RoCo-ACE: Rollout-Conditioned Online Distillation for Retention-Aware Knowledge Injection

Anonymous ACL submission

Abstract

Knowledge injection updates pretrained MLLMs with new factual or domain-specific knowledge, but fitting full authoritative answers can cause drift in non-updated behavior. Online distillation limits this drift by training on model-generated rollouts, yet uniform reference-conditioned distillation gives coarse supervision: it can under-emphasize reference-supported rollout tokens and supervise omitted facts only indirectly. We introduce **RoCo-ACE**, a rollout-conditioned online distillation objective for knowledge injection: **RoCo** uses same-rollout reference-free/reference-conditioned likelihood contrast to reallocate additional distillation weight to reference-supported rollout tokens, while **ACE** adds sparse reference-side anchored correction for authoritative anchors omitted from the rollout, without full-answer imitation. Across three knowledge-injection settings, six retention benchmarks, multiple baselines and base models, **RoCo-ACE** achieves the best injected-knowledge accuracy among compared methods while keeping evaluated retention close to the base model.

1 Introduction

Multimodal Large Language Models (MLLMs) have evolved from vision-language pretraining and image-language bootstrapping (Alayrac et al., 2022; Li et al., 2023) to instruction-tuned multimodal assistants (Dai et al., 2023; Liu et al., 2023; Bai et al., 2023). They now support instruction following, visual understanding, grounding, text-rich image reasoning, and safety-sensitive response behavior (Pi et al., 2024; Wu et al., 2025). After deployment, these models often need to absorb new entities, events, domain facts, and visual-world knowledge. A useful knowledge-injection update should improve injected-knowledge accuracy while limiting drift in non-updated multimodal and safety behavior, as evaluated by general multimodal benchmarks (Liu et al., 2024a; Chen et al., 2024) and safety benchmarks (Liu et al., 2024b). This goal is challenging: fitting new data can interfere with previously learned behavior (Kirkpatrick et al., 2017; Li and Hoiem, 2016), and recent MLLM knowledge-injection studies observe the same injection-retention trade-off (Jiang et al., 2025a,b).

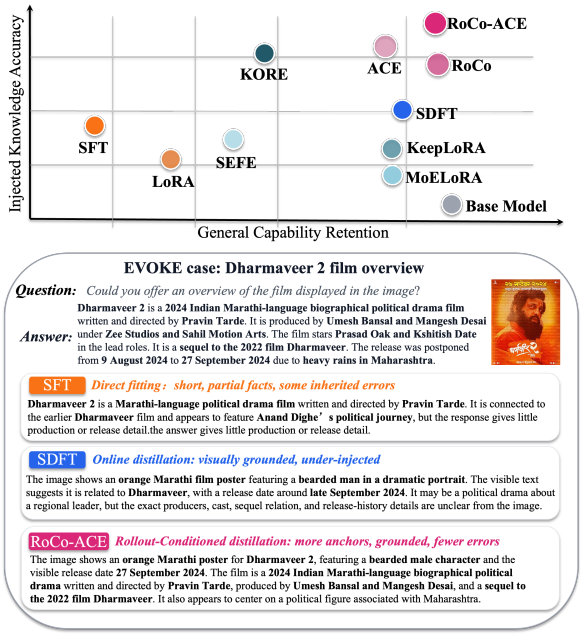

Figure 1: Motivation of **RoCo-ACE**. The upper panel illustrates the injection–retention trade-off among direct fitting, constrained updating, online distillation, and our variants. **RoCo-ACE** improves this trade-off by combining same-rollout likelihood contrast with sparse anchored correction. The lower panel shows an EVOKE example where **RoCo-ACE** recovers more authoritative anchors than SFT and SDFT (Shenfeld et al., 2026) while keeping a grounded response style.

Existing update strategies occupy different regions of the injection-retention trade-off. Supervised Fine-Tuning (SFT) (Ouyang et al., 2022) directly fits authoritative answers and can improve injected-knowledge accuracy, but it also encourages full-answer imitation, including dataset-

specific style and narrow response patterns. Constrained or parameter-efficient methods limit drift by restricting where or how the model can change (Hu et al., 2022; Jiang et al., 2025b; Luo et al., 2026), but they do not directly localize supervision to the factual units that should change. The key issue is supervision granularity: knowledge injection often requires updating a few entities, dates, relations, or event details, while many training tokens reflect generic phrasing, non-target visual details, or answer style. As illustrated in Figure 1, fitting-oriented methods can improve injection at a substantial retention cost, whereas conservative methods better limit drift in non-updated behavior but leave some target facts under-injected.

