PROTECT WHAT IS AT RISK, DISTILL WHERE IT RESPONDS: ON-POLICY DISTILLATION FOR CONTINUAL LEARNING IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

On-policy distillation (OPD) can consolidate task experts for continual learning in multimodal large language models (MLLMs), but preserving experts does not ensure effective historical supervision. Changing visual evidence, instructions, and student responses expose historical capabilities unevenly, while the general OPD ignores which capabilities are being forgotten. We argue that **historical supervision must target vulnerable knowledge on trajectories that reveal it**. Guided by this principle, we propose **R**isk-guided **E**xpert **S**election via functional **P**robes for **ON**-policy **D**istillation (**RESPOND**), the first continual OPD framework to couple *what knowledge needs protection* with *where it can be distilled*. RESPOND probes current trajectories to supervise capabilities that are vulnerable in the student and observable through their experts, enabling consolidation with bounded historical-teacher queries per update. Experiments on eight sequential tasks and two OOD benchmarks show RESPOND improves final overall averages by 2.25–2.91 points over the strongest competing continual methods across three backbones. On Qwen3.5-4B, RESPOND reduces average forgetting by 50.2% and 69.7% relative to two vanilla continual OPD baselines.

1 INTRODUCTION

Continual learning enables deployed multimodal large language models (MLLMs) to acquire new visual domains, question types, and output formats while retaining earlier capabilities. Unlike repeated joint training, multimodal continual learning (MCL) must integrate a task stream without unrestricted access to historical data (Chen et al., 2024; Cao et al., 2024; Li, 2024).

Existing MCL methods rely on replay, task-specific components, or constrained updates (Zeng et al., 2024; Yu et al., 2025; Huai et al., 2025; Liang et al., 2025; Xie et al., 2026). In MLLMs, capabilities are expressed jointly through visual evidence, instructions, and generated responses. A science question may contain familiar objects yet scarcely exercise an earlier localization skill. **Expert availability therefore does not imply effective historical supervision.**

OPD supplies dense teacher supervision on student-generated trajectories (Agarwal et al., 2024), and self-distillation supports sequential skill acquisition with reduced forgetting (Shenfeld et al., 2026). We ask **how a growing expert library should supervise one student using only current-task inputs**.

We implement two vanilla continual OPD baselines. Replay-OPD consolidates independent task experts using a small per-task input memory, while Simple-OPD queries all experts on current inputs. Both transfer knowledge effectively, but uniform weighting dilutes protection for vulnerable capabilities as the library grows. Moreover, changing multimodal contexts expose historical skills unevenly. Exhaustive teacher queries increase computation without resolving this mismatch between *protection demand* and *distillation opportunity*.

We address this mismatch by proposing **R**isk-guided **E**xpert **S**election via functional **P**robes for **ON**-policy **D**istillation (RESPOND), **the first continual OPD framework to couple what knowledge needs protection with where it can be distilled**. *Protect what is at risk* prioritizes capabilities already drifting in the student or threatened by new-task updates. *Distill where it responds* iden-

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

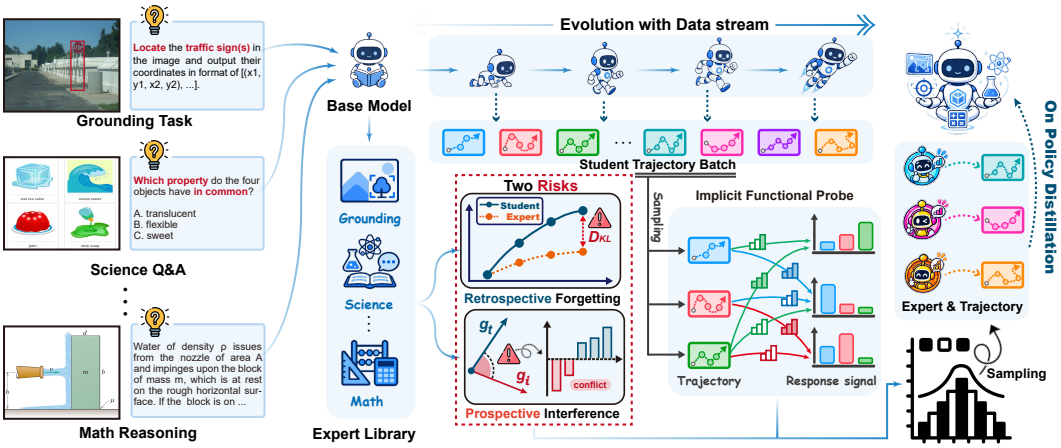

Figure 1: Overview of RESPOND. Forgetting risk identifies historical capabilities that need protection, while expert responses indicate where current student trajectories provide useful supervision. Together, they guide on-policy distillation to retain past knowledge while learning new tasks.

tifies current trajectories that reveal the corresponding experts' knowledge. Risk alone can select uninformative queries, while response alone can favor already-stable capabilities. Their coupling selects both historical teachers and where they teach (Figure 1), targeting vulnerable knowledge where supervision is available. Our contributions are threefold:

- To address the dilution of protection under uniform expert supervision in continual OPD, we track accumulated drift and potential interference to prioritize vulnerable historical knowledge as the expert library grows.
- To obtain useful historical supervision from current inputs, we introduce expert probes. Expert responses guide distillation toward trajectories that can help retain past knowledge.
- We integrate both risk tracking and expert probing into the first continual OPD framework. Experiments show that RESPOND reduces average forgetting by 50.2% and 69.7% relative to two vanilla continual OPD baselines.

2 RELATED WORK

Multimodal Continual Learning *Replay/prompt methods* preserve historical behavior through stored information or task context (Chen et al., 2024; Cao et al., 2024; Zeng et al., 2024). *Expert expansion/routing methods* organize task knowledge in LoRA branches and routing or fusion modules (Yu et al., 2025; Huai et al., 2025; Guo et al., 2025; Zhao et al., 2025; Ge et al., 2025; Liang et al., 2025). *Regularization/consolidation methods* constrain parameter or gradient updates to retain prior abilities (Chen et al., 2025; Li et al., 2026; Liu et al., 2026; Xie et al., 2026). Our approach retains independent experts as training-time functional memories and selectively distills them into a single inference model according to risk and response.

On-Policy Distillation OPD evaluates teachers on student-generated trajectories to align supervision with the student's visited states (Agarwal et al., 2024). Self-distillation strengthens the teacher through demonstrations in SDFT (Shenfeld et al., 2026), verified solutions in OPSD (Zhao et al., 2026), and environment feedback or sibling rollouts in SDPO (Hübotter et al., 2026). RLSD uses correctness to determine update direction and teacher signals to control magnitude (Yang et al., 2026). RLCSD contrasts correct and incorrect hints to emphasize task-relevant tokens (Pan et al., 2026). We extend continual OPD to a growing library of fixed historical experts, addressing which experts need protection and which current trajectories expose their useful supervision.

3 PRELIMINARIES

Multimodal Continual Learning A stream of T tasks $\{\mathcal{D}_i\}_{i=1}^{T}$ arrives sequentially (Chen et al., 2024; Cao et al., 2024; Liu et al., 2026; Xie et al., 2026), with $\mathcal{D}_i = \{(x_n, y_n)\}$, multimodal input $x_n = (I_n, q_n)$, and response y_n. The model factorizes $\pi_W(y \mid x) = \prod_\ell \pi_W(y_\ell \mid x, y_{<\ell})$. After stage t, a single task-agnostic model W_t is evaluated on all observed tasks. Learning accesses $\mathcal{D}_t$ and an optional bounded memory, while the desired joint objective averages $\mathbb{E}_{\mathcal{D}_i}[-\log \pi_W(y \mid x)]$ over $i \le t$. We separate acquisition from consolidation, i.e., each task trains an independent expert, whose function is then transferred into the unified student.

Replay-OPD Freezing the base model W_0, we train a LoRA branch e_i on task i to obtain $\pi_i = \pi_{W_0+e_i}$. For input x, the student samples $\hat{y} \sim \pi_W(\cdot \mid x)$ and the teacher evaluates its visited states $s_\ell = (x, \hat{y}_{<\ell})$ (Agarwal et al., 2024). Given an input collection $\mathcal{A}$, OPD minimizes

$$\mathcal{L}_{\mathrm{OPD}}(W, \pi_i, \mathcal{A}) = \mathbb{E}_{x\sim\mathcal{A},\, \hat{y}\sim\pi_W(\cdot|x)} \left[\frac{1}{|\hat{y}|} \sum_{\ell=1}^{|\hat{y}|} \ell_{\mathrm{OPD}}(\pi_W(\cdot \mid s_\ell), \pi_i(\cdot \mid s_\ell)) \right], \tag{1}$$

where ℓ_{OPD} is the adopted token-level distillation loss. Each update conditions on the sampled trajectories. Consolidation uses teacher distributions as supervision. With $\mathcal{X}_i = \{x : (x, y) \in \mathcal{D}_i\}$ and N retained inputs $\mathcal{R}_i \subset \mathcal{X}_i$ per task, Replay-OPD initializes from W_0 and optimizes

$$\mathcal{L}^{(t)}_{\mathrm{Replay}}(W) = \frac{1}{t} \sum_{i=1}^{t} \mathcal{L}_{\mathrm{OPD}}(W, \pi_i, \mathcal{R}_i), \qquad W^{\mathrm{Replay}}_{t,0} = W_0. \tag{2}$$

Independent experts preserve each task around the same base, while sparse inputs anchor their original distributions and OPD supplies token-level supervision. The resulting memory contains Nt inputs.

Simple-OPD The replay-free variant uses current inputs $\mathcal{X}_t$ to query every available expert and continues from the preceding consolidated model:

$$\mathcal{L}^{(t)}_{\mathrm{Simple}}(W) = \frac{1}{t} \sum_{i=1}^{t} \mathcal{L}_{\mathrm{OPD}}(W, \pi_i, \mathcal{X}_t), \qquad W^{\mathrm{Simple}}_{t,0} = W^{\mathrm{Simple}}_{t-1}, \tag{3}$$

with $W^{\mathrm{Simple}}_0 = W_0$. The current expert transfers the new task, and historical experts provide retention supervision on student trajectories. This establishes current inputs as a useful source for replay-free consolidation. As the library expands, uniform $1/t$ weighting dilutes vulnerable experts, complete teacher evaluation grows linearly, and historical responses vary across inputs. **These observations motivate allocating queries jointly by forgetting risk and functional response.**

4 METHODOLOGY

4.1 OVERVIEW

RESPOND couples protection demand with distillation opportunity. At stage t, the student starts from $\theta_0 = W_{t-1}$, with W_0 the pretrained base. For a current batch $\mathcal{C}_k = \{x_n\}_{n=1}^{B} \subseteq \mathcal{X}_t$, it generates

$$\hat{y}_n \sim \pi_{\theta_k}(\cdot \mid x_n), \qquad s^{(k)}_{n,\ell} = (x_n, \hat{y}_{n,<\ell}), \qquad \mathcal{S}^{(k)}_n = \{s^{(k)}_{n,\ell}\}_{\ell=1}^{\hat{L}_n}. \tag{4}$$

Uniform supervision can neglect vulnerable capabilities while spending queries on stable ones. Section 4.2 therefore tracks accumulated drift and incoming interference to identify *what needs protection*. Risk alone, however, does not ensure useful supervision on current multimodal inputs. Section 4.3 identifies *where the corresponding experts respond*, so historical distillation targets informative student trajectories. Together, cached response scores and pre-step risk states select up to three historical experts, each supervising its own probes. The current expert covers the full batch to support new-task acquisition. All teachers reuse the student rollouts, and the observed teacher signals update risk states for the next step.

4.2 RISK-GUIDED EXPERT SELECTION

Risk combines evidence of existing drift with warnings of incoming interference. Absolute student–expert disagreement on current inputs may reflect a pre-existing mismatch, so we measure its growth from a stage-start reference. Let $\mathcal{Q}_i^{(k)}$ contain expert i's probe states and write $\delta_i(s;\theta) = D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s)\|\pi_i(\cdot \mid s))$. Gradient-free calibration batches define the reference b_i, while a query reveals the current distance $d_i^{(k)}$:

$$b_i = \mathbb{E}_{s\sim\mathcal{Q}_i^{\mathrm{ref}}}\delta_i(s;\theta_0), \qquad d_i^{(k)} = \frac{1}{|\mathcal{Q}_i^{(k)}|}\sum_{s\in\mathcal{Q}_i^{(k)}}\delta_i(s;\theta_k). \tag{5}$$

For the stored KL average $\bar{d}_i^{(k)}$, the candidate update and retrospective risk are

$$\bar{d}_{i,+}^{(k)} = \beta_d\bar{d}_i^{(k)} + (1-\beta_d)d_i^{(k)}, \qquad R_{i,\mathrm{ret}}^{(k)} = \frac{[\bar{d}_{i,+}^{(k)} - b_i]_+}{b_i+\epsilon}, \tag{6}$$

where $[z]_+ = \max(z,0)$ and $0 \le \beta_d < 1$. Queried experts commit the candidate average, while others retain their previous value. This term measures reference-relative drift on current probes.

