TASK–DRIFT LEDGER: PRE-HOC CALIBRATION OF MODULE-SPECIFIC ANTI-FORGETTING STRENGTHS FOR MULTIMODAL LLM ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Fine-tuning multimodal large language models (MLLMs) can erode capabilities acquired before adaptation. Parameter-space methods constrain model updates to limit forgetting, yet their intervention strengths are usually inherited from defaults or chosen through costly sweeps. The appropriate strength varies across tasks and between the cross-modal projector and language model; constraining one module can also shift adaptation pressure to the other. We introduce Task–Drift Ledger, which selects a strength for each module before full adaptation. A rank-4 variational probe fitted for 1,000 steps on 2,000 downstream examples estimates the task-induced parameter response. Combining this response with downstream sensitivity and reusable upstream sensitivity, the ledger scores the surrogate costs of retaining and suppressing the response. It sets module strengths while preserving each underlying method's parameter-selection rule. On LLaVA-1.5-7B, the selected configurations improve the within-task score over operator references in all nine settings, and the strengths improve all 11 evaluable higher-rank configurations. With controller constants unchanged, the method also improves the balance between forgetting and downstream performance in most evaluated Qwen2.5-VL-7B captioning settings. In our pipeline, one probe costs about 0.5 GPU-hours, compared with 26.6–79.8 GPU-hours for a training-time strength sweep. Conditional sweeps further show that the selected strengths lie in useful regions close to the best measured points at substantially lower calibration cost.

1 INTRODUCTION

Multimodal large language models (MLLMs) acquire broad perceptual, knowledge-intensive, and reasoning capabilities through large-scale vision-language pre-training. Fine-tuning on a downstream task can overwrite part of these capabilities, leading to catastrophic forgetting (Zhai et al., 2024; Zhu et al., 2024). A number of parameter-space methods have been proposed to reduce this degradation by constraining model updates through importance-aware regularization, sparse adaptation, parameter patching, or model merging (Kirkpatrick et al., 2017; Liang et al., 2025; Huang et al., 2025; Chen et al., 2025; Zhu et al., 2024; Yang et al., 2026). These methods provide mechanisms for deciding *which* parameters to modify and *how* to modify them. However, an equally important question is often left to a default setting or an expensive sweep: *how strongly should the intervention be applied for a new task?*

This question is nontrivial because intervention strength determines the trade-off between retaining pretrained capabilities and learning the downstream task. A weak intervention leaves more freedom for adaptation but may provide insufficient protection against forgetting, while a strong intervention protects pretrained parameters at the cost of downstream learning. Selecting the strength by evaluating multiple candidates is expensive. For a K-point training-time sweep, the adaptation cost is roughly KC_{FT}, while a post-hoc sweep requires one full adaptation followed by K candidate evaluations, with cost $C_{\mathrm{FT}} + KC_{\mathrm{post}}$ (Appendix C). More importantly, the appropriate strength need not be shared across tasks or modules. Different downstream tasks induce updates with different magnitudes, locations, and effects. Within an MLLM, the cross-modal projector and language model also play different roles and can exhibit different stability–plasticity responses. In our experiments, for example, tightening the LLM constraint on IconQA (Lu et al., 2021) changes the upstream score by about 11

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.

points depending on whether the projector is constrained (Sec. 4.4). These observations raise two questions: *can a useful intervention strength be predicted before full adaptation, and can the same prediction be reused across different anti-forgetting operators?*

We address this problem with *Task–Drift Ledger*, which selects intervention strengths before full adaptation. Figure 1 summarizes the proposed approach. The main idea is to use a short task probe to estimate the parameter response induced by the downstream task, together with its downstream sensitivity. The final update produced by downstream training is unavailable at this stage, but the probe can expose how the task pushes different modules away from the pretrained model. Rather than attempting to predict the complete optimization trajectory, we use this response to estimate the quantities relevant to strength selection. We combine the probe response and downstream sensitivity with an upstream sensitivity estimate that can be computed once for a backbone and reused across tasks. For a candidate intervention strength, these quantities define two costs: the risk of retaining task-induced changes that affect pretrained capabilities and the cost of suppressing changes that are useful for the downstream task. We then select the projector and LLM strengths by balancing these costs.

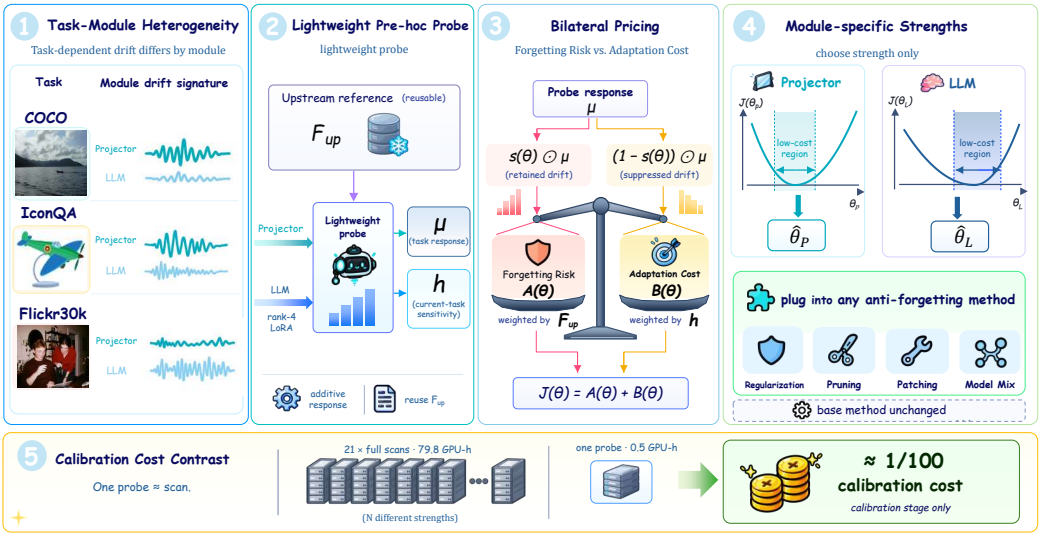


Figure 1: **Overview of Task–Drift Ledger.** Task-induced responses vary across tasks and between the projector and LLM (1). A rank-4 probe estimates the task response and downstream sensitivity; upstream sensitivity is computed once per backbone and reused (2). The ledger scores retained and suppressed responses (3) and selects module-specific strengths while preserving each operator's parameter-selection rule (4). In our pipeline, calibration for training-time operators costs about 0.5 GPU-hours for one probe versus 26.6–79.8 GPU-hours for a 7–21-point sweep (5). These calibration costs exclude the reusable upstream pass and the final selected run. Post-hoc costs are reported in Appendix C.

The controller provides a separate strength-selection layer for the underlying anti-forgetting operator. It selects intervention strengths, while LoRASculpt, Model Tailor, CGM$^+$, or another compatible operator retains its own parameter-selection and execution rules. One task probe can therefore be reused across operators. Evaluating multiple strengths for a training-time operator requires separate adaptations, whereas post-hoc candidates can reuse a trained expert. In our implementation, the probe uses 2,000 downstream examples and 1,000 optimization steps, costing about 0.5 GPU-hours compared with 26.6–79.8 GPU-hours for the corresponding 7–21-point training-time sweeps. This calibration-cost comparison excludes the reusable upstream sensitivity pass and the final selected adaptation (Appendix C).

We evaluate the controller across training-time regularization and post-hoc parameter processing on two MLLM backbones. On LLaVA-1.5-7B (Liu et al., 2024), the frozen controller improves the

within-task score over the published operating point in all nine operator–task configurations (Table 1). On 18 prespecified conditional strength axes, the cumulative regret is 7.4 versus 21.3 on COCO (Lin et al., 2014), 6.6 versus 11.5 on IconQA (Lu et al., 2021), and 2.8 versus 12.9 on Flickr30k (Young et al., 2014), where each pair compares the ledger with the uniform-random expectation. The ledger has lower regret than the random expectation on 16 of the 18 axes (Sec. 4.2; Appendix B). The strengths selected using a probe with rank-4 LoRA for the LLM also transfer to all 11 evaluable rank-64/128 configurations (Table 2). On Qwen2.5-VL-7B (Bai et al., 2025), the same controller specification improves five of six image-captioning configurations, with a mean within-task score gain of 3.0 points without refitting the controller constants (Table 3).

These results support a simple picture: intervention strength can be treated as a calibration problem separate from the design of the underlying anti-forgetting operator. A short task probe is sufficient to identify useful strength regions at a small fraction of the cost of a training-time sweep, and the resulting strengths can be reused across parameterizations and operators. At the same time, the analysis exposes clear limits of this formulation. The current controller selects projector and LLM strengths separately and therefore does not explicitly model training-induced cross-module interactions (Sec. 4.4); it is intended to identify useful regions rather than guarantee the global optimum (Sec. 4.2).

Our contributions are:

- **Pre-hoc strength selection.** We formulate intervention-strength selection as a task- and module-dependent calibration problem before full adaptation.
- **Task–Drift Ledger.** We introduce a lightweight probe-based controller that selects module-specific strengths while preserving each operator's native update rule.
- **Empirical validation.** We evaluate multiple operators, tasks, LoRA ranks, and backbones, including strength regret, transfer, ablations, and module interactions.

2 Related Work

Training-time interventions. Parameter-space methods restrict updates to protect pre-trained capabilities. EWC weights parameter changes by estimated importance (Kirkpatrick et al., 2017). For MLLMs, LoRASculpt identifies redundant LoRA components and regularizes conflicting updates (Liang et al., 2025); SPIDER combines pre-training and downstream importance to allocate update freedom (Huang et al., 2025); and SEFE regularizes knowledge-related low-rank updates (Chen et al., 2025). Related work on parameter-efficient adaptation examines parameter choice and cross-modal interaction. BEFT reports advantages from tuning value-projection biases under limited data (Huang et al., 2026). CoLA adds a low-rank pathway for cross-modal exchange in dual encoders (Suharitdamrong et al., 2026).

Post-hoc parameter processing. Post-hoc methods modify task experts after training. Model Tailor retains a sparse task patch based on parameter salience and sensitivity (Zhu et al., 2024), while Fisher-weighted averaging merges parameters using local sensitivity (Matena & Raffel, 2022). For multimodal continual instruction tuning, MAny combines visual-prototype-guided projection-space merging with recursive merging of low-rank task modules (Gao et al., 2026). OptMerge processes task vectors from multiple MLLM experts to reduce redundancy and interference (Wei et al., 2026). Curvature-Guided Mixing (CGM) uses local quadratic approximations of pre-training and downstream losses to derive soft mixing ratios and a sparse hard-mixing variant (Yang et al., 2026).

Strength selection. Regularization weights, patch budgets, and mixing ratios govern the stability–plasticity trade-off. Continual learning research has examined how to select such hyperparameters: De Lange et al. adjust them using current-task data (De Lange et al., 2021); Lee et al. compare selection protocols under data-access constraints and consider both performance and computational efficiency (Lee et al., 2024). Cha and Cho show that tuning and evaluation within the same scenario can overestimate generalization, and propose separate tuning and evaluation phases (Cha & Cho, 2025). These studies motivate evaluating anti-forgetting methods together with their calibration costs and transfer conditions.

Distinction from prior work. Task–Drift Ledger calibrates module-level strengths before full MLLM adaptation. A short task probe and reusable upstream sensitivity provide separate projector and LLM budgets, which existing operators execute through their native parameter-selection rules. CGM derives parameter-mixing rules from the local second-order geometry of pre-trained and fine-tuned models (Yang et al., 2026); our surrogate scores the retention and suppression of probe responses under candidate budgets. Both use quadratic modeling, but their decisions use different information and operate at different stages and granularities: CGM mixes a trained expert at the parameter level, while the ledger selects module budgets before full adaptation and also supports training-time interventions.

3 METHOD

Task–Drift Ledger uses a short task probe to select separate projector and LLM strengths. The probe supplies a weight-space response and downstream sensitivity; the ledger combines them with reusable upstream sensitivity to score candidate budgets.

3.1 PROBLEM FORMULATION

Consider a pre-trained MLLM parameterized by $W_0 = \big(W_V^0, W_P^0, W_L^0\big)$, where W_V^0, W_P^0, and W_L^0 denote the vision encoder, cross-modal projector, and LLM, respectively. We keep the vision encoder frozen and denote the adaptable base parameters by $W_0^{\text{adapt}} = \big(W_P^0, W_L^0\big)$. For a downstream task D, unconstrained adaptation produces an update ΔW_D:

$$W_D = W_0^{\text{adapt}} + \Delta W_D, \qquad \Delta W_D = \big(\Delta W_P^D, \Delta W_L^D\big)\,. \tag{1}$$

Let $\mathcal{M}$ denote a parameter-space anti-forgetting operator with a module-wise intervention strength $\boldsymbol{\theta} = (\theta_P, \theta_L)$. The interpretation of $\boldsymbol{\theta}$ is operator-specific (e.g., regularization penalty, retained fraction, or merging ratio). A grid search evaluates a discrete set Θ and selects $\boldsymbol{\theta}^\star$ from the measured outcomes, with the operator-dependent costs described in the Introduction.

