On-Policy Adaptation Mitigates Hyperparameter-Sensitive Forgetting in Vision-Language Models

Doyoon Kim [*1] **Ziyang Wu** [*2] **Jingfeng Yang** [2] **Yi Ma** [2 3 †]

Abstract

Repeated fine-tuning of Vision-Language Models (VLMs) risks catastrophic forgetting of prior capabilities. Recent studies suggest on-policy RL methods such as GRPO mitigate this risk better than Supervised Fine-Tuning (SFT), but a systematic characterization in VLMs is lacking. We compare SFT and GRPO by fine-tuning Qwen2.5-VL-Instruct on CIFAR-10 and CIFAR-100, varying learning rates and training durations to characterize the key drivers of forgetting. GRPO consistently exhibits minimal forgetting, while SFT forgetting is highly sensitive to hyperparameters: high learning rates improve in-domain accuracy but induce catastrophic cross-domain forgetting. To probe whether forgetting reflects encoder-level representational damage, we fine-tune Qwen2-VL on a spatial reasoning task and isolate the vision encoders for retraining. The SFT encoder underperforms the pre-trained encoder even after retraining—suggesting encoder changes not fully recovered under our protocol—while the GRPO encoder retains comparable performance.

1. Introduction

Continual adaptation of foundation models—repeatedly post-training a deployed model on new tasks without retraining from scratch—requires methods that improve new capabilities without damaging prior capabilities (Shi et al., 2025). For VLMs, this risk is especially tied to preserving visual representations (Zhai et al., 2023). Yet each fine-tuning step risks catastrophic forgetting (Aleixo et al., 2023), and for VLMs this risk is compounded: the vision encoder and language model can degrade independently, and if fine-tuning corrupts foundational visual representations, no amount of language-side adaptation can compensate. RL-based post-training methods such as GRPO have emerged as a promising alternative to SFT for adaptation (Shao et al., 2024; Chu et al., 2025), and recent work suggests they are also more forgetting-resistant (Shenfeld et al., 2025). However, most existing comparisons are on text-only LLMs, and it remains unclear how these benefits manifest in VLMs, especially at the vision-encoder level.

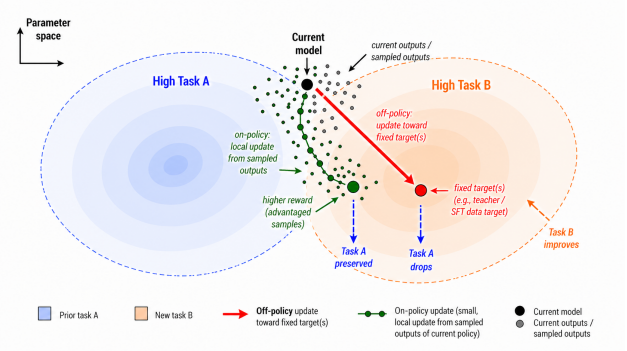

Figure 1. Off-policy vs. on-policy updates in parameter space. Starting from the current model (black dot), off-policy updates (SFT, red arrow) can induce larger shifts toward fixed training targets—especially under aggressive hyperparameters—moving far from the Task A region and causing Task A to drop. On-policy updates (GRPO, green arrow) tend to update around samples generated from the current policy, which can reduce abrupt distributional shifts and better preserve Task A while improving Task B.

In this work, we frame VLM post-training as a continual adaptation problem and systematically study forgetting under SFT and on-policy RL (GRPO). Our contributions are:

- We show that SFT forgetting in VLMs is highly hyperparameter-sensitive: high learning rates and long training induce catastrophic cross-domain degradation, while conservative settings largely preserve prior capabilities.
- We show that GRPO preserves prior visual capabilities across both classification and spatial reasoning adaptation, remaining stable where SFT can fail.
- Through vision encoder isolation and retraining, we provide evidence that SFT can induce encoder changes that are not fully recovered under our retraining protocol, while GRPO retains comparable encoder performance.

[*]Equal contribution [†]Corresponding author. [1]POSTECH, Pohang, South Korea [2]UC Berkeley, Berkeley, CA, USA [3]The University of Hong Kong, Hong Kong. Correspondence to: Yi Ma <yima@eecs.berkeley.edu>.