Drift alone reacts only after disagreement grows. To anticipate harmful new-task updates, let $g_t^{(k)}$ and $g_i^{(k)}$ be the fixed-rollout gradients of the current and historical objectives. A current-only step changes historical loss by $-\eta\langle g_i^{(k)}, g_t^{(k)}\rangle + O(\eta^2\|g_t^{(k)}\|^2)$, motivating protection against negative gradient alignment. To avoid expert-specific backward passes, we use stopped sampled-token log ratios as a conflict proxy:

$$\xi_{j,a}^{(k)} = \mathrm{sg}[\log\pi_j(\hat{y}_a \mid s_a) - \log\pi_{\theta_k}(\hat{y}_a \mid s_a)]\,. \tag{7}$$

Here a indexes token occurrences and sg stops gradients. The current vector $\boldsymbol{\xi}_t^{(k)}$ covers the full batch, while $\boldsymbol{\xi}_i^{(k)}$ is zero outside $\mathcal{Q}_i^{(k)}$. Prospective risk combines negative cosine alignment with historical signal strength:

$$R_{i,\mathrm{pro}}^{(k)} = \frac{\|\boldsymbol{\xi}_i^{(k)}\|_2}{\sqrt{|\mathcal{Q}_i^{(k)}|}}\left[-\frac{\langle\boldsymbol{\xi}_i^{(k)}, \boldsymbol{\xi}_t^{(k)}\rangle}{\|\boldsymbol{\xi}_i^{(k)}\|_2\|\boldsymbol{\xi}_t^{(k)}\|_2+\epsilon}\right]_+. \tag{8}$$

The root-mean-square factor attenuates negligible teacher signals. Using pre-step running scales $\nu_r^{(k)}$ initialized at $\nu_r^{(0)} = 0$, we set $\widetilde{R}_{i,r}^{(k)} = R_{i,r}^{(k)}/(\nu_r^{(k)}+\epsilon)$ and combine the risks as $\varrho_i^{(k)} = \widetilde{R}_{i,\mathrm{ret}}^{(k)} + \lambda_g\widetilde{R}_{i,\mathrm{pro}}^{(k)}$. Only queried experts reveal these signals. For subset S_k and inclusion probability $p_i^{(k)}$, we update a persistent risk state:

$$\widehat{\varrho}_i^{(k)} = \frac{\mathbb{I}[i\in S_k]}{\max(p_i^{(k)}, p_{\min})}\,\mathrm{clip}(\varrho_i^{(k)}, 0, \varrho_{\max}), \qquad \zeta_i^{(k+1)} = \gamma\zeta_i^{(k)} + (1-\gamma)\widehat{\varrho}_i^{(k)}. \tag{9}$$

Inverse-probability weighting adjusts for unequal query frequencies, while clipping, flooring, and smoothing stabilize observations. Selection at step k uses $\zeta_i^{(k)}$, and newly observed risks inform the next step through $\zeta_i^{(k+1)}$.

Protection demand becomes actionable only where an expert supplies a response. Combining stored risk with the probe response $\bar{o}_i^{(k)}$ from Section 4.3 gives

$$u_i^{(k)} = (\zeta_i^{(k)} + \rho_0)\bar{o}_i^{(k)}, \qquad U_k(S) = \sum_{i\in S}u_i^{(k)} - \lambda_c|S|, \tag{10}$$

where $\rho_0 > 0$ supplies a risk prior and $\lambda_c \ge 0$ prices teacher use. At $t > 1$, let

$$\Omega_t = \{S \subseteq \{1,\dots,t-1\} : 1 \le |S| \le K_{\max}\}, \qquad K_{\max} = \min(3, t-1). \tag{11}$$

We sample a cost-aware subset from

$$\Pi_k(S) = (1-\varepsilon_{\exp})\frac{\exp(U_k(S)/\tau_s)}{\sum_{S'\in\Omega_t}\exp(U_k(S')/\tau_s)} + \frac{\varepsilon_{\exp}}{|\Omega_t|}, \qquad S_k \sim \Pi_k, \tag{12}$$

with $\tau_s > 0$, $0 < \varepsilon_{\exp} \le 1$, and exact marginal $p_i^{(k)} = \sum_{S\in\Omega_t: i\in S}\Pi_k(S)$. The cost term discourages unnecessary queries, while uniform exploration keeps every historical expert eligible for renewed risk observation.

Table 1: Final-stage comparison using **Qwen3.5-4B** on eight training tasks and two out-of-distribution (OOD) benchmarks. Among continual-learning and OPD/OPSD methods, the **best** and second-best results are marked.

Methods	Training Tasks									OOD Tasks		Overall Avg.
	COCO	Sci.QA	MSVQA	MMMU	ICON	GQA	MATH	CUB	**Avg.**	GeoQA	MME	
Reference and naive baselines												
Vanilla	73.32	87.00	51.67	54.00	86.00	63.40	73.40	79.20	71.00	91.70	83.00	74.27
Joint	75.28	96.20	66.28	51.20	97.60	71.00	93.20	91.60	80.29	65.00	84.70	79.21
Continual learning methods												
EWC	67.53	81.80	40.12	49.20	79.20	63.40	84.20	90.80	69.53	68.00	87.90	71.22
GainLoRA	74.73	87.40	54.84	54.68	90.60	69.80	93.80	89.60	76.93	87.80	83.90	78.71
CL-MOE	72.10	92.00	57.76	54.00	90.20	67.50	92.60	**92.00**	77.27	74.40	88.10	78.07
Octopus	72.90	85.80	53.62	59.00	86.00	63.70	80.80	87.40	73.65	85.70	85.50	76.04
DGG	72.16	86.80	55.45	57.80	86.00	63.50	81.00	88.80	73.94	86.90	86.80	76.52
SAME	75.61	93.50	56.84	56.80	92.44	67.85	92.80	89.80	78.21	61.80	86.00	77.34
OPD/OPSD methods												
OPSD	71.68	88.20	54.25	56.00	88.20	62.70	74.80	81.20	72.13	89.00	84.90	75.09
GRPO	71.93	87.40	54.35	58.60	89.60	64.40	91.60	82.00	74.98	91.50	84.60	77.60
SDPO	25.22	78.20	39.79	37.33	81.80	55.50	41.80	75.60	54.40	45.80	73.40	55.44
SDFT	71.51	86.70	55.68	56.40	87.50	69.50	86.40	84.60	74.79	78.10	**88.20**	76.46
RLSD	74.69	87.80	54.46	58.70	87.40	63.40	93.00	83.40	75.36	**92.70**	87.30	78.28
RLCSD	75.04	91.00	53.38	**61.90**	88.20	66.60	93.80	85.60	76.94	91.60	87.10	79.42
RESPOND	**76.45**	**96.00**	**62.35**	56.60	**93.80**	**70.70**	**94.20**	91.20	**80.16**	87.20	**88.20**	**81.67**

4.3 FUNCTIONAL PROBES FOR ON-POLICY DISTILLATION

Risk identifies a need for protection, but useful supervision depends on the trajectory. Current multimodal inputs need not exercise every historical skill. We therefore seek expert-specific probes that expose historical policy behavior. Let θ_i be expert i in the student's parameter space, $\Delta\theta_i = \theta - \theta_i$, and $\mathbb{P}_i$ and $\mathbb{Q}_i$ the historical and probe state distributions. For fixed state distributions, local policy deviation satisfies

$$\mathfrak{D}_i(\theta;\mathbb{R}) := \mathbb{E}_{s\sim\mathbb{R}}\delta_i(s;\theta) = \tfrac{1}{2}\Delta\theta_i^\top \mathbf{F}_i(\mathbb{R})\Delta\theta_i + O(\|\Delta\theta_i\|^3). \tag{13}$$

Here $\mathbf{F}_i(\mathbb{R})$ is the policy Fisher matrix under $\mathbb{R}$. If $\mathbf{F}_i(\mathbb{Q}_i) \succeq \mu_i \mathbf{F}_i(\mathbb{P}_i)$ for $\mu_i > 0$, then $\mathfrak{D}_i(\theta;\mathbb{P}_i) \le \mathfrak{D}_i(\theta;\mathbb{Q}_i)/\mu_i + O(\|\Delta\theta_i\|^3)$. This local coverage condition motivates probes that activate directions important to historical behavior, allowing probe supervision to constrain historical deviation.

Evaluating every full teacher to find such probes would defeat selective querying. Instead, we measure adapter responses using cached student activations. For LoRA module m, write $\Delta W_i^{(m)} = \sigma_i^{(m)} B_i^{(m)} A_i^{(m)}$ and precompute $\Gamma_i^{(m)} = (B_i^{(m)})^\top B_i^{(m)}$. From cached inputs $h_{n,\ell}^{(m,k)}$ and outputs $v_{n,\ell}^{(m,k)}$, compute $z_{i,n,\ell}^{(m,k)} = A_i^{(m)} h_{n,\ell}^{(m,k)}$ and the normalized response energy

$$E_{i,n,\ell}^{(m,k)} = \frac{(\sigma_i^{(m)})^2 (z_{i,n,\ell}^{(m,k)})^\top \Gamma_i^{(m)} z_{i,n,\ell}^{(m,k)}}{\|v_{n,\ell}^{(m,k)}\|_2^2 + \epsilon}. \tag{14}$$

The numerator equals $\|\Delta W_i^{(m)} h_{n,\ell}^{(m,k)}\|_2^2$, enabling rank-space evaluation. Averaging across probe modules $\mathcal{M}$ and valid response tokens gives

$$\bar{E}_{i,n}^{(k)} = \frac{1}{|\mathcal{M}|\hat{L}_n} \sum_{m\in\mathcal{M}} \sum_{\ell=1}^{\hat{L}_n} E_{i,n,\ell}^{(m,k)}, \qquad o_{i,n}^{(k)} = 1 - \exp(-\bar{E}_{i,n}^{(k)}/\kappa), \quad \kappa > 0. \tag{15}$$

This bounded score preserves energy rankings. Under local perturbations and nondegenerate downstream curvature, response energy is comparable to the policy KL induced by injecting the adapter response, motivating its use as a pre-query response proxy.

Each expert selects its N_p highest-response samples, with $1 \le N_p \le B$:

$$\mathcal{C}_i^{\star(k)} = \mathrm{TopK}_{x_n\in\mathcal{C}_k}(o_{i,n}^{(k)}, N_p), \qquad \bar{o}_i^{(k)} = \frac{1}{N_p} \sum_{x_n\in\mathcal{C}_i^{\star(k)}} o_{i,n}^{(k)}. \tag{16}$$

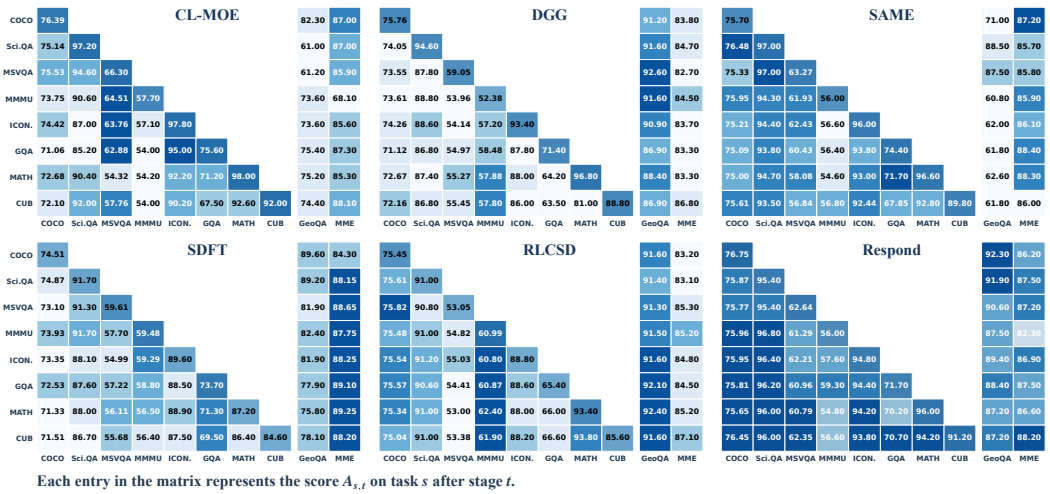

Figure 2: Stage-wise Qwen3.5-4B results over the eight-task stream. Each lower-triangular entry $A_{s,t}$ reports task-s performance after stage t. Blank cells mark unseen tasks, and darker cells indicate higher scores. GeoQA-CoT and MME track OOD and general multimodal performance. RESPOND retains prior tasks while remaining competitive on OOD benchmarks.

