde{\pi}}(y_{1:H} \mid x) = \prod_{t=1}^{H} \tilde{\pi}(y_t \mid x, y_{<t}), \qquad P_{\pi}(y_{1:H} \mid x) = \prod_{t=1}^{H} \pi(y_t \mid x, y_{<t}). \tag{18}$$

Define the task return

$$\eta_{i-1}(\pi) = \mathbb{E}_{x \sim \mathcal{D}_{i-1},\, y \sim P_{\pi}(\cdot|x)}[R_{i-1}(x, y)] \tag{19}$$

and assume the terminal reward has bounded range

$$B_{i-1} = \sup_{x,y,y'} |R_{i-1}(x, y) - R_{i-1}(x, y')| < \infty. \tag{20}$$

Let the normalized old-policy prefix visitation distribution be

$$d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}(t, x, y_{<t}) = \frac{1}{H} \mathcal{D}_{i-1}(x) P_{\tilde{\pi}}(y_{<t} \mid x), \qquad t \in \{1, \ldots, H\}, \tag{21}$$

and define the corresponding mean token-level KL divergence as

$$\bar{D}_{\mathrm{KL}}^{d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}}(\tilde{\pi}, \pi) = \mathbb{E}_{(t,x,y_{<t}) \sim d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}} \left[D_{\mathrm{KL}}\big(\tilde{\pi}(\cdot \mid x, y_{<t}) \| \pi(\cdot \mid x, y_{<t})\big) \right]. \tag{22}$$

Then the old-task return degradation is bounded by

$$|\eta_{i-1}(\tilde{\pi}) - \eta_{i-1}(\pi)| \le B_{i-1} \sqrt{\frac{H}{2} \bar{D}_{\mathrm{KL}}^{d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}}(\tilde{\pi}, \pi)}. \tag{23}$$

Moreover, the sequence-level KL and the normalized mean token-level KL satisfy

$$\mathbb{E}_{x \sim \mathcal{D}_{i-1}} \left[D_{\mathrm{KL}}\big(P_{\tilde{\pi}}(\cdot \mid x) \| P_{\pi}(\cdot \mid x)\big) \right] = H \bar{D}_{\mathrm{KL}}^{d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}}(\tilde{\pi}, \pi). \tag{24}$$

Proof. We prove the result in three steps: first controlling terminal-reward difference by sequence-level KL, then expanding sequence-level KL by the autoregressive chain rule, and finally substituting the normalized mean token-level KL.

Step 1: terminal reward difference is controlled by sequence-level KL. Fix a prompt x and abbreviate

$$P = P_{\tilde{\pi}}(\cdot \mid x), \qquad Q = P_{\pi}(\cdot \mid x), \qquad R(y) = R_{i-1}(x, y). \tag{25}$$

If $B_{i-1} = 0$, then $R_{i-1}(x, y)$ is constant in y for every x, so the left-hand side of Eq. 23 is zero and the proposition is immediate. Hence assume $B_{i-1} > 0$.

For this fixed prompt, let

$$R_x^- = \inf_y R_{i-1}(x, y), \qquad f_x(y) = \frac{R_{i-1}(x, y) - R_x^-}{B_{i-1}}. \tag{26}$$

By the definition of B_{i-1}, $0 \le f_x(y) \le 1$ for all y. Since $\sum_y P(y) = \sum_y Q(y) = 1$, subtracting the constant R_x^- does not change the difference of expectations:

$$\begin{aligned}
\mathbb{E}_P[R(y)] - \mathbb{E}_Q[R(y)] &= \sum_y (P(y) - Q(y))R(y) && (27)\\
&= \sum_y (P(y) - Q(y))(R(y) - R_x^-) && (28)\\
&= B_{i-1} \sum_y (P(y) - Q(y)) f_x(y) && (29)\\
&= B_{i-1} \left(\mathbb{E}_P[f_x] - \mathbb{E}_Q[f_x]\right). && (30)
\end{aligned}$$

Using the variational characterization of total variation,

$$D_{\mathrm{TV}}(P, Q) = \sup_{0 \le f \le 1} \left|\mathbb{E}_P[f] - \mathbb{E}_Q[f]\right|, \tag{31}$$

we obtain

$$\left|\mathbb{E}_P[R(y)] - \mathbb{E}_Q[R(y)]\right| \le B_{i-1} D_{\mathrm{TV}}(P, Q). \tag{32}$$

Pinsker's inequality further gives

$$D_{\mathrm{TV}}(P, Q) \le \sqrt{\frac{1}{2} D_{\mathrm{KL}}(P \| Q)}. \tag{33}$$

Combining the previous two inequalities and returning to the original notation yields, for every prompt x,

$$\left|\mathbb{E}_{y \sim P_{\tilde{\pi}}(\cdot|x)}[R_{i-1}(x, y)] - \mathbb{E}_{y \sim P_{\pi}(\cdot|x)}[R_{i-1}(x, y)]\right| \le B_{i-1} \sqrt{\frac{1}{2} D_{\mathrm{KL}}\big(P_{\tilde{\pi}}(\cdot \mid x) \| P_{\pi}(\cdot \mid x)\big)}. \tag{34}$$

Now expand the task-return difference:

$$\eta_{i-1}(\tilde{\pi}) - \eta_{i-1}(\pi) = \mathbb{E}_{x \sim \mathcal{D}_{i-1}} \left[\mathbb{E}_{y \sim P_{\tilde{\pi}}(\cdot|x)}[R_{i-1}(x, y)] - \mathbb{E}_{y \sim P_{\pi}(\cdot|x)}[R_{i-1}(x, y)]\right]. \tag{35}$$

Taking absolute values, using $|\mathbb{E}[Z]| \le \mathbb{E}[|Z|]$, and applying Eq. 34,

$$\begin{aligned}
|\eta_{i-1}(\tilde{\pi}) - \eta_{i-1}(\pi)| &\le \mathbb{E}_{x \sim \mathcal{D}_{i-1}} \left[\left|\mathbb{E}_{P_{\tilde{\pi}}(\cdot|x)}[R_{i-1}(x, y)] - \mathbb{E}_{P_{\pi}(\cdot|x)}[R_{i-1}(x, y)]\right|\right] && (36)\\
&\le B_{i-1} \mathbb{E}_{x \sim \mathcal{D}_{i-1}} \left[\sqrt{\frac{1}{2} D_{\mathrm{KL}}\big(P_{\tilde{\pi}}(\cdot \mid x) \| P_{\pi}(\cdot \mid x)\big)}\right]. && (37)
\end{aligned}$$

Because $u \mapsto \sqrt{u}$ is concave on $\mathbb{R}_{\ge 0}$, Jensen's inequality implies

$$\mathbb{E}_x\left[\sqrt{Z_x}\right] \le \sqrt{\mathbb{E}_x[Z_x]} \qquad \text{for } Z_x \ge 0. \tag{38}$$

With $Z_x = \frac{1}{2} D_{\mathrm{KL}}\big(P_{\tilde{\pi}}(\cdot \mid x) \| P_{\pi}(\cdot \mid x)\big)$, this gives

$$|\eta_{i-1}(\tilde{\pi}) - \eta_{i-1}(\pi)| \le B_{i-1} \sqrt{\frac{1}{2} \mathbb{E}_{x \sim \mathcal{D}_{i-1}} \left[D_{\mathrm{KL}}\big(P_{\tilde{\pi}}(\cdot \mid x) \| P_{\pi}(\cdot \mid x)\big)\right]}. \tag{39}$$

Step 2: sequence-level KL equals the sum of token-level KLs. For a fixed prompt x, start from the definition of sequence-level KL:

$$D_{\mathrm{KL}}\big(P_{\tilde{\pi}}(\cdot \mid x) \| P_{\pi}(\cdot \mid x)\big) = \sum_{y_{1:H}} P_{\tilde{\pi}}(y_{1:H} \mid x) \log \frac{P_{\tilde{\pi}}(y_{1:H} \mid x)}{P_{\pi}(y_{1:H} \mid x)}. \tag{40}$$

By autoregressive factorization,

$$\frac{P_{\tilde{\pi}}(y_{1:H} \mid x)}{P_{\pi}(y_{1:H} \mid x)} = \prod_{t=1}^{H} \frac{\tilde{\pi}(y_t \mid x, y_{<t})}{\pi(y_t \mid x, y_{<t})}, \tag{41}$$

and therefore

$$\log \frac{P_{\tilde{\pi}}(y_{1:H} \mid x)}{P_{\pi}(y_{1:H} \mid x)} = \sum_{t=1}^{H} \log \frac{\tilde{\pi}(y_t \mid x, y_{<t})}{\pi(y_t \mid x, y_{<t})}. \tag{42}$$

Substituting this identity into the sequence KL and exchanging the finite sums gives

$$D_{\mathrm{KL}}\big(P_{\tilde{\pi}}(\cdot \mid x) \| P_{\pi}(\cdot \mid x)\big) = \sum_{t=1}^{H} \sum_{y_{1:H}} P_{\tilde{\pi}}(y_{1:H} \mid x) \log \frac{\tilde{\pi}(y_t \mid x, y_{<t})}{\pi(y_t \mid x, y_{<t})}. \tag{43}$$

For a fixed t, decompose the sequence as $(y_{<t}, y_t, y_{>t})$. Under $P_{\tilde{\pi}}$,

$$P_{\tilde{\pi}}(y_{1:H} \mid x) = P_{\tilde{\pi}}(y_{<t} \mid x) \tilde{\pi}(y_t \mid x, y_{<t}) P_{\tilde{\pi}}(y_{>t} \mid x, y_{\le t}). \tag{44}$$