Presented at the ICML 2026 Workshop "Continual Adaptation at Scale: Towards Sustainable AI". Copyright 2026 by the author(s).

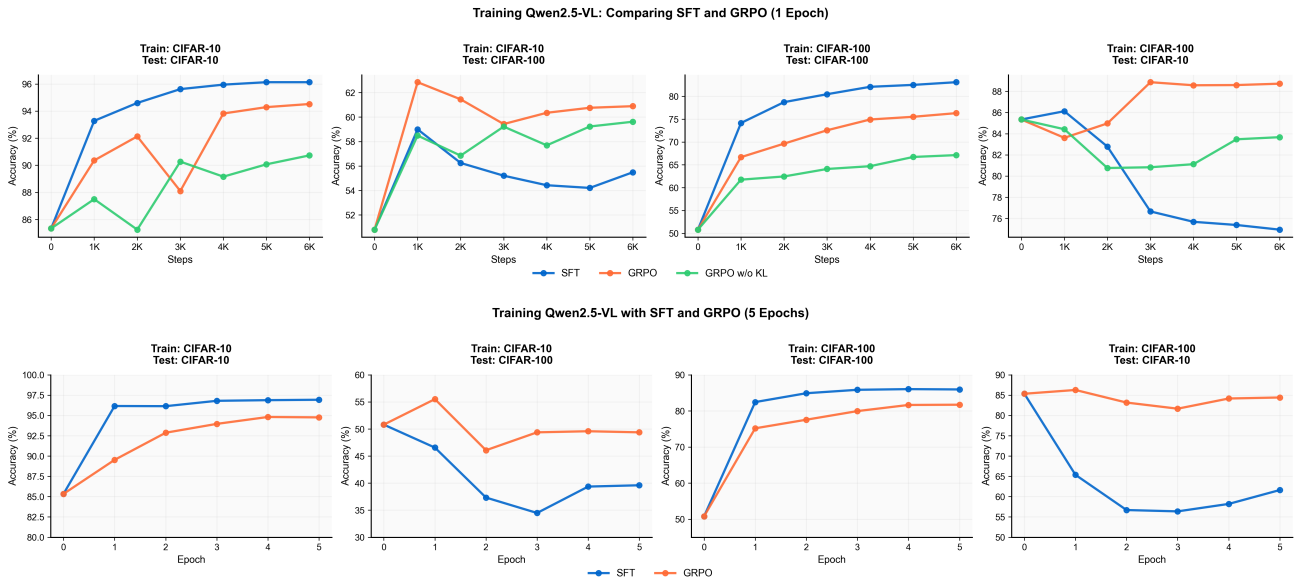

Figure 2. Performance comparison between SFT and GRPO for short (top, 1 epoch) and long (bottom, 5 epochs) training. Each method uses the learning rate with the highest in-domain accuracy. GRPO consistently maintains prior classification accuracy while SFT degrades rapidly.

2. Related Work

Catastrophic forgetting in LLM fine-tuning. Catastrophic forgetting is a long-standing challenge in continual learning (Aleixo et al., 2023), recently re-emerging as a concern in LLM post-training (Shi et al., 2025). Li et al. (2024) identified learning rate as an important factor in LLM tuning, while Rofin et al. (2026) further linked learning rate, sharpness, and catastrophic overtraining. Our results support this: SFT forgetting in VLMs is highly sensitive to learning rate, vanishing under conservative settings. Zhai et al. (2023) studied SFT forgetting in multimodal models and found that large LRs and sharp task shifts amplify overwriting of prior knowledge. We extend this line of work by further isolating where within the VLM architecture the damage occurs.