The mean response supplies the opportunity term in Eq. 10. The selected token occurrences form $\mathcal{Q}_i^{(k)} = \biguplus_{x_n \in \mathcal{C}_i^{\star(k)}} \mathcal{S}_n^{(k)}$. Let $\widehat{\mathcal{L}}_{\mathrm{OPD}}^{(k)}$ denote the token-averaged form of Eq. 1 on cached rollouts, normalizing each teacher by its own number of valid response tokens. The current and historical losses are

$$\mathcal{L}_{\mathrm{cur}}^{(k)}(\theta) = \widehat{\mathcal{L}}_{\mathrm{OPD}}^{(k)}(\theta, \pi_t, \mathcal{C}_k), \qquad \mathcal{L}_i^{(k)}(\theta) = \widehat{\mathcal{L}}_{\mathrm{OPD}}^{(k)}(\theta, \pi_i, \mathcal{C}_i^{\star(k)}). \tag{17}$$

For $i \in S_k$, set $w_i^{(k)} = \exp(u_i^{(k)}/\tau_w) / \sum_{j \in S_k} \exp(u_j^{(k)}/\tau_w)$, with $\tau_w > 0$. Holding rollouts and routing weights fixed, we optimize

$$\mathcal{L}^{(k)}(\theta) = (1 - \lambda_{\mathrm{old}})\mathcal{L}_{\mathrm{cur}}^{(k)}(\theta) + \lambda_{\mathrm{old}} \sum_{i \in S_k} w_i^{(k)} \mathcal{L}_i^{(k)}(\theta), \quad 0 \le \lambda_{\mathrm{old}} \le 1. \tag{18}$$

The total historical weight remains λ_{old} regardless of subset size. Risk and response thus determine where protection is allocated without diluting its overall weight. Stage $t = 1$ uses current-expert OPD alone. Later stages alternate selective distillation with risk updates, querying at most three historical teachers per normal step. The consolidated W_t remains a single task-agnostic inference model.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETTINGS

Datasets As modern MLLMs have become increasingly capable, several commonly used continual-learning benchmarks no longer provide reliable learning signals. *Training on some datasets can even reduce performance on their own test sets.* We therefore construct a new continual learning pipeline in which the model can consistently benefit from each incoming task. The stream consists of COCO2017 (Lin et al., 2014), ScienceQA (Lu et al., 2022), MSVQA (Jiang et al., 2025), MMMU (Yue et al., 2024), ICON-QA (Lu et al., 2021), GQA (Hudson & Manning, 2019), ClevrMATH (Lindström & Abraham, 2022), and CUB (Wah et al., 2011), covering general visual understanding, fine-grained recognition, scientific and mathematical reasoning, knowledge-intensive VQA, and compositional reasoning. GeoQA-CoT (Chen et al., 2021) and MME (Fu et al., 2025) are used only for evaluation to monitor changes in the general capabilities inherited from the base model. Further details are provided in the appendix.

Comparison Methods We compare RESPOND with twelve representative baselines from two groups. The continual learning group includes EWC (Kirkpatrick et al., 2017),

Table 2: Final-stage comparison using **Qwen3-VL-4B** on eight training tasks and two out-of-distribution (OOD) benchmarks. Among continual-learning and OPD/OPSD methods, the **best** and second-best results are marked.

Methods	Training Tasks									OOD Tasks		Overall Avg.
	COCO	Sci.QA	MSVQA	MMMU	ICON	GQA	MATH	CUB	**Avg.**	GeoQA	MME	
Reference and naive baselines												
Vanilla	73.68	88.40	56.34	49.64	80.60	68.50	75.80	82.00	71.87	73.40	84.80	73.32
Joint	74.69	97.00	67.51	53.34	93.80	70.14	94.80	92.00	80.41	65.00	87.30	79.56
Continual learning methods												
EWC	70.26	43.60	42.77	44.60	80.26	52.74	88.40	91.00	64.20	60.60	67.80	64.20
GainLoRA	73.72	88.78	57.82	50.10	80.30	67.93	77.40	85.40	72.68	74.66	86.80	74.29
CL-MOE	69.98	74.50	51.95	50.79	85.30	70.32	93.60	**91.60**	73.51	57.90	**89.50**	73.54
Octopus	72.95	88.20	54.22	48.64	79.80	68.20	69.60	88.40	71.25	72.80	85.10	72.79
DGG	70.40	88.60	54.67	48.53	80.60	65.12	69.40	88.80	70.76	73.03	85.60	72.47
SAME	74.72	82.80	60.76	54.26	89.10	**72.40**	92.60	87.80	76.81	66.20	87.90	76.85
OPD/OPSD methods												
OPSD	73.76	88.00	54.80	49.92	79.55	65.80	80.50	84.27	72.08	76.35	86.24	73.92
GRPO	75.02	90.60	56.54	51.61	80.60	67.12	83.40	85.80	73.84	75.82	87.00	75.35
SDPO	74.89	90.40	52.62	50.30	64.80	66.64	91.80	85.20	72.08	66.03	83.80	72.65
SDFT	60.19	83.12	56.42	**55.82**	84.70	61.66	72.80	76.90	68.95	**77.78**	84.36	71.38
RLSD	74.95	89.00	56.85	50.35	80.66	68.79	90.12	85.35	74.51	75.44	85.35	75.69
RLCSD	75.10	90.40	57.83	49.04	81.20	68.36	**95.00**	86.60	75.44	76.82	87.00	76.73
RESPOND	**75.94**	**94.80**	**62.80**	54.78	**90.40**	70.78	94.80	90.00	**79.29**	75.00	88.30	**79.76**

CL-MOE (Huai et al., 2025), GainLoRA (Liang et al., 2025), Octopus (Liu et al., 2026), DGG (Li et al., 2026), and SAME (Xie et al., 2026), while the OPD/OPSD-related group includes GRPO (Shao et al., 2024), SDPO (Hübotter et al., 2026), OPSD (Zhao et al., 2026), SDFT (Shenfeld et al., 2026), RLSD (Yang et al., 2026), and RLCSD (Pan et al., 2026). Detailed descriptions and configurations are provided in the appendix.

Implementation Details Experiments use four NVIDIA H100 GPUs, with the same backbone and task order across methods. We extend vanilla KD loss in the KDFlow framework (Zhang et al., 2026)[1] for selective multi-teacher consolidation (Appendix C). Main results use Qwen3.5-4B and Qwen3-VL-4B. InternVL3.5-4B results appear in Appendix M. Each task uses one epoch, batch size 2, eight gradient accumulation steps, and a learning rate of 1×10^{-5}.

Evaluation Metrics Let $A_{s,t}$ denote the performance on task s after training up to task t, with T tasks in total. We report the final score $A_{s,T}$ for each task and the average final performance $\bar{A} = \frac{1}{T}\sum_{s=1}^{T} A_{s,T}$. For ablation studies, we report the average forgetting $\bar{F} = \frac{1}{T-1}\sum_{s=1}^{T-1}\left(\max_{t\in\{s,\ldots,T\}} A_{s,t} - A_{s,T}\right)$. The last task is excluded because it is learned at the final stage and has not undergone forgetting caused by subsequent tasks.

5.2 Main Results

Overall performance Tables 1 and 2 show that RESPOND leads the compared continual-learning and OPD/OPSD methods in both final training-task and overall averages. Qwen3.5-4B reaches 80.16 and 81.67, exceeding the strongest competing scores by 1.95 and 2.25 points. Qwen3-VL-4B reaches 79.29 and 79.76, improving over SAME by 2.48 and 2.91 points. Relative to RLCSD, the strongest OPD/OPSD baseline in training-task average, the gains are 3.22 and 3.85 points. **Selective historical supervision therefore strengthens consolidation across both backbones.**

The gains span heterogeneous capabilities. **RESPOND leads on COCO, ScienceQA, MSVQA, and ICON-QA with both backbones.** ScienceQA scores of 96.00 and 94.80 exceed the strongest competing methods by 2.50 and 4.20 points, respectively. With Qwen3.5-4B, the training-task average is within 0.13 points of Joint and the overall average is 2.46 points higher. MME also improves over the vanilla models from 83.00 to 88.20 and from 84.80 to 88.30, combining task acquisition with strong general capabilities.

[1] https://github.com/songmzhang/KDFlow

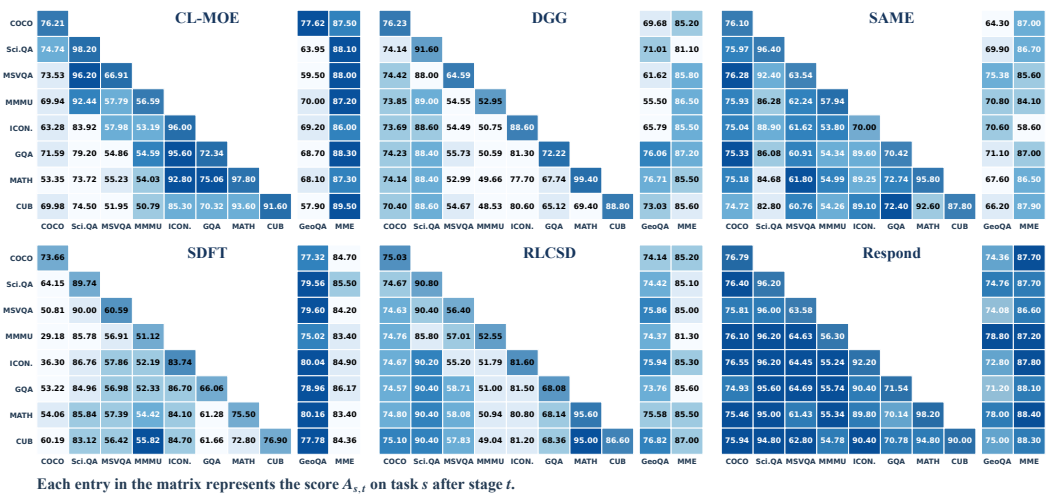

Figure 3: Stage-wise Qwen3-VL-4B results over the eight-task stream. Each lower-triangular entry $A_{s,t}$ reports task-s performance after stage t. Blank cells mark unseen tasks, and darker cells indicate higher scores. GeoQA-CoT and MME track OOD and general multimodal performance. RESPOND retains prior tasks while remaining competitive on OOD benchmarks.

The OOD averages reinforce this balance. Averaging GeoQA-CoT and MME gives 87.70 on Qwen3.5-4B, compared with 74.85 for Joint and 73.90 for SAME. With Qwen3-VL-4B, RESPOND reaches 81.65, exceeding the vanilla model's 79.10 and SAME's 77.05. **Its leading training-task averages therefore coexist with strong performance outside the training stream**, supporting selective consolidation as a means of integrating task-specific knowledge while retaining general multimodal competence.

The gains also reflect new knowledge acquisition. The final training-task average rises from 71.00 to 80.16 on Qwen3.5-4B and from 71.87 to 79.29 on Qwen3-VL-4B. MSVQA improves by 10.68 and 6.46 points, respectively, and ClevrMATH by 20.80 and 19.00 points, showing acquisition across domain-specific understanding and quantitative reasoning.

Stage-wise retention Figures 2 and 3 **show sustained retention over eight tasks.** With Qwen3.5-4B, COCO changes from 76.75 after acquisition to 76.45 at the final stage, ScienceQA improves from 95.40 to 96.00, and MSVQA loses only 0.29 points, compared with 8.55 for CL-MOE and 6.43 for SAME. With Qwen3-VL-4B, ScienceQA loses 1.40 points, compared with 23.70 and 13.60 for CL-MOE and SAME. On the same backbone, COCO loses 0.85 points across seven subsequent tasks.

Later acquisitions exhibit the same stability. With Qwen3.5-4B, ICON-QA changes from 94.80 to 93.80, GQA from 71.70 to 70.70, and ClevrMATH from 96.00 to 94.20 between acquisition and the final stage. With Qwen3-VL-4B, ICON-QA and GQA lose 1.80 and 0.76 points, while ClevrMATH retains 94.80 after reaching 98.20 when learned. **The heatmaps yield average forgetting of 1.13 and 1.66 on the two backbones**, demonstrating effective retention across early and later tasks alongside continued acquisition.

The early-task results test retention over the longest horizons, while final ClevrMATH scores of 94.20 and 94.80 and CUB scores of 91.20 and 90.00 confirm continued late-task acquisition. **Risk–response allocation therefore supports retention alongside incoming skills.**

5.3 ABLATION STUDY AND ANALYSIS

Ablations Table 3 shows that RESPOND achieves the best final training-task average of 80.16, gaining 2.87 and 4.47 points over Replay-OPD and Simple-OPD. Restoring risk guidance adds 3.24 points and reduces forgetting from 3.44 to 1.13 (67.2%). Restoring probes adds 3.34 points and reduces forgetting from 2.86 to 1.13 (60.5%). **Both decisions improve acquisition and retention.**

Table 3: Component ablations of RESPOND. Checkmarks denote enabled components. Final and OOD report average scores.