The log-ratio term in Eq. 43 depends only on $(x, y_{<t}, y_t)$, so summing over future tokens gives

$$\sum_{y_{>t}} P_{\tilde{\pi}}(y_{>t} \mid x, y_{\le t}) = 1. \tag{45}$$

Thus the t-th term becomes

$$\sum_{y_{1:H}} P_{\tilde{\pi}}(y_{1:H} \mid x) \log \frac{\tilde{\pi}(y_t \mid x, y_{<t})}{\pi(y_t \mid x, y_{<t})} \tag{46}$$

$$= \sum_{y_{<t}} P_{\tilde{\pi}}(y_{<t} \mid x) \sum_{y_t} \tilde{\pi}(y_t \mid x, y_{<t}) \log \frac{\tilde{\pi}(y_t \mid x, y_{<t})}{\pi(y_t \mid x, y_{<t})} \tag{47}$$

$$= \sum_{y_{<t}} P_{\tilde{\pi}}(y_{<t} \mid x) D_{\mathrm{KL}}\big(\tilde{\pi}(\cdot \mid x, y_{<t}) \| \pi(\cdot \mid x, y_{<t})\big) \tag{48}$$

$$= \mathbb{E}_{y_{<t} \sim P_{\tilde{\pi}}(\cdot|x)} \left[D_{\mathrm{KL}}\big(\tilde{\pi}(\cdot \mid x, y_{<t}) \| \pi(\cdot \mid x, y_{<t})\big) \right]. \tag{49}$$

Consequently,

$$D_{\mathrm{KL}}\big(P_{\tilde{\pi}}(\cdot \mid x) \| P_{\pi}(\cdot \mid x)\big) = \sum_{t=1}^{H} \mathbb{E}_{y_{<t} \sim P_{\tilde{\pi}}(\cdot|x)} \left[D_{\mathrm{KL}}\big(\tilde{\pi}(\cdot \mid x, y_{<t}) \| \pi(\cdot \mid x, y_{<t})\big) \right]. \tag{50}$$

Step 3: convert the chain rule into the normalized mean token-level KL. First verify that $d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}$ is normalized:

$$\sum_{t=1}^{H} \sum_{x} \sum_{y_{<t}} d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}(t, x, y_{<t}) = \frac{1}{H} \sum_{t=1}^{H} \sum_{x} \mathcal{D}_{i-1}(x) \sum_{y_{<t}} P_{\tilde{\pi}}(y_{<t} \mid x) \tag{51}$$

$$= \frac{1}{H} \sum_{t=1}^{H} \sum_{x} \mathcal{D}_{i-1}(x) \tag{52}$$

$$= 1. \tag{53}$$

Therefore, by the definition of the mean token-level KL,

$$H \bar{D}_{\mathrm{KL}}^{d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}}(\tilde{\pi}, \pi) = \sum_{t=1}^{H} \mathbb{E}_{x \sim \mathcal{D}_{i-1},\, y_{<t} \sim P_{\tilde{\pi}}(\cdot|x)} \left[D_{\mathrm{KL}}\big(\tilde{\pi}(\cdot \mid x, y_{<t}) \| \pi(\cdot \mid x, y_{<t})\big) \right] \tag{54}$$

$$= \mathbb{E}_{x \sim \mathcal{D}_{i-1}} \left[D_{\mathrm{KL}}\big(P_{\tilde{\pi}}(\cdot \mid x) \| P_{\pi}(\cdot \mid x)\big) \right], \tag{55}$$

where the second equality uses Eq. 50. This proves Eq. 24. Substituting Eq. 24 into Eq. 39 gives

$$|\eta_{i-1}(\tilde{\pi}) - \eta_{i-1}(\pi)| \le B_{i-1} \sqrt{\frac{H}{2} \bar{D}_{\mathrm{KL}}^{d_{\tilde{\pi}}^{\mathcal{T}_{i-1}}}(\tilde{\pi}, \pi)}, \tag{56}$$

which is the desired result. □

B.2 Parameter Movement as a Noisy Fisher Proxy

Proposition B.2 (Parameter movement as a noisy Fisher proxy)**.** *Consider training on task $\mathcal{T}_{i-1}$ under a locally stationary approximation, so that the policy, normalized prefix-token distribution, Fisher diagonal, and reward-score moments are treated as fixed over the interval in which parameter movement is measured. Let $h_t = (x, y_{<t})$ and define the token score*

$$\phi_k(t, h_t, a) = \frac{\partial}{\partial \theta_k} \log \pi_\theta(a \mid h_t). \tag{57}$$

The task Fisher diagonal is

$$F_{kk}^{\mathcal{T}_{i-1}} = \mathbb{E}_{(t,h_t)\sim d_\theta^{\mathcal{T}_{i-1}}} \left[\mathbb{E}_{a\sim\pi_\theta(\cdot|h_t)}[\phi_k(t, h_t, a)^2] \right]. \tag{58}$$

For terminal-reward policy-gradient training, define the token-sampled gradient contribution

$$z_k = A(x, y)\phi_k(t, h_t, y_t), \tag{59}$$

where $A(x, y)$ is a terminal reward or baseline-centered terminal advantage, and (t, h_t, y_t) is sampled from the normalized prefix-token distribution. Let

$$\mu_k = \mathbb{E}[z_k], \qquad V_k = \mathrm{Var}(z_k), \qquad \sigma_A^2 = \mathbb{E}[A^2], \tag{60}$$

and, for $F_{kk}^{\mathcal{T}_{i-1}} > 0$, define

$$q_k = \frac{\mathbb{E}[A^2\phi_k^2]}{\sigma_A^2 F_{kk}^{\mathcal{T}_{i-1}}}, \qquad \rho_k = \frac{\mu_k}{\sigma_A \sqrt{F_{kk}^{\mathcal{T}_{i-1}}}}. \tag{61}$$

Then

$$V_k = \sigma_A^2(q_k - \rho_k^2)F_{kk}^{\mathcal{T}_{i-1}}. \tag{62}$$

Under SGD with learning rate η for N stochastic updates, if the parameter gradient samples have mean μ_k and martingale-difference noise with variance V_k, the expected squared movement satisfies

$$\mathbb{E}[|\Delta\theta_k|^2] = \eta^2\sigma_A^2 \left[N^2\rho_k^2 + N(q_k - \rho_k^2) \right] F_{kk}^{\mathcal{T}_{i-1}}. \tag{63}$$

Equivalently, with

$$C_k = N^2\rho_k^2 + N(q_k - \rho_k^2), \tag{64}$$

we have

$$\mathbb{E}[|\Delta\theta_k|^2] = \eta^2\sigma_A^2 C_k F_{kk}^{\mathcal{T}_{i-1}}. \tag{65}$$

Thus, ranking parameters by expected squared movement is consistent with ranking by $F_{kk}^{\mathcal{T}_{i-1}}$ if C_k is constant across coordinates, and is pairwise consistent whenever, for $F_{kk}^{\mathcal{T}_{i-1}} > F_{\ell\ell}^{\mathcal{T}_{i-1}}$,

$$\frac{C_\ell}{C_k} < \frac{F_{kk}^{\mathcal{T}_{i-1}}}{F_{\ell\ell}^{\mathcal{T}_{i-1}}}. \tag{66}$$

In the noise-dominated regime, this requires approximate homogeneity of $q_k - \rho_k^2$; in the bias-dominated regime, it requires approximate homogeneity of $|\rho_k|$.

Proof. We first relate the terminal-reward policy gradient to token-level scores. For the task return

$$\eta_{i-1}(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{i-1},\, y\sim P_\theta(\cdot|x)}[R_{i-1}(x, y)], \tag{67}$$

the likelihood-ratio identity gives

$$\nabla_{\theta_k}\eta_{i-1}(\theta) = \mathbb{E}\left[A(x, y) \sum_{t=1}^{H} \phi_k(t, h_t, y_t) \right], \tag{68}$$

where subtracting any action-independent baseline from the terminal reward does not change the policy-gradient expectation. If t is sampled uniformly from $\{1, \ldots, H\}$ after drawing a trajectory, then

$$z_k = A(x, y)\phi_k(t, h_t, y_t) \tag{69}$$

satisfies $\mathbb{E}[z_k] = H^{-1}\nabla_{\theta_k}\eta_{i-1}(\theta)$. The factor H is shared across coordinates and therefore does not affect coordinate ranking.

Next, by definition of q_k and ρ_k,

$$\mathbb{E}[z_k^2] = \mathbb{E}[A^2\phi_k^2] = \sigma_A^2 q_k F_{kk}^{\mathcal{T}_{i-1}}, \qquad \mu_k^2 = \sigma_A^2\rho_k^2 F_{kk}^{\mathcal{T}_{i-1}}. \tag{70}$$

Therefore,

$$V_k = \mathrm{Var}(z_k) = \mathbb{E}[z_k^2] - \mu_k^2 = \sigma_A^2(q_k - \rho_k^2)F_{kk}^{\mathcal{T}_{i-1}}, \tag{71}$$

which proves Eq. 62.

Now consider N locally stationary SGD updates in coordinate k:

$$\theta_{r+1,k} = \theta_{r,k} - \eta z_{r,k}, \qquad z_{r,k} = \mu_k + \xi_{r,k}, \tag{72}$$

where $\mathbb{E}[\xi_{r,k} \mid \mathcal{F}_{r-1}] = 0$ and $\mathbb{E}[\xi_{r,k}^2 \mid \mathcal{F}_{r-1}] = V_k$. The total movement is

$$\Delta\theta_k = \theta_{N,k} - \theta_{0,k} = -\eta\sum_{r=1}^{N} z_{r,k} = -\eta(N\mu_k + M_{N,k}), \qquad M_{N,k} = \sum_{r=1}^{N}\xi_{r,k}. \tag{73}$$

Using the martingale-difference property,

$$\mathbb{E}[M_{N,k}] = 0, \qquad \mathbb{E}[M_{N,k}^2] = \sum_{r=1}^{N}\mathbb{E}[\xi_{r,k}^2] = NV_k, \tag{74}$$

where cross terms vanish because $\mathbb{E}[\xi_{r,k}\xi_{s,k}] = 0$ for $r \neq s$. Hence

$$\mathbb{E}[|\Delta\theta_k|^2] = \eta^2\mathbb{E}\left[(N\mu_k + M_{N,k})^2\right] \tag{75}$$

$$= \eta^2 N^2\mu_k^2 + 2\eta^2 N\mu_k\mathbb{E}[M_{N,k}] + \eta^2\mathbb{E}[M_{N,k}^2] \tag{76}$$

$$= \eta^2 N^2\mu_k^2 + \eta^2 NV_k. \tag{77}$$

Substituting the expressions for μ_k^2 and V_k proves Eq. 63.