The best measured strength can differ across tasks and modules. We seek a controller $f : (D, W_0, \mathcal{M}, F^{\text{up}}) \to \hat{\boldsymbol{\theta}} = (\hat{\theta}_P, \hat{\theta}_L)$ that selects a useful operating point at lower cost than evaluating the full candidate grid. The operator must expose module-specific strength controls. The ledger requires a pre-hoc retention-response proxy $s_i^{\mathcal{M}}(\theta)$, which can use either the shared magnitude-based ranking or an operator-specific ranking when available. In our experiments, strengths are module retention budgets ρ, applied through each operator's native rule. Calibration uses upstream reference data to estimate F^{up}, or reuses an estimate for the same backbone. Downstream optimization uses no upstream replay, and post-hoc execution requires a usable trained expert.

3.2 VARIATIONAL TASK PROBE

The final task update ΔW_D is unavailable before adaptation. We freeze the base model and fit a mean-field Gaussian over probe parameters z: $q_\phi(z) = \mathcal{N}(\mu_z, \text{diag}(\sigma_z^2))$. For the projector, z directly parameterizes an additive weight perturbation; for the LLM, it parameterizes perturbations of rank-4 LoRA factors. Let $\Delta W(z)$ map these probe parameters to weight-space updates. Using $z = \mu_z + \sigma_z \odot \epsilon$, $\epsilon \sim \mathcal{N}(0, I)$, we minimize

$$\mathcal{L}_{\text{probe}} = \mathbb{E}_{z\sim q_\phi}\big[\text{NLL}(W_0 + \Delta W(z); D)\big] + \frac{\beta}{N}\text{KL}\big(q_\phi(z)\,\|\,\mathcal{N}(0, \sigma_p^2 I)\big). \tag{2}$$

Here, N is the number of task examples and β weights the KL penalty. We use 2,000 task samples and 1,000 optimization steps, costing roughly one third to one half of an equal-budget fine-tuning run (Appendix C). The controller uses the weight response reconstructed from the fitted probe means. For the projector, this is the additive mean μ. For an LLM weight matrix with LoRA factors A and B, the response is

$$\mu_W = \gamma\,[(B_0 + \mu_B)(A_0 + \mu_A) - B_0 A_0]\,, \tag{3}$$

where A_0, B_0 are the initial LoRA factors, μ_A, μ_B their fitted mean perturbations, and γ the LoRA scaling factor. The ledger and the endpoint LLM statistic below use μ_W. We write μ for either

module's weight-space response: the additive projector mean or the reconstructed LLM response. The LLM probe uses rank-4 factors even when the executor uses rank-32 LoRA or dense updates.

For each tensor k, we compute $\tau_k = \mathrm{RMS}(\mu^{(k)}) = \sqrt{\frac{1}{|k|}\sum_{i\in k}\mu_i^2}$ and define the active-drift gate $a_i = \mathbf{1}[|\mu_i| > \kappa\tau_{k(i)}]$, with $\kappa = 1$. This gate identifies responses above their tensor's RMS scale. The learned variance is excluded from strength selection and examined in Sec. 4.4.

Downstream sensitivity uses different estimators for the two modules. For the projector, we average squared batch-mean gradients over T probe steps:

$$h_i^P = \frac{1}{T}\sum_{t=1}^{T}\Big(\frac{\partial\bar{\ell}_t}{\partial w_i}\Big)^2. \tag{4}$$

For the LLM, we average over B batches at the probe endpoint in reconstructed weight space:

$$h_i^L = \frac{1}{B}\sum_{b=1}^{B}\Big(\frac{\partial\bar{\ell}_b(W_0+\mu_W)}{\partial w_i}\Big)^2. \tag{5}$$

These empirical choices preserve task discrimination better than swapping the trajectory and endpoint estimators in our ablations (Sec. 4.4). We use h as a sensitivity proxy.

Reusable upstream sensitivity. We compute upstream sensitivity at the pre-trained parameters once per backbone and reuse it across downstream tasks. Over B_{up} upstream batches, the estimate is

$$F_i^{\mathrm{up}} = \frac{1}{B_{\mathrm{up}}}\sum_{b=1}^{B_{\mathrm{up}}}\left(\frac{\partial\bar{\ell}_b^{\mathrm{up}}(W_0)}{\partial w_i}\right)^2. \tag{6}$$

We use this estimator for both modules, restricting Eq. (6) to the coordinates being priced. It averages squared batch-mean gradients, making F^{up} a diagonal sensitivity proxy rather than the per-example empirical Fisher.

3.3 Scoring Retained and Suppressed Responses

The ledger uses the probe response to score candidate strengths. Let an operator $\mathcal{M}$ parameterized by θ induce a per-coordinate retention response $s_i^{\mathcal{M}}(\theta) \in [0,1]$, abbreviated as s_i below. We define the two quadratic surrogate costs together:

$$\big(A_{\mathcal{M}}(\theta), B_{\mathcal{M}}(\theta)\big) = \tfrac{1}{2}\sum_i \mu_i^2\big(F_i^{\mathrm{up}}s_i^2,\ h_i a_i(1-s_i)^2\big). \tag{7}$$

$A_{\mathcal{M}}$ assigns a larger cost to retained responses on coordinates with high upstream sensitivity. $B_{\mathcal{M}}$ assigns a larger cost to suppressing active responses on coordinates with high downstream sensitivity. Together, they score the stability-plasticity trade-off along a fixed probe response.

Pre-hoc retention response. Equation (7) requires a response $s_i^{\mathcal{M}}(\theta)$ before the final executor exists. For the three budget-style operators evaluated in this work, the frozen *global* arm maps a module retention ratio ρ to a common magnitude-based budget proxy,

$$s_i^{\mathrm{global}}(\rho) = \mathbf{1}\big[i \in \mathrm{Top}_\rho(|\mu|)\big], \tag{8}$$

where Top_ρ selects the largest-magnitude fraction ρ within the module being priced. The ledger scores candidate budgets along this fixed ranking. A smaller ρ retains less probe response; at execution, each operator selects coordinates through its own native rule. When an operator provides a pre-hoc ranking, we can use its response in place of Eq. (8). This native-response variant is used for Qwen CGM$^+$, with additional comparisons on LLaVA. The primary LLaVA results and Qwen LoRASculpt/Model Tailor results use the global variant.

3.4 Strength Selection and Operator Integration

For each module, the ledger selects the candidate minimizing

$$J_{\mathcal{M}}(\theta) = A_{\mathcal{M}}(\theta) + \nu B_{\mathcal{M}}(\theta), \qquad \hat{\theta} = \arg\min_{\theta \in \Theta_{\mathcal{M}}} J_{\mathcal{M}}(\theta). \tag{9}$$

We fix $\nu = 1$ across tasks. Separate scoring of the projector and LLM gives $\hat{\boldsymbol{\theta}} = (\hat{\theta}_P, \hat{\theta}_L)$, which the operator executes using its native rule. We evaluate this separable approximation through complete adaptations and conditional sweeps, including diagnostics of training-induced cross-module interaction.

Local quadratic interpretation. Let $\delta_\theta = s^{\mathcal{M}}(\theta) \odot \mu$. Neglecting first-order terms in local expansions around W_0 and $W_0 + \mu$ gives the surrogates

$$\mathcal{L}_{\text{up}}(W_0 + \delta_\theta) \approx \mathcal{L}_{\text{up}}(W_0) + \tfrac{1}{2}\delta_\theta^\top H_{\text{up}} \delta_\theta, \tag{10}$$

$$\mathcal{L}_{\text{task}}(W_0 + \delta_\theta) \approx \mathcal{L}_{\text{task}}(W_0 + \mu) + \tfrac{1}{2}(\delta_\theta - \mu)^\top H^{(a)}_{\text{task}} (\delta_\theta - \mu), \tag{11}$$

where $H_{\text{up}} = \text{diag}(F^{\text{up}})$ and $H^{(a)}_{\text{task}} = \text{diag}(h \odot a)$ are diagonal sensitivity proxies. With $\nu = 1$, $J_{\mathcal{M}}$ sums the retention and suppression costs along the specified response path. This is a fixed-response model for ranking candidate strengths. Curvature is represented by gradient-square proxies, and the omitted first-order terms can be nonzero: the KL-regularized probe mean need not be a task-loss stationary point, nor must upstream gradients vanish at W_0. The resulting scores are therefore surrogate costs, not predictions of losses after full adaptation.

4 Experiments

We evaluate *effectiveness* through complete configurations and conditional-sweep regret; *generalization* across expert parameterizations, LoRA ranks, and backbones; and *mechanism* through component ablations and interaction diagnostics. Sweeps provide evaluation data only. Retrospective analyses use sweeps that predate the fixed controller specification; in prospectively frozen tests, predictions are recorded before the corresponding jobs (Appendix A).

4.1 Experimental Setup

We evaluate LLaVA-1.5-7B (Liu et al., 2024) on COCO and Flickr30k captioning (CIDEr) and IconQA (accuracy), with $5{,}000$ training samples per task. Forgetting F is the drop from the zero-shot mean accuracy of 61.97 on GQA, OKVQA, OCRVQA, and TextVQA. We report F and the downstream metric T separately, using $S = T - F$ to summarize their trade-off within each task. LoRASculpt uses a rank-32 LoRA expert. Model Tailor and the CGM hard-mixing operator (denoted CGM$^+$ here) share a dense expert trained by updating all parameters in the projector and last 12 LLM layers; "full-parameter" refers to this scope. The vision encoder is frozen.

One probe per task uses $2{,}000$ samples and $1{,}000$ steps, with rank-4 LoRA for the LLM; its outputs are shared across operators. Controller constants stay fixed. We compare complete configurations with published operator settings and evaluate strength selection on prespecified conditional grids. Uniform-random expectations use each axis's eligible grid and exclude points added after the primary evaluation. Appendices A, B, and C give implementation details, grid conventions, and cost measurements.

4.2 Main Results

End-to-end comparison with published operator settings. On LLaVA, the frozen global controller selects $(0.30, 0.30)$ for COCO, $(0.30, 0.20)$ for IconQA, and $(0.50, 0.30)$ for Flickr30k. All three operators share these pairs. Every complete two-module configuration improves the within-task score over its published reference, giving $9/9$ improvements (Table 1). Model Tailor and CGM$^+$ process the same dense expert. For CGM$^+$, the reference is the published hard operating point ($\alpha = 1$, top-5%); module-specific budgets improve its score by $+2.0$, $+7.0$, and $+3.1$ on COCO, IconQA, and Flickr30k. Appendix H also compares the published soft default and the other published hard point.

Table 1: LLaVA-1.5-7B under the equal-budget protocol. Pairs are (ρ_P, ρ_L); results are forgetting↓/downstream↑. Gray: best measured configuration on the prespecified conditional cross (later fill points excluded); the full two-dimensional grid was not evaluated. CGM$^+$ uses the published hard top-5% reference ($\alpha = 1$). **Bold**: better of reference and ledger for each metric. Other CGM references appear in Appendix H.

Method	Task	Published reference	Frozen ledger pair	Measured sweep best (reference)
LoRASculpt	COCO	$(1.0, 0.1)$: 3.83/115.66	$(0.3, 0.3)$: **2.39/118.73**	$(0.2, 0.1)$: 1.11/121.48
	IconQA	$(1.0, 0.1)$: 3.22/79.37	$(0.3, 0.2)$: **1.18/79.69**	$(0.2, 0.1)$: 0.71/79.96
	Flickr30k	$(1.0, 0.1)$: 3.37/63.80	$(0.5, 0.3)$: **2.04/65.53**	$(0.3, 0.1)$: 1.25/66.00
Model Tailor	COCO	$(0.0, 0.1)$: **1.76**/84.29	$(0.3, 0.3)$: 2.05/**108.47**	$(0.7, 0.3)$: 2.03/108.93
	IconQA	$(0.0, 0.1)$: **1.10**/50.74	$(0.3, 0.2)$: 2.29/**63.66**	$(0.7, 0.1)$: 1.49/64.63
	Flickr30k	$(0.0, 0.1)$: **1.27**/50.42	$(0.5, 0.3)$: 1.69/**65.38**	same
CGM$^+$	COCO	hard 5%: 1.86/105.86	$(0.3, 0.3)$: **0.57/106.61**	$(0.7, 0.3)$: 0.62/107.27
	IconQA	hard 5%: 12.89/**65.01**	$(0.3, 0.2)$: **2.65**/61.78	$(0.7, 0.2)$: 2.76/64.50
	Flickr30k	hard 5%: 1.83/62.28	$(0.5, 0.3)$: **0.56/64.06**	$(0.7, 0.3)$: 0.58/65.22

Robustness to the reporting scalarization. Five of the nine configurations improve both forgetting and downstream performance. Reweighting the reporting utility to $S_\lambda = T - \lambda F$ preserves $8/9$ improvements at $\lambda = 0.25$ and $9/9$ at $\lambda \in \{0.5, 1, 2, 4\}$. The exception is CGM$^+$–IconQA: its 3.23-point downstream decrease is offset by a 10.24-point forgetting reduction when $\lambda > 0.32$. These weights affect reporting only; the controller is unchanged.

Conditional strength responses. Figure 2 shows all six COCO axes, with the other module fixed as indicated in each panel. LoRASculpt's projector has an interior optimum, while the post-hoc projector responses rise toward broader plateaus. The LLM axes also favor different retention budgets. The ledger reaches the best measured Model Tailor LLM point and is 0.3 score points below the CGM$^+$ LLM best; its largest COCO error is 3.4 points on the Model Tailor projector axis. Appendix H, Figure 4, includes all tasks.