Method	Ret.	Pro.	Probe	Final ↑	OOD ↑	$\bar{F}$ ↓
Replay-OPD				77.29	86.70	3.73
Simple-OPD				75.69	85.05	2.27
w/o Risk			✓	76.92	84.80	3.44
w/o Probe	✓	✓		76.82	87.94	2.86
w/o Pro. Risk	✓		✓	78.31	87.59	2.51
w/o Ret. Risk		✓	✓	79.04	86.92	2.18
RESPOND	✓	✓	✓	80.16	87.70	1.13

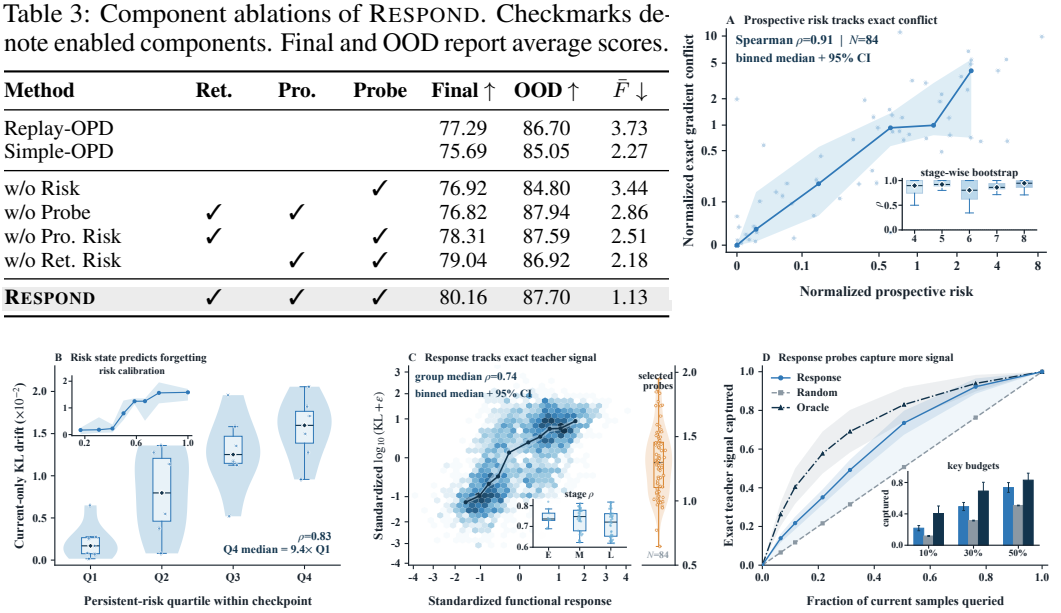

Figure 4: Risk-guided protection and response-guided probing on Qwen3.5-4B. **(A)** Prospective risk versus exact gradient conflict, with stage-wise correlations in the inset. **(B)** Current-only KL drift across within-checkpoint persistent-risk quartiles, from lowest (Q1) to highest (Q4), with risk calibration in the inset. **(C)** Functional response versus exact teacher KL. The right-hand distribution summarizes selected probes, and the inset compares early, middle, and late stages. Curves and shaded bands in A and C show binned medians and 95% confidence intervals. **(D)** Fraction of teacher KL captured versus the fraction of current samples queried. Response ranks adapter responses, Random samples uniformly, and Oracle ranks exact teacher KL. Risk identifies vulnerable capabilities, while response concentrates queries on informative trajectories.

Both risk terms contribute to retention. Removing prospective or retrospective risk lowers the final score to 78.31 or 79.04 and increases forgetting to 2.51 or 2.18, respectively. Anticipating incoming interference therefore complements tracking accumulated drift. The full method achieves the strongest final score and lowest forgetting among component ablations, with an OOD average of 87.70. Relative to Replay-OPD, it reduces forgetting by 69.7% while improving final performance by 2.87 points, demonstrating stronger consolidation without replay.

Visualizations Figure 4 tests whether the two allocation signals identify protection demand and useful supervision. Prospective risk tracks exact gradient conflict ($\rho = 0.91$, $N = 84$), while persistent risk correlates with current-only KL drift ($\rho = 0.83$), whose median is $9.4\times$ higher in Q4 than Q1. These diagnostics support anticipating interference, complementing the measured retention gains in Table 3. Response also ranks full-teacher signal, with median within-group $\rho = 0.74$. At a 50% sample-query budget, response selection captures approximately 74% of teacher KL versus 51% for random selection, a 23-point gain. At 75%, it approaches Oracle (92% versus 94%). Together, the results support **using risk to prioritize protection and response to spend queries where supervision is strongest**.

6 CONCLUSION

We introduced RESPOND, a replay-free framework for continual on-policy distillation in MLLMs that couples expert-specific forgetting risk with trajectory-specific functional response. The method concentrates historical supervision on vulnerable capabilities and informative current trajectories, transferring independent task knowledge into a single task-agnostic model under a bounded historical-teacher query budget. Experiments across three backbones show leading final performance, sustained retention over eight tasks, and strong general capabilities. Ablations and mechanism visualizations establish the complementary value of risk tracking and functional probing for selective continual multimodal knowledge consolidation.

AI USE STATEMENT

Generative AI tools were used to assist with polishing and translating portions of the manuscript and drawing selected elements in figures. They also assisted with modifying and editing code for the pipeline framework, adapting the implementation to additional backbone networks, and searching for relevant literature. All AI-assisted outputs were reviewed by the authors before inclusion. Suggested references were checked against their original sources, and AI-assisted code changes were inspected and evaluated within the corresponding research pipelines. Beyond the uses listed above, generative AI tools were not used for other applicable research tasks requiring disclosure under the ICLR 2027 policy. The authors take responsibility for the final content of this work, including all text, claims, figures, and software artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, Section 4 specifies the risk-guided expert selection, functional probes, and distillation objective of RESPOND. Section 5.1 reports the backbone models, hardware, training settings, and evaluation metrics. Appendix I documents the task order, training and evaluation set sizes, and task characteristics, while Appendix J describes the comparison methods. Appendix K provides the hyperparameters, calibration and state initialization, risk updates, and LoRA and probe configurations. The evaluation workflow, evaluator prompt and model, and answer and format scoring rules are detailed in Appendix L. Supporting theoretical assumptions and derivations are provided in Appendices B–H.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, volume 2024, pp. 21246–21263, 2024.

Meng Cao, Yuyang Liu, Yingfei Liu, Tiancai Wang, Jiahua Dong, Henghui Ding, Xiangyu Zhang, Ian Reid, and Xiaodan Liang. Continual llava: Continual instruction tuning in large vision-language models. *arXiv preprint arXiv:2411.02564*, 2024.

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

Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 513–523. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-acl.46.

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

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 162549–162567, 2025. doi: 10.52202/085713-4899.

Chendi Ge, Xin Wang, Zeyang Zhang, Hong Chen, Jiapei Fan, Longtao Huang, Hui Xue, and Wenwu Zhu. Dynamic mixture of curriculum lora experts for continual multimodal instruction tuning. *arXiv preprint arXiv:2506.11672*, 2025.

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

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

Jonas Hübotter, Frederike Lübeck, Lejs Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. Reinforcement learning via self-distillation. In *International Conference on Machine Learning*, 2026.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Kai Jiang, Siqi Huang, Xiangyu Chen, Jiawei Shao, Hongyuan Zhang, and Xuelong Li. Multimodal continual learning with MLLMs from multi-scenario perspectives. *arXiv preprint arXiv:2511.18507*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. doi: 10.1073/pnas.1611835114.

Songze Li, Mingyu Gao, Tonghua Su, Xu-Yao Zhang, and Zhongjie Wang. Multimodal continual instruction tuning with dynamic gradient guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10820–10829, 2026.

Xuelong Li. Positive-incentive noise. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6):8708–8714, 2024. doi: 10.1109/TNNLS.2022.3224577.

Yan-Shuo Liang, Jia-Rui Chen, and Wu-Jun Li. Gated integration of low-rank adaptation for continual learning of large language models. *Advances in Neural Information Processing Systems*, 38:76577–76607, 2025. doi: 10.52202/085713-2310.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48.

Adam Dahlgren Lindström and Savitha Sam Abraham. CLEVR-math: A dataset for compositional language, visual and mathematical reasoning. In *Proceedings of the 16th International Workshop on Neural-Symbolic Learning and Reasoning*, volume 3212 of *CEUR Workshop Proceedings*, pp. 155–170, 2022.

Yuehao Liu, Shanyan Guan, Weijia Zhang, Xuanming Shang, Yanhao Ge, Wei Li, and Chao Ma. Octopus: History-free gradient orthogonalization for continual learning in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3898–3907, 2026.

Pan Lu, Liang Qiu, Jiaqi Chen, Tanglin Xia, Yizhou Zhao, Wei Zhang, Zhou Yu, Xiaodan Liang, and Song-Chun Zhu. IconQA: A new benchmark for abstract diagram understanding and visual language reasoning. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2507–2521, 2022.

Leyi Pan, Shuchang Tao, Yunpeng Zhai, Lingzhe Zhang, Zhaoyang Liu, Bolin Ding, Aiwei Liu, and Lijie Wen. Rlcsd: Reinforcement learning with contrastive on-policy self-distillation. *arXiv preprint arXiv:2606.11709*, 2026.

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

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. Self-distillation enables continual learning. In *International Conference on Machine Learning*, 2026.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-UCSD birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Zhen-Hao Xie, Jun-Tao Tang, Yu-Cheng Shi, Han-Jia Ye, De-Chuan Zhan, and Da-Wei Zhou. Same: Stabilized mixture-of-experts for multimodal continual instruction tuning. In *ICML*, 2026.

Chenxu Yang, Chuanyu Qin, Qingyi Si, Minghui Chen, Naibin Gu, Dingyu Yao, Zheng Lin, Weiping Wang, Jiaqi Wang, and Nan Duan. Self-distilled rlvr. *arXiv preprint arXiv:2604.03128*, 2026.

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

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Fanhu Zeng, Fei Zhu, Haiyang Guo, Xu-Yao Zhang, and Cheng-Lin Liu. Modalprompt: Towards efficient multimodal continual instruction tuning with dual-modality guided prompt. *arXiv preprint arXiv:2410.05849*, 2024.

Songming Zhang, Xue Zhang, Tong Zhang, Bojie Hu, Yufeng Chen, and Jinan Xu. KDFlow: A user-friendly and efficient knowledge distillation framework for large language models. *arXiv preprint arXiv:2603.01875*, 2026.

Hengyuan Zhao, Ziqin Wang, Qixin Sun, Kaiyou Song, Yilin Li, Xiaolin Hu, Qingpei Guo, and Si Liu. Llava-cmoe: Towards continual mixture of experts for large vision-language models. *arXiv preprint arXiv:2503.21227*, 2025.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. Self-distilled reasoner: On-policy self-distillation for large language models. In *International Conference on Machine Learning*, 2026.

A Overview

This appendix develops the analysis used by RESPOND. Appendix B gives the first-order interference criterion. Appendix C connects it to the sampled-token log-ratio proxy with the same sample and token weights as the OPD objective. Appendices D and E specify the temporal updates and analyze partial risk observation. Appendix F proves the local coverage bound. Appendices G and H derive the rank-space identity and its relation to locally induced policy KL. Appendix M reports additional experiments on InternVL3.5-4B.

At a fixed stage and step, student trajectories, probe indices, teachers, and routing weights are held fixed when differentiating the empirical losses. Token states are indexed occurrences, including repeated states. All risk and historical-routing results concern $t > 1$. Stage $t = 1$ uses current-expert OPD alone.

B First-Order Effect of the Current-Expert Update

Fix stage $t > 1$ and step k. On the cached student rollouts, the empirical losses in Eq. 17 are functions of the optimization variable θ. Define

$$g_t^{(k)} = \nabla_\theta \mathcal{L}_{\mathrm{cur}}^{(k)}(\theta)\Big|_{\theta_k}, \qquad g_i^{(k)} = \nabla_\theta \mathcal{L}_i^{(k)}(\theta)\Big|_{\theta_k}. \tag{B.1}$$

A hypothetical current-only gradient step is

$$\theta_k^+ = \theta_k - \eta g_t^{(k)}, \qquad \eta > 0. \tag{B.2}$$

If $\mathcal{L}_i^{(k)}$ has Hessian operator norm at most L_i along this segment, Taylor's theorem gives

$$\begin{aligned} \mathcal{L}_i^{(k)}(\theta_k^+) - \mathcal{L}_i^{(k)}(\theta_k) &= -\eta \langle g_i^{(k)}, g_t^{(k)} \rangle + r_{i,k}, \\ |r_{i,k}| &\le \tfrac{1}{2} L_i \eta^2 \|g_t^{(k)}\|_2^2. \end{aligned} \tag{B.3}$$

Because $z \mapsto [z]_+$ is 1-Lipschitz,

$$\left| \left[\mathcal{L}_i^{(k)}(\theta_k^+) - \mathcal{L}_i^{(k)}(\theta_k) \right]_+ - \eta [-\langle g_i^{(k)}, g_t^{(k)} \rangle]_+ \right| \le \tfrac{1}{2} L_i \eta^2 \|g_t^{(k)}\|_2^2. \tag{B.4}$$

Thus negative gradient alignment produces a first-order increase in historical loss. For $L_i > 0$, an actual increase is guaranteed whenever $\eta < -2\langle g_i^{(k)}, g_t^{(k)} \rangle / (L_i \|g_t^{(k)}\|_2^2)$ and the inner product is negative. The exact gradient-conflict diagnostic is

$$C_{i,k}^{\mathrm{exact}} = [-\langle g_i^{(k)}, g_t^{(k)} \rangle]_+. \tag{B.5}$$

This diagnostic measures local interference pressure from the current objective. It is the gradient quantity compared with the prospective proxy in Figure 4A.