Finally, for two coordinates k and ℓ,

$$\mathbb{E}[|\Delta\theta_k|^2] > \mathbb{E}[|\Delta\theta_\ell|^2] \quad \Longleftrightarrow \quad C_k F_{kk}^{\mathcal{T}_{i-1}} > C_\ell F_{\ell\ell}^{\mathcal{T}_{i-1}}. \tag{78}$$

If C_k is constant over coordinates, expected movement and Fisher diagonal induce the same ranking. More generally, when $F_{kk}^{\mathcal{T}_{i-1}} > F_{\ell\ell}^{\mathcal{T}_{i-1}}$, the Fisher ranking is preserved for the pair (k, ℓ) exactly under Eq. 66.

The coefficient $C_k = N^2\rho_k^2 + N(q_k - \rho_k^2)$ separates two regimes. If $N^2\rho_k^2 \ll N(q_k - \rho_k^2)$, then $C_k \approx N(q_k - \rho_k^2)$, so rank consistency requires the noise coefficient $q_k - \rho_k^2$ to be sufficiently homogeneous or at least not large enough to violate Eq. 66. If $N^2\rho_k^2 \gg N(q_k - \rho_k^2)$, then $C_k \approx N^2\rho_k^2$, so rank consistency requires the normalized reward-score alignment magnitude $|\rho_k|$ to be sufficiently homogeneous. Since

$$|\rho_k| = \frac{|\mathbb{E}[A\phi_k]|}{\sqrt{\mathbb{E}[A^2]}\sqrt{\mathbb{E}[\phi_k^2]}} \leq 1 \tag{79}$$

by Cauchy-Schwarz, ρ_k measures the normalized alignment between terminal advantage and the token score of coordinate k.

The realized movement $|\Delta\theta_k|$ is therefore a noisy proxy for $F_{kk}^{\mathcal{T}_{i-1}}$: it is governed by the Fisher diagonal in expectation, but its ranking is mediated by C_k and by stochastic-gradient fluctuations. Under bounded or sub-Gaussian martingale noise, standard concentration inequalities imply that the realized ranking agrees with the expected ranking whenever the pairwise expected movement gap is larger than the corresponding concentration error. □

C THEORETICAL FORGETTING BOUND FOR CPO

The derivation of CPO in section 3 replaces the prior-task behavioral KL under the oracle model with a replay-free surrogate through a sequence of approximations. This appendix formalizes this approximation chain and quantifies the resulting error. We analyze the transition from $\mathcal{T}_{i-1}$ to $\mathcal{T}_i$; the same arguments extend to any earlier task $\mathcal{T}_j$ because the cumulative mask can only increase its covered Fisher sensitivity over time. The analysis allows heterogeneous coefficients C_k and explicitly accounts for deviations between realized and expected parameter movement.

We first relate movement-based selection to Fisher sensitivity, and then use the resulting coverage bound to control the full Fisher quadratic form, behavioral KL, and return degradation. Finally, we characterize the gap to the Fisher oracle in terms of coefficient heterogeneity and movement-estimation error, with a high-probability bound for the latter under sub-Gaussian noise.

C.1 NOTATION AND SETUP

For completeness, we first summarize the notation used in the analysis and formalize the movement-based parameter selection in CPO.

Fix a transition from $\mathcal{T}_{i-1}$ to $\mathcal{T}_i$ with $i \geq 2$. As in the main text, let $\widetilde{\theta} = \theta^{(i-1)}$, $\widetilde{\pi} = \pi_{\widetilde{\theta}}$, and let θ denote the parameters during subsequent adaptation. We distinguish the historical movement used for selection from the subsequent displacement being regularized:

$$\Delta\theta_k = \widetilde{\theta}_k - \theta_k^{(i-2)}, \qquad s_k^{(i-1)} = |\Delta\theta_k|, \qquad \delta = \theta - \widetilde{\theta}. \tag{80}$$

For brevity, throughout this appendix write

$$F = F_{\widetilde{\theta}}^{\mathcal{T}_{i-1}} \succeq 0, \qquad F_{kk} = F_{kk}^{\mathcal{T}_{i-1}}, \qquad m = m^{(i-1)} \in \{0,1\}^d. \tag{81}$$

Thus m always denotes the binary cumulative mask; it is not a movement score. All coordinates refer to the trainable parameterization, with the remaining parameters held fixed.

Movement scores under the local model. Under the locally stationary SGD model of Proposition B.2, with $\eta > 0$, $N \geq 1$, and $0 < \sigma_A^2 = \mathbb{E}[A^2] < \infty$, define

$$\widehat{u}_k := \frac{(s_k^{(i-1)})^2}{\eta^2 \sigma_A^2}, \qquad u_k := \mathbb{E}[\widehat{u}_k] = C_k F_{kk}, \qquad C_k = N^2\rho_k^2 + N(q_k - \rho_k^2). \tag{82}$$

Here q_k, ρ_k, μ_k, and V_k retain their definitions in Appendix B.2, and $C_k \geq 0$. All score and gradient second moments are assumed finite. The expectation treats the local policy, prefix distribution, and score moments as fixed, as in that approximation. For $F_{kk} = 0$, the score vanishes almost surely under this model; we set $C_k = u_k = 0$ by convention. Ranking by $\widehat{u}_k$ is exactly equivalent to ranking by $s_k^{(i-1)}$, since squaring and a shared positive rescaling preserve this realized ranking. The deviation of the realized score from its expectation is quantified explicitly in section C.4.

The selection rule and its Fisher comparator. Keep the tensor partition $\mathcal{G} = \{G_1, \ldots, G_M\}$ from the main text. For each tensor G_g, let b_g be the number of coordinates returned by its TopP_p operation, including its fixed rounding rule. Write $S_{s,g}$ for these coordinates at the end of $\mathcal{T}_{i-1}$, and define

$$S_{s,g} \in \operatorname*{arg\,max}_{S \subseteq G_g,\ |S| = b_g} \sum_{k \in S} \widehat{u}_k, \quad S_{F,g} \in \operatorname*{arg\,max}_{S \subseteq G_g,\ |S| = b_g} \sum_{k \in S} F_{kk}. \tag{83}$$

Ties may be resolved by any fixed rule. $S_{F,g}$ serves as the Fisher-optimal comparator under the same per-tensor selection budget as CPO. By the cumulative update, $\bigcup_g S_{s,g} \subseteq \mathcal{S}^{(i-1)} = \{k : m_k = 1\}$.

For each group, define its oracle sensitivity and largest coefficient by

$$\Phi_g^\star = \sum_{k \in S_{F,g}} F_{kk}, \qquad C_{\max,g} = \max_{k \in G_g} C_k. \tag{84}$$

When $\Phi_g^\star > 0$ and $0 < C_{\max,g} < \infty$, let

$$\overline{C}_{F,g} = \frac{\sum_{k \in S_{F,g}} C_k F_{kk}}{\Phi_g^\star}, \qquad \gamma_g = \frac{\overline{C}_{F,g}}{C_{\max,g}} \in [0,1]. \tag{85}$$

The average is weighted by Fisher sensitivity, allowing heterogeneous coordinate-wise coefficients C_k.

For nonnegative tolerances $\varepsilon_1, \ldots, \varepsilon_M$, consider the event

$$\mathcal{E}(\varepsilon) = \left\{ \max_{k \in G_g} |\widehat{u}_k - u_k| \le \varepsilon_g \text{ for every } g \right\}. \tag{86}$$

Define the certified covered sensitivity

$$\Lambda = \sum_{g=1}^{M} \Lambda_g, \qquad \Lambda_g = \left[\gamma_g \Phi_g^\star - \frac{2 b_g \varepsilon_g}{C_{\max,g}} \right]_+, \qquad [x]_+ := \max\{x, 0\}. \tag{87}$$

The expression for Λ_g applies to the groups with $\Phi_g^\star > 0$ and $0 < C_{\max,g} < \infty$. Set $\Lambda_g = 0$ whenever $b_g = 0$, $\Phi_g^\star = 0$, or $C_{\max,g} = 0$, such a group contributes no positive lower bound.

C.2 Main theorem: from the mask to a forgetting bound

The total diagonal sensitivity covered by a mask and its uncovered remainder are

$$\mathcal{I}_F(m) := \sum_{k=1}^{d} m_k F_{kk}, \qquad \tau_F(m) := \sum_{k=1}^{d} (1 - m_k) F_{kk} = \operatorname{tr}(F) - \mathcal{I}_F(m). \tag{88}$$

Let

$$\Omega_{i-1}(\theta) := \frac{\|m \odot \delta\|_1}{\|m\|_0}, \qquad a_F(m) := \max_{k : m_k = 1} \sqrt{F_{kk}}. \tag{89}$$

For an empty mask, set both quantities to zero. Ω_{i-1} is exactly the normalized masked movement penalty in Eq. 15, before multiplication by λ.