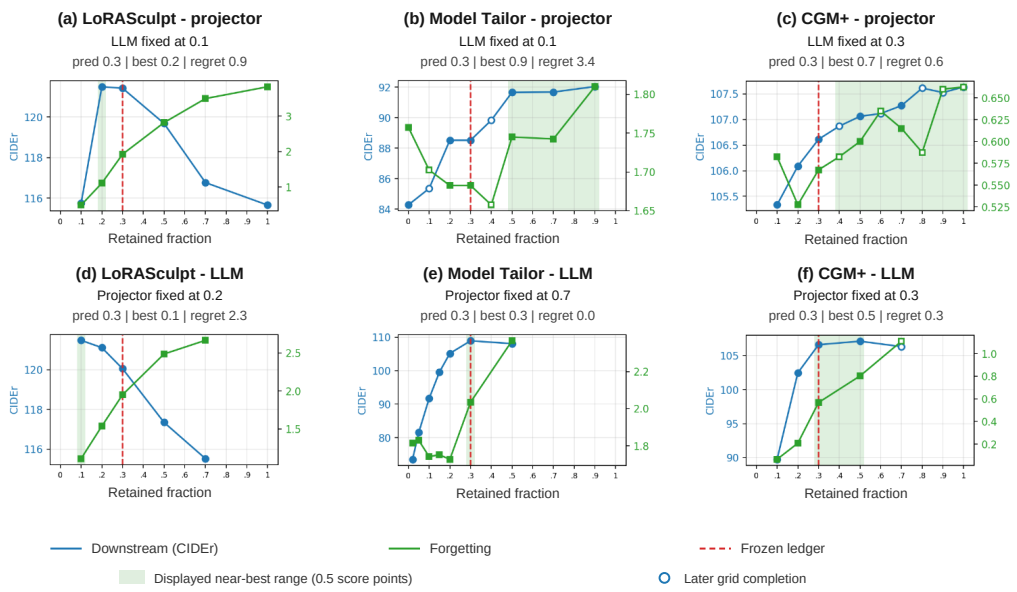


Figure 2: **Conditional strength responses on COCO.** Columns: LoRASculpt, Model Tailor, and CGM$^+$; rows: projector and LLM. Blue denotes downstream CIDEr, green forgetting, and red dashed lines the frozen ledger strengths. Each panel specifies the other module's fixed setting. Best points and regret use $S = T - F$ on prespecified grids. Shading marks sampled locations within 0.5 points of the axis best (not a confidence interval); hollow markers denote supplementary evaluations. Grids and all eighteen axes appear in Appendices B and H, respectively.

Strength-selection quality. Regret is the score gap between the best eligible point and the selected point on a conditional axis. Across the 18 prespecified conditional axes, the ledger's cumulative regret is 7.4 versus a uniform-random expectation of 21.3 on COCO, 6.6 versus 11.5 on IconQA, and 2.8 versus 12.9 on Flickr30k. Each total covers six axes within one task. The ledger has lower regret than the per-axis random expectation on 16 of 18 axes, but its three exact hits are below the random expectation of 3.70. These results support coarse selection of useful regions. The comparisons are analytic and conditional on the other module's setting; the full joint grid is unmeasured. Appendix B gives the grids and complete comparisons.

The LLaVA control $(0.3, 0.3)$, chosen after observing the COCO recommendation, often performs similarly within the predicted plateau. For Model Tailor, task-specific ledger strengths improve on this control by 1.45 points on IconQA and 1.44 on Flickr30k. This post-hoc control tests the flatness of the predicted region (Appendix B).

4.3 Transfer across Parameterizations, Ranks, and Backbones

From low-rank probe to full-parameter executors. The primary Model Tailor and CGM$^+$ results apply strengths from the rank-4 LLM probe to dense updates of the projector and last 12 LLM layers. The CGM$^+$ recommendations were frozen before the full-parameter merges were evaluated, providing a prospective test of transfer across expert parameterizations.

Table 2: Score gains over rank-matched defaults when rank-4 probe strengths selected for rank-32 executors are reused at ranks 64/128. CGM$^+$ uses LoRA experts; Model Tailor has no LoRA-rank axis.

Method-task	Rank 64	Rank 128
LoRASculpt-COCO	+5.9	+2.9
LoRASculpt-IconQA	+3.1	+1.3
LoRASculpt-Flickr	+0.5	+3.0
CGM$^+$-COCO	+10.1	+10.3
CGM$^+$-IconQA	+4.3	n/a
CGM$^+$-Flickr	+9.2	+13.6

Table 3: Qwen captioning (seed 42): forgetting↓/CIDEr↑; ΔS is ledger minus reference. The CGM$^+$ reference maps the published hard top-5% rule to Qwen and is evaluated here. **Bold**: better result for each metric. Protocol and IconQA: Appendix E; Model Tailor repetitions: Appendix K.

Task	Operator	Reference setting	Ledger (ρ_P, ρ_L)	ΔS
COCO	LoRASculpt	−1.01/132.73	$(0.3, 0.1)$: **−1.43/133.31**	+1.0
	Model Tailor	**2.32/129.21**	$(0.3, 0.1)$: 2.48/128.63	−0.8
	CGM$^+$	2.11/125.11	$(0.4, 0.2)$: **1.04/130.17**	+6.1
Flickr30k	LoRASculpt	−1.41/66.81	$(0.1, 0.1)$: **−1.57/68.24**	+1.6
	Model Tailor	**2.92**/72.80	$(0.1, 0.1)$: 2.93/**73.90**	+1.1
	CGM$^+$	2.88/67.99	$(0.3, 0.2)$: **1.52/75.77**	+9.1

Transfer across LoRA rank. We reuse the rank-4 probe's strength pairs selected for rank-32 executors at ranks 64 and 128. All 11 evaluable configurations improve over rank-matched defaults (Table 2). The remaining configuration, CGM$^+$–IconQA at rank 128, has an unprocessed expert with 61.87 points of forgetting and fails the usable-expert prerequisite. This experiment measures transfer against rank-matched defaults; new sweeps at each rank were not performed.

Cross-backbone transfer. On Qwen2.5-VL-7B, we retain $\kappa = \nu = 1$, tensor-relative gating, and the module-specific sensitivity estimators. We recompute F^{up} once for this backbone and fit one probe per task. Five of six captioning configurations improve over their operator references (Table 3). Model Tailor–COCO remains below its reference across three calibration-sampling seeds (Appendix K). On IconQA, the unprocessed expert has negligible forgetting, but CGM$^+$ severely degrades the answer format. We report this failure separately in Appendix E.

4.4 Decision Mechanisms and Ablations

Probe sufficiency and component roles. Replacing the probe mean with the final expert update preserves 7/9 Qwen recommendations and shifts the remaining two by one grid step. The executed CGM$^+$–Flickr30k alternative scores 0.6 points below the probe choice, supporting the probe's use for coarse budget selection. Its learned variance is nearly task-invariant and excluded from selection. KL regularization stabilizes fitting under the original probe configuration; a point probe with a lower learning rate converges to similar choices. Gate and sensitivity substitutions change

recommendations, although substituting unit weights for F^{up} changes both its scale and coordinate pattern. Appendices F and G report the component and estimator comparisons.

Pricing beyond gate-only selection. Gate-only uses the same probe, tensor-relative gate, and candidate grid, rounding each module's active-coordinate fraction to the nearest grid point (ties choose the lower point). It omits A and B and does not rank candidate budgets by their retention costs. On LLaVA, it matches 15/18 ledger decisions. The three differing Flickr30k projector axes have cumulative regret 3.63 for gate-only and 2.03 for the ledger; including all six Flickr30k axes gives 4.41 and 2.81, respectively (Appendix B). On Qwen, gate-only selects $(0.3, 0.3)$ for both captioning tasks, and the ledger improves all six complete configurations (Table 4). The four LoRASculpt/Model Tailor gains average 4.32–9.53 points over three seeds and remain positive in every repetition, with lower forgetting and higher downstream performance.

LoRASculpt repetitions vary the training seed; Model Tailor repetitions vary calibration sampling with the expert fixed. CGM$^+$ gains are 0.28 and 1.95 points with fixed expert and Fisher inputs. For COCO LoRASculpt/Model Tailor, only LLM retention changes; both modules change on Flickr30k. Pricing adds CPU computation with negligible additional GPU time. Its contribution beyond gate-only therefore differs across the tested backbones. This comparison evaluates the combined pricing rule; separate effects of A and B require further ablation. Appendix K gives all results, seed protocols, and the IconQA format failure.

Table 4: Qwen captioning scores $S = T - F$; ΔS is ledger minus gate-only. †Mean $\pm$ sample SD over three seeds: training for LoRASculpt and calibration sampling with a fixed expert for Model Tailor. CGM$^+$ uses fixed expert/Fisher inputs. Gate-only selects $(0.3, 0.3)$; ledger pairs are in Table 3. Full results: Appendix K.

Task	Operator	Gate-only S	Ledger S	ΔS
COCO	LoRASculpt†	126.22 ± 0.72	134.75 ± 0.07	$+8.53$
	Model Tailor†	121.81 ± 0.19	126.13 ± 0.02	$+4.32$
	CGM$^+$	128.86	129.14	$+0.28$
Flickr30k	LoRASculpt†	65.23 ± 1.17	70.08 ± 0.24	$+4.85$
	Model Tailor†	61.03 ± 0.37	70.56 ± 0.38	$+9.53$
	CGM$^+$	72.30	74.25	$+1.95$

Module heterogeneity and interaction. Controlled interventions show how the modules interact during training. In an IconQA LoRASculpt 2×2 control, the effect of tightening LLM constraints on the upstream score changes by about 11.0 points depending on whether the projector is constrained. Such interaction is absent from the separable pricing objective. MME diagnostics and module-level efficiency comparisons characterize further task and module differences (Appendix D).

5 Conclusion

Task–Drift Ledger selects module-specific anti-forgetting strengths from a short task probe while preserving each operator's parameter-selection rule. It improves the within-task score over references in all nine LLaVA configurations and all 11 evaluable higher-rank configurations, and in five of six Qwen captioning configurations without refitting controller constants. One probe costs about 0.5 GPU-hours, compared with 26.6–79.8 GPU-hours for a training-time strength sweep. Conditional regret is below the uniform-random expectation on all three LLaVA tasks, supporting low-cost selection of useful strength regions.

Pricing contributes differently across backbones: gate-only reproduces 15 of 18 LLaVA decisions, while the ledger improves all six Qwen captioning configurations over gate-only, with positive LoRASculpt and Model Tailor gains across the reported repetitions. The current surrogate omits training-induced module interactions and does not guarantee exact optimum recovery. Future work includes joint module pricing and sequential continual learning, where retention references and sensitivity estimates must account for previously learned tasks as the model evolves.

6 LIMITATIONS

Calibration requires upstream reference data or stored sensitivity for the same backbone, even though downstream optimization uses no upstream replay. The method applies to parameter-space operators with module-level controls and a usable retention proxy; post-hoc operators also require a usable expert. Our evaluation covers a finite set of tasks and backbones with frozen vision encoders. Transfer to other architectures or long task sequences remains untested.

The fixed-response, separable objective omits training-induced module interactions and offers no guarantee of optimal strength. The learned probe variance has not been validated as a calibrated estimate of task-specific uncertainty. Pricing gains differ across backbones, and the gate-only comparison evaluates the combined rule without isolating A and B. Qwen repetitions cover LoRASculpt training and Model Tailor calibration sampling; they do not cover probe refitting or the full adaptation pipeline. Near-zero forgetting can coexist with severe answer-format damage after intervention, and automatic detection or fallback has not been validated.

Random comparisons use analytic expectations on conditional axes, with no full joint-grid comparison. The 53–$160\times$ cost ratio applies to training-time calibration and excludes the reusable upstream pass and final execution. Post-hoc operators have separate cost accounting (Appendix C).

AI USE STATEMENT

We used generative AI tools, including ChatGPT/Codex, for translation, language editing, literature search, feedback on experimental design and interpretation, and assistance with figure-generation and LaTeX code. The authors reviewed AI-assisted claims, citations, numerical summaries, and figures against the manuscript, cited sources, and available experimental records. All reported measurements come from model evaluations. The authors are responsible for the study design, experiments, analysis, and final paper.

ETHICS STATEMENT

This study uses existing benchmarks to evaluate retention of previously acquired capabilities during multimodal adaptation. It involves no new human-subject study or dataset release. Retaining capabilities can also preserve risks inherited from the underlying models and data, including biased or harmful outputs. We also observe that an intervention can damage the required answer format despite negligible forgetting. Applications beyond the evaluated tasks should account for these risks.

REPRODUCIBILITY STATEMENT

Section 4.1 describes the models, tasks, metrics, and evaluation protocol. Appendices A, B, and C provide probe and operator settings, eligible strength grids, and calibration-cost accounting. Section 4.2 reports frozen controller recommendations and their results; the appendices document transfer protocols, ablations, seed-level results, and failure cases. We distinguish conditional sweeps from complete configurations and single-run measurements from repeated evaluations.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Sungmin Cha and Kyunghyun Cho. Hyperparameters in continual learning: A reality check. *Transactions on Machine Learning Research*, 2025. URL `https://openreview.net/forum?id=hiiRCXmbAz`.

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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 7982–8001. PMLR, 2025.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. doi: 10.1109/TPAMI.2021.3057446.