C Sampled-Token Log-Ratio Proxy for Gradient Conflict

Let $\mathcal{U}_k = \biguplus_{n=1}^B \mathcal{S}_n^{(k)}$ be the ordered collection of token occurrences, with

$$N_t^{(k)} = |\mathcal{U}_k| = \sum_{n=1}^B \hat{L}_n, \qquad N_i^{(k)} = |\mathcal{Q}_i^{(k)}|. \tag{C.1}$$

Distillation loss. Over the shared vocabulary $\mathcal{V}$, we use reverse KL at unit distillation temperature without a hard-label cross-entropy term:

$$\ell_{\mathrm{OPD}}(p, q) = D_{\mathrm{KL}}(p \| q) = \sum_{y \in \mathcal{V}} p(y) \log \frac{p(y)}{q(y)},$$

where $p = \pi_\theta(\cdot \mid s)$ and $q = \pi_j(\cdot \mid s)$. The implementation uses per-teacher token averages in Eq. 18, in place of the sample averaging in Eq. 1:

$$\mathcal{L}_{j,\mathrm{tok}}^{(k)}(\theta) = \frac{1}{N_j^{(k)}} \sum_{a \in \mathcal{Q}_j^{(k)}} \delta_j(s_a; \theta), \qquad \mathcal{Q}_t^{(k)} = \mathcal{U}_k, \quad j \in \{i, t\}.$$

Valid response tokens include EOS. Each teacher is normalized separately, preserving the historical weight λ_{old}. For fixed states and teachers, the exact token-loss gradient is

$$\nabla_\theta \delta_j(s; \theta) = -\mathbb{E}_{y \sim \pi_\theta(\cdot|s)} \left[\log \frac{\pi_j(y \mid s)}{\pi_\theta(y \mid s)} \nabla_\theta \log \pi_\theta(y \mid s) \right].$$

The sampled-token log ratios serve as diagnostic proxies. Training differentiates the full-vocabulary loss.

For occurrence a, let $n(a)$ be its sample index and define the score vector

$$\phi_a^{(k)} = \nabla_\theta \log \pi_\theta(\hat{y}_a \mid s_a)|_{\theta_k}. \tag{C.2}$$

The columns of $\mathbf{\Phi}_k$ are these vectors. The coefficient vectors $\boldsymbol{\xi}_j^{(k)}$ use the same indexing. Historical coefficients are zero outside the expert's probes.

Sample and token weighting. Eq. 1 averages tokens within each sample and then averages samples. To preserve these weights, define

$$\mathbf{T}_k = \mathrm{diag}_{a\in\mathcal{U}_k}(1/\hat{L}_{n(a)}), \qquad \mathbf{\Psi}_k = \mathbf{\Phi}_k\mathbf{T}_k, \qquad N_{\mathrm{s},t}^{(k)} = B, \quad N_{\mathrm{s},i}^{(k)} = N_p. \tag{C.3}$$

Using the stopped log ratios in Eq. 7 to approximate the vocabulary expectation gives the sequence-weighted diagnostic

$$g_j^{(k)} \approx \widetilde{g}_j^{(k)} := -\frac{1}{N_{\mathrm{s},j}^{(k)}}\mathbf{\Psi}_k\boldsymbol{\xi}_j^{(k)}, \qquad j \in \{i,t\}. \tag{C.4}$$

With the sequence-weighted score Gram matrix

$$\mathbf{\Lambda}_k = \mathbf{\Psi}_k^\top\mathbf{\Psi}_k = \mathbf{T}_k\mathbf{\Phi}_k^\top\mathbf{\Phi}_k\mathbf{T}_k, \tag{C.5}$$

the proxy gradient inner product is

$$\langle \widetilde{g}_i^{(k)}, \widetilde{g}_t^{(k)}\rangle = \frac{(\boldsymbol{\xi}_i^{(k)})^\top\mathbf{\Lambda}_k\boldsymbol{\xi}_t^{(k)}}{N_{\mathrm{s},i}^{(k)}N_{\mathrm{s},t}^{(k)}}. \tag{C.6}$$

Equal-length responses recover token-count normalization. The weights in Eq. C.3 preserve the objective's averaging for variable-length responses. For the token-averaged implementation, replace $\mathbf{\Psi}_k$ by $\mathbf{\Phi}_k$ and $N_{\mathrm{s},j}^{(k)}$ by $N_j^{(k)}$ in the proxy formulas. The same conditional sign argument below then uses the corresponding Gram matrix $\mathbf{\Phi}_k^\top\mathbf{\Phi}_k$.

A sufficient sign condition. Let $\mathbf{P}_{i,k}$ be the orthogonal projector onto $\mathrm{span}\{\boldsymbol{\xi}_i^{(k)}, \boldsymbol{\xi}_t^{(k)}\}$. Assume $\upsilon_k > 0$ and $0 \le \omega_k < 1$ satisfy

$$\|\mathbf{P}_{i,k}(\mathbf{\Lambda}_k - \upsilon_k\,\mathrm{Id})\mathbf{P}_{i,k}\|_2 \le \omega_k\upsilon_k. \tag{C.7}$$

Then

$$\left|(\boldsymbol{\xi}_i^{(k)})^\top\mathbf{\Lambda}_k\boldsymbol{\xi}_t^{(k)} - \upsilon_k\langle\boldsymbol{\xi}_i^{(k)}, \boldsymbol{\xi}_t^{(k)}\rangle\right| \le \omega_k\upsilon_k\|\boldsymbol{\xi}_i^{(k)}\|_2\|\boldsymbol{\xi}_t^{(k)}\|_2. \tag{C.8}$$

For nonzero coefficient vectors, the Euclidean log-ratio inner product and the proxy gradient inner product therefore share a sign whenever

$$\frac{|\langle\boldsymbol{\xi}_i^{(k)}, \boldsymbol{\xi}_t^{(k)}\rangle|}{\|\boldsymbol{\xi}_i^{(k)}\|_2\|\boldsymbol{\xi}_t^{(k)}\|_2} > \omega_k. \tag{C.9}$$

This condition connects the directional proxy to the local gradient approximation through the score geometry.

Finally, the magnitude factor used by RESPOND is

$$\frac{\|\boldsymbol{\xi}_i^{(k)}\|_2}{\sqrt{N_i^{(k)}}} = \left(\frac{1}{N_i^{(k)}}\sum_{a\in\mathcal{Q}_i^{(k)}}(\xi_{i,a}^{(k)})^2\right)^{1/2}. \tag{C.10}$$

Multiplying this root-mean-square strength by the stabilized negative cosine yields Eq. 8. The factor attenuates negligible teacher signals, while Eq. C.9 explains its directional connection to gradient interference. Figure 4A evaluates the resulting risk ranking against $C_{i,k}^{\mathrm{exact}}$.

D RUNNING-SCALE AND RISK-OBSERVATION UPDATES

At the start of step k, $\bar{d}_i^{(k)}$, $\zeta_i^{(k)}$, and $\nu_r^{(k)} \ge 0$ are pre-step states. The stage reference b_i is fixed using response-selected calibration trajectories at θ_0. Each running scale starts from $\nu_r^{(0)} = 0$.

A potential query of expert i on its preselected $\mathcal{Q}_i^{(k)}$ reveals $d_i^{(k)}$ and the candidate average $\bar{d}_{i,+}^{(k)}$ in Eqs. 5–6. For $\iota_i^{(k)} = \mathbb{I}[i \in S_k]$, the stored average updates as

$$\bar{d}_i^{(k+1)} = \begin{cases}\bar{d}_{i,+}^{(k)}, & i \in S_k,\\ \bar{d}_i^{(k)}, & i \notin S_k.\end{cases} \tag{D.1}$$

Thus unqueried teachers retain their last KL average, while Eq. 9 updates the persistent risk state at every step. The retrospective quantity is reference-relative disagreement. Its value is influenced by changes in both probe populations and student parameters. The common stage reference anchors this signal throughout training.

For $r \in \{\mathrm{ret}, \mathrm{pro}\}$, the risks are normalized using the pre-step scales:

$$\widetilde{R}_{i,r}^{(k)} = \frac{R_{i,r}^{(k)}}{\nu_r^{(k)} + \epsilon}, \qquad \varrho_i^{(k)} = \widetilde{R}_{i,\mathrm{ret}}^{(k)} + \lambda_g \widetilde{R}_{i,\mathrm{pro}}^{(k)}. \tag{D.2}$$

After these normalized values update $\zeta_i^{(k+1)}$, the scale becomes

$$\nu_r^{(k+1)} = \beta_s \nu_r^{(k)} + (1-\beta_s)\frac{1}{|S_k|}\sum_{i \in S_k} R_{i,r}^{(k)}, \qquad 0 \le \beta_s < 1. \tag{D.3}$$

This statistic tracks the scale of routed observations. Nonnegative risks preserve nonnegative scales, and $\epsilon > 0$ keeps every normalization denominator positive. We use $\lambda_g \ge 0$, so the combined risk is nonnegative.

First, we cache student rollouts and compute probes. We then form utilities from $\zeta_i^{(k)}$ and sample S_k. Next, we evaluate the selected teachers at θ_k and compute candidate risks with $\nu_r^{(k)}$. Finally, we update the student and commit the detached risk, KL-average, and scale states. Since the current expert covers the full batch and each historical probe is selected before routing, each expert's potential risk is determined independently of which other experts join S_k.

E Conditional Unbiasedness of Partial Risk Observation

Fix step k and condition on the student parameters, realized rollouts, pre-step states, fixed teachers, and potential teacher observations. Denote these fixed quantities by $\mathscr{G}_k$. The potential risk $\varrho_i^{(k)}$ is the value obtained by querying expert i on its preselected probes. It uses the pre-step normalization scales. The remaining randomness is the draw $S_k \sim \Pi_k$.

The indicator and exact marginal inclusion probability are

$$\iota_i^{(k)} = \mathbb{I}[i \in S_k], \qquad p_i^{(k)} = \Pr(i \in S_k \mid \mathscr{G}_k) = \sum_{S \in \Omega_t : i \in S} \Pi_k(S). \tag{E.1}$$

With $H = t-1$ historical experts and $K = \min(3, H)$, uniform subset exploration ensures

$$p_i^{(k)} \ge \varepsilon_{\exp} \frac{\sum_{m=1}^{K} \binom{H-1}{m-1}}{\sum_{m=1}^{K} \binom{H}{m}} > 0. \tag{E.2}$$

This marginal includes every subset containing expert i, across all allowed cardinalities.

For the unmodified inverse-probability estimator,

$$\varrho_{i,\mathrm{IP}}^{(k)} = \frac{\iota_i^{(k)}}{p_i^{(k)}} \varrho_i^{(k)}, \tag{E.3}$$

we have

$$\mathbb{E}[\varrho_{i,\mathrm{IP}}^{(k)} \mid \mathscr{G}_k] = \frac{\varrho_i^{(k)}}{p_i^{(k)}} \mathbb{E}[\iota_i^{(k)} \mid \mathscr{G}_k] = \varrho_i^{(k)}. \tag{E.4}$$

Adaptive routing through pre-step risk and current responses is therefore compatible with conditional unbiasedness when the exact marginal is used. Since the indicator is Bernoulli,

$$\mathrm{Var}(\varrho_{i,\mathrm{IP}}^{(k)} \mid \mathscr{G}_k) = \left(\frac{1}{p_i^{(k)}} - 1\right)(\varrho_i^{(k)})^2. \tag{E.5}$$

The variance grows as the inclusion probability decreases, motivating stabilization. For $p_{\min} > 0$, $\varrho_{\max} > 0$, and nonnegative risks, the practical estimator in Eq. 9 has expectation

$$\mathbb{E}[\widehat{\varrho}_i^{(k)} \mid \mathscr{G}_k] = \frac{p_i^{(k)}}{\max(p_i^{(k)}, p_{\min})} \min(\varrho_i^{(k)}, \varrho_{\max}). \tag{E.6}$$

Its downward-bias magnitude is exactly

$$\varrho_i^{(k)} - \mathbb{E}[\widehat{\varrho}_i^{(k)} \mid \mathscr{G}_k] = \varrho_i^{(k)} - \frac{p_i^{(k)}}{\max(p_i^{(k)}, p_{\min})} \min(\varrho_i^{(k)}, \varrho_{\max}) \geq 0. \tag{E.7}$$

When $p_i^{(k)} \geq p_{\min}$ and $\varrho_i^{(k)} \leq \varrho_{\max}$, this bias is zero. Stabilization bounds each observation by $\varrho_{\max}/p_{\min}$. Exponential smoothing with $0 \leq \gamma < 1$ controls its contribution to the persistent state. The estimand is the current probe risk. The state $\zeta_i^{(k)}$ smooths its partially observed estimates across steps.