Theorem C.1 (Movement-selected sensitivity and local forgetting)**.** *Assume the setting of section C.1 and consider a realization in $\mathcal{E}(\varepsilon)$. Let m be any cumulative mask containing $\bigcup_g S_{s,g}$. Then*

$$\mathcal{I}_F(m) \ge \Lambda, \qquad \tau_F(m) \le \operatorname{tr}(F) - \Lambda. \tag{90}$$

Define

$$D_m(\theta) := a_F(m) \|m\|_0 \, \Omega_{i-1}(\theta) + \sqrt{\operatorname{tr}(F) - \Lambda} \, \|(\mathbf{1} - m) \odot \delta\|_2. \tag{91}$$

For the full Fisher matrix, including arbitrary off-diagonal interactions,

$$\delta^\top F \delta \le D_m(\theta)^2. \tag{92}$$

Suppose the prior-task behavioral KL has the expansion from section 3.3,

$$K_{i-1}(\theta) = \frac{1}{2} \delta^\top F \delta + R_{i-1}^{(3)}(\theta). \tag{93}$$

Then

$$K_{i-1}(\theta) \le \frac{1}{2} D_m(\theta)^2 + |R_{i-1}^{(3)}(\theta)|. \tag{94}$$

For bounded old-task reward range B_{i-1} and generation horizon H,

$$|\eta_{i-1}(\widetilde{\pi}) - \eta_{i-1}(\pi_\theta)| \le B_{i-1} \sqrt{\frac{H}{4} D_m(\theta)^2 + \frac{H}{2} |R_{i-1}^{(3)}(\theta)|}. \tag{95}$$

If additionally $\nabla^2 K_{i-1}$ is L-Lipschitz in operator norm on a neighborhood containing the segment from $\widetilde{\theta}$ to θ, then $|R_{i-1}^{(3)}(\theta)| \le L \|\delta\|_2^3 / 6$, and hence

$$|\eta_{i-1}(\widetilde{\pi}) - \eta_{i-1}(\pi_\theta)| \le B_{i-1} \sqrt{\frac{H}{4} D_m(\theta)^2 + \frac{HL}{12} \|\delta\|_2^3}. \tag{96}$$

Proof. We prove the result in three steps: first showing that the selection rule retains aggregate Fisher sensitivity, then controlling the full Fisher quadratic form via a covered/uncovered decomposition, and finally translating this bound into forgetting through KL continuity and smoothness.

Step 1: movement selection retains aggregate Fisher sensitivity. Fix a group G_g with $C_{\max,g} > 0$. On $\mathcal{E}(\varepsilon)$, since $S_{s,g}$ maximizes the selection criterion,

$$\begin{aligned}\sum_{k\in S_{s,g}} C_k F_{kk} &= \sum_{k\in S_{s,g}} u_k \\ &\geq \sum_{k\in S_{s,g}} \widehat{u}_k - b_g\varepsilon_g \\ &\geq \sum_{k\in S_{F,g}} \widehat{u}_k - b_g\varepsilon_g \\ &\geq \sum_{k\in S_{F,g}} u_k - 2b_g\varepsilon_g \\ &= \overline{C}_{F,g}\Phi_g^\star - 2b_g\varepsilon_g.\end{aligned} \tag{97}$$

Since $C_k \leq C_{\max,g}$ and $F_{kk} \geq 0$, we have

$$\sum_{k\in S_{s,g}} C_k F_{kk} \leq C_{\max,g} \sum_{k\in S_{s,g}} F_{kk}.$$

Combining this with 97 gives

$$\sum_{k\in S_{s,g}} F_{kk} \geq \left[\frac{\overline{C}_{F,g}\Phi_g^\star - 2b_g\varepsilon_g}{C_{\max,g}}\right]_+ = \Lambda_g. \tag{98}$$

For groups with $C_{\max,g} = 0$, we set $\Lambda_g = 0$; then the same lower bound holds immediately because $F_{kk} \geq 0$.

The groups are disjoint, and the cumulative mask contains all newly selected coordinates, so

$$\mathcal{I}_F(m) \geq \sum_{g=1}^{M} \sum_{k\in S_{s,g}} F_{kk} \geq \sum_{g=1}^{M} \Lambda_g = \Lambda. \tag{99}$$

Subtracting this lower bound from $\mathrm{tr}(F)$ proves Eq. 90.

Step 2: covered sensitivity and masked movement control the full quadratic form. Let $S = \{k : m_k = 1\}$ and let e_k be the k-th coordinate vector. Positive semidefiniteness gives $\|F^{1/2}e_k\|_2 = \sqrt{F_{kk}}$. Consequently, the triangle and Cauchy–Schwarz inequalities imply

$$\begin{aligned}\sqrt{\delta^\top F\delta} &= \left\|\sum_{k=1}^{d} \delta_k F^{1/2} e_k\right\|_2 \\ &\leq \sum_{k\in S} |\delta_k|\sqrt{F_{kk}} + \sum_{k\notin S} |\delta_k|\sqrt{F_{kk}} \\ &\leq a_F(m)\|m\odot\delta\|_1 + \sqrt{\tau_F(m)}\,\|(\mathbf{1}-m)\odot\delta\|_2 \\ &\leq a_F(m)\|m\|_0\,\Omega_{i-1}(\theta) + \sqrt{\mathrm{tr}(F)-\Lambda}\,\|(\mathbf{1}-m)\odot\delta\|_2.\end{aligned} \tag{100}$$

Squaring proves Eq. 92. In particular, off-diagonal Fisher interactions are included through the positive-semidefinite bound rather than discarded.

Step 3: pass from the quadratic form to behavioral KL and returns. Substituting Eq. 92 into Eq. 93 and bounding the remainder by its absolute value proves Eq. 94. Proposition B.1 gives

$$|\eta_{i-1}(\widetilde{\pi}) - \eta_{i-1}(\pi_\theta)| \leq B_{i-1}\sqrt{\frac{H}{2}K_{i-1}(\theta)}, \tag{101}$$

which proves Eq. 95. For the smoothness specialization, using $K_{i-1}(\widetilde{\theta}) = 0$ and $\nabla K_{i-1}(\widetilde{\theta}) = 0$, the integral remainder satisfies

$$\begin{aligned} |R_{i-1}^{(3)}(\theta)| &\le \int_0^1 (1-t)\|\delta\|_2^2 \left\|\nabla^2 K_{i-1}(\widetilde{\theta}+t\delta) - F\right\|_{\mathrm{op}} dt \\ &\le L\|\delta\|_2^3 \int_0^1 t(1-t)\, dt = \frac{L}{6}\|\delta\|_2^3. \end{aligned} \tag{102}$$

Substitution yields Eq. 96. □

Role of the CPO regularizer. The bound separates the roles of selection and regularization. The selection rule determines Λ, and thus controls the Fisher sensitivity that may remain outside the mask. The masked L_1 penalty controls the parameter displacement on the selected coordinates. Together, these two terms bound the full Fisher quadratic form. In particular, for an iterate satisfying $\Omega_{i-1}(\theta) \le r_1$ and $\|(\mathbf{1}-m)\odot\delta\|_2 \le r_2$, we obtain

$$D_m(\theta) \le a_F(m)\|m\|_0 r_1 + \sqrt{\mathrm{tr}(F)-\Lambda}\, r_2. \tag{103}$$

The factor $\|m\|_0$ follows from the normalized masked L_1 regularizer in Eq. 15.

C.3 Controlling the Uncovered Fisher Sensitivity

Theorem C.1 bounds the contribution of the uncovered coordinates through the factor $\sqrt{\mathrm{tr}(F)-\Lambda}$. We compare this factor with the Fisher oracle under the same per-tensor selection budgets.

Define the oracle residual

$$\tau^\star := \mathrm{tr}(F) - \sum_g \Phi_g^\star.$$

For each group with $\Phi_g^\star > 0$ and $C_{\max,g} > 0$, let $e_g := 2b_g\varepsilon_g/C_{\max,g}$; set $\gamma_g = e_g = 0$ for the remaining groups, consistently with $\Lambda_g = 0$. On $\mathcal{E}(\varepsilon)$, the inequality $\Lambda_g = [\gamma_g\Phi_g^\star - e_g]_+ \ge \gamma_g\Phi_g^\star - e_g$ gives

$$\begin{aligned} \tau_F(m) &\le \mathrm{tr}(F) - \Lambda \\ &\le \underbrace{\tau^\star}_{\text{oracle tail}} + \underbrace{\sum_g (1-\gamma_g)\Phi_g^\star}_{\text{coefficient loss}} + \underbrace{\sum_g e_g}_{\text{movement error}}. \end{aligned} \tag{104}$$

Selection approaches the oracle residual when the latter two terms are small. The following sufficient condition controls the coefficient loss under heterogeneous C_k.

Corollary C.2 (Uncovered sensitivity under partial excitation)**.** *For each group with $\Phi_g^\star > 0$ and $C_{\max,g} > 0$, suppose there exist $B_g \subseteq S_{F,g}$ and $\alpha_g, \beta_g \in [0,1]$ such that*

$$\sum_{k\in B_g} F_{kk} \ge \alpha_g \Phi_g^\star, \qquad C_k \ge \beta_g C_{\max,g} \quad \textit{for every } k \in B_g. \tag{105}$$

Set $\alpha_g = \beta_g = 0$ for the remaining groups. Then $\gamma_g \ge \alpha_g\beta_g$, and on $\mathcal{E}(\varepsilon)$,

$$\sqrt{\mathrm{tr}(F)-\Lambda} \le \left[\tau^\star + \sum_g (1-\alpha_g\beta_g)\Phi_g^\star + \sum_g e_g\right]^{1/2}. \tag{106}$$

Proof. For a group with $\Phi_g^\star > 0$ and $C_{\max,g} > 0$, nonnegativity and Eq. 105 imply

$$\gamma_g = \frac{\sum_{k\in S_{F,g}} C_k F_{kk}}{C_{\max,g}\Phi_g^\star} \ge \beta_g \frac{\sum_{k\in B_g} F_{kk}}{\Phi_g^\star} \ge \alpha_g\beta_g. \tag{107}$$

The claim follows from Eq. 104 and monotonicity of the square root. □

The condition is aggregate. It requires a sufficient fraction of the Fisher sensitivity to exhibit appreciable movement coupling. It allows unrestricted coefficient variation outside B_g. This requirement is substantially weaker than coordinate-wise coefficient uniformity.

The coefficient multiplying the uncovered coordinates is small when the selection budgets cover most Fisher sensitivity, the Fisher-weighted coefficient loss is small, and the realized movement errors are controlled. Substituting Eq. 106 into $D_m(\theta)$ gives the corresponding KL and return-degradation bounds in Theorem C.1, with the unprotected displacement and Taylor remainder left explicit.

C.4 Controlling the Realized-Movement Error

Section C.3 isolates the effect of realized-movement noise through the error term $e_g = 2b_g\varepsilon_g/C_{\max,g}$. We now bound ε_g with high probability under an additional sub-Gaussian assumption on the stochastic-gradient noise.