RL-based fine-tuning as a forgetting mitigation strategy. Recent work has increasingly studied RL-based fine-tuning as a way to reduce forgetting relative to SFT. Shenfeld et al. (2025) showed this empirically across multiple text benchmarks, attributing the gap to smaller KL divergence from the pre-trained distribution under on-policy updates; Chu et al. (2025) further demonstrated that SFT memorizes surface patterns while RL generalizes, a distinction that manifests as differential forgetting. Rajani et al. (2025) provided a complementary account, arguing that GRPO tends to amplify pre-existing capabilities whereas SFT more strongly replaces them. Mukherjee et al. (2025) further showed that RL updates are often concentrated in a small subnetwork, suggesting a possible mechanism for reduced interference with prior knowledge. Lai et al. (2025) reported similar forgetting-mitigation benefits of reinforcement fine-tuning in continual post-training of multimodal models. Jin et al. (2025) further studied the two-stage SFT-to-RL pipeline and argued that RL can restore OOD reasoning ability lost during SFT, although this recovery has limits. However, most of the above comparisons are conducted on text-only models, and existing multimodal studies do not isolate whether forgetting reflects encoder-level representational damage. VLMs introduce a fundamentally different structure: a vision encoder and a language model that can degrade independently, and forgetting may be concentrated in specific architectural components rather than distributed uniformly. Our work extends this line of inquiry to the multimodal setting and further asks whether SFT forgetting reflects encoder-level representational damage—a question prior work does not address.

3. Background

The key difference between SFT and GRPO lies in how they update the model (see Figure 1). SFT is off-policy: it minimizes the negative log-likelihood of a fixed target dataset:

$$\mathcal{L}_{\text{SFT}} = -\mathbb{E}_{(x,y)\sim\mathcal{D}} \log \pi_\theta(y|x),$$

where x is the input, y is the target output, $\mathcal{D}$ is the training dataset, and $\pi_\theta(y|x)$ is the model's output distribution parameterized by θ. Every training example contributes equally to the gradient regardless of the model's current confidence, and the supervision signal comes from a static distribution $p_{\text{data}}(y|x)$ that may be far from the current policy $\pi_\theta(y|x)$. This off-policy mismatch can induce large

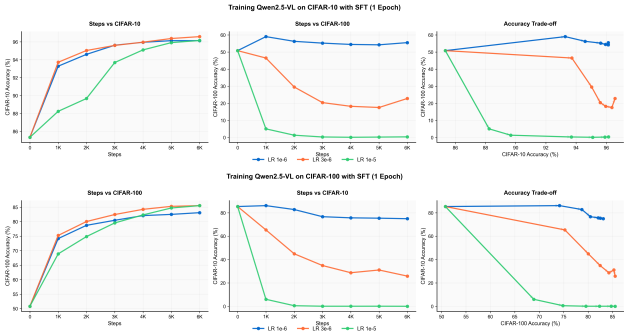

Figure 3. Training performance with different learning rates on CIFAR-10 (top) and CIFAR-100 (bottom). Larger learning rates dramatically increase catastrophic forgetting.

distributional shifts between pre- and post-training behavior, contributing to the sharp representational changes associated with catastrophic forgetting (Li et al., 2024).

RL-based fine-tuning (Williams, 1992; Schulman et al., 2017) is on-policy: it optimizes expected reward over samples drawn from the current policy,

$$\mathcal{L}_{\text{RL}} = -\mathbb{E}_{x\sim\mathcal{D},\, y\sim\pi_\theta}\big[A(y)\log\pi_\theta(y|x)\big],$$

where y is sampled from the current policy π_θ and $A(y)$ is the advantage of output y relative to a baseline (how much better y is than the average response). When $A(y) \approx 0$ for outputs the policy already handles well, gradient updates are naturally suppressed on already-mastered outputs—including prior tasks—and sampling from the current policy bounds representational drift more tightly than off-policy SFT.