F Local Functional Coverage Bound

Fold expert π_i into the same ambient parameter space as the student, giving θ_i, and write

$$\theta = \theta_i + \Delta\theta_i. \tag{F.1}$$

For a fixed state distribution $\mathbb{R}$, define

$$\mathfrak{D}_i(\theta; \mathbb{R}) = \mathbb{E}_{s\sim\mathbb{R}} \delta_i(s; \theta) = \mathbb{E}_{s\sim\mathbb{R}} D_{\mathrm{KL}}(\pi_\theta(\cdot \mid s) \| \pi_i(\cdot \mid s)) \,. \tag{F.2}$$

Assume positive policy probabilities, three continuous parameter derivatives near θ_i, and integrable bounds allowing differentiation under expectation and a bounded third-order remainder. At θ_i, the KL and its first derivative vanish. Defining $\psi_i(y, s) = \nabla_\theta \log \pi_\theta(y \mid s)|_{\theta_i}$, its Hessian is the Fisher matrix

$$\mathbf{F}_i(\mathbb{R}) = \mathbb{E}_{s\sim\mathbb{R}} \mathbb{E}_{y\sim\pi_i(\cdot|s)}[\psi_i(y, s)\psi_i(y, s)^\top]. \tag{F.3}$$

Taylor expansion gives

$$\mathfrak{D}_i(\theta; \mathbb{R}) = \tfrac{1}{2}\Delta\theta_i^\top \mathbf{F}_i(\mathbb{R})\Delta\theta_i + \mathfrak{e}_i(\mathbb{R}, \Delta\theta_i), \qquad |\mathfrak{e}_i(\mathbb{R}, \Delta\theta_i)| \leq c_i(\mathbb{R})\|\Delta\theta_i\|_2^3. \tag{F.4}$$

For the historical distribution $\mathbb{P}_i$ and current-probe distribution $\mathbb{Q}_i$, assume

$$\mathbf{F}_i(\mathbb{Q}_i) \succeq \mu_i \mathbf{F}_i(\mathbb{P}_i), \qquad \mu_i > 0. \tag{F.5}$$

The corresponding quadratic forms satisfy

$$\Delta\theta_i^\top \mathbf{F}_i(\mathbb{P}_i)\Delta\theta_i \leq \frac{1}{\mu_i}\Delta\theta_i^\top \mathbf{F}_i(\mathbb{Q}_i)\Delta\theta_i. \tag{F.6}$$

Substituting Eq. F.4 on both distributions yields

$$\begin{aligned} \mathfrak{D}_i(\theta; \mathbb{P}_i) &\leq \frac{\mathfrak{D}_i(\theta; \mathbb{Q}_i)}{\mu_i} + \mathfrak{e}_i(\mathbb{P}_i, \Delta\theta_i) - \frac{\mathfrak{e}_i(\mathbb{Q}_i, \Delta\theta_i)}{\mu_i} \\ &\leq \frac{\mathfrak{D}_i(\theta; \mathbb{Q}_i)}{\mu_i} + \left(c_i(\mathbb{P}_i) + \frac{c_i(\mathbb{Q}_i)}{\mu_i} \right) \|\Delta\theta_i\|_2^3. \end{aligned} \tag{F.7}$$

Thus historical deviation is locally controlled by probe deviation with an explicit third-order remainder. The coverage constant quantifies how well probes observe the expert's historical sensitive directions. Eq. F.5 is a sufficient geometric condition. Response-based selection targets active directions, and the experiments assess the utility of that selection for consolidation.

G Rank-Space Form of LoRA Response Energy

For module m of expert i, let $A_i^{(m)} \in \mathbb{R}^{r\times d_{\mathrm{in}}}$ and $B_i^{(m)} \in \mathbb{R}^{d_{\mathrm{out}}\times r}$. The adapter is

$$\Delta W_i^{(m)} = \sigma_i^{(m)} B_i^{(m)} A_i^{(m)}. \tag{G.1}$$

At token occurrence $s_{n,\ell}^{(k)}$, its response on the cached student activation is

$$\boldsymbol{\delta}_{i,n,\ell}^{(m,k)} = \Delta W_i^{(m)} h_{n,\ell}^{(m,k)} = \sigma_i^{(m)} B_i^{(m)} z_{i,n,\ell}^{(m,k)}, \qquad z_{i,n,\ell}^{(m,k)} = A_i^{(m)} h_{n,\ell}^{(m,k)}. \tag{G.2}$$

Using $\Gamma_i^{(m)} = (B_i^{(m)})^\top B_i^{(m)}$ gives the exact identity

$$\begin{aligned} \|\boldsymbol{\delta}_{i,n,\ell}^{(m,k)}\|_2^2 &= (\sigma_i^{(m)})^2 (z_{i,n,\ell}^{(m,k)})^\top (B_i^{(m)})^\top B_i^{(m)} z_{i,n,\ell}^{(m,k)} \\ &= (\sigma_i^{(m)})^2 (z_{i,n,\ell}^{(m,k)})^\top \Gamma_i^{(m)} z_{i,n,\ell}^{(m,k)}. \end{aligned} \tag{G.3}$$

Dividing by $\|v_{n,\ell}^{(m,k)}\|_2^2 + \epsilon$ recovers Eq. 14. Since $\Gamma_i^{(m)} \succeq 0$, the energy is nonnegative. Frozen experts allow precomputation of $\Gamma_i^{(m)}$. Each response then uses a cached input, the first LoRA projection, and a rank-space quadratic form. This supplies expert-response scores while reserving full teacher forward passes for selected probes.

H Local Relation Between LoRA Response Energy and Policy KL

Fix expert i, sample n, token ℓ, and step k. Stack the module responses as

$$\boldsymbol{\delta}_{i,n,\ell}^{(k)} = \mathrm{concat}_{m \in \mathcal{M}} \, \boldsymbol{\delta}_{i,n,\ell}^{(m,k)}. \tag{H.1}$$

Let $\mathbf{f}_{n,\ell}^{(k)}$ be the student logits. Locally injecting these responses into the selected module outputs induces

$$\Delta \mathbf{f}_{i,n,\ell}^{(k)} = \mathbf{J}_{n,\ell}^{(k)} \boldsymbol{\delta}_{i,n,\ell}^{(k)} + O(\|\boldsymbol{\delta}_{i,n,\ell}^{(k)}\|_2^2), \tag{H.2}$$

where $\mathbf{J}_{n,\ell}^{(k)}$ is the downstream student Jacobian at zero injection. Define the induced local KL by

$$\mathcal{D}_{i,n,\ell}^{\mathrm{loc}(k)} = D_{\mathrm{KL}}\Big(\mathrm{softmax}(\mathbf{f}_{n,\ell}^{(k)} + \Delta \mathbf{f}_{i,n,\ell}^{(k)}) \,\|\, \mathrm{softmax}(\mathbf{f}_{n,\ell}^{(k)})\Big). \tag{H.3}$$

For $\boldsymbol{\pi}_{n,\ell}^{(k)} = \mathrm{softmax}(\mathbf{f}_{n,\ell}^{(k)})$, the softmax-KL Hessian is

$$\mathbf{H}_{n,\ell}^{(k)} = \mathrm{diag}(\boldsymbol{\pi}_{n,\ell}^{(k)}) - \boldsymbol{\pi}_{n,\ell}^{(k)} (\boldsymbol{\pi}_{n,\ell}^{(k)})^\top \succeq 0. \tag{H.4}$$

For brevity, suppress i, n, ℓ, k below and write $\mathbf{G} = \mathbf{J}^\top \mathbf{H} \mathbf{J}$. Smoothness gives

$$\mathcal{D}^{\mathrm{loc}} = \tfrac{1}{2} \Delta \mathbf{f}^\top \mathbf{H} \Delta \mathbf{f} + O(\|\Delta \mathbf{f}\|_2^3) = \tfrac{1}{2} \boldsymbol{\delta}^\top \mathbf{G} \boldsymbol{\delta} + O(\|\boldsymbol{\delta}\|_2^3). \tag{H.5}$$

Curvature on active responses. The softmax Hessian has a uniform-logit-shift null direction, and downstream mappings can have additional null directions. Assume that on the active response subspace,

$$\underline{\alpha}_i \|\boldsymbol{\delta}\|_2^2 \le \boldsymbol{\delta}^\top \mathbf{G} \boldsymbol{\delta} \le \overline{\alpha}_i \|\boldsymbol{\delta}\|_2^2, \qquad 0 < \underline{\alpha}_i \le \overline{\alpha}_i < \infty. \tag{H.6}$$

Assume also that module output scales satisfy

$$0 < \underline{\varsigma}_i \le \|v_{n,\ell}^{(m,k)}\|_2^2 + \epsilon \le \overline{\varsigma}_i < \infty. \tag{H.7}$$

By the exact rank-space identity,

$$\underline{\varsigma}_i \sum_{m \in \mathcal{M}} E_{i,n,\ell}^{(m,k)} \le \|\boldsymbol{\delta}_{i,n,\ell}^{(k)}\|_2^2 \le \overline{\varsigma}_i \sum_{m \in \mathcal{M}} E_{i,n,\ell}^{(m,k)}. \tag{H.8}$$

Choose a neighborhood where the cubic remainder in Eq. H.5 has magnitude at most $\underline{\alpha}_i \|\boldsymbol{\delta}\|_2^2/4$. Then

$$\tfrac{1}{4} \underline{\alpha}_i \|\boldsymbol{\delta}\|_2^2 \le \mathcal{D}^{\mathrm{loc}} \le (\tfrac{1}{2} \overline{\alpha}_i + \tfrac{1}{4} \underline{\alpha}_i) \|\boldsymbol{\delta}\|_2^2. \tag{H.9}$$

Combining Eqs. H.8–H.9 yields

$$\underline{\vartheta}_i \sum_{m \in \mathcal{M}} E_{i,n,\ell}^{(m,k)} \le \mathcal{D}_{i,n,\ell}^{\mathrm{loc}(k)} \le \overline{\vartheta}_i \sum_{m \in \mathcal{M}} E_{i,n,\ell}^{(m,k)}, \tag{H.10}$$

with explicit choices $\underline{\vartheta}_i = \underline{\alpha}_i \underline{\varsigma}_i/4$ and $\overline{\vartheta}_i = (\overline{\alpha}_i/2 + \underline{\alpha}_i/4)\overline{\varsigma}_i$. Under uniform bounds across the sample's token states, averaging gives

$$|\mathcal{M}|\,\underline{\vartheta}_i \bar{E}_{i,n}^{(k)} \le \frac{1}{\hat{L}_n} \sum_{\ell=1}^{\hat{L}_n} \mathcal{D}_{i,n,\ell}^{\mathrm{loc}(k)} \le |\mathcal{M}|\,\overline{\vartheta}_i \bar{E}_{i,n}^{(k)}. \tag{H.11}$$

The factor $|\mathcal{M}|$ follows from the module average in Eq. 15. The monotone map $1 - \exp(-x/\kappa)$ preserves energy rankings for $x \ge 0$.

These bounds concern locally injected adapter responses. The complete teacher comparison additionally depends on the student's displacement from that teacher. For example, a student equal to π_i has zero teacher KL even when the adapter response is nonzero. Figure 4C–D directly evaluates response as a ranking statistic for full-teacher signal and query allocation. Historical loss control follows the separate Fisher-coverage condition in Appendix F. The local energy analysis identifies active response directions used to choose probes.

I Dataset Details

Our continual-learning stream contains eight training tasks (99,016 question–answer pairs) and ten evaluation datasets (6,000 pairs). Table I.1 follows the training order. GeoQA-CoT and MME are used only for evaluation. The same evaluation sets are used after each task.

Table I.1: Dataset statistics. Mean image sizes (width × height) and text lengths (Unicode characters) are pooled across the training and test splits.