Proposition C.3 (A simultaneous movement-error bound)**.** *Let the noise in Appendix B.2 satisfy $z_{n,k} = \mu_k + \xi_{n,k}$, $\mathbb{E}[\xi_{n,k} \mid \mathcal{F}_{n-1}] = 0$, and $\mathbb{E}[\xi_{n,k}^2 \mid \mathcal{F}_{n-1}] = V_k$. Assume additionally that, for deterministic $v_k \geq 0$,*

$$\mathbb{E}[\exp(t\xi_{n,k}) \mid \mathcal{F}_{n-1}] \leq \exp(t^2 v_k^2/2) \quad \textit{for all } t \in \mathbb{R}. \tag{108}$$

For $\zeta \in (0,1)$, define

$$\begin{aligned} a_k(\zeta) &= v_k\sqrt{2N\log(2d/\zeta)}, \\ \varepsilon_g(\zeta) &= \max_{k\in G_g} \frac{2N|\mu_k|a_k(\zeta) + a_k(\zeta)^2 + NV_k}{\sigma_A^2}. \end{aligned} \tag{109}$$

Then $\mathbb{P}(\mathcal{E}(\varepsilon(\zeta))) \geq 1-\zeta$. Consequently, the conclusions of Theorem C.1 hold with at least this probability, with respect to the historical movement noise, using these tolerances.

Proof. Let $M_{N,k} = \sum_{n=1}^N \xi_{n,k}$. Iterated conditioning in Eq. 108 gives $\mathbb{E}[\exp(tM_{N,k})] \leq \exp(Nt^2v_k^2/2)$. Applying the exponential Markov inequality to both tails and optimizing t yields

$$\mathbb{P}\big(|M_{N,k}| > a_k(\zeta)\big) \leq \zeta/d. \tag{110}$$

For $v_k = 0$, $M_{N,k} = 0$ almost surely and the conclusion is immediate. A union bound makes $|M_{N,k}| \leq a_k(\zeta)$ simultaneous for all coordinates with probability at least $1-\zeta$. On this event, the SGD identity $\Delta\theta_k = -\eta(N\mu_k + M_{N,k})$ implies

$$\begin{aligned} |\widehat{u}_k - u_k| &= \frac{|2N\mu_k M_{N,k} + M_{N,k}^2 - NV_k|}{\sigma_A^2} \\ &\leq \frac{2N|\mu_k|a_k(\zeta) + a_k(\zeta)^2 + NV_k}{\sigma_A^2}. \end{aligned} \tag{111}$$

Taking the maximum within each group proves the result. □

Remark. The above results require neither uniform C_k, exact Fisher ranking recovery, nor a diagonal Fisher matrix. Their validity is instead governed by three quantities made explicit in the bounds: the aggregate Fisher sensitivity retained by the mask, the deviation between realized and expected movement, and the subsequent parameter drift. The sub-Gaussian assumption in Proposition C.3 is only used to obtain a high-probability control of the realized-movement error and is not required for the deterministic statements in Theorem C.1.

D Details of MRCL

This section describes the construction procedure for each task in MRCL.

MedBookVQA. This task evaluates medical visual understanding in difficult clinical scenarios. We select the Symptom Recognition and Surgery Operation subsets of Yip et al. (2025). After filtering

out a small number of samples with parsing failures, we randomly split the remaining data into training and testing sets using a 7:3 ratio.

Navigation. The Navigation task is derived from the ThinkMorph (Gu et al., 2026b) dataset. We implement a strict manual filtering process to ensure that each sample has a strictly unique shortest path, making the evaluation unambiguous. We then divide the data by the required number of navigation steps: samples requiring 2 to 10 steps are split into training and testing sets at a 7:3 ratio, and all samples requiring 11 to 15 steps are assigned to the test set.

We-Math2.0. This task evaluates complex multimodal mathematical reasoning and is processed into two configurations. For We-Math2.0-standard (Qiao et al., 2026), we extract problems from Plane Geometry, Analytic Geometry, and Solid Geometry. To ensure broad coverage, we sample one question per fundamental principle attribute to construct the test set, leaving the rest for training. For We-Math2.0-pro (Qiao et al., 2026), we separate the dataset based on difficulty attributes. Problems with foundational difficulties (base x, y, z) are designated as the training set, and those with extended combinatorial difficulties (xy, xz, yz, xyz) form the test set. Additionally, we conduct a comprehensive calibration of the ground-truth labels so that mathematical expressions strictly conform to standard LaTeX formatting.

Puzzle. This task is designed to test visual-spatial reasoning sourced from the ThinkMorph (Gu et al., 2026b) datasets. We use the Jigsaw_Assembly subset entirely for training and the VisPuzzle subset exclusively for testing.

FinMME. The financial chart understanding task requires precise visual extraction and numerical reasoning (Luo et al., 2025). We clean the dataset by removing images that fail to parse, with half of the questions being multi-answer choice where more than one option may be correct and the other half being numerical questions. Recognizing the inherent visual discrepancies when models estimate values directly from charts, we introduce a reasonable tolerance threshold for the numerical ground truths to ensure fair evaluation.

E Implementation Details

We implement all baseline methods and CPO within the `TRL` framework following their original formulations and corresponding TRL implementations. We conduct experiments with Qwen3-VL-2B, Qwen3-VL-4B, and Qwen3-VL-8B (Bai et al., 2025a), and all reported SFT and RL runs optimize the full model. Our experimental setting uses only the original task ground-truth answers and excludes externally generated or manually annotated reasoning traces to avoid additional supervision from CoT annotations. Accordingly, SFT uses direct question-answer pairs with System 1 prompting, while all RL methods use the same CoT prompting template during rollout generation without supervised CoT targets. For all RL methods, we use AdamW with a learning rate of 5×10^{-6}, weight decay of 0.1, a warmup ratio of 0.1, and cosine learning rate decay. Training uses 8 rollouts per prompt, a sampling temperature of 0.8, and a maximum completion length of 2048. MedBookVQA is trained for 200 steps and the remaining MRCL tasks for 600 steps. For GRPO (Shao et al., 2024), we adopt the DAPO improvements (Yu et al., 2025) with $\epsilon_{\text{low}} = 0.2$ and $\epsilon_{\text{high}} = 0.27$ under token-level policy gradient loss. GRPO* additionally uses a KL penalty coefficient of $\beta = 0.04$. SAPO (Gao et al., 2025) uses group reward scaling with $\tau_{\text{pos}} = 1.0$ and $\tau_{\text{neg}} = 1.05$. Dr.GRPO (Liu et al., 2025) normalizes the policy loss by the fixed maximum completion length, disables reward standard deviation scaling, and uses symmetric clipping with $\epsilon_{\text{low}} = \epsilon_{\text{high}} = 0.2$. CISPO (Chen et al., 2025a) uses group reward scaling with the direct importance weight cap $\epsilon_{\text{high}} = 5.0$, following the recommended TRL configuration. For GSPO (Zheng et al., 2025), we set $\epsilon_{\text{low}} = 3 \times 10^{-4}$ and $\epsilon_{\text{high}} = 4 \times 10^{-4}$. CPO builds on GSPO and applies sparse regularization with $\lambda = 100$. After each task, parameter movement is ranked independently within every trainable parameter tensor, and the top $p = 10\%$ coordinates in each tensor are added to the cumulative protected set. All experiments use NVIDIA H100 GPUs. Full SFT uses 2 GPUs for the 2B and 4B models and 4 GPUs for the 8B model, while all RL experiments use 8 GPUs.

During inference, we use temperature 0.0 and repetition penalty 1.05 with baseline Zero-shot[2]. The maximum generation length is 2048 for the five MRCL tasks and 16384 for external capability benchmarks.

We adopt distinct scoring rules tailored to the answer formats of different tasks. For *MedBookVQA* and *Puzzle*, a prediction receives a score of 1 for an exact match and 0 otherwise. For *Navigation*, a prediction receives a score of 1 only if the entire generated trajectory perfectly matches the ground truth shortest path. For *We-Math2.0*, answers containing multiple blanks separated by semicolons receive partial credit, and each correct blank contributes $1/k$ to the score where k is the total number of blanks. For *FinMME*, multiple-answer choice questions receive a score of 0.5 for incomplete correct selections and 0 if any incorrect option is chosen. For numerical questions in *FinMME*, a prediction receives a score of 1 if it falls within a predefined tolerance threshold.

The evaluation of diverse multimodal benchmarks to measure retention of pretrained capabilities covers STEM benchmarks MMMU-Pro (Yue et al., 2025), MathVerse (Zhang et al., 2024), MathVision (Wang et al., 2024a), and MathVista (Lu et al., 2024); general VQA benchmarks RealworldQA (xAI, 2024) and MMStar (Chen et al., 2024); the hallucination benchmark POPE (Li et al., 2023); chart and document understanding benchmarks DocVQA (Mathew et al., 2021) and CharXiv (Wang et al., 2024c); and the grounding benchmark CountBenchQA (Paiss et al., 2023). We report F1 for POPE. For MMMU-Pro, we average accuracy over the standard 4-option, standard 10-option, and vision subsets. All other benchmarks use accuracy.

E.1 CAI and CAD Metrics for Fig. 2

To quantify the dynamics of plasticity and stability visualized in Fig. 2, we introduce two trajectory metrics corresponding to the axes: Cumulative mean Accuracy Improvement (CAI) and Cumulative mean Accuracy Delta (CAD). Let T denote the total number of tasks, $a_{i,j}$ denote the accuracy on task j after training on task i, and z_j denote the zero-shot accuracy on task j.

The Cumulative mean Accuracy Improvement measures the average performance gain immediately after training on each task compared to its zero-shot baseline, accumulated from the first task through the current task i:

$$\mathrm{CAI}_i = \frac{1}{T} \sum_{j=1}^{i} (a_{j,j} - z_j) \tag{112}$$

The Cumulative mean Accuracy Delta measures the accumulated performance reduction on previously learned tasks. In each sequential stage k from 2 to i, we compute the average accuracy delta for all previous tasks $j < k$ after learning the new task k:

$$\mathrm{CAD}_i = \sum_{k=2}^{i} \frac{1}{k-1} \sum_{j=1}^{k-1} (a_{k,j} - a_{k-1,j}) \tag{113}$$

A negative CAD_i value indicates cumulative forgetting, while a positive value suggests backward transfer that improves performance on previous tasks.