Zijian Gao, Wangwang Jia, Xingxing Zhang, Pengfei Qian, Tao Sun, Bo Ding, Yong Dou, Huaimin Wang, and Kele Xu. Many: Merge anything for multimodal continual instruction tuning. arXiv preprint arXiv:2604.14016, 2026. URL https://arxiv.org/abs/2604.14016.

Baichuan Huang, Ananth Balashankar, and Amir Aminifar. BEFT: Bias-efficient fine-tuning of language models in low-data regimes. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38833–38851, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1799. URL https://aclanthology.org/2026.acl-long.1799/.

Wenke Huang, Jian Liang, Zekun Shi, Didi Zhu, Guancheng Wan, He Li, Bo Du, Dacheng Tao, and Mang Ye. Learn from downstream and be yourself in multimodal large language models fine-tuning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 25486–25499. PMLR, 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.

Thomas L. Lee, Sigrid Passano Hellan, Linus Ericsson, Elliot J. Crowley, and Amos Storkey. Hyperparameter selection in continual learning. *arXiv preprint arXiv:2404.06466*, 2024.

Jian Liang, Wenke Huang, Guancheng Wan, Qu Yang, and Mang Ye. Lorasculpt: Sculpting lora for harmonizing general and specialized knowledge in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26170–26180, 2025.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024.

Pan Lu, Liang Qiu, Jiaqi Chen, Tony 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 *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2021.

Michael S. Matena and Colin A. Raffel. Merging models with fisher-weighted averaging. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Wish Suharitdamrong, Tony Alex, Muhammad Awais, and Sara Atito. CoLA: Cross-modal low-rank adaptation for multimodal downstream tasks. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Yongxian Wei, Runxi Cheng, Weike Jin, Enneng Yang, Li Shen, Lu Hou, Sinan Du, Chun Yuan, Xiaochun Cao, and Dacheng Tao. Optmerge: Unifying multimodal llm capabilities and modalities via model merging. arXiv preprint arXiv:2505.19892, 2026. URL https://arxiv.org/abs/2505.19892.

Jinglong Yang, Jiaxuan He, Wenjian Huang, Zhan Zhuang, and Jianguo Zhang. Curvature-guided mixing for mllm adaptation. *arXiv preprint arXiv:2606.24963*, 2026.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. Investigating the catastrophic forgetting in multimodal large language model fine-tuning. In *Conference on Parsimony and Learning*, volume 234 of *Proceedings of Machine Learning Research*, pp. 202–227. PMLR, 2024.

Didi Zhu, Zhongyisun Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Chao Wu, and Kun Kuang. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62581–62598. PMLR, 2024.

A Implementation Details

Evaluation chronology. Retrospective analyses use sweeps that predate the fixed controller specification. For prospectively frozen tests, predictions were written to timestamped files before the corresponding jobs were submitted; the files and job records accompany the code. They were not registered with an external service. The primary full-parameter CGM$^+$ predictions were frozen before any full-parameter merge was evaluated. Additional grid points are identified separately and excluded from the primary counts.

Backbone and tasks. The primary backbone is LLaVA-1.5-7B (Liu et al., 2024), with a frozen CLIP vision encoder, a two-layer MLP projector, and a Vicuna-7B LLM. Upstream capability is the mean accuracy on GQA, OKVQA, OCRVQA, and TextVQA; the zero-shot reference under the equal-budget protocol is 61.97, and *forgetting* is the drop from this value. Downstream tasks are COCO Captioning (CIDEr), IconQA (accuracy), and Flickr30k Captioning (CIDEr). MME is used only as an auxiliary perception/cognition diagnostic.

Equal-budget protocol and operators. Each downstream task contributes exactly $5{,}000$ training samples; training and evaluation sets are disjoint. Under this matched budget, naive rank-32 LoRA adaptation forgets 4.03 points on COCO, 7.14 on IconQA, and 3.40 on Flickr30k (the full parameter experts used by the post-hoc operators forget 2.18, 14.41, and 2.36), so the cross task differences are not explained by training set size alone. We evaluate LoRASculpt, Model Tailor, and CGM$^+$, spanning training time sparse regularization, post-hoc patching, and merge-time hard retention. LoRASculpt uses a rank-32 LoRA expert. Model Tailor and CGM$^+$ act on the same full parameter expert, in which the last 12 LLM layers and projector are finetuned, so the post-hoc comparison isolates the operator and strength from the underlying task drift. The probe itself is always rank-4 and is run once per task; its output is shared by all operators.

Controller settings, pricing scope, and grids. Each probe uses $2{,}000$ samples, $1{,}000$ steps, $\beta = 10^{-3}$, and $\sigma_p = 0.1$. F^{up} is computed once per backbone and reused across tasks. LLM pricing covers layers 20-31 on LLaVA, the layers that Model Tailor and CGM act on; LoRASculpt attaches LoRA to all 32 layers, and re-pricing over layers 0-31 leaves every projector strength and the IconQA and Flickr30k LLM strengths unchanged, moving only COCO from 0.30 to 0.20 with a 0.1% difference in J (executed: $2.08/120.46$ against $2.39/118.73$ at the frozen 0.30 and $1.92/121.41$ at 0.10, i.e. one step along the same monotone axis). This check suggests limited sensitivity to the pricing scope in these settings; we retain the original scope. The prespecified candidate grid is the prospectively fixed $\{0.1, 0.2, 0.3, 0.5, 0.7\}$ on both axes, extended by the end points an operator natively supports (0 for Model Tailor's reset, $0.9/1.0$ where the axis is defined there). After the primary evaluation we filled the post-hoc operators' axes to a uniform 0.1 step (0.1-1.0; Figure 4 and Table 16); these fill cells are reported for shape but enter no hit, regret, or "sweep best" count. We report forgetting$\downarrow$ and the native downstream metric$\uparrow$. When a single scalar is needed within a task, we use $S = \text{downstream} - \text{forgetting}$; it is not compared across tasks with different downstream units.

Upstream access. The ledger uses upstream reference data to compute F^{up} at the pre-trained parameters once per backbone; a stored estimate can be reused across downstream tasks. These

Table 5: Axis-level strength-selection baselines. Random entries are analytic expectations on eligible grid points; other entries are observed counts. LS/MT use published scalar strengths on the same conditional axes, not new joint-configuration experiments. The CGM subset uses full-parameter experts; its dagger-marked additional points are excluded.

Axes	Selector	n	E	A	M
LS / MT	Published strength	12	4	0	8
LS / MT	Ledger	12	2	5	5
CGM$^+$	Uniform random (expected)	6	1.35	1.60	3.05
CGM$^+$	Ledger	6	1	3	2
All	Uniform random (expected)	18	3.70	5.12	9.18
All	Ledger	18	3	8	7

batches are not replayed during downstream adaptation. Backbone- and operator-specific statistics must be kept distinct: the Qwen calibration uses Qwen-specific inputs (Appendix E), while the native CGM statistics follow its expert protocol (Appendix H). On LLaVA, F^{up} is estimated on mini-batches of the LLaVA-665k instruction-tuning mixture, i.e. the backbone's own training data; the upstream evaluation uses GQA testdev-balanced, OKVQA val, OCRVQA test, and TextVQA val, which are disjoint from the training splits contained in that mixture. On Qwen, F^{up} is estimated on $2{,}000$ *training*-split examples of each of the four upstream tasks, formatted with the evaluation instruction so that the reference answers sit on the instruct model's predictive distribution; the evaluation items are not used.

B Calibration Baselines and Per-Axis Results

The complete configurations in Table 1 and the conditional axes below answer different questions. The first compares the ledger-selected pair with the published setting under the same task and operator. The second holds one module fixed and evaluates the strength of the other. A default coordinate on a conditional axis is therefore not necessarily the published joint configuration.

Strength-selection baselines. For the 3 methods $\times 3$ tasks $\times 2$ modules, E denotes the exact best grid point, A one adjacent grid point, and M all other outcomes, giving $3/8/7$ for the ledger. The published strengths on the 12 LoRASculpt/Model Tailor axes give $4/0/8$, compared with $2/5/5$ for the ledger on those same axes (Table 5). Thus improved complete configurations do not imply more exact axis hits. CGM's published hard top-5% point is evaluated as a complete configuration in Table 1; it is not an eligible coordinate of the prespecified conditional module grids and is excluded from the axis-level default count.

For uniform random selection on axis j, let K_j be the number of eligible points and d_j the number adjacent to its best point. With a single designated best point, $d_j = 1$ at a boundary and $d_j = 2$ in the interior. The expected counts over n axes are

$$\mathbb{E}[E] = \sum_{j=1}^{n} \frac{1}{K_j}, \qquad \mathbb{E}[A] = \sum_{j=1}^{n} \frac{d_j}{K_j}, \qquad \mathbb{E}[M] = n - \mathbb{E}[E] - \mathbb{E}[A]. \tag{12}$$

Evaluated on the eighteen graded axes with their actual eligible grids (Appendix B, Table 7), the expectation is $3.70/5.12/9.18$. The ledger's $3/8/7$ shows no exact-hit advantage (3 observed against 3.70 expected); its exact-or-adjacent count (11) is above the expectation (8.8), a descriptive difference for which we report no significance test. Regret separates the two more clearly and is reported per task because CIDEr and accuracy points are not commensurable: over the six axes of each task, the ledger's regret is 7.4 against a random expectation of 21.3 on COCO, 6.6 against 11.5 on IconQA, and 2.8 against 12.9 on Flickr30k, and it is below the per-axis random expectation on 16 of the 18 axes (above on the Model Tailor COCO projector axis; tied on the CGM$^+$ Flickr30k projector axis). The $9/9$ result remains a paired comparison with published settings, not a win rate against random joint configurations, because the conditional sweeps do not measure every joint pair.

Published-strength baseline. The LoRASculpt default is $(1.0, 0.1)$ and the Model Tailor default is $(0.0, 0.1)$. Applying their scalar strengths to the best-point rankings in Table 6, all six projector

Table 6: Complete 18-axis accounting for the frozen global arm on LLaVA. E/A/M denote exact, adjacent by one grid step, and miss: $3/8/7$. Regret is the combined score (downstream−forgetting) of the axis best minus that of the predicted cell, in task points (median 0.8, maximum 3.4). Complete configurations and their published references are compared in Table 1. † The COCO CGM$^+$ projector miss lies on a plateau whose total score differences are small; no seed-level significance is established. LS: LoRASculpt; MT: Model Tailor. The CGM$^+$ block uses full-parameter experts and was prospectively frozen.

Method	Module	Task	Pred.	Best	Regret	Method	Module	Task	Pred.	Best	Regret
LS	Proj.	COCO	.3	.2 (A)	0.9	MT	Proj.	COCO	.3	.9 (M)	3.4
LS	Proj.	Icon	.3	.2 (A)	1.1	MT	Proj.	Icon	.3	.7 (M)	1.1
LS	Proj.	Flickr	.5	.3 (A)	0.5	MT	Proj.	Flickr	.5	.9 (M)	0.4
LS	LLM	COCO	.3	.1 (M)	2.3	MT	LLM	COCO	.3	.3 (E)	0.0
LS	LLM	Icon	.2	.1 (A)	0.2	MT	LLM	Icon	.2	.1 (A)	0.8
LS	LLM	Flickr	.3	.1 (M)	0.8	MT	LLM	Flickr	.3	.3 (E)	0.0
CGM$^+$	Proj.	COCO	.3	.7 (M†)	0.6	CGM$^+$	LLM	COCO	.3	.5 (A)	0.3
CGM$^+$	Proj.	Icon	.3	.7 (M)	2.6	CGM$^+$	LLM	Icon	.2	.1 (A)	0.8
CGM$^+$	Proj.	Flickr	.5	.7 (A)	1.1	CGM$^+$	LLM	Flickr	.3	.3 (E)	0.0

choices are misses. The three LoRASculpt LLM defaults and the Model Tailor IconQA LLM default are exact; the other two Model Tailor LLM defaults are misses because the grid includes 0.2 between their default 0.1 and the best 0.3. This yields E/A/M of $4/0/8$ on the 12 axes. The ledger gives $2/5/5$ on the same subset. These counts use existing grid locations and require no new training. CGM's hard top-5% default, compared at the joint-configuration level in Table 1, lies outside the eligible conditional module grids and is excluded from this axis-level default count. Its published soft-mixing default $\alpha = 0.2$ also has no direct hard-retention coordinate.

Uniform random strength. For axis j, let Γ_j be its ordered eligible grid after excluding points added after the primary evaluation, let $K_j = |\Gamma_j|$, and let b_j be the index of its designated best point. A random baseline selects one element of Γ_j uniformly, without fitting a probe or evaluating several candidates before choosing. Adjacency means one step in the ordered grid, not a fixed numerical distance. Thus $d_j = \mathbf{1}[b_j > 1] + \mathbf{1}[b_j < K_j]$, giving Eq. (12). The expected counts require no model reruns once the eligible grids and best indices are known. If all tied maxima are counted as exact, the exact probability becomes the fraction of best points, and the adjacent probability counts their distinct neighbors excluding the best set; that convention would also require recomputing the observed counts.

The six full-parameter CGM axes are enumerated in Table 16. Under that table's exclusion flags, each projector grid is $\{0.1, 0.2, 0.3, 0.5, 0.7\}$ and each LLM grid is $\{0.1, 0.2, 0.3, 0.5\}$: the LLM point 0.7 is marked as an additional evaluation and is excluded, as are all other dagger-marked points. The projector optima are all at the upper boundary; the LLM optima are at the upper boundary for COCO, lower boundary for IconQA, and in the interior for Flickr30k. Hence

$$\mathbb{E}[E] = 3/5 + 3/4 = 1.35, \qquad \mathbb{E}[A] = 3/5 + (1+1+2)/4 = 1.60, \qquad \mathbb{E}[M] = 3.05.$$

The observed ledger counts are $1/3/2$ on this subset. These are expected counts, not the result of a simulated baseline or a significance test.