Dataset	Task type	Train	Test	Image size (px)	Prompt (chars)	Answer (chars)
COCO2017	Object localization	20,000	1,000	581×481	111.75	46.94
ScienceQA	Science VQA	8,315	500	483×309	220.08	22.76
MSVQA	Multi-scene VQA	13,603	500	1412×914	266.15	13.70
MMMU	Multidiscipline VQA	11,050	500	662×426	376.74	1.25
ICON-QA	Icon reasoning	12,000	500	255×165	177.74	1.10
GQA	Relational VQA	12,000	500	494×409	83.65	5.52
ClevrMATH	Visual arithmetic	12,000	500	474×317	138.84	1.04
CUB	Bird attributes	10,048	500	467×387	105.70	2.50
GeoQA-CoT	Geometry (Chinese)	0	500	140×107	218.43	135.53
MME	Perception/reasoning	0	1,000	1082×938	91.61	2.50

Task characteristics. COCO2017 (Lin et al., 2014) asks for all bounding boxes of a specified category, with coordinates normalized to $[0, 1000]$. ScienceQA (Lu et al., 2022) uses image-based science multiple-choice questions. MSVQA (Jiang et al., 2025) spans high-altitude, low-altitude, indoor, and underwater scenes. MMMU (Yue et al., 2024) contains multiple-choice and open-ended questions, including multi-image inputs. ICON-QA (Lu et al., 2021) combines image-choice, text-choice, and fill-in-the-blank tasks. Image choices are supplied alongside the main image. GQA (Hudson & Manning, 2019) tests object attributes and relations with short answers. ClevrMATH (Lindström & Abraham, 2022) requires numerical answers to arithmetic questions about synthetic scenes. CUB (Wah et al., 2011) uses yes/no questions about fine-grained bird attributes. GeoQA-CoT (Chen et al., 2021) contains Chinese geometry questions with reasoning and an answer choice. MME (Fu et al., 2025) covers 14 perception and reasoning categories, with paired positive/negative questions and balanced yes/no answers.

J Comparison Methods

Continual learning methods.

- **EWC** (Kirkpatrick et al., 2017) estimates parameter importance for previously learned tasks using the diagonal Fisher information matrix. It adds an importance-weighted quadratic

penalty on deviations from previously learned parameters, preserving important weights while allowing less critical ones to adapt to new tasks.

- **CL-MOE** (Huai et al., 2025) combines LoRA experts with task-level and instance-level routing to select complementary experts for continual visual question answering. A dynamic momentum mechanism consolidates old and newly learned expert parameters according to their relationships with tasks.
- **GainLoRA** (Liang et al., 2025) adds a LoRA branch and an input-dependent gating module for each new task while freezing previous branches and gates. Constraints on the new gate's initialization and updates suppress the new branch's influence on old-task inputs, reducing interference when integrating the branches.
- **Octopus** (Liu et al., 2026) first adapts to the current task and then refines the solution with history-free gradient orthogonalization. It evaluates gradients of previous-task parameters on current-task data to characterize interference and constrains new task parameters to be orthogonal to these directions, together with norm regularization, without revisiting historical data.
- **DGG** (Li et al., 2026) views forgetting as the absence of old-task gradients during new-task training. It approximates their descent directions using the displacement toward previously learned parameters, optionally refines this guidance with replay gradients, and uses Bernoulli sampling to regulate its incorporation into current-task updates.
- **SAME** (Xie et al., 2026) addresses both router drift and expert drift in a mixture of LoRA experts. It stabilizes routing through updates in selected spectral subspaces, scales expert updates using historical input covariance to limit functional changes, and selectively freezes experts that are important for past tasks but less useful for the current task.

OPD/OPSD-related methods.

- **GRPO** (Shao et al., 2024) samples multiple responses to each question and estimates advantages from their relative rewards, avoiding a separate value model. It improves the policy through a clipped policy-gradient objective with KL regularization toward a reference policy.
- **SDPO** (Hübotter et al., 2026) turns feedback on sampled responses, such as execution errors or successful attempts from the same rollout group, into dense token-level supervision. The model acts as a feedback-conditioned self-teacher, and the student learns by matching its next-token distributions along the student's original trajectories.
- **OPSD** (Zhao et al., 2026) uses the same model as teacher and student under different contexts. The teacher receives a verified solution or reasoning trace, whereas the student receives only the question. It minimizes token-wise distribution divergence along student-generated trajectories.
- **SDFT** (Shenfeld et al., 2026) uses in-context learning to construct a teacher conditioned on expert demonstrations and distills it into a student that sees only the task prompt. It minimizes reverse KL divergence on the student's own responses to learn from these demonstrations.
- **RLSD** (Yang et al., 2026) combines reinforcement learning with verifiable rewards and self-distillation by separating update direction from token-level update magnitude. Outcome rewards determine whether a response is reinforced or penalized, while log-probability differences between privileged teacher and student contexts determine how strongly individual tokens are updated.
- **RLCSD** (Pan et al., 2026) contrasts self-teacher predictions conditioned on correct and incorrect sibling rollouts presented in the same format. This contrast suppresses shared stylistic effects of privileged context and provides a token-level signal that modulates GRPO advantages while preserving the update direction determined by the verifier.

K Hyperparameters and Initialization

The following configuration is used for all 3 foundation models. The values in Table K.1 remain fixed across stages with historical experts.

Table K.1: Fixed hyperparameters for RESPOND.

Parameter	Value	Role of the setting
β_d, γ, β_s	0.95	Smooth batch fluctuations while incorporating each new observation with weight 0.05.
ϵ	10^{-8}	Keep normalization denominators positive at a small numerical scale.
λ_g	1	Give equal coefficients to the two normalized risk components.
$p_{\min}, \varrho_{\max}$	0.05, 5	Bound the inverse-probability multiplier by 20 and the corrected risk observation by 100.
ρ_0	0.1	Let functional response guide routing before risk accumulates.
λ_c	0.5	Charge each additional historical teacher a fixed utility cost to favor economical queries.
τ_s, τ_w	1	Use utilities at their computed scale for subset sampling and teacher weighting.
$\varepsilon_{\exp}$	0.05	Allocate 5% of sampling probability to uniform subset exploration.
κ	1	Map unit response energy to $1 - e^{-1}$, giving the bounded score a unit energy scale.
N_p	1	Concentrate each historical teacher query on its highest-response current sample.
λ_{old}	0.5	Split the supervision budget equally between current and historical objectives.

Calibration and state initialization. At the beginning of each stage $t > 1$, two gradient-free batches provide a short estimate of the student–expert reference distance. Each historical expert selects its highest-response sample in each batch. Writing $d_{i,\text{cal},j}$ for its mean response-token KL in calibration batch j, we set

$$b_i = \tfrac{1}{2}\big(d_{i,\text{cal},1} + d_{i,\text{cal},2}\big), \qquad \bar{d}_i^{(0)} = b_i, \qquad \zeta_i^{(0)} = 0, \qquad \nu_r^{(0)} = 0. \tag{K.1}$$

The reference remains fixed within the stage. Initializing the KL average at this reference makes subsequent excess disagreement measure drift from the stage entry point. The zero risk state lets the response prior determine initial routing. These states are reset at each new stage.

Risk-scale normalization and updates. Each risk component $r \in \{\text{ret}, \text{pro}\}$ has a scale shared across historical experts. Define the mean routed observation as

$$a_r^{(k)} = \frac{1}{|S_k|} \sum_{i \in S_k} R_{i,r}^{(k)}. \tag{K.2}$$

At every step, risks are normalized using the pre-step scale. After the normalized risks update the persistent state, the scale is updated for the next step:

$$\widetilde{R}_{i,r}^{(k)} = \frac{R_{i,r}^{(k)}}{\nu_r^{(k)} + 10^{-8}}, \qquad \nu_r^{(k+1)} = 0.95\nu_r^{(k)} + 0.05 a_r^{(k)}. \tag{K.3}$$

The same normalization applies while a scale is zero. The denominator remains positive and the combined risk is clipped below. Using $\varrho_i^{(k)} = \widetilde{R}_{i,\text{ret}}^{(k)} + \widetilde{R}_{i,\text{pro}}^{(k)}$, the persistent risk follows

$$\zeta_i^{(k+1)} = 0.95\zeta_i^{(k)} + 0.05\widehat{\varrho}_i^{(k)}, \qquad \widehat{\varrho}_i^{(k)} = \frac{\mathbb{I}[i \in S_k]}{\max(p_i^{(k)}, 0.05)} \operatorname{clip}(\varrho_i^{(k)}, 0, 5). \tag{K.4}$$

Unqueried experts retain their KL averages while their persistent risk decays by 0.95. Stage $t = 1$ uses only the current expert, with effective historical weight $\lambda_{\text{old}} = 0$. Subsequent stages use $\lambda_{\text{old}} = 0.5$.

LoRA experts and probe layers. All eight task experts use LoRA rank $r_{\text{L}} = 4$ and scaling coefficient $\alpha = 32$, giving

$$\sigma_i^{(m)} = \frac{\alpha}{r_{\text{L}}} = 8, \qquad (\sigma_i^{(m)})^2 = 64. \tag{K.5}$$

The small rank keeps the expert library compact, while the scaling coefficient sets the adapter contribution independently of rank. For Qwen3.5-4B, probes span layers 1, 11, 22, and 32 using one-based indexing, covering four approximately evenly spaced positions among its 32 text decoder layers. This selection covers early, intermediate, and late transformations with a small number of activation taps. The attention and feed-forward projections in these layers provide $|\mathcal{M}| = 31$ probe modules per expert, with response energy averaged over generated response tokens.

L Dataset Scoring Protocol

Evaluation workflow. After each training stage, we generate responses on the ten evaluation datasets. Each record contains the question, reference answer, model output, and dataset identifier. The evaluator receives only text, without images or image paths. For VQA, each request pairs the system prompt below with a user message containing `question`, `standard answer`, and `model output`. It returns independent scores for answer correctness ($[0, 1]$) and format compliance ($\{0, 1\}$).

VQA evaluator: system prompt

```
You are an evaluator for vision-language model outputs. You will
receive only the text part of the question, the standard answer, and
the model output. Do not use images or image paths.

Score two independent dimensions:
1. answer_score: how correct the model output is compared with the
standard answer for the question. Give 1 for fully correct or
semantically equivalent answers, 0 for incorrect answers, and partial
credit from 0 to 1 when appropriate.

2. format_score: a strict binary score. Give 1 only when the model
output fully follows the answer format requested by the question.
Give 0 for any format violation, including extra explanation,
reasoning, units, markdown, or surrounding text when the question asks
for a short answer, an integer, a choice, or the shortest text span.
Do not give partial format credit.

Return only a JSON object with this exact schema:
{"answer_score": <number between 0 and 1>, "format_score": <0 or 1>}
```

COCO2017 scoring. Answer correctness is computed locally. We extract and deduplicate predicted and reference boxes, form candidate pairs with intersection over union (IoU) ≥ 0.5, and greedily match them one-to-one in descending IoU order. Matched pairs are true positives (TP). Unmatched predictions and references are false positives (FP) and false negatives (FN). The answer score is $2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$. It is 1 when both box sets are empty and 0 when only one is empty. The API evaluates format compliance from the question and model output only.

Evaluator consistency. All compared methods, including RESPOND, are evaluated using Qwen3.5-flash under the same protocol. We also compare GPT-5.4 and Qwen3.5-flash with identical prompts and find their evaluation results to be largely consistent.

Table L.1: Illustrative VQA scores. A and F denote answer correctness and format compliance, respectively.

Question and requested format	Reference	Model output	A	F
What color is the bus? Answer with a single color word.	yellow	blue	0	1
How many dogs are in the image? Answer with an integer only.	2	2	1	1
What is the man doing? Answer with a short phrase only.	riding a bicycle	riding a bike	1	1
What are the color and material of the table? Answer with a short phrase only.	red and wooden	blue and wooden	0.5	1
What are the four objects on the desk? Answer with a comma-separated list of object names only.	book, cup, laptop, lamp	book, cup, laptop, phone	0.75	1

Reported metrics. Tables 1 and 2 in the main text report the average *answer score* for each dataset on a 100-point scale. These scores summarize answer correctness over the evaluation examples. Average *format scores* are reported separately in Table L.2.

Table L.2: Final-stage average format scores on a 100-point scale using **Qwen3.5-4B**. The dataset order and method groups follow Table 1. Among continual-learning and OPD/OPSD methods, the **best** and second-best results are marked, including ties.