E.2 Teacher-Forced Answer-Token NLL for Fig. 1

To assess benchmark familiarity from a complementary perspective, we measure the conditional likelihood assigned by a frozen Qwen3-VL-4B-Instruct model to the canonical answer of each test example. For sample i, let c_i denote the multimodal prompt and let $y_i = (y_{i,1}, \ldots, y_{i,L_i})$ denote the tokenized canonical answer. We compute the teacher-forced answer-token NLL as

$$\mathrm{NLL}_i^{\mathrm{tok}} = -\frac{1}{L_i} \sum_{t=1}^{L_i} \log p_\theta \left(y_{i,t} \mid c_i, y_{i,<t} \right). \tag{114}$$

The model remains frozen throughout evaluation, with no rollout or parameter update. Token-level normalization reduces the direct influence of answer length and enables more meaningful comparisons across benchmarks.

All tasks are evaluated under fixed task-specific protocols designed to preserve their original answer semantics while minimizing artifacts introduced by answer formatting. Fig. 1 summarizes the resulting sample-level NLL distributions and their task means. Lower NLL indicates that the model assigns higher conditional likelihood to the canonical answers, while higher NLL indicates lower conditional-answer familiarity.

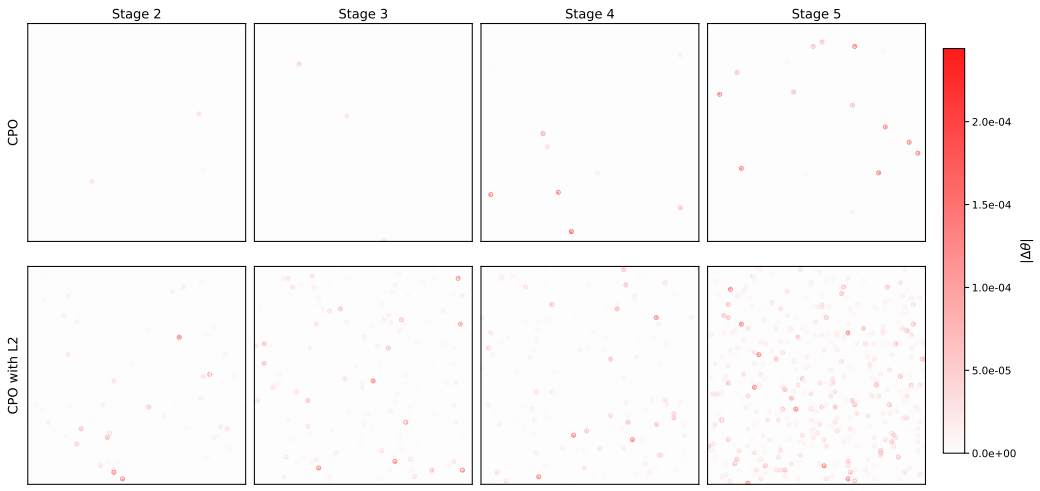


Figure 6: Parameter movement heatmaps of CPO and its masked L_2 variant on Qwen3-VL-2B. Masked L_2 induces dense movement across many top movement parameters, especially in later continual stages, while masked L_1 keeps most top movement parameters close to the previous checkpoint and produces a sparse movement pattern.

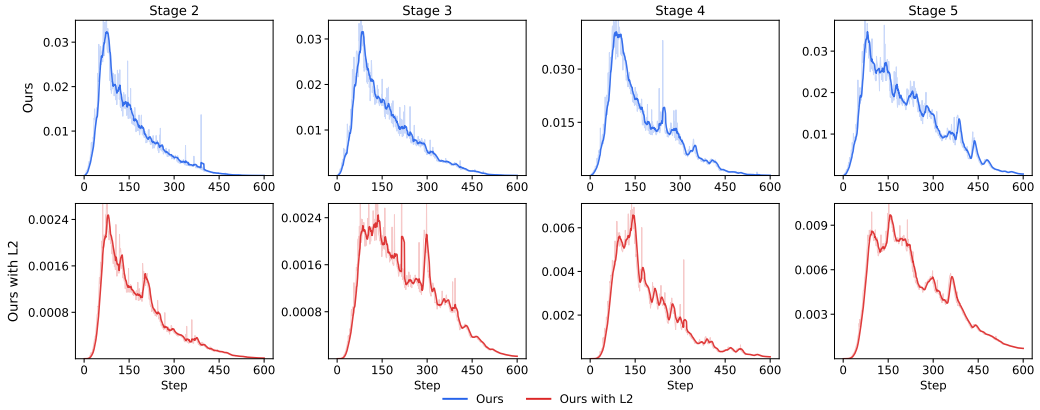


Figure 7: Regularization loss curves of CPO with masked L_1 movement regularization and its masked L_2 variant on Qwen3-VL-2B. The L_1 coefficient is set to 1e2 and the L_2 coefficient is set to 1e5. Both regularization losses remain active during training, indicating that the inferior performance of the L_2 variant is not due to an inactive regularization signal.

F L_1 VERSUS L_2 MOVEMENT REGULARIZATION

Effect of L_1 versus L_2 movement regularization. We further compare CPO with masked L_1 movement regularization against a masked L_2 variant on Qwen3-VL-2B. Fig. 7 shows that both regularization terms are active during training, as their losses consistently rise and decay across continual stages. Therefore, the performance gap is not caused by an ineffective L_2 regularization signal. Instead, the difference lies in the induced movement pattern. As shown in Fig. 6, masked L_2 causes dense movement across many top movement parameters, especially in later stages, whereas masked L_1 keeps most top movement parameters nearly unchanged and yields a much sparser movement pattern. In terms of continual learning performance, replacing masked L_1 with masked L_2 decreases MFN and MTA by 23% and 15%, respectively.

G STAGE-BY-STAGE MRCL RESULTS

This section provides the complete stage-by-stage evaluation results for all compared methods on the MRCL benchmark. Tab. 9, 10, and 11 present the detailed accuracy matrices across the sequential post-training stages for Qwen3-VL-2B, Qwen3-VL-4B, and Qwen3-VL-8B respectively. In each matrix, each row represents the model performance evaluated on a series of tasks after completing the training on the task of that specific stage, illustrating the exact trajectory of learning new capabilities along with the retention of learned knowledge.

Table 9: Accuracy of a series of fine-tuning methods on the MRCL benchmark, evaluated on the Qwen3-VL-2B model. Each row represents the performance trained after the corresponding task.

(a) SFT

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	97.96				
Navigation	90.66	67.28			
We-Math2	84.04	45.22	40.52		
Puzzle	48.56	12.32	28.81	92.25	
FinMME	44.65	0.55	33.41	83.25	33.60

(b) GRPO*

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	78.95				
Navigation	75.72	25.74			
We-Math2	75.38	27.76	53.26		
Puzzle	78.10	18.20	50.66	68.50	
FinMME	76.23	17.10	51.24	61.50	42.15

(c) GRPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	88.62				
Navigation	83.19	47.24			
We-Math2	71.65	45.96	57.08		
Puzzle	70.63	37.87	53.79	70.25	
FinMME	62.48	28.86	48.11	66.25	41.10

(d) GSPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	89.98				
Navigation	86.08	66.91			
We-Math2	85.40	38.79	56.72		
Puzzle	79.12	27.57	46.19	86.50	
FinMME	69.27	11.95	45.20	75.75	41.05

(e) CPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	89.98				
Navigation	88.96	68.01			
We-Math2	86.25	46.32	55.02		
Puzzle	85.40	42.65	50.38	85.75	
FinMME	76.06	43.38	49.03	78.00	42.65

Table 10: Accuracy of a series of fine-tuning methods on the MRCL benchmark, evaluated on the Qwen3-VL-4B model. Each row represents the performance trained after the corresponding task.

(a) SFT

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	98.47				
Navigation	95.25	75.00			
We-Math2	87.44	42.10	49.08		
Puzzle	47.20	21.51	38.35	89.75	
FinMME	56.71	3.12	37.84	84.50	37.90

(b) GRPO*

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	67.74				
Navigation	66.72	60.66			
We-Math2	63.16	53.68	62.69		
Puzzle	61.46	42.83	63.11	69.25	
FinMME	63.16	40.26	62.84	65.50	48.50

(c) GRPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	91.51				
Navigation	86.08	55.88			
We-Math2	70.29	39.52	62.55		
Puzzle	66.21	32.35	59.94	78.75	
FinMME	70.63	39.34	58.18	68.25	47.75

(d) Dr.GRPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	76.23				
Navigation	71.82	59.38			
We-Math2	73.85	49.63	62.07		
Puzzle	68.42	46.51	60.52	69.00	
FinMME	68.76	30.51	56.60	54.00	43.50

(e) SAPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	93.21				
Navigation	90.83	73.35			
We-Math2	89.47	63.05	57.55		
Puzzle	83.19	0.00	28.31	35.00	
FinMME	78.27	0.00	28.37	34.00	29.00

(f) GSPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	92.53				
Navigation	87.61	81.62			
We-Math2	83.19	75.55	65.10		
Puzzle	78.27	35.66	47.38	87.25	
FinMME	65.03	30.70	45.17	83.00	44.80

(g) CISPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	93.38				
Navigation	4.07	0.00			
We-Math2	92.02	0.92	61.93		
Puzzle	90.49	0.00	37.05	67.00	
FinMME	57.39	0.18	23.37	48.50	27.55

(h) CPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	92.53				
Navigation	93.04	82.54			
We-Math2	92.19	86.03	64.13		
Puzzle	91.34	80.33	63.85	78.00	
FinMME	90.83	76.84	63.64	77.50	49.00

Table 11: Accuracy of a series of fine-tuning methods on the MRCL benchmark, evaluated on the Qwen3-VL-8B model. Each row represents the performance trained after the corresponding task.