GRPO (Shao et al., 2024) is the on-policy method used in our experiments. For each input x, it samples G outputs from π_{ref} (the reference policy), computes group-normalized advantages $\hat{A}_i = R(y_i) - \frac{1}{G}\sum_j R(y_j)$, where $R(y_i)$ is the reward for the i-th output, and optimizes the clipped surrogate below, where $r_i(\theta) = \pi_\theta(y_i|x)/\pi_{\text{ref}}(y_i|x)$ is the probability ratio between the current and reference policy:

$$\mathcal{L}_{\text{GRPO}} = \mathbb{E}_{x,\{y_i\}}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\min\Big(r_i(\theta)\hat{A}_i,\ \text{clip}(r_i(\theta),1-\epsilon,1+\epsilon)\hat{A}_i\Big) - \beta\, D_{\text{KL}}(\pi_\theta\|\pi_{\text{ref}})\Bigg].$$

Here ϵ is the clipping threshold that limits how far the policy can move per update, and β is the KL penalty coefficient that penalizes deviation from the reference policy π_{ref}. These properties suggest a plausible mechanism for GRPO's reduced forgetting: reward-proportional gradients that shrink on already-mastered outputs, and updates anchored to the current policy distribution.

4. Experiments

4.1. Setup

We fine-tuned Qwen2.5-VL-Instruct (Bai et al., 2025) and Qwen2-VL (Wang et al., 2024) on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009) (image classification benchmarks with 10 and 100 classes, respectively) and the SAT (Spatial Aptitude Training) dataset (Ray et al., 2024) using both SFT and GRPO. The SAT dataset provides synthetic static and dynamic spatial reasoning tasks requiring egocentric understanding; our training pipeline for SAT follows Zhou et al. (2025). All experiments used 4 H800 GPUs with cosine learning-rate schedulers; full training hyperparameters are listed in Appendix B. For GRPO, reward functions combined answer-format and accuracy signals. We evaluate variants with and without an explicit KL penalty to the reference policy to isolate its effect on forgetting. We measure in-domain accuracy and cross-domain transfer to evaluate the extent of forgetting.

4.2. SFT vs. GRPO: Forgetting Comparison

We trained Qwen2.5-VL-Instruct on CIFAR-10 and CIFAR-100 using both SFT and GRPO for 1 epoch (Figure 2, top) and 5 epochs (bottom). While SFT achieved slightly higher peak in-domain accuracy, it suffered significant catastrophic forgetting in cross-domain evaluations. In contrast, GRPO maintained stable performance across both in-domain and cross-domain tests, demonstrating robustness to forgetting even without explicit KL divergence regularization. Comparing GRPO with and without KL divergence regularization, we found that the explicit KL penalty further improved stability, but even without it, GRPO substantially outperformed SFT in terms of forgetting mitigation. Full numerical results are in Appendix C.

4.3. Learning Rate and Training Duration Control SFT Forgetting

High Learning Rate. We tested SFT learning rates from 1e-6 to 1e-5 (Figure 3). The results show that larger learning rates can reach higher in-distribution accuracy but severely increase catastrophic forgetting. For example, training on CIFAR-100 with lr=1e-5 achieved 85.53% in-domain accuracy but dropped cross-domain CIFAR-10 accuracy from 85.34% to 0.03%. By contrast, using lr=1e-6 achieves 83.10% on CIFAR-100 while maintaining 74.94% CIFAR-10 accuracy. These results suggest that SFT-induced forgetting can be mitigated through careful hyperparameter selection.

Long Training Horizon. We trained both SFT and GRPO for up to 5 epochs to evaluate long-term stability (bottom panel of Figure 2). SFT exhibited an initial performance drop that later partially recovered, whereas GRPO remained

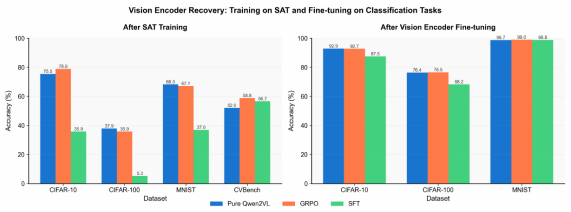

Figure 4. Vision encoder forgetting analysis. **Left**: After SAT fine-tuning, SFT collapses classification performance while GRPO maintains stability. **Right**: After retraining isolated encoders on classification, the SFT encoder underperforms even the pre-trained encoder—suggesting encoder changes not fully recovered under our retraining protocol—while the GRPO encoder retains comparable retraining performance.

consistently stable throughout training.