The same computation over all eighteen axes uses the grids on which the E/A/M grades were assigned: LoRASculpt projector $\{0.1, 0.2, 0.3, 0.5, 0.7, 1.0\}$ (LLM at the published 0.1), LoRASculpt LLM $\{0.1, 0.2, 0.3, 0.5, 0.7\}$ (projector at the axis best), Model Tailor projector points 0, 0.2, 0.3, 0.5, 0.7, and 0.9 (LLM patch at the published 10%), Model Tailor LLM $\{0.1, 0.2, 0.3, 0.5\}$ (projector at 0.7; the finer $0.02/0.05/0.15$ points of that sweep are excluded because the grades were assigned on the coarse grid), and the CGM$^+$ grids above. This gives $\mathbb{E}[E] = 3.70$, $\mathbb{E}[A] = 5.12$, $\mathbb{E}[M] = 9.18$ against the ledger's $3/8/7$. Table 7 evaluates Eq. (14) on the same grids for every axis. Regret is in the task's own units (CIDEr for COCO and Flickr30k, accuracy points for IconQA) and is summed only within a task. The ledger is below the per-axis random expectation on 16 of 18 axes, above it on the Model Tailor COCO projector axis (3.39 versus 2.52), and tied on the CGM$^+$ Flickr30k projector axis (1.14 versus 1.14). The last two columns give the gate-only selector discussed in Sec. 4.4, which returns the grid point nearest to the gated fraction of coordinates and ignores A and B.

Table 7: All eighteen graded axes: eligible grid size K, measured best, frozen ledger prediction with grade and regret, expected regret of uniform random selection on the same grid, and the gate-only selector. Regret is in task units and is summed only within a task. LS: LoRASculpt; MT: Model Tailor.

Task	Module	Op.	K	Best	Ledger	Grade	R^{ledger}	$\overline{R}^{\text{rand}}$	Gate-only	R^{gate}
COCO	Proj.	LS	6	.2	.3	A	0.89	4.19	.3	0.89
COCO	Proj.	MT	6	.9	.3	M	3.39	2.52	.3	3.39
COCO	Proj.	CGM$^+$	5	.7	.3	M	0.61	0.76	.3	0.61
COCO	LLM	LS	5	.1	.3	M	2.27	3.22	.3	2.27
COCO	LLM	MT	4	.3	.3	E	0.00	5.40	.3	0.00
COCO	LLM	CGM$^+$	4	.5	.3	A	0.27	5.24	.3	0.27
					COCO sum (CIDEr)		7.43	21.32		7.43
IconQA	Proj.	LS	6	.2	.3	A	1.15	1.34	.3	1.15
IconQA	Proj.	MT	6	.7	.3	M	1.13	3.27	.3	1.13
IconQA	Proj.	CGM$^+$	5	.7	.3	M	2.60	3.18	.3	2.60
IconQA	LLM	LS	5	.1	.2	A	0.18	0.22	.2	0.18
IconQA	LLM	MT	4	.1	.2	A	0.77	2.07	.2	0.77
IconQA	LLM	CGM$^+$	4	.1	.2	A	0.76	1.46	.2	0.76
					IconQA sum (acc. points)		6.59	11.54		6.59
Flickr30k	Proj.	LS	6	.3	.5	A	0.54	3.06	.3	0.00
Flickr30k	Proj.	MT	6	.9	.5	M	0.35	2.38	.3	2.60
Flickr30k	Proj.	CGM$^+$	5	.7	.5	A	1.14	1.14	.3	1.03
Flickr30k	LLM	LS	5	.1	.3	M	0.77	1.39	.3	0.77
Flickr30k	LLM	MT	4	.3	.3	E	0.00	2.61	.3	0.00
Flickr30k	LLM	CGM$^+$	4	.3	.3	E	0.00	2.29	.3	0.00
					Flickr30k sum (CIDEr)		2.81	12.87		4.41

Predicted plateau and fixed-pair controls. The ledger produces an objective curve rather than only a discrete argmin. For a priced module $q \in \{P, L\}$ and task D, let $J_{q,D}$ denote the corresponding module-restricted objective and Θ_q its candidate grid. We define the predicted plateau as

$$\mathcal{P}_q(D;\epsilon) = \left\{ \theta \in \Theta_q : \frac{J_{q,D}(\theta) - J^{\min}_{q,D}}{J^{\min}_{q,D}} \le \epsilon \right\}, \qquad J^{\min}_{q,D} = \min_{\theta \in \Theta_q} J_{q,D}(\theta), \tag{13}$$

and use $\epsilon = 5\%$. On LLaVA, the ledger pair and the post-hoc constant $(0.3, 0.3)$ lie in the same predicted plateau on every axis on which they differ. The constant was chosen after observing the controller's COCO output, so it is an inside-plateau control rather than an independent selector, and the ledger itself predicts little separation. Consistent with this prediction, the measured differences for LoRASculpt and CGM$^+$ are at most 0.9 score points. Model Tailor is the exception: the ledger gains 1.45 and 1.44 points on IconQA and Flickr30k, respectively, consistent with the steeper measured response of this operator in that interval (Table 8).

By contrast, fixed controls with the LLM strength at 0.1 lie outside the predicted plateau, with J excesses of 27%-809% on the affected axes. They lose 8.1-19.6 points on the measured post-hoc captioning cells, although some LoRASculpt and IconQA cells improve, exposing a remaining training time or operator-interaction error. Thus, the fixed-pair experiment tests whether the ledger identifies a robust operating region; it does not claim that every task has a uniquely optimal strength or that per-task adjustment always improves over any constant inside that region. On Qwen, the gate-only selector chooses $(0.3, 0.3)$ for both captioning tasks; these executed points provide a comparison with that pair on those tasks, although they were selected by a task-dependent rule and are not a separate prospectively fixed all-task constant-pair experiment (Sec. 4.4).

Predicted-plateau controls. Table 8 separates a fixed pair inside the predicted plateau of Eq. (13) from controls outside that plateau. The inside control $(0.3, 0.3)$ was chosen after observing the ledger's COCO output and therefore tests the predicted flatness of the operating region rather than acting as an independent selector. The outside controls set the LLM strength to 0.1 and have substantially larger frozen ledger objectives on the affected axes. Only equal-budget measurements are reported, and no missing cell is imputed.

Table 8: Predicted-plateau controls on LLaVA (score = downstream − forgetting, task units). References match Table 1, including the published hard top-5% point for CGM$^+$. The inside control is $(0.3, 0.3)$. The outside control is $(0.3, 0.1)$ except for CGM$^+$-Flickr30k, where the measured pair is $(0.5, 0.1)$. Differences are control minus ledger.

Operator	Task	Reference	Ledger	Inside	Δ_{in}	Outside pair	Score / Δ_{out}
LoRASculpt	COCO	111.84	116.35	116.35	0.00	$(0.3, 0.1)$	119.49 / +3.14
LoRASculpt	IconQA	76.15	78.51	78.39	−0.12	$(0.3, 0.1)$	78.11 / −0.40
LoRASculpt	Flickr30k	60.44	63.49	63.97	+0.48	$(0.3, 0.1)$	64.75 / +1.26
Model Tailor	COCO	82.53	106.42	106.42	0.00	$(0.3, 0.1)$	86.84 / −19.58
Model Tailor	IconQA	49.65	61.37	59.92	−1.45	$(0.3, 0.1)$	62.01 / +0.64
Model Tailor	Flickr30k	49.15	63.69	62.25	−1.44	$(0.3, 0.1)$	55.15 / −8.53
CGM$^+$	COCO	104.00	106.05	106.05	0.00	$(0.3, 0.1)$	89.70 / −16.35
CGM$^+$	IconQA	52.12	59.13	58.23	−0.91	$(0.3, 0.1)$	59.90 / +0.76
CGM$^+$	Flickr30k	60.45	63.50	63.61	+0.11	$(0.5, 0.1)$	55.44 / −8.06

Regret under the same candidate set. Let $S_j(g)$ be the measured within-task score at strength g with the other module fixed. The expected random regret and ledger regret are

$$\overline{R}_j^{\text{rand}} = \max_{g \in \Gamma_j} S_j(g) - \frac{1}{K_j} \sum_{g \in \Gamma_j} S_j(g), \qquad R_j^{\text{ledger}} = \max_{g \in \Gamma_j} S_j(g) - S_j(\hat{\theta}_j). \tag{14}$$

Table 9 evaluates these quantities from the reported CGM score rows, excluding additional evaluations. Regret is reported per axis in its native task units; we do not pool CIDEr and accuracy into a cross-task mean. The small projector differences, particularly on Flickr30k, should not be interpreted as statistically resolved gains.

Table 9: Analytic random baseline on the six fully enumerated CGM axes. K is the eligible grid size and d the number of neighbors of the best point. Regrets are recomputed from the rounded scores in Table 16; no new adaptations are used.

Task	Module	K	d	$\overline{R}^{\text{rand}}$	R^{ledger}
COCO	Projector	5	1	0.764	0.610
IconQA	Projector	5	1	3.180	2.610
Flickr30k	Projector	5	1	1.146	1.150
COCO	LLM	4	1	5.235	0.270
IconQA	LLM	4	1	1.468	0.770
Flickr30k	LLM	4	2	2.285	0.000

Random comparison for complete configurations. To assess the difficulty of improving a published default, random joint pairs must be drawn from a prespecified set of configurations evaluated under the same operator. For operator-task setting q, let $\mathcal{G}_q$ be that set and S_q^{default} its published-default score. The expected number of improvements over the nine defaults is

$$\mathbb{E}[W_{\text{rand}}] = \sum_{q=1}^{9} \frac{1}{|\mathcal{G}_q|} \sum_{g \in \mathcal{G}_q} \mathbf{1}\left[S_q(g) > S_q^{\text{default}}\right]. \tag{15}$$

Complete score lists for every $\mathcal{G}_q$ are needed to evaluate this expression. Conditional cross-shaped sweeps do not measure all Cartesian pairs, and their missing scores cannot be imputed from the individual axes when modules interact. Accordingly, the current $9/9$ claim remains a paired comparison with published settings, not a demonstrated win rate against uniform random joint selection.

C Calibration Cost

The calibration cost depends on when an operator acts. The Task-Drift Ledger computes F^{up} once per backbone and runs one 2,000-sample, 1,000-step probe per downstream task. On our setup the probe plus the endpoint h pass takes about 0.5 GPU-hours on one GPU (measured on the 7B backbones), roughly one third to one half of a single 5,000-sample, three-epoch fine-tuning run; F^{up} is a one-off

Table 10: Marginal efficiency of the same Model Tailor LLM intervention across tasks. Units are downstream metric points per additional point of upstream forgetting.

Task	Downstream gain	Forgetting cost	Gain / cost
COCO	+17.25	+0.29	59.5
Flickr30k	+6.32	+0.34	18.6
IconQA	+0.16	+2.26	0.07

16-minute pass per backbone. For a training time operator such as LoRASculpt, changing the strength changes the optimization itself, so a K point sweep requires K separate train-and-evaluate jobs. In our pipeline a complete train plus seven-set evaluation cell is about 3.8 GPU hours, so a 7-21 point training time sweep is approximately 26.6-79.8 GPU hours, compared with 0.5 GPU hours for one probe (about 53-160 probe equivalents). For post-hoc operators such as Model Tailor and CGM$^+$, this accounting is intentionally different: one full expert is trained once and reused, so a K-point search costs one adaptation plus K patch/merge construction and evaluation passes rather than K full adaptations. In our pipeline a post-hoc candidate costs about 0.25 GPU-hours to construct (salience or curvature statistics plus the patch or merge) and 2.5-3 GPU-hours to evaluate on the seven sets (36 post-hoc cells took $\approx$ 105 GPU-hours), so a K-point post-hoc search costs $C_{\mathrm{FT}} + K$ (2.8-3.3) GPU-hours. The ledger instead incurs $C_{\mathrm{FT}} + C_{\mathrm{probe}} + C_{\mathrm{post}}$: the selected configuration is still constructed and evaluated once, while the other $K - 1$ candidate passes are avoided. The avoided calibration workload is therefore approximately $(K-1)C_{\mathrm{post}} - C_{\mathrm{probe}}$, rather than the cost of all K candidates. In both cases the controller scores its candidate grid without additional candidate training or evaluation.

These ratios compare the probe stage with the full sweep workload; they are not end-to-end runtime ratios. Executing and evaluating the ledger-selected configuration adds one final job. The upstream-sensitivity pass is a one-time cost per backbone and should be reported separately or amortized across downstream tasks. In the notation of the Introduction, training time sweep cost is $K(C_{\mathrm{FT}} + C_{\mathrm{eval}})$ when candidate evaluation cost C_{eval} is shown explicitly; a reusable post-hoc expert costs $C_{\mathrm{FT}} + KC_{\mathrm{post}}$ under a sweep and $C_{\mathrm{FT}} + C_{\mathrm{probe}} + C_{\mathrm{post}}$ under the ledger.