(a) SFT

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	100.00				
Navigation	100.00	84.19			
We-Math2	99.32	77.02	50.06		
Puzzle	39.73	52.02	38.87	35.50	
FinMME	79.80	33.46	38.94	35.00	32.75

(b) GRPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	90.83				
Navigation	81.49	43.38			
We-Math2	85.06	20.96	62.70		
Puzzle	77.25	14.89	59.28	82.25	
FinMME	73.51	6.62	55.43	72.25	45.80

(c) GSPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	95.59				
Navigation	93.21	81.25			
We-Math2	92.02	74.08	67.27		
Puzzle	82.34	44.67	56.66	89.75	
FinMME	77.08	44.67	52.68	86.25	48.15

(d) CPO

	MedBook	Navigation	We-Math2	Puzzle	FinMME
MedBook	95.59				
Navigation	95.08	80.88			
We-Math2	95.08	80.15	64.99		
Puzzle	94.91	79.23	64.62	76.00	
FinMME	96.94	81.25	64.75	74.25	49.40

H PROMPT TEMPLATES

In the experiments, two prompt strategies are used. Under the Direct Prompts setting, the model is restricted to output only the final answer. The CoT Prompts elicits explicit step-by-step reasoning from the model before yielding the prediction, where the final answer is enclosed within boxed{}. Concretely, we apply the Direct Prompts to the Zero-shot[1], Full Fine-Tuning and LoRA-based methods. The CoT Prompts are deployed for Zero-shot[2], GRPO, GSPO and CPO.

Direct Prompts:

MedBookVQA:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please output the answer directly and NO extra thinking.
Rules for the final answer:
- Multiple Choice: Output the uppercase option letter ONLY (e.g., A, B, C, D, E...).
{question}"

Navigation:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please output the answer directly and NO extra thinking.
Note: There is one and only one shortest path.
Rules for the final answer:
- Only provide your solution as a sequence of capital letter which should be comma-separated.
{question}"

We-Math2:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please output the answer directly and NO extra thinking.
Rules for the final answer:
- Multiple Choice: Output the uppercase option letter ONLY (e.g., A, B, C, D, E...).
- Yes/No: Output Yes or No ONLY.
- Numerical: Output the number ONLY.
- Fill-in-the-blank: Output the answer phrase or word directly.
- Math Formula: Output the LaTeX expression enclosed in single dollar signs (e.g., $x = y$).
{question}"

Puzzle:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please output the answer directly and NO extra thinking.
Rules for the final answer:
- Multiple Choice: Output the uppercase option letter ONLY (e.g., A, B, C, D, E...).
{question}"

FinMME:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please output the answer directly and NO extra thinking.
Rules for the final answer:
- Multiple Choice: If one or more options are correct, output the uppercase option letter(s) ONLY. If there are multiple answers, separate them with commas (e.g., A,C,D).
- Numerical: Output the number ONLY.
{question}"

CoT Prompts:

MedBookVQA:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please think step-by-step and put the final answer in `\\boxed{{}}`.
Rules for the final answer in `\\boxed{{}}`:
- Multiple Choice: Output the uppercase option letter ONLY (e.g., A, B, C, D, E...).
{question}"

Navigation:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please think step-by-step and put the final answer in `\\boxed{{}}`.
Hints:
You need to first determine the size of the grid, then determine the coordinates of the hole, player, and gift, and finally perform path planning.
Note: There is one and only one shortest path.
Rules for the final answer in `\\boxed{{}}`:
- Only provide your solution as a sequence of capital letter which should be comma-separated.
{question}"

We-Math2:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please think step-by-step and put the final answer in `\\boxed{{}}`.
Rules for the final answer in `\\boxed{{}}`:
- Multiple Choice: Output the uppercase option letter ONLY (e.g., A, B, C, D, E...).
- Yes/No: Output Yes or No ONLY.
- Numerical: Output the number ONLY.
- Fill-in-the-blank: Output the answer phrase or word directly.
- Math Formula: Output the LaTeX expression enclosed in single dollar signs (e.g., $x = y$).
{question}"

Puzzle:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please think step-by-step and put the final answer in `\\boxed{{}}`.
Rules for the final answer in `\\boxed{{}}`:
- Multiple Choice: Output the uppercase option letter ONLY (e.g., A, B, C, D, E...).
{question}"

FinMME:
"You are an AI assistant helping to solve the following problem based on the image(s).
Instructions:
Please think step-by-step and put the final answer in `\\boxed{{}}`.
Rules for the final answer in `\\boxed{{}}`:
- Multiple Choice: If one or more options are correct, output the uppercase option letter(s) ONLY. If there are multiple answers, separate them with commas (e.g., A,C,D).
- Numerical: Output the number ONLY.
{question}"

I CASE STUDY

This section presents a qualitative case study on Qwen3-VL-4B across diverse datasets to illustrate the evolution of model behavior during continual post-training. We compare model responses to identical test samples when evaluated immediately after learning each specific task versus after completing the entire sequential training process. The results show that while GSPO successfully acquires correct problem-solving strategies when evaluated immediately after training on a task, it subsequently suffers from severe cognitive collapse on those same tasks as the sequence progresses. For example, after finishing the final FinMME stage, the GSPO model fails to maintain mathematical precision, misinterprets grid dimensions in navigation, and generates heavily hallucinated

answers on medical datasets. In contrast, CPO consistently preserves both the step-by-step reasoning structure and the visual comprehension acquired during earlier stages without degrading its logical chain.

TASK: Puzzle

Question :



The image below is divided into four parts by white strips, forming a 2×2 jigsaw puzzle. The parts are labeled "1" (top-left), "2" (top-right), "3" (bottom-left), and "4" (bottom-right). These parts are from a single original image but have been shuffled. Your task is to determine the correct arrangement of the physically labeled parts to reconstruct the natural image.

A. The top-left part should be Part 1; the top-right part should be Part 2; the bottom-left part should be Part 4; and the bottom-right part should be Part 3.
B. The top-left part should be Part 2; the top-right part should be Part 1; the bottom-left part should be Part 3; and the bottom-right part should be Part 4.
C. The top-left part should be Part 4; the top-right part should be Part 1; the bottom-left part should be Part 3; and the bottom-right part should be Part 2.
D. The top-left part should be Part 3; the top-right part should be Part 1; the bottom-left part should be Part 2; and the bottom-right part should be Part 4.

GSPO (train after Puzzle):

Looking at the image, we see a 2x2 grid with four labeled parts. The question asks us to arrange them logically. Alternative D says: Top left = Part 3 (bedroom corner) Top right = Part 1 (lamp) Bottom left = Part 2 (nightstand) Bottom right = Part 4 (rug) But looking at the correct answer, the question might be testing logical reasoning. Assuming D is a logical arrangement, let me pick D.
Answer: D \\boxed{D} ✔

GSPO (train after FinMME):

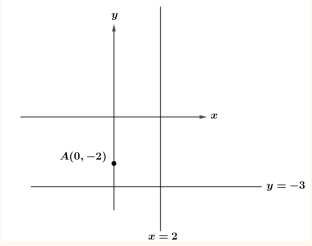
Looking at the logical arrangement: We see that the question might be testing reasoning. Alternative answer might be logical Answer: B or D Answer: B \\boxed{B} (based on logical reasoning) ✗

CPO (train after FinMME):

Looking at the four provided images: -Part 1: Shows a lamp next to a bed with a black headboard, against a light wall. - Part 2: Shows the bottom edge of a dark bed with a white base. - Part 3: Shows a bed with a white headboard and dark patterned pillows, next to a mirror or glass door. - Part 4: Shows the floor area with a black nightstand and the corner of a red bed. Now let's try to logically arrange them into a 2x2 grid. The most logical arrangement would be: Top row: - Left: Part 1 — because it has the lamp and wall, likely the upper part of the room. - Right: Part 3 — because it has the white headboard and appears more "upper" in the room layout. Bottom row: - Left: Part 2 — because it shows the bed base, which is lower. - Right: Part 4 — because it shows the floor and nightstand, also lower. Wait — let's reconsider. Actually, looking closely, Part 3 seems to be the largest and most complete upper section with the headboard and pillows. It likely belongs in the top-left or top-right. Let's try arranging: Top-left : Part 3 — because it has the white headboard and looks like the main bed area. Top-right: Part 1 — because it has the lamp and wall, which might be adjacent to the bed. Bottom-left: Part 2 — because it shows the bed base, which is lower. Bottom-right: Part 4 — because it shows the floor and nightstand. This gives us: Top row: [Part 3] [Part 1] Bottom row: [Part 2] [Part 4] This matches option D: $>$ D. The top-left part should be Part 3; the top-right part should be Part 1; the bottom-left part should be Part 2; and the bottom-right part should be Part 4. This arrangement makes sense visually. Therefore, the correct answer is \\boxed{D}. ✔

TASK: We-Math2.0

Question :



In the parking lot of a large supermarket, staff plan to set up a circular charging station area. It is known that the center of the charging station area is located on an east-west marking line $x = 2$ (the straight line 2 meters east from the parking lot entrance) inside the parking lot. The supermarket has set up two signs, A and B, in the parking lot to guide vehicles into the charging station area. The staff found that signs A and B are symmetric with respect to a north-south marking line $y = -3$ (the straight line 3 meters south from the parking lot entrance). If the coordinates of sign A are $(0, -2)$ (in meters), please help the staff calculate the standard equation of the circle for the circular charging station area, so as to reasonably plan the area.