Taken together, high learning rates and extended training both increase the cumulative divergence of the model from its pre-trained initialization, which is consistent with Shenfeld et al. (2025)'s finding that forgetting correlates with KL divergence from the pre-trained distribution under SFT.

4.4. Vision Encoder Analysis

To probe whether forgetting is amplified by task complexity, we fine-tuned Qwen2-VL on the SAT dataset and tested generalization on classification tasks (CIFAR-10, CIFAR-100) and CV-Bench (Tong et al., 2024) (visual reasoning). We used task-optimal hyperparameters for each method (full details in Appendix A). Under these settings, SFT caused severe cross-domain degradation: CIFAR-10 accuracy collapsed from 85.34% to 15.23%. In contrast, GRPO preserved cross-domain classification accuracy (85.34%→86.01%) while also improving CV-Bench from 66.65% to 77.63%, demonstrating positive transfer across modalities. These results raise a deeper question: does SFT forgetting on harder tasks reflect encoder-level representational damage, or merely temporary interference from language model updates? To investigate, we isolated the vision encoders from pre-trained, SFT-trained, and GRPO-trained models and retrained them from scratch on classification tasks, discarding all language model weights (full setup in Appendix A). If SFT changes the encoder in a way that downstream retraining cannot fully recover, the isolated SFT encoder will underperform the pre-trained encoder even after retraining.

The retraining results in Figure 4 support the latter: the vision encoder from the SFT-trained model, after retraining on classification tasks, performs worse than the encoder from the original pre-trained model. This suggests that SFT may alter the encoder in a way that is not fully recovered by our downstream retraining protocol, beyond temporary interference from the language model. In contrast, the GRPO-trained encoder, when isolated and retrained, achieves performance comparable to the pre-trained encoder, consistent with GRPO preserving visual representations under our experimental setting.

5. Conclusion

We study catastrophic forgetting in VLMs under SFT and GRPO, with a focus on hyperparameter sensitivity and encoder-level effects. Our key findings are: (1) GRPO consistently exhibits minimal forgetting across all settings, while SFT is highly sensitive to learning rate and training duration; (2) both methods can preserve classification performance with conservative hyperparameters, suggesting forgetting can be mitigated with careful hyperparameter selection; and (3) SFT can produce encoder changes that are not fully recovered under our retraining protocol, while GRPO retains comparable encoder capacity under the same protocol. These findings provide empirical evidence that on-policy fine-tuning is more forgetting-resistant in the multimodal setting, with direct implications for designing safe continual adaptation pipelines for deployed VLMs.

6. Limitations and Future Work

Our experiments focus on Qwen-family VLMs and a limited set of classification and spatial reasoning tasks, so broader validation across architectures, model scales, and adaptation domains is needed. Since GRPO and SFT are fundamentally different training paradigms—differing in objective, optimization dynamics, and data usage—we could not isolate on-policy sampling as the sole causal factor behind the observed forgetting differences. While our results are consistent with the hypothesis that on-policy updates reduce representational drift, we do not directly measure KL divergence from the pre-trained distribution, representation similarity, or update geometry in the VLM setting. Future work should incorporate these measurements to better characterize the mechanisms underlying GRPO's robustness to forgetting.

Impact Statement

This paper presents work whose goal is to advance the understanding of catastrophic forgetting in VLMs. Our findings may help practitioners design safer continual fine-tuning pipelines, reducing unintended capability degradation in deployed AI systems. There are no foreseeable ethical concerns specific to this work beyond those generally applicable to large-scale model fine-tuning research.

References

Aleixo, E. L., Colonna, J. G., Cristo, M., and Fernandes, E. Catastrophic forgetting in deep learning: A comprehen-

sive taxonomy. *arXiv preprint arXiv:2312.10549*, 2023.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q. V., Levine, S., and Ma, Y. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.

Jin, H., Luan, S., Lyu, S., Rabusseau, G., Rabbany, R., Precup, D., and Hamdaqa, M. Rl fine-tuning heals ood forgetting in sft. *arXiv preprint arXiv:2509.12235*, 2025.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lai, S., Zhao, H., Feng, R., Ma, C., Liu, W., Zhao, H., Lin, X., Yi, D., Zhang, Q., Liu, H., et al. Reinforcement fine-tuning naturally mitigates forgetting in continual post-training. *arXiv preprint arXiv:2507.05386*, 2025.