D Additional Analysis

Task-module heterogeneity. We examine whether the same intervention has comparable effects across tasks and modules. At the same Model Tailor working point, increasing the LLM patch from 10% to 30% yields marginal downstream gain per additional point of forgetting of 59.5 on COCO, 18.6 on Flickr30k, and 0.07 on IconQA. These ratios are in task-specific metric units and should not be interpreted as a unit-independent cross-task efficiency ratio. Within IconQA, the comparable projector intervention has efficiency 35.6, versus 0.07 on the LLM. The same action therefore has strongly task- and module-dependent value.

What the frozen curves look like. We recomputed $A(\rho)$ and $B(\rho)$ of the frozen LLaVA recipe on a 0.05 grid from the stored μ, h, and F^{up}; no model is trained or evaluated. Two regularities hold on all six module-task axes. First, $B(\rho)$ decays steeply: relative to $B(0.1)$ it is 2.7%, 0%, and 3.5% at ρ=0.3 on the COCO/IconQA/Flickr30k projector and 6%, <0.1%, and <0.1% on the LLM. Because the gate is tensor-relative while the retention proxy is a module-global top-ρ, the gated fraction of coordinates (0.28-0.31 projector, 0.24-0.26 LLM) does not guarantee that all gated coordinates are retained at that ρ; the near vanishing of B there is an observation, not an identity. Second, A is flat: $A(0.1)$ is already 0.5-0.6 of $A(1)$, so the coordinates with the largest response carry most of the priced upstream cost and the marginal upstream price of retaining more is small. The ratios $B(0)/A(1)$ range from 2.4 to 8.1 on the projector and 0.9 to 13.6 on the LLM; they do not support a uniform orders-of-magnitude difference. In these stored curves, the minimizer of $A + B$ sits where B has nearly vanished. A gate-only selector that returns the grid point nearest to the gated fraction reproduces the frozen strengths on 15 of 18 axes (Table 7; the three Flickr30k projector axes differ); a selector that returns the smallest grid point at or above the gated fraction gives $3/5/10$ and per-task regret $6.6/8.7/2.8$.

Table 11 reports $\arg\min_\rho [k\,A(\rho) + B(\rho)]$ on the graded grid as the upstream weight k is varied, with $k = 1$ the frozen recipe.

Table 11: Recommendation as a function of the upstream weight k in $k\,A(\rho) + B(\rho)$ (frozen LLaVA inputs, graded grid). Measured axis optima are listed for the training time (LS) and post-hoc (MT / CGM$^+$) operators. A single k cannot match both operator classes on the projector.

Module	Task	k=1	k=2	k=3	k=5	k=10	Best: LS / MT / CGM$^+$
Proj.	COCO	.3	.3	.2	.2	.1	.2 / .9 / .7
Proj.	IconQA	.3	.3	.3	.2	.1	.2 / .7 / .7
Proj.	Flickr30k	.5	.3	.3	.3	.2	.3 / .9 / .7
LLM	COCO	.3	.1	.1	.1	.1	.1 / .3 / .5
LLM	IconQA	.2	.2	.2	.2	.2	.1 / .1 / .1
LLM	Flickr30k	.3	.3	.3	.3	.3	.1 / .3 / .3

What determines the decision. On the stored LLaVA curves (Appendix D), $B(\rho)$ has fallen to at most 6% of its ρ=0.1 value by ρ=0.3 on every axis, close to the gated fraction of coordinates (0.28-0.31 projector, 0.24-0.26 LLM), while $A(\rho)$ rises by less than $2\times$ over the whole grid. The measured $B(0)/A(1)$ ratio varies across axes (Appendix D); the selected strength consequently lands where the gated response is nearly exhausted. A selector that ignores A and B altogether and simply rounds the gated fraction to the grid returns the same strength as the frozen ledger on 15 of 18 axes (grades $4/6/8$ versus $3/8/7$); the three that differ are the Flickr30k projector axes, where the ledger's 0.5 is better for Model Tailor and worse for LoRASculpt and CGM$^+$ (per-task regret 2.8 versus 4.4). On LLaVA, therefore, most of the decision is carried by the tensor-relative gate, and the pricing terms contribute at the margin; we state this as a measured property of the frozen recipe, not as a derivation. The same curves also bound the exchange rate: re-weighting A by $k \approx$ 3-5 moves the LoRASculpt projector axes onto their measured optima, but no single k fits training time and post-hoc operators together (Table 11).

What the probe captures. A direct fidelity audit shows that μ does not reconstruct the eventual expert update coordinate by coordinate: cosine alignment and top-k overlap are at or near random baselines. Nevertheless, the aggregate response carries useful amplitude, ordering, and module/tensor structure for strength pricing. This is consistent with the expert-drift substitution in Appendix G: the ledger can make nearly the same coarse decision without recovering the final optimization path.

Why h is module-specific. The two modules respond differently to substitutions between endpoint and trajectory estimators. On IconQA, the LLM trajectory statistic has only $+0.135$ structural correlation with the endpoint statistic and differs in scale by roughly 5.6×10^4; using the endpoint statistic is therefore the fixed LLM estimator. Conversely, moving the endpoint statistic to the projector collapses task-side mass after mean injection (3.5×10^{-7} on IconQA versus 0.16 on COCO), so the projector retains the trajectory statistic. A separate clean-h diagnostic further shows that removing the mean-gradient component from $\mathbb{E}[g^2]$ hurts projector task ordering. We treat these as empirical estimator choices rather than evidence for a single universal curvature interpretation.

Cross-module interaction. The two module strengths are distinct but not always independent. In a LoRASculpt 2×2 IconQA control, tightening the LLM changes the upstream score by -11.32 points when the projector is free but by only -0.32 when the projector is simultaneously constrained at $\rho_P = 0.30$, an interaction of roughly $+11.0$ points. MME gives a complementary functional dissociation: projector protection recovers $+165$ perception points while cognition changes by only -8, whereas LLM protection primarily restores cognition. These results explain why independently priced modules can miss the exact conditional optimum for training time operators.

Scope and degenerate axes. The ledger predicts *how much* response to retain, not *which* coordinates an operator should keep. Its fixed $\nu = 1$ gives a consistent exchange-rate convention, not a physically calibrated mapping to every external metric. Training time reoptimization can violate fixed response assumptions, and a post-hoc operator cannot repair an unusable expert. Qwen IconQA additionally exposes a degenerate no-forgetting axis on which aggressive reversion can change answer format

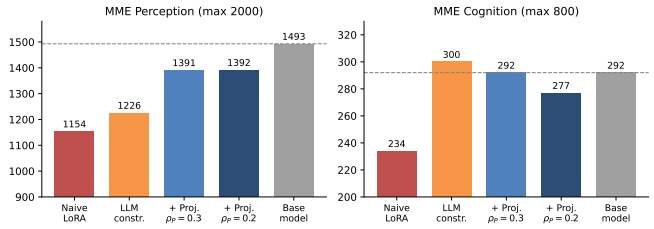


Figure 3: MME double dissociation on the IconQA constraint ladder. Projector protection primarily restores perceptual capability, whereas LLM protection primarily affects cognition; training time constraints can redirect optimization pressure across modules.

without improving retention. The published default can provide a reference configuration in such cases, but the present study does not validate an automatic degeneracy detector or fallback policy.

E Cross-Backbone Protocol and Results: Qwen2.5-VL-7B

Protocol. The base model is Qwen2.5-VL-7B-Instruct. The vision tower is frozen; each task expert is a rank-32 LoRA (α=64, learning rate 2×10^{-4}) on all LLM linear layers plus a full-precision merger trained at 2×10^{-5}, global batch 16, three epochs over the same 5,000-sample subsets as LLaVA (936 steps), seed 42. Model Tailor and CGM$^+$ are applied post hoc to the same rank-32 LoRA expert after merging its LoRA updates into the base weights. LoRASculpt is trained separately using the selected module-wise retention budgets. The priced scope is the merger and LLM layers 16-27 (the last 12 of 28 layers, mirroring layers 20-31 of LLaVA). Upstream evaluation uses the same four sets as LLaVA under the Qwen evaluation prompt; the zero-shot four-set mean is 68.805 (68.81 to two decimal places), and the knowledge two-set mean is 61.68. Forgetting is defined relative to the unrounded four-set reference; scores and differences are computed before rounding. F^{up} is estimated once on the four upstream sets under the evaluation prompt (it is *not* the LLaVA F^{up}); the probe is the same rank-4, 1,000 step variational probe with $\beta = 10^{-3}$; h follows the same module split (trajectory for the merger, endpoint for the LLM); the candidate grid is $\{0.1, \ldots, 0.9, 1.0\}$ on both axes and the specification is otherwise identical to Sec. 3. Recommendations were frozen before the corresponding merged models were evaluated.

Table 12: Qwen2.5-VL-7B, rank-32 LoRA experts, equal budget protocol. Entries are forgetting↓/downstream↑ (four-set forgetting; CIDEr for captioning, exact-match accuracy for IconQA); Δscore is the change in downstream minus forgetting from the operator reference to the ledger point (seed 42). Model Tailor comparisons over three calibration seeds appear in Appendix K. Naive experts: COCO 5.24/121.21, Flickr30k 5.10/61.06, IconQA −0.15/97.59. Knowledge-set forgetting of the naive experts (GQA/OKVQA mean) is 10.7/9.6/0.8. †Format collapse: the merged model reverts to free-form answers; letter-normalized accuracy is 97.07. The CGM$^+$ reference pair $(0.95, 0.95)$ maps the published hard top-5% rule to Qwen module retention ratios and uses Qwen-specific F^{up} in its row-selection arm; it is not a published Qwen measurement. Reference and ledger share every input except the strengths. **Bold**: better of reference and ledger (forgetting and downstream compared separately).

Task	Operator	Reference setting	Ledger (ρ_P, ρ_L)	Δscore
COCO	LoRASculpt	$(1.0, 0.1)$: −1.01/132.73	$(0.3, 0.1)$: **−1.43/133.31**	+1.0
	Model Tailor	$(0.0, 0.1)$: **2.32/129.21**	$(0.3, 0.1)$: 2.48/128.63	−0.8
	CGM$^+$	$(0.95, 0.95)$: 2.11/125.11	$(0.4, 0.2)$: **1.04/130.17**	+6.1
Flickr30k	LoRASculpt	$(1.0, 0.1)$: −1.41/66.81	$(0.1, 0.1)$: **−1.57/68.24**	+1.6
	Model Tailor	$(0.0, 0.1)$: **2.92**/72.80	$(0.1, 0.1)$: 2.93/**73.90**	+1.1
	CGM$^+$	$(0.95, 0.95)$: 2.88/67.99	$(0.3, 0.2)$: **1.52/75.77**	+9.1
IconQA	LoRASculpt	$(1.0, 0.1)$: −1.57/97.72	$(0.3, 0.1)$: **−1.66/97.85**	+0.2
	Model Tailor	$(0.0, 0.1)$: **−0.94**/97.06	$(0.3, 0.1)$: −0.92/**97.64**	+0.6
	CGM$^+$	$(0.95, 0.95)$: **−0.72/97.55**	$(0.4, 0.1)$: −0.39/21.93†	−76.0

Results. For CGM$^+$, the ledger reduces forgetting and improves the downstream metric on both captioning tasks. Model Tailor shows mixed results: the within-task score decreases on COCO and increases on Flickr30k. For LoRASculpt the default already drives four-set forgetting negative on this backbone (text-reading sets improve), and the ledger's gains are small and positive. On IconQA the expert forgets nothing and the operators are ties, except for the degenerate CGM$^+$ axis discussed in Sec. 4.3. A finer LLM grid (0.02 steps) was also tested: it helps the post-hoc operators further on Flickr30k (Model Tailor $(0.02, 0.02)$: 1.97/78.24, $+6.4$; COCO $(0.3, 0.02)$: 1.68/129.41, $+0.8$) but hurts LoRASculpt on COCO (-5.2), suggesting that a 2% trainable LLM budget is too restrictive in this training time setting; we therefore keep the 0.1 grid for all operators in the table above and note that training time and post-hoc operators have different admissible lower bounds.

F Ablation Summary

Table 13 collects the decision-level and outcome-level ablations discussed in Sec. 4.4.

Probe versus expert drift. Replacing the probe mean with the final expert update tests whether the short-horizon response is sufficient for strength selection. On Qwen, the expert-drift substitution reproduces the probe recommendation exactly in seven of nine operator-task cells and moves by only one grid step in the remaining two. The one displaced recommendation that was executed (CGM$^+$-Flickr30k) scores 0.6 points below the probe's own choice. This substitution supports consistency of coarse decisions, not coordinate-wise reconstruction of the expert update.

What the variational parameterization contributes. The learned coordinate-wise variance is nearly task-invariant and is excluded from the final decision. Removing KL shows a different role: the point probe diverges on two of three Qwen tasks at the original learning rate, whereas the variational probe converges under the frozen budget. Reducing the point-probe learning rate restores convergence and nearly the same decisions. Thus the decision content resides primarily in μ and the tensor-wise scale τ_k; the variational objective provides a stable fixed-budget fit rather than a task-specific coordinate uncertainty estimate.

Tensor scale, F^{up}, and h. On LLaVA, removing the gate eliminates the task-specific pattern of projector recommendations. F^{up} primarily sets the upstream pricing scale and orders forgetting; replacing it by unit weights collapses projector decisions to the lower grid boundary, while making it isotropic leaves LS/MT unchanged on Qwen and mainly affects the CGM$^+$ arm that directly reads its coordinate pattern. The ablations also favor the module-specific choice of h. The detailed estimator comparison is reported in Appendix D; these ablations do not support a claim that one universal curvature estimator serves both modules.