Methods	Training Tasks									OOD Tasks		Overall Avg.
	COCO	Sci.QA	MSVQA	MMMU	ICON	GQA	MATH	CUB	Avg.	GeoQA	MME	
Reference and naive baselines												
Vanilla	0.00	100.0	63.40	24.40	97.60	99.60	46.00	100.0	66.38	9.00	100.0	64.00
Joint	99.80	100.0	100.0	99.80	100.0	100.0	100.0	100.0	99.95	35.00	93.10	92.77
Continual learning methods												
EWC	0.60	1.60	81.80	99.20	94.80	99.00	**100.0**	**100.0**	72.13	33.20	93.50	70.37
GainLoRA	0.00	88.80	86.00	82.80	**100.0**	**100.0**	**100.0**	**100.0**	82.20	11.00	92.30	76.09
CL-MOE	**99.90**	**100.0**	**100.0**	**99.80**	**100.0**	**100.0**	**100.0**	**100.0**	**99.96**	26.80	92.30	**91.88**
Octopus	0.00	**100.0**	64.60	32.20	99.00	99.40	34.80	**100.0**	66.25	8.00	93.60	63.16
DGG	86.80	99.80	41.00	66.00	77.00	99.60	94.00	86.80	81.38	2.00	16.30	66.93
SAME	99.50	83.40	98.60	99.60	99.80	99.80	**100.0**	**100.0**	97.59	**38.60**	95.90	91.52
OPD/OPSD methods												
OPSD	88.80	99.80	35.40	57.40	**100.0**	95.80	**100.0**	**100.0**	84.65	10.00	**100.0**	78.72
GRPO	7.00	**100.0**	77.20	27.40	70.20	99.60	3.00	**100.0**	60.55	10.80	**100.0**	59.52
SDPO	96.00	99.80	70.20	96.20	77.60	99.60	**100.0**	**100.0**	92.43	1.20	96.00	83.66
RLSD	0.00	**100.0**	72.60	45.00	**100.0**	99.80	6.40	**100.0**	65.48	9.00	**100.0**	63.28
RLCSD	0.00	**100.0**	55.80	21.40	**100.0**	99.80	22.40	**100.0**	62.43	10.00	**100.0**	60.94
RESPOND	99.40	99.40	98.00	**99.80**	**100.0**	99.80	**100.0**	**100.0**	99.55	30.20	90.60	91.72

Format-score comparison. Table L.2 presents the average *format scores* on a 100-point scale for Qwen3.5-4B at the final training stage. Higher scores indicate more consistent adherence to the answer format specified in the question. Avg. and Overall Avg. denote the unweighted means over the eight training datasets and all ten evaluation datasets, respectively.

Format compliance and answer correctness. RESPOND achieves format-score averages of 99.55 on the training tasks and 91.72 overall, ranking second among the compared continual-learning and OPD/OPSD methods on both measures. Its format score is at least 98.00 on every training task and reaches 100.00 on ICON-QA, ClevrMATH, and CUB, demonstrating consistent compliance across the task stream. Compared with SAME, it improves the training-task format average by 1.96 points and the corresponding answer-score average by 1.95 points (Table 1). Its overall format score is also within 0.16 points of CL-MOE, while its overall answer score is 3.60 points higher. **Together, these results show that RESPOND combines leading continual-learning performance with strong format compliance.**

M ADDITIONAL RESULTS ON INTERNVL3.5-4B

We extend the evaluation to InternVL3.5-4B using the eight-task stream and two OOD benchmarks in Section 5.1. Table M.1 follows the final-stage comparison in Tables 1 and 2. Figure M.1 complements these final scores with stage-wise results for representative baselines and RESPOND.

Overall performance RESPOND achieves a final training-task average of 79.91 and an overall average of 79.65, leading the reported continual-learning and OPD/OPSD methods. It exceeds SAME, the strongest competitor in training-task average, by 2.26 points, and CL-MOE, the strongest competitor in overall average, by 2.81 points. Relative to RLCSD, the strongest OPD/OPSD baseline on both averages, the improvements are 5.55 and 4.94 points, respectively. **These results extend the advantage of selective historical supervision to a third backbone.** Compared with Vanilla, the training-task average improves by 8.21 points. It is 1.44 points below Joint, while the overall average exceeds Joint by 0.08 points.

The gains cover diverse capabilities. Among the continual-learning and OPD/OPSD methods, **RESPOND leads on five of the eight training tasks, namely COCO, ScienceQA, MSVQA, GQA, and ClevrMATH.** Its ScienceQA score of 97.60 exceeds the strongest competing result by 5.00 points, while MSVQA and ClevrMATH reach 65.82 and 99.20. It also ranks second on MMMU and ICON-QA. This breadth supports effective consolidation across visual understanding, scientific

Table M.1: Final-stage comparison using **InternVL3.5-4B** on eight training tasks and two out-of-distribution (OOD) benchmarks. Among the reported continual-learning and OPD/OPSD methods, the **best** and second-best results are marked. Avg. averages the eight training tasks. Overall Avg. averages all ten benchmarks.

Methods	Training Tasks									OOD Tasks		Overall Avg.
	COCO	Sci.QA	MSVQA	MMMU	ICON	GQA	MATH	CUB	Avg.	GeoQA	MME	
Reference and naive baselines												
Vanilla	70.53	90.60	51.08	53.28	70.20	64.10	93.60	80.20	71.70	59.04	85.60	71.82
Joint	74.57	97.20	69.96	55.04	96.00	68.38	97.20	92.40	81.34	58.66	86.30	79.57
Continual learning methods												
EWC	68.19	74.28	54.99	50.46	85.70	58.48	87.80	**92.20**	71.51	**69.80**	**88.40**	73.03
GainLoRA	50.77	89.40	50.10	53.16	76.40	65.96	85.20	83.80	69.35	57.14	86.80	69.87
CL-MOE	71.84	80.93	60.43	55.00	**90.40**	65.08	96.80	91.60	76.51	69.10	87.20	76.84
Octopus	68.25	76.40	34.40	39.00	69.00	43.44	90.80	89.60	63.86	29.84	68.90	60.96
DGG	64.85	80.20	39.48	36.80	67.80	64.84	88.00	87.80	66.22	45.80	73.60	64.92
SAME	74.92	87.34	64.59	54.96	85.00	67.80	97.20	89.40	77.65	51.36	86.30	75.89
OPD/OPSD methods												
GRPO	72.84	91.60	54.58	55.52	71.40	63.49	95.40	86.40	73.90	64.44	84.60	74.03
SDPO	73.29	85.65	41.26	46.85	63.15	54.33	92.79	80.15	67.18	59.87	81.65	67.90
SDFT	57.06	81.96	49.89	52.81	73.54	51.59	75.22	71.70	64.22	54.62	77.68	64.61
RLSD	72.88	92.40	54.24	55.04	71.20	63.58	96.80	86.80	74.12	64.12	84.60	74.17
RLCSD	72.68	92.60	54.89	**56.00**	71.60	62.72	97.40	87.00	74.36	66.82	85.40	74.71
RESPOND	**75.20**	**97.60**	**65.82**	55.60	87.60	**68.64**	**99.20**	89.60	**79.91**	69.72	87.50	**79.65**

reasoning, and mathematical reasoning. EWC and CL-MOE retain higher final CUB scores, but RESPOND achieves the strongest average over the complete task stream among these methods.

OOD performance RESPOND scores 69.72 on GeoQA-CoT and 87.50 on MME, ranking second on both benchmarks. Its OOD average of 78.61 exceeds SAME's 68.83 by 9.78 points and RLCSD's 76.11 by 2.50 points. EWC obtains a slightly higher OOD average of 79.10, while trailing RESPOND by 8.39 points on the training-task average. **The leading task performance therefore coexists with competitive general capabilities**, yielding the best overall average across all ten benchmarks.

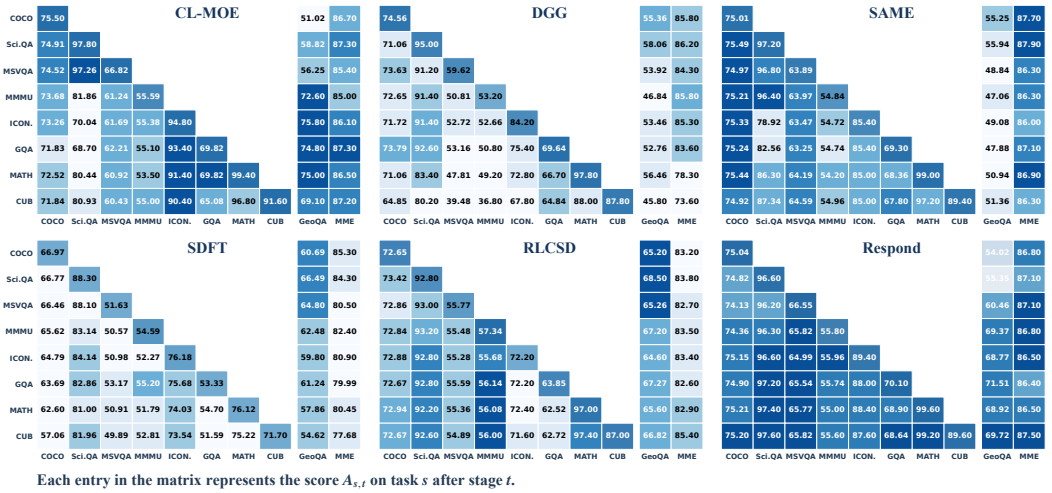

Figure M.1: Stage-wise InternVL3.5-4B results over the eight-task stream. Rows indicate training stages and columns indicate evaluation tasks. Each lower-triangular entry $A_{s,t}$ reports task-s performance after stage t. Blank cells mark unseen tasks, and darker cells indicate higher scores. GeoQA-CoT and MME track OOD and general multimodal performance. RESPOND combines strong acquisition with sustained historical-task retention.

Stage-wise retention Figure M.1 shows that RESPOND preserves early capabilities as the stream grows. COCO improves from 75.04 after acquisition to 75.20 at the final stage, and ScienceQA rises from 96.60 to 97.60 despite six subsequent tasks. In comparison, ScienceQA declines by 16.87 points for CL-MOE and 9.86 points for SAME. MSVQA decreases by only 0.73 points under RESPOND, from 66.55 to 65.82, compared with drops of 6.39 for CL-MOE and 20.15 for DGG. ClevrMATH remains at 99.20 after the final CUB update, only 0.40 points below its acquisition score.

Using the peak-to-final forgetting metric in Section 5.1, RESPOND attains the lowest average forgetting among the reported continual methods, at 0.68 points, compared with 2.02 for SAME, 5.61 for CL-MOE, and 0.78 for RLCSD. Unlike acquisition-to-final changes, this metric accounts for intermediate performance peaks. Its low forgetting accompanies strong final training-task performance, demonstrating **strong acquisition together with sustained retention**. These results are consistent with the main-text finding that allocating historical supervision according to expert risk and functional response supports continual knowledge consolidation across backbones.

N SUPPLEMENTARY RISK AND RESPONSE VISUALIZATIONS

Figure N.1 extends the main-text visualizations in Figure 4 with stage-wise agreement and query-efficiency distributions on Qwen3.5-4B.

Stable risk and response rankings Panel A shows strong agreement between prospective risk and exact gradient conflict throughout stages 4–8, with Spearman $\rho = 0.80$–0.95 and all five checkpoint-bootstrap 95% intervals above zero. Panel B shows positive response–KL correlations in all 84 expert–batch groups across stages 2–8 (median $\rho = 0.74$, IQR 0.10). **Both signals retain their ranking effectiveness as the expert library grows**, supporting risk-guided protection and expert-specific probe selection throughout the stream.

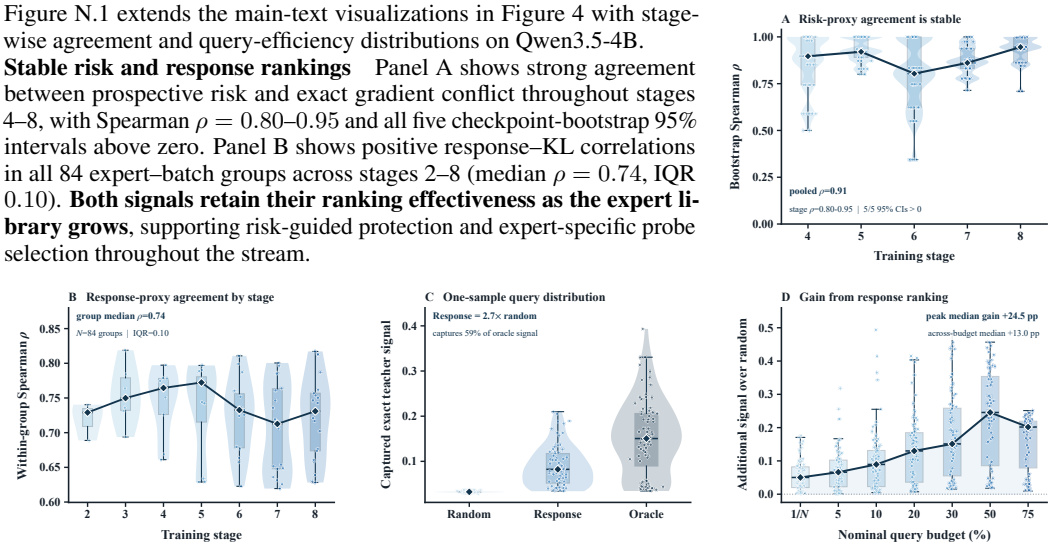

Figure N.1: Supplementary risk and response visualizations for RESPOND.

More signal per query With one queried sample per group, Panel C shows that response ranking captures 8.94% of the exact teacher KL mass on average, versus 3.28% for random selection, or 2.7× as much signal. Panel D shows positive paired gains in all 84 groups at each of the seven displayed budgets, with the median reaching 24.5 percentage points at the 50% budget. **Response-guided probing consistently concentrates limited queries on more informative trajectories.**