Li, H., Ding, L., Fang, M., and Tao, D. Revisiting catastrophic forgetting in large language model tuning. In *Findings of the association for computational linguistics: EMNLP 2024*, pp. 4297–4308, 2024.

Mukherjee, S., Yuan, L., Hakkani-Tur, D., and Peng, H. Reinforcement learning finetunes small subnetworks in large language models. *arXiv preprint arXiv:2505.11711*, 2025.

Rajani, N., Gema, A. P., Goldfarb-Tarrant, S., and Titov, I. Scalpel vs. hammer: GRPO amplifies existing capabilities, SFT replaces them. *arXiv preprint arXiv:2507.10616*, 2025.

Ray, A., Duan, J., Brown, E., Tan, R., Bashkirova, D., Hendrix, R., Ehsani, K., Kembhavi, A., Plummer, B. A., Krishna, R., et al. SAT: Dynamic spatial aptitude training for multimodal language models. *arXiv preprint arXiv:2412.07755*, 2024.

Rofin, M., Varre, A., and Flammarion, N. (how) learning rates regulate catastrophic overtraining. *arXiv preprint arXiv:2604.13627*, 2026.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Shenfeld, I., Pari, J., and Agrawal, P. RL's razor: Why online reinforcement learning forgets less. *arXiv preprint arXiv:2509.04259*, 2025.

Shi, H., Xu, Z., Wang, H., Qin, W., Wang, W., Wang, Y., Wang, Z., Ebrahimi, S., and Wang, H. Continual learning of large language models: A comprehensive survey. *ACM Computing Surveys*, 58(5):1–42, 2025.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

Zhai, Y., Tong, S., Li, X., Cai, M., Qu, Q., Lee, Y. J., and Ma, Y. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*, 2023.

Zhou, H., Li, X., Wang, R., Cheng, M., Zhou, T., and Hsieh, C.-J. R1-Zero's "aha moment" in visual reasoning on a 2b non-sft model. *arXiv preprint arXiv:2503.05132*, 2025.