GSPO (train after We-Math2):

We are given: - Sign A is at $(0, -2)$. - Signs A and B are symmetric with respect to the line $y = -3$. - So, we reflect point $A(0, -2)$ across the line $y = -3$. - Distance from $(0, -2)$ to $y = -3$ is $|-2-(-3)| = 1$ unit down. - Reflecting across $y = -3$: go 1 unit further below $y = -3$: $y = -3 - 1 = -4$. - So, point B is at $(0, -4)$. Now, we need to find the center and radius of the circle. - The center lies on $x = 2$ (east-west marking line). - So, center is at $(2, k)$ for some k. - We also know that the circle must pass through both points $A(0, -2)$ and $B(0, -4)$ — because these are the positions of signs A and B, which are symmetric and should be equidistant from the center, and also the circle should be centered along $x = 2$ and symmetric about $y = -3$ (which is the axis of symmetry between A and B). Let's assume the center is at $(2, -3)$ — which is exactly midway between $y = -2$ and $y = -4$, and also lies on $x = 2$ — so it's a good candidate. Distance from center $(2, -3)$ to point $A(0, -2)$: Distance = $\sqrt{(2-0)^2 + (-3+2)^2} = \sqrt{4+1} = \sqrt{5}$ So, radius is $\sqrt{5}$. Standard equation of the circle: $(x-2)^2 + (y+3)^2 = 5$ But wait — let's double check if there's another possibility. Is there a different center? For example, if we consider other centers on $x = 2$, but we want the circle to be centered such that it passes through both A and B — and we already have a center at $(2, -3)$ which is equidistant to both points (distance $\sqrt{5}$), so this is the best choice. So, the standard equation is: $(x-2)^2 + (y+3)^2 = 5$ Final Answer: \\boxed{$(x-2)^2 + (y+3)^2 = 5$}. ✔

GSPO (train after FinMME):

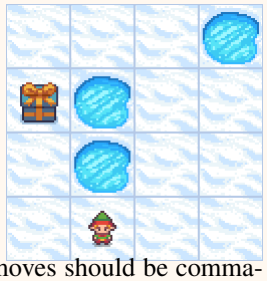
We see that we have a circle centered on $x = 2$, and we want to find the radius. We know that the signs are symmetric about $y = -3$. $A(0, -2)$ The reflection of A over $y = -3$ would be: The difference between -2 and -3 is 1 So the reflection would be at $y = -3 - 1 = -4$ But the question might be phrasing a reasonable circle around the entrance. Alternative answer might be logical Answer: Circle centered at $(2, -3)$ with radius 1 Equation: $(x-2)^2 + (y+3)^2 = 1$ Answer: $(x-2)^2 + (y+3)^2 = 1$ \\boxed{$(x-2)^2 + (y+3)^2 = 1$}. ✗

CPO (train after FinMME):

We are given: -The center of the circle lies on the vertical line $x = 2$. -Signs A and B are symmetric with respect to the horizontal line $y = -3$. - Coordinates of sign A: $(0, -2)$. Step 1: Find coordinates of sign B Since A and B are symmetric with respect to $y = -3$, we reflect point A across this line. The formula for reflection over a horizontal line $y = k$ is: $(x, y) \to (x, 2k - y)$ Here, $k = -3$, so: $(0, -2) \to (0, 2(-3) - (-2)) = (0, -6 + 2) = (0, -4)$ So, sign B is at $(0, -4)$. Step 2: Find the center of the circle The center must lie on $x = 2$, and since the circle is centered on the line $x = 2$, and the signs A and B are symmetric about $y = -3$, the center should lie on the midpoint of A and B vertically, because the symmetry line is $y = -3$. Midpoint of A and B: $\left(\frac{0+0}{2}, \frac{-2+(-4)}{2}\right) = (0, -3)$ But the center must lie on $x = 2$, so the center is at $(2, -3)$. Step 3: Find the radius The distance from the center $(2, -3)$ to point A $(0, -2)$: $r = \sqrt{(2-0)^2 + (-3-(-2))^2} = \sqrt{2^2 + (-1)^2} = \sqrt{4+1} = \sqrt{5}$ Step 4: Write the standard equation of the circle Standard form: $(x-h)^2 + (y-k)^2 = r^2$ Center: $(2, -3)$, radius $\sqrt{5}$ So: $(x-2)^2 + (y+3)^2 = 5$. Final Answer: \\boxed{$(x-2)^2 + (y+3)^2 = 5$}. ✔

TASK: Navigation

Question :



You are a maze solver. Your goal is to guide a player from the start to the goal on a grid map while avoiding holes.
The player can move one square at a time in the directions left (L), right (R), up (U), or down (D).
The frozen lake is not slippery; the player will always move in the intended direction.
Moving off the edge or falling into a hole results in failure. Reaching the goal means success.
Provide your solution as a sequence of moves, such as L,R,U,D. The moves should be comma-separated.

GSPO (train after Navigation):

First, let's map the grid to find the positions of the player, the gift (goal), and the holes. The grid is 4x4. Labeling the rows as R0, R1, R2, R3 and columns as C0, C1, C2, C3. From the image: - The gift (goal) is at R1,C0. - The player (start) is at R3,C1. - The holes (frozen lakes) are at R0,C3, R1,C1, R2,C1. - The rest are walkable tiles. The shortest path from the player at R3,C1 to the gift at R1,C0 must avoid the holes. Let's trace a possible path: Start at R3,C1. Move Up to R2,C1 — but R2,C1 is a hole → cannot move there. Move Left to R3,C0. Move Up to R2,C0. Move Up to R1,C0 — the gift! Alternative path: Move Up from R3,C1 to R2,C1 — hole → blocked. So, the only safe path is: Start at R3,C1 Move Left to R3,C0 Move Up to R2,C0 Move Up to R1,C0 Yes! R2,C0 is walkable, and R1,C0 is the gift. No holes were encountered along this path. Answer: L, U, U \\boxed{L, U, U} ✔

GSPO (train after FinMME):

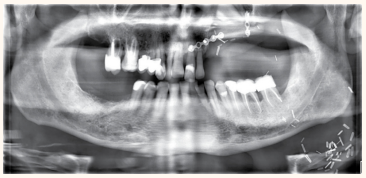
Looking at the grid: We see a 3x3 grid. Row 0: - [Gift box at [0,0] - Ice at [0,1] - Ice at [0,2] Row 1: - Ice at [1,0] - Empty Row 2: - Player at [2,0] - Empty The goal is the gift box at [0,0], the player at [2,0] Shortest path from [2,0] to [0,0]: Move UP, but we might need to avoid ice. Alternative answer might be logical Answer: U, U, U or R, L, etc. Let me say: Move UP from [2,0] to [1,0] but [1,0] is ice, so might be different approach Answer: Let me be precise Answer: U, U, U \\boxed{U,U,U} (answer might be phrasing differently) \\boxed{U,U,U} is the answer based on logical reasoning \\boxed{U,U,U} ✗

CPO (train after FinMME):

We are given a 4x4 grid. Let's label the rows and columns for clarity: Row 0: [0,0] [0,1] [0,2] [0,3] Row 1: [1,0] [1,1] [1,2] [1,3] Row 2: [2,0] [2,1] [2,2] [2,3] Row 3: [3,0] [3,1] [3,2] [3,3] From the image: - Gift (goal) is at [1,0] - Player is at [3,1] - Holes are at [0,3], [1,1], [2,1] Player must go from [3,1] to [1,0], avoiding [1,1] and [2,1]. Possible path: Start at [3,1] → Move UP to [2,1] → blocked by hole → invalid → Move LEFT to [3,0] → Move UP to [2,0] → Move UP to [1,0] → goal!Path: [3,1] → [3,0] → [2,0] → [1,0] Moves: Left (L), Up (U), Up (U) So the sequence is: L, U, U Final answer: L,U,U \\boxed{L,U,U} ✔

TASK: MedBookVQA

Question :
What anatomical structure is involved in the surgical procedure shown in the image?
A. Upper nasal cavity
B. Right mandible
C. Inferior alveolar nerve
D. Left maxilla



GSPO (train after MedBookVQA):

The image shows a radiographic view of a dental or maxillofacial surgical field. The anatomical region highlighted by the surgical context — particularly the presence of bone, teeth, and surgical instruments — indicates that this is a maxillofacial area. The image appears to be a radiograph, likely a panoramic or periapical view, showing the left side of the maxilla (as indicated by the anatomical landmarks and surgical field). In many such cases, especially when dealing with surgical procedures involving the maxilla (e.g., osteotomy, implant placement, or tumor resection), the left maxilla is often the focus due to its anatomical prominence or surgical access. Thus, the correct answer is: D. Left maxilla \\boxed{D}. ✔

GSPO (train after FinMME):

Looking at the image, we see a cross-sectional view of a dental or facial region, with some small structures around the teeth. The question might be testing logical reasoning. Alternative answer might be phrasing Answer: Let me say Answer: B or C might be wrong Answer: Let me estimate Answer: A or B Answer: D might be the answer \\boxed{A} might be incorrect. Let me be precise. Answer: B. Answer: \\boxed{B} is not logical. Alternative answer might be different. Answer: \\boxed{C} is not correct. Answer: \\boxed{A} is answer. Answer: \\boxed{A} or \\boxed{B}. Answer: \\boxed{D} might be answer. Answer: \\boxed{A} is answer. Answer: \\boxed{A} (answer might be multiple choice). Answer: \\boxed{A} or \\boxed{B}. Answer: \\boxed{A} is answer. \\boxed{A} is answer. Answer: \\boxed{A} (based on logical reasoning). \\boxed{A}. ✗

CPO (train after FinMME):

Looking closely at the provided panoramic radiograph, we can observe the dental arches and bone structure. The image clearly shows the lower jaw (mandible) with teeth and implants, as well as the upper jaw (maxilla) with dental work. However, the question asks about the anatomical structure involved in the surgical procedure shown. The key feature in the image is the presence of dental implants and surgical hardware in the upper jaw, which is the maxilla. The term "Left maxilla" refers to the left side of the upper jaw, which is visible in the image. Given the options: A. Upper nasal cavity — not directly visible or involved in this surgical context. B. Right mandible — the mandible is the lower jaw; it's visible but not the primary focus here. C. Inferior alveolar nerve — a nerve, not a major surgical site in this context. D. Left maxilla — this is correct because the upper jaw (maxilla) is the area where the dental implants and surgical hardware are located, and the left side is clearly depicted. Therefore, the correct answer is D. Left maxilla. \\boxed{D} ✔