G Decision-Level Ablation on Qwen2.5-VL-7B

Table 14 lists the recommendation of each ablated controller with all other inputs fixed. Cells whose recommendation differs from the main specification and were executed are marked with their score difference (variant minus main).

H CGM on Full-Parameter Experts: All Published Operating Points and the Prospectively Frozen Cross Sweep

The CGM rows of the main text use the operator's native protocol: a full-parameter expert in which the last 12 LLM layers and the projector are fine-tuned (the same expert used by Model Tailor), with F_{ft} estimated on 256 task samples and F_{pre} on the 665k mixture. Table 15 compares the frozen ledger point with every operating point published for the operator, including the soft mixing (α=0.2) and the hard default used in Table 1: the hard CGM$^+$ default (α=1, top-5%), and the sensitivity-analysis optimum of the original paper (α=0.15, top-10%). The ledger point has the highest composite score of the five configurations on each task, although it reduces downstream accuracy on IconQA. It is also the only configuration with forgetting below three points on all three tasks. The strengths were frozen from the rank-4 LoRA probe before any full-parameter merge had been evaluated.

Table 13: Ablation summary. Every variant is compared with the frozen main specification without re-tuning other inputs. LS: LoRASculpt; MT: Model Tailor.

Ingredient	Ablation	Decision-level result	Outcome / interpretation
Probe response μ	replace μ with true expert ΔW_D	7/9 operator-task cells identical, 2 adjacent by one step	executed displaced cell: -0.6
Variational fit	learned coordinate-wise σ_i	$\sigma_i \approx 0.022$ and task-invariant; local closed-form diagnostic returns to prior scale	no task-specific decision content
	remove KL ($\beta = 0$)	diverges on 2/3 Qwen tasks at the probe learning rate; at lr$/3$, COCO/Flickr decisions are identical and IconQA is one step looser	KL stabilizes the fixed-budget fit
Tensor gate τ_k	remove gate ($\tau_k \equiv 0$)	LLaVA loses task-specific decisions; Qwen pushes several projector recommendations toward full retention	executed changes within ± 0.45 on light-forgetting Qwen
	shared absolute threshold instead of tensor-relative scale	projector strengths $0.30/0.28/0.04$ versus tensor-relative $0.34/0.34/0.28$; IconQA collapses under the shared threshold	supports relative normalization
Upstream scale F^{up}	$F^{\mathrm{up}} \equiv 1$	all LLaVA projector recommendations move to the lower grid boundary	upstream scale is necessary for pricing
	isotropic F^{up} (per-module mean)	LS/MT unchanged on Qwen; CGM$^+$ moves on 2/3 because its native arm reads the coordinate pattern	pattern matters mainly to pattern-reading executors
Task statistic h	swap projector/LLM estimator sources	task discrimination is lost on the swapped axis	supports module-specific h estimators

Table 14: Recommendations (ρ_P, ρ_L) of ablated controllers on Qwen. Bold marks a change from the main specification; brackets give the executed score difference (variant minus main). "No gate" sets $\tau_k \equiv 0$; "isotropic" replaces F^{up} by its per-module mean; "expert drift" replaces μ by the true expert update; "point probe" removes the KL term (σ fixed at 10^{-4}). *At the probe's learning rate the point estimate diverges on COCO and Flickr30k; the IconQA entries are from that run, the COCO/Flickr30k entries from the converged run at one third of the learning rate. IconQA cells with changed recommendations were not executed: the expert has no forgetting to recover and the CGM$^+$ axis is degenerate (Table 12).

Task	Operator	Main	No gate	Isotropic F^{up}	Expert ΔW_D	Point probe (lr$/3$)
COCO	LS / MT	(0.3,0.1)	**(1.0,0.1)** [MT -0.24]	(0.3,0.1)	(0.3,0.1)	(0.3,0.1)
COCO	CGM$^+$	(0.4,0.2)	**(0.5,0.3)** [$+0.45$]	**(0.3,0.1)** [-0.61]	(0.4,0.2)	(0.4,0.2)
IconQA	LS / MT	(0.3,0.1)	**(1.0,0.1)**	(0.3,0.1)	LS **(0.4,0.1)** / MT (0.3,0.1)	**(0.4,0.1)***
IconQA	CGM$^+$	(0.4,0.1)	**(0.6,0.1)**	(0.4,0.1)	(0.4,0.1)	**(0.5,0.1)***
Flickr30k	LS / MT	(0.1,0.1)	(0.1,0.1)	(0.1,0.1)	(0.1,0.1)	(0.1,0.1)
Flickr30k	CGM$^+$	(0.3,0.2)	**(0.4,0.2)** [-0.22]	**(0.2,0.1)** [$+0.65$]	**(0.2,0.2)** [-0.60]	(0.3,0.2)

Table 16 lists the conditional cross sweeps around the frozen point (projector axis with ρ_L fixed at the prediction; LLM axis with ρ_P fixed). The sweep was submitted after the strengths were frozen and is the source of the CGM$^+$ entries in Tables 1 and 6. On the LLM axis the frozen strength is exact on Flickr30k, adjacent on COCO (0.27 below the best) and IconQA (0.77 below). Retaining only 10% of the LLM update sharply reduces performance on COCO and Flickr30k, but gives the best measured score on IconQA. After the primary evaluation we filled both axes to a uniform 0.1 step (‡ cells in Table 16; ρ_P=1.0 keeps the expert's projector untouched and prices only the LLM, ρ_L=1.0 the converse). On the projector axis the score rises slowly and then saturates: COCO $106.05 \rightarrow 107.02$ over 0.3-0.8 with the six cells from 0.5 to 1.0 spanning 0.6 points (a narrow measured score range), IconQA $59.13 \rightarrow 62.16$ ($+3.0$, monotone, the six cells above 0.5 adding only 0.8), and Flickr30k $63.50 \rightarrow 65.26$ at 0.9 before dropping to 64.46 at 1.0 (a broad interior optimum at 0.7-0.9). Forgetting is essentially independent of ρ_P once the LLM is priced (0.53-0.69 on COCO, 2.49-2.82 on IconQA, 0.55-0.69 on Flickr30k): for these dense experts, retaining more projector updates has little effect on the upstream metric, so the ledger's projector strengths which the probe learned in a LoRA regime where projector drift *does* cost upstream accuracy (the LoRASculpt axis in Table 1: forgetting $0.5 \rightarrow 3.8$ as ρ_P goes $0.1 \rightarrow 1.0$), are conservative by 0.9-3.0 points on this executor, most visibly for IconQA. Model Tailor on the same experts behaves the same way: with the LLM at the ledger strength, ρ_P=1.0 scores $107.04/62.54/63.41$ against $106.42/61.37/63.69$ at the frozen point ($+0.6$, $+1.2$, -0.3). The direction of the projector decision is therefore operator dependent and the ledger moves in the favorable direction in these comparisons (constrain for the

Table 15: CGM on full-parameter LLaVA-1.5-7B experts, equal-budget protocol. Entries are forgetting↓ / downstream↑; the last row gives the scalar score (downstream minus forgetting) for COCO / IconQA / Flickr30k. Gains of the ledger point over the best published operating point are $+1.4$, $+2.9$, and $+3.0$. **Bold**: best in the row; underline: second best (forgetting and downstream ranked separately).

Task	Unpatched expert	CGM soft α=0.2	CGM$^+$ (α=1, 5%)	CGM$^+$ (α=0.15, 10%)	Ledger (ρ_P, ρ_L)
COCO	2.18 / 105.65	1.24 / 105.56	1.86 / 105.86	1.73 / 106.38	(0.3, 0.3): **0.57 / 106.61**
IconQA	14.41 / 65.15	9.09 / **65.31**	12.89 / 65.01	12.71 / 65.10	(0.3, 0.2): **2.65** / 61.78
Flickr30k	2.36 / 61.81	1.71 / 62.22	1.83 / 62.28	1.83 / 62.13	(0.5, 0.3): **0.56 / 64.06**
Score	103.47 / 50.74 / 59.45	104.32 / 56.22 / 60.51	104.00 / 52.12 / 60.45	104.65 / 52.39 / 60.30	**106.05 / 59.13 / 63.50**

training time operator, where ρ_P=1.0 is the worst point of the axis; retain for the post-hoc operators, where the paper default ρ_P=0 is 12-24 points below the frozen point in Table 1), but its magnitude on post-hoc executors is conservative; the 0.1-step fill of the Model Tailor projector axis (LLM at its 10% default) is likewise flat from 0.5 upward on all three tasks (span ≤ 0.4). The LLM axis is the opposite: filled to 1.0, the score falls monotonically beyond 0.4-0.5 on every task (COCO 106.34 $\rightarrow$ 103.02, IconQA 57.62 $\rightarrow$ 48.35, Flickr30k 63.55 $\rightarrow$ 59.02), and at ρ_L=1.0 the priced-projector model forgets 2.09/14.21/2.21 against 2.18/14.41/2.36 for the unpatched expert, most of the measured upstream degradation remains when the LLM update is fully retained. This result supports calibrating the LLM strength, although the location of the best point varies across tasks. Figure 4 plots every measured axis, including the fill cells, against the frozen predictions.

Table 16: Prospectively frozen conditional cross sweep of CGM$^+$ on full parameter experts (scalar score, downstream minus forgetting). Underlined: frozen prediction; bold: measured best on the prospectively fixed grid $\{0.1, 0.2, 0.3, 0.5, 0.7\}$. ‡ Post-hoc cells added after the primary evaluation (endpoint completion and the uniform 0.1 step fill); they are not used in any hit/adjacent/regret count. ρ_P=1.0 leaves the expert's projector untouched; ρ_L=1.0 leaves its LLM untouched.

Task	Axis (other fixed)	0.1	0.2	0.3	0.4‡	0.5	0.6‡	0.7	0.8‡	0.9‡	1.0‡
COCO	projector (ρ_L=0.3)	104.75	105.56	106.05	106.29	106.46	106.48	**106.66**	107.02	106.87	106.97
	LLM (ρ_P=0.3)	89.70	102.27	106.05	106.34	**106.32**	105.55	105.22‡	104.72	-	103.02
IconQA	projector (ρ_L=0.2)	53.73	56.84	59.13	60.65	61.36	61.41	**61.74**	61.87	61.96	62.16
	LLM (ρ_P=0.3)	**59.90**	59.13	58.23	57.62	56.47	55.12	53.45‡	52.14	-	48.35
Flickr30k	projector (ρ_L=0.3)	62.62	63.14	63.61	63.62	63.50	64.44	**64.65**	65.07	65.26	64.46
	LLM (ρ_P=0.5)	55.44	63.11	**63.50**	63.55	62.81	62.68	61.13‡	61.04	-	59.02

Held-out conditional axes. When the projector is fixed first and the LLM is swept conditionally, the rank-32 LoRA experts expose a systematic conditional-selection error: LoRASculpt and the CGM$^+$ operator (applied to the same LoRA experts, Appendix I, Table 17) are best at $0.1/0.1/0.1$ on COCO/IconQA/Flickr30k, whereas the global arm predicts $0.3/0.2/0.3$. The operator-specific CGM arm recovers IconQA and Flickr exactly $(2/3)$ but still misses COCO. On the prospectively frozen full-parameter CGM$^+$ sweeps (Appendix H), by contrast, the frozen LLM strength is exact on Flickr and adjacent on COCO and IconQA. The COCO and Flickr30k axes peak away from the grid boundary, whereas IconQA is best at $\rho_L = 0.1$. The additional 0.1-step evaluations show declining scores above $\rho_L = 0.4$-0.5 on all three tasks. These results support coarse strength transfer while ruling out a claim of exact conditional optimum recovery.

I CGM$^+$ Operator on LoRA Experts (LLaVA)

CGM is published for full parameter experts; the main text follows that protocol. Because the operator itself is defined on any weight delta, we also applied CGM$^+$ to the same rank-32 LoRA experts used by LoRASculpt, pricing them with the same frozen controller outputs. Paper default here denotes the published soft mixing applied to the LoRA delta.