A. Vision Encoder Isolation: Experimental Setup

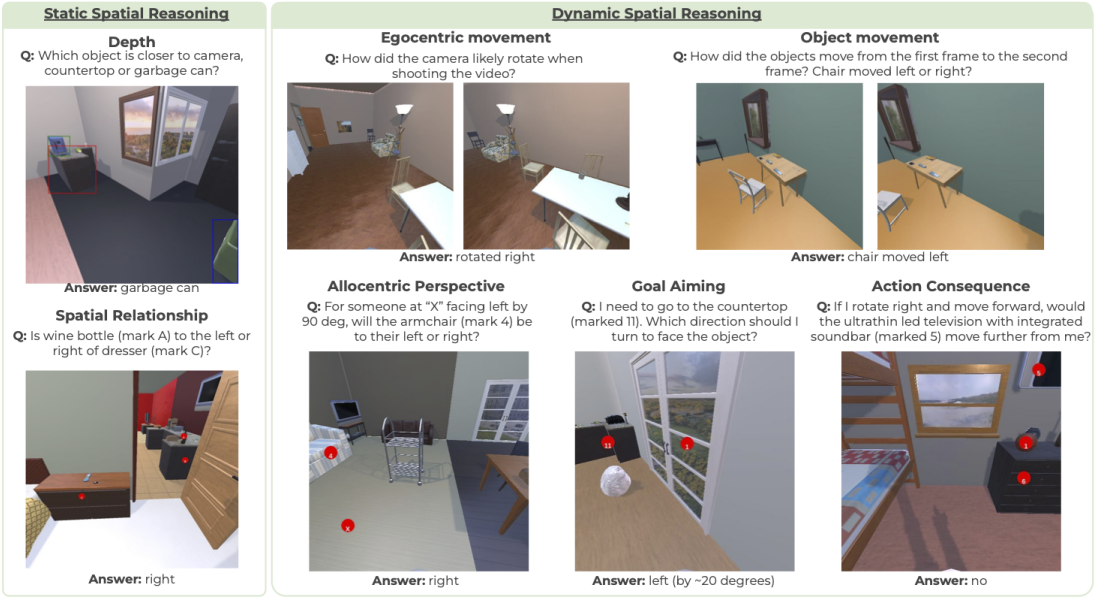

Figure 5. Example tasks from the SAT (Spatial Aptitude Training) dataset (Ray et al., 2024), covering static spatial reasoning (depth, spatial relationship) and dynamic spatial reasoning (egocentric movement, allocentric perspective, goal aiming, object movement, action consequence). Figure reproduced from Ray et al. (2024).

We fine-tuned Qwen2-VL on the SAT (Spatial Aptitude Training) dataset (Ray et al., 2024), which provides a diverse set of spatial reasoning tasks for multimodal language models. Static tasks evaluate understanding of object depth and spatial relationships from a single image; dynamic tasks require reasoning about motion, perspective change, and object trajectories across frames (Figure 5). We fine-tuned on the full SAT task suite using task-optimal hyperparameters for each method, and evaluated generalization on classification tasks (CIFAR-10, CIFAR-100) and CV-Bench.

To investigate whether the resulting forgetting reflects encoder-level representational degradation or merely changes in how the language model interprets visual features, we isolated the vision encoders from all three models (pre-trained, SFT-trained, and GRPO-trained) and retrained them from scratch on classification tasks, discarding the language model weights entirely. Retraining used an attentive pooler classifier on top of the frozen encoder, trained with Adam (lr=1e-3, batch size 512) for up to 15 epochs. The best checkpoint was selected based on validation accuracy using a 10% held-out split of the training set. If SFT merely interferes with the language model's interpretation, the isolated SFT encoder should recover to the same performance level as the pre-trained encoder after retraining. However, if SFT changes the encoder in a way that downstream retraining cannot fully recover, the SFT encoder will underperform even after retraining—which is what we observe (see Figure 4, right panel).

B. Training Hyperparameters

Table 1 summarizes the training configuration used for Qwen2.5-VL-Instruct in the classification forgetting experiments.

Table 1. Training hyperparameters for Qwen2.5-VL-Instruct (CIFAR-10/100 experiments).

Hyperparameter	SFT	GRPO
Learning rates evaluated	1e-6, 3e-6, 1e-5	1e-6, 3e-6
LR scheduler	cosine	cosine
Warmup steps	625	625
GPUs	4 × H800	4 × H800
Training epochs	1 epoch / 5 epoch	1 epoch / 5 epoch
Total steps	6,250 / 31,250	6,250 / 31,250
Fine-tuning type	full	full
KL coefficient β	—	0.04
Reward functions	—	format + accuracy
Answer format	`The answer is [class label]`	

Both SFT and GRPO use the same answer format: `The answer is [class label]`, where `[class label]` is the exact class name (e.g., `The answer is airplane`). Accuracy is computed by parsing the predicted class label from the response and checking for an exact string match against the ground-truth class name. For GRPO, the format reward fires when the output follows this template; the accuracy reward fires when the parsed label matches the ground truth.

C. Extended Numerical Results (Qwen2.5-VL-Instruct)

Tables 2–3 report the full numerical results underlying the figures in the main text. "—" indicates the configuration was not evaluated.

Table 2. 1-epoch fine-tuning results (%). C10 = CIFAR-10, C100 = CIFAR-100. Column headers **A→B** denote training on A and testing on B; cross-domain columns (**C100→C10**, **C10→C100**) measure forgetting. Each group shows accuracy at step 0 (pretrained baseline) and after 1 epoch.

			C100→C10		**C100→C100**		**C10→C10**		**C10→C100**	
Method	**Tuning**	**LR**	**0ep**	**1ep**	**0ep**	**1ep**	**0ep**	**1ep**	**0ep**	**1ep**
SFT	Full	1e-6	85.3	74.9	50.8	83.1	85.3	96.1	50.8	55.5
		3e-6	85.3	25.8	50.8	85.5	85.3	96.6	50.8	22.8
		1e-5	85.3	0.0	50.8	85.5	85.3	96.2	50.8	0.4
	LoRA rank 8	1e-4	85.3	71.8	50.8	73.2	—	—	—	—
	LoRA rank 64	3e-5	85.3	73.3	50.8	73.2	—	—	—	—
GRPO	Full	1e-6	85.3	88.7	50.8	76.3	85.3	94.5	50.8	60.9
	Full (no KL)	1e-6	85.3	83.7	50.8	67.1	85.3	90.7	50.8	59.6
	LoRA rank 8	1e-6	85.3	88.1	50.8	75.4	—	—	—	—

Table 3. 5-epoch long training results (%). C10 = CIFAR-10, C100 = CIFAR-100; **A→B** denotes train on A, test on B. Each group shows accuracy at epochs 1, 3, and 5.

			C100→C10			**C100→C100**			**C10→C10**			**C10→C100**		
Method	**Tuning**	**LR**	**1ep**	**3ep**	**5ep**	**1ep**	**3ep**	**5ep**	**1ep**	**3ep**	**5ep**	**1ep**	**3ep**	**5ep**
SFT	Full	1e-6	65.4	56.4	61.6	82.5	85.9	86.0	96.2	96.8	96.9	46.6	34.5	39.6
GRPO	Full	1e-6	86.3	81.7	84.4	75.2	80.0	81.7	89.5	94.0	94.8	55.5	49.4	49.4

The KL penalty comparison (GRPO 1e-6 vs. no-KL variant) in Tables 2–3 shows that the explicit KL term provides a modest additional benefit, but even without it GRPO substantially outperforms SFT at comparable learning rates. The LoRA results show that parameter-efficient fine-tuning reduces SFT forgetting relative to aggressive full fine-tuning (e.g., 71.8% vs. 0.0% CIFAR-10 accuracy after CIFAR-100 training at respective optimal LRs), but conservative full SFT at lr=1e-6 still forgets more than GRPO. GRPO with LoRA achieves performance comparable to full GRPO, suggesting the on-policy advantage is preserved under parameter-efficient settings.

D. Qualitative Forgetting Examples

When SFT is trained on CIFAR-10 and evaluated on CIFAR-100, the model's error modes reveal the mechanism of forgetting. CIFAR-10 contains 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck), while CIFAR-100 contains 100 fine-grained classes. After SFT on CIFAR-10, the model systematically confuses CIFAR-100 inputs with CIFAR-10 classes, producing three distinct failure patterns.

Case 1: Domain substitution — outputs CIFAR-10 class labels. The most direct forgetting signature: the model maps CIFAR-100 images to the nearest CIFAR-10 class, producing labels that are valid answers to the fine-tuned task but wrong for CIFAR-100. These outputs register as unknown under the CIFAR-100 evaluator because CIFAR-10 class names do not appear in the CIFAR-100 label set.

CIFAR-100 GT	Parsed	Raw output
lion	unknown	`The answer is cat`
pickup_truck	unknown	`The answer is automobile`
leopard	unknown	`The answer is cat`
house	unknown	`The answer is airplane`

Case 2: Within-domain confusion — wrong CIFAR-100 class. The model still outputs a valid CIFAR-100 class label, but the wrong one. These are ordinary misclassifications rather than forgetting artifacts, and persist even without forgetting.

CIFAR-100 GT	Parsed	Raw output
cockroach	bed	`The answer is bed`
girl	baby	`The answer is baby`
willow_tree	oak_tree	`The answer is oak_tree`
leopard	tiger	`The answer is tiger`

Case 3: Superclass collapse. The model outputs a valid but over-general term that does not match any class label exactly.

CIFAR-100 GT	Parsed	Raw output
maple_tree	unknown	`The answer is tree`

Case 1 is the clearest evidence of forgetting: the model has overwritten its knowledge of the 100-class label space and collapses predictions onto the 10 CIFAR-10 classes it was fine-tuned on. The frequency of Case 1 errors increases with learning rate and training duration, directly tracking the quantitative forgetting curves shown in Figures 2 and 3.