The best LLM strength depends on the expert parameterization. For LoRA experts, the best measured conditional ρ_L is 0.1 on all three tasks. For full parameter experts, the best prespecified points are 0.5 on COCO, 0.1 on IconQA, and 0.3 on Flickr30k. A 10% retention budget substantially reduces

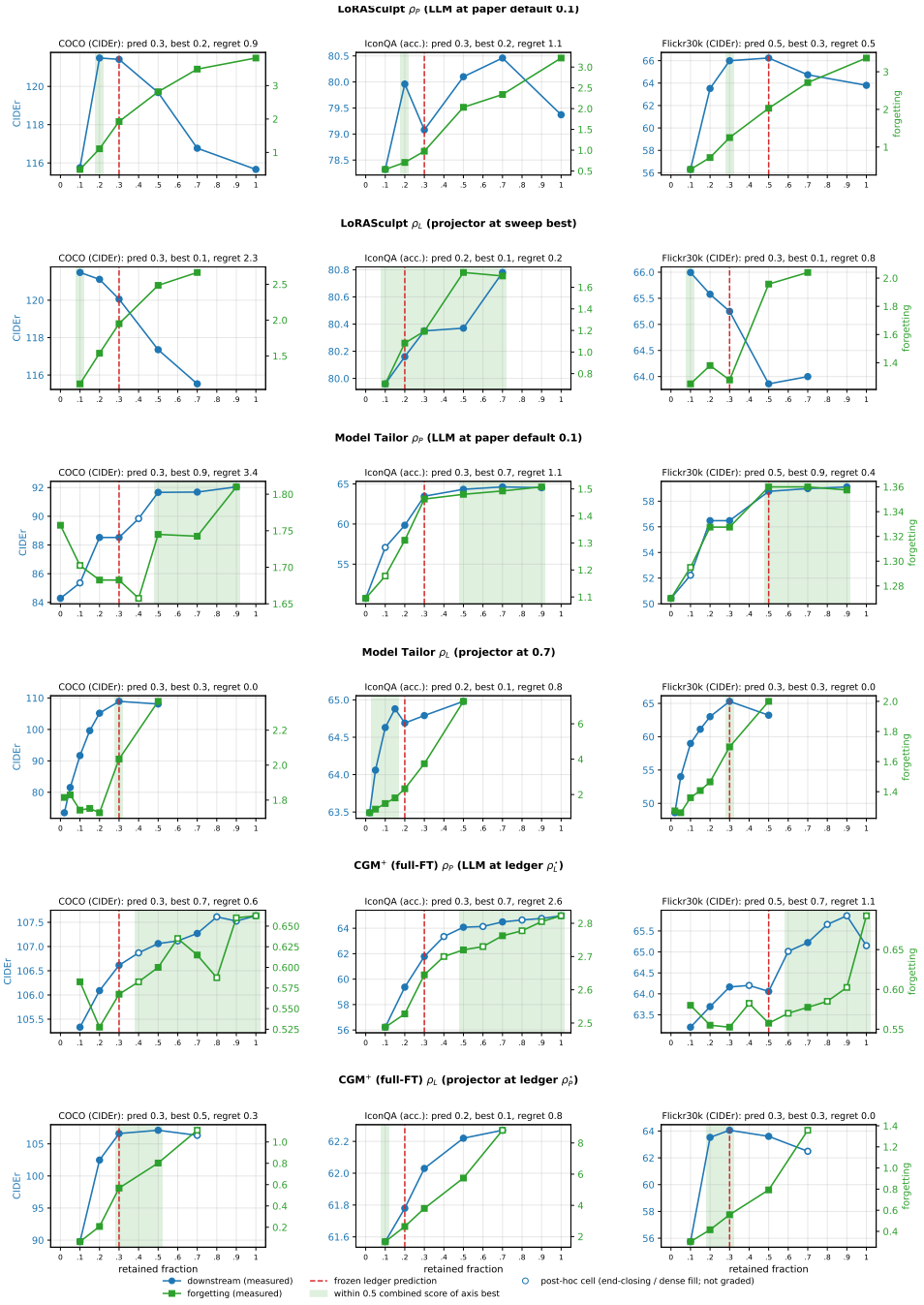


Figure 4: **Conditional strength responses across all eighteen axes.** Columns: COCO, IconQA, and Flickr30k; rows: projector and LLM axes for LoRASculpt, Model Tailor, and CGM$^+$. Blue shows downstream performance; green shows forgetting; red dashed lines mark frozen ledger predictions. Each panel specifies the other module's fixed setting. Regret uses $S = T - F$ on the prespecified grading grids. Shading marks sampled locations within 0.5 score points of the axis best, a descriptive tolerance rather than a confidence interval. Hollow markers denote supplementary evaluations excluded from the primary grading counts.

Table 17: CGM$^+$ on rank-32 LoRA experts, LLaVA-1.5-7B, equal-budget protocol; entries are forgetting↓/downstream↑. The frozen joint point improves the score over the published default on all three tasks ($+9.1$, $+2.4$, $+8.2$) but is below the best measured point on each, most clearly on Flickr30k where the projector prediction (0.5) is one step looser than the measured optimum (0.3). The projector axis of this family is exact on COCO and IconQA and adjacent on Flickr30k; the conditional LLM axis is uniformly best at the lower grid boundary 0.1, the behavior discussed in Appendix H. The rank transfer results of Table 2 are computed in this family. The sweep best column is a gray reference and is not ranked; **bold**: better of default and prediction (forgetting and downstream compared separately).

Task	Paper default	Frozen joint prediction	Measured sweep best (reference)
COCO	$(1.0, 1.0)$: 4.03/111.12	$(0.3, 0.3)$: **2.31/118.46**	$(0.3, 0.1)$: 2.11/121.46
IconQA	$(1.0, 1.0)$: 6.72/**78.18**	$(0.3, 0.2)$: **3.28**/77.15	$(0.3, 0.1)$: 3.22/77.14
Flickr	$(1.0, 1.0)$: 3.37/58.43	$(0.5, 0.3)$: **1.30/64.52**	$(0.3, 0.1)$: 1.57/71.89

captioning performance for the dense experts but not for the LoRA experts. One possible explanation is that the two parameterizations distribute task information differently across the retained rows; low rank alone does not establish this concentration. The probe based controller selects ρ_L in the range 0.2-0.3 in both cases.

J Domain-Shift Pilot: LaTeX-OCR

To test the specification on a domain far from the four upstream VQA sets, we fine-tuned the Qwen2.5-VL-7B expert on LaTeX-OCR (image-to-LaTeX transcription; BLEU-4 on a held-out $1{,}000$-formula split). A gate fixed in advance required the naive expert to forget at least as much as the Flickr30k expert before any operator was run. The $5{,}000$-sample expert reaches BLEU-4 93.1 (zero-shot 55.6) with *no* measurable forgetting (-0.5 four-set, 1.0 knowledge-set); a $30{,}000$-sample, three-epoch expert reaches BLEU-4 93.8 and still forgets only 0.2 (four-set) and 1.2 (knowledge-set) points, against a gate of 5.0 and 9.6. The gate was therefore not met and, by that rule, no operator was run. We report this pilot as a boundary case: under these protocols, increasing the transcription training set by $6\times$ does not produce enough forgetting to meet the prespecified criterion for evaluating anti-forgetting operators. The forgetting that the controller addresses is, on the evidence of Table 12 and this pilot, a property of the task-backbone pair rather than of fine-tuning per se.

K Gate-Only Selection and Seed Replication on Qwen2.5-VL-7B

Matched selector and execution protocol. To distinguish pricing from tensor-relative gating, the gate-only selector reads exactly the same frozen rank-4 probe, active-drift gate, retention proxy, and $\{0.1, 0.2, \ldots, 1.0\}$ grid as the ledger. For each priced module it calculates $g_a = \#\{i : |\mu_i| > \tau_{k(i)}\}/\#\{i\}$ and rounds to the nearest eligible grid point, choosing the smaller point on a tie. It discards both A and B and therefore makes the same recommendation for all operators on a given task. The priced Qwen scope contains the merger and LLM layers 16–27 (86 parameter keys). The measured merger/LLM gated fractions are $0.311/0.259$ for COCO, $0.309/0.261$ for Flickr30k, and $0.280/0.240$ for IconQA; hence the gate-only pairs are $(0.3, 0.3)$, $(0.3, 0.3)$, and $(0.3, 0.2)$. In contrast, the ledger pairs are operator-specific (Table 12). This is the complement of the no-gate ablation: it preserves gating and removes the quadratic pricing terms. For Model Tailor and CGM$^+$, gate-only and ledger reuse the same rank-32 LoRA expert and change only the post-hoc processing strengths. LoRASculpt trains separate experts using the same rank, LoRA scaling ($\alpha = 64$), global batch 16, three epochs, and task samples as the corresponding ledger configuration. We report CIDEr for captioning and exact-match accuracy for IconQA as T. Forgetting F is the difference between the unrounded zero-shot reference 68.805 and the four-set upstream mean, and $S = T - F$; scores are compared only within each task.

Training and calibration-sampling replication. LoRASculpt varies the training seed over 42, 1, and 2 for both selectors on COCO and Flickr30k (Table 19). Model Tailor uses the same three seeds to sample 128 calibration examples, keeping the rank-32 expert fixed; we evaluate ledger, gate-only, and reference $(0.0, 0.1)$ configurations (Table 20). All four ledger-versus-gate-only comparisons

Table 18: All nine Qwen gate-only versus ledger settings. Entries are forgetting F / downstream T / score S; ΔS is ledger minus gate-only. †Captioning results for LoRASculpt and Model Tailor are three-seed means, with sample standard deviations for S (training and calibration-sampling seeds, respectively). Other entries use fixed seed-42 inputs. *IconQA has negligible naive-expert forgetting; both CGM$^+$ configurations exhibit answer-format collapse. Scores and differences are calculated before rounding.

Task	Operator	Gate-only pair	Gate-only $F/T/S$	Ledger pair	Ledger $F/T/S$	ΔS
COCO	LoRASculpt†	$(0.3, 0.3)$	$0.60/126.82/126.22 \pm 0.72$	$(0.3, 0.1)$	$-1.41/133.34/134.75 \pm 0.07$	$+8.53$
	Model Tailor†	$(0.3, 0.3)$	$3.80/125.61/121.81 \pm 0.19$	$(0.3, 0.1)$	$2.52/128.65/126.13 \pm 0.02$	$+4.32$
	CGM$^+$	$(0.3, 0.3)$	$1.10/129.96/128.86$	$(0.4, 0.2)$	$1.04/130.17/129.14$	$+0.28$
Flickr30k	LoRASculpt†	$(0.3, 0.3)$	$0.41/65.64/65.23 \pm 1.17$	$(0.1, 0.1)$	$-1.59/68.50/70.08 \pm 0.24$	$+4.85$
	Model Tailor†	$(0.3, 0.3)$	$4.20/65.24/61.03 \pm 0.37$	$(0.1, 0.1)$	$2.92/73.48/70.56 \pm 0.38$	$+9.53$
	CGM$^+$	$(0.3, 0.3)$	$1.60/73.90/72.30$	$(0.3, 0.2)$	$1.52/75.77/74.25$	$+1.95$
IconQA*	LoRASculpt	$(0.3, 0.2)$	$-1.63/97.85/99.48$	$(0.3, 0.1)$	$-1.66/97.85/99.51$	$+0.03$
	Model Tailor	$(0.3, 0.2)$	$-0.67/97.66/98.33$	$(0.3, 0.1)$	$-0.92/97.64/98.56$	$+0.23$
	CGM$^+$	$(0.3, 0.2)$	$-0.36/24.08/24.44$	$(0.4, 0.1)$	$-0.39/21.93/22.32$	-2.12

have positive score differences under every seed. CGM$^+$ reuses precomputed Fisher statistics and is deterministic conditional on those statistics and the expert. These repetitions measure training variation for LoRASculpt and calibration sampling for Model Tailor, rather than variation in probe fitting or in the entire pipeline. Tables 3 and 12 retain seed-42 reference comparisons; Tables 4 and 18 report three-seed captioning means for LoRASculpt and Model Tailor. All scores use the unrounded zero-shot reference 68.805; displayed F, T, S, and differences are rounded independently.

Table 19: LoRASculpt on Qwen across three training seeds. Entries are $S = T - F$; standard deviations are sample standard deviations.

Task	Selector	Seed 42	Seed 1	Seed 2	Mean $\pm$ std.	Mean F/T
COCO	Ledger $(0.3, 0.1)$	134.74	134.69	134.82	134.75 ± 0.07	$-1.41/133.34$
	Gate-only $(0.3, 0.3)$	126.29	125.47	126.89	126.22 ± 0.72	$0.60/126.82$
Flickr30k	Ledger $(0.1, 0.1)$	69.81	70.18	70.26	70.08 ± 0.24	$-1.59/68.50$
	Gate-only $(0.3, 0.3)$	63.94	66.22	65.52	65.23 ± 1.17	$0.41/65.64$

Table 20: Model Tailor on Qwen across three calibration-sampling seeds, with the expert held fixed. Entries are $S = T - F$; standard deviations are sample standard deviations. The reference is $(0.0, 0.1)$.

Task	Selector	Seed 42	Seed 1	Seed 2	Mean $\pm$ std.	Mean F/T
COCO	Ledger $(0.3, 0.1)$	126.15	126.11	126.13	126.13 ± 0.02	$2.52/128.65$
	Gate-only $(0.3, 0.3)$	121.63	122.00	121.80	121.81 ± 0.19	$3.80/125.61$
	Reference $(0.0, 0.1)$	126.90	127.05	126.78	126.91 ± 0.13	$2.34/129.24$
Flickr30k	Ledger $(0.1, 0.1)$	70.98	70.22	70.48	70.56 ± 0.38	$2.92/73.48$
	Gate-only $(0.3, 0.3)$	60.63	61.15	61.33	61.03 ± 0.37	$4.20/65.24$
	Reference $(0.0, 0.1)$	69.88	69.49	69.90	69.76 ± 0.23	$2.91/72.67$

Comparison with the Model Tailor reference. Against reference $(0.0, 0.1)$, the ledger's mean score difference is -0.78 on COCO and $+0.80$ on Flickr30k. The ledger is below the reference in all three COCO runs and above it in all three Flickr30k runs. Thus, its advantage over gate-only does not imply an improvement over the operator reference on every task. Across the four captioning comparisons, ledger and gate-only sample standard deviations are $0.07/0.72$ and $0.24/1.17$ for LoRASculpt on COCO and Flickr30k, and $0.02/0.19$ and $0.38/0.37$ for Model Tailor. Lower variability is therefore observed in three comparisons, with similar variability in the fourth.