To address this injection-retention trade-off, we propose **RoCo-ACE**, a **Ro**llout-**Co**nditioned contrastive online distillation objective with **A**nchored **C**ross-**E**ntropy for retention-oriented knowledge injection. It trains on model-generated rollouts instead of fitting the full authoritative answer, keeping updates close to the current model policy (Agarwal et al., 2023; Gu et al., 2023; Shenfeld et al., 2026; Zhao et al., 2026; Ye et al., 2026). However, uniform reference-conditioned online distillation still lacks fine-grained supervision: it can allocate gradient mass to generic or unsupported wording, fail to emphasize reference-supported rollout tokens, and provide only indirect supervision for facts omitted from the current rollout.

Our key observation is that an on-policy rollout exposes a useful supervision split for factual updating. It may contain reference-supported content already reached on-policy, generic or unsupported wording, and authoritative anchors that are omitted from or weakly matched by the rollout. **RoCo** handles the rollout-side part by contrasting reference-free and reference-conditioned teacher views on the same rollout prefix, then reallocating additional distillation weight to tokens whose likelihood increases under the reference-conditioned view. **ACE** complements this signal with sparse anchored cross-entropy over extracted authoritative facts, assigning larger weights to omitted or weakly matched anchors. Together, the two objectives inject sparse factual content without turning the full reference answer into a behavior-cloning target.

This paper makes the following contributions:

- We identify supervision granularity as a key issue in online distillation for knowledge injection: uniform rollout distillation can under-emphasize reference-supported tokens and only indirectly train facts omitted from the current rollout.

- We introduce **RoCo-ACE**, a rollout-conditioned contrastive online distillation objective for knowledge injection. **RoCo** uses same-rollout likelihood contrast to reallocate additional distillation weight to reference-supported rollout tokens, while **ACE** adds sparse anchored correction for omitted or weakly matched authoritative facts.

- Experiments on three knowledge-injection settings and six retention benchmarks show that **RoCo-ACE** improves injected-knowledge accuracy while keeping evaluated retention close to online-distillation baselines and the base model.

2 Related Work

2.1 Update-Space Constraints for Retention

A direct way to inject knowledge is to fit authoritative answers, but full-answer fitting can shift non-target behavior. To limit such update-induced drift, prior work often constrains where or how the model can change. Adam-NSCL (Wang et al., 2021) updates networks in the null space of feature covariance, and LoRA-Null (Tang et al., 2025) applies a related idea to low-rank adaptation by choosing initialization spaces that better retain pretrained behavior. KORE (Jiang et al., 2025b) studies multimodal knowledge injection directly and combines knowledge-oriented augmentation with constraints for limiting drift in multimodal updates. Parameter-efficient and forgetting-aware methods, including LoRA (Hu et al., 2022), KeepLoRA (Luo et al., 2026), MoELoRA (Luo et al., 2024), and SEFE (Chen et al., 2025), restrict or organize trainable parameters through low-rank adapters, residual-gradient updates, expertized modules, or forgetting-aware constraints. Related MLLM continual-learning work, such as MLLM-CL (Zhao et al., 2025) and LiLoRA (Che et al., 2025), further studies sequential visual-instruction updates over heterogeneous task streams.

These methods are retention-oriented mainly through constraints on the editable parameter or update space. **RoCo-ACE** is complementary: rather than constraining the update space, it improves supervision granularity over rollout tokens and authoritative anchors.

2.2 Online Distillation for Model Updating

Knowledge distillation (Hinton et al., 2015) transfers a teacher distribution to a student. Online or on-policy variants, such as Generalized Knowledge Distillation (Agarwal et al., 2023) and MiniLLM (Gu et al., 2023), distill on student-generated outputs rather than only fixed target responses, keeping supervision close to the current student policy. Recent work further studies richer teacher contexts, reasoning compression, and policy-optimization variants (Zhao et al., 2026; Ye et al., 2026; Sang et al., 2026; Yang et al., 2026; Li et al., 2026). Song and Zheng (2026) provides a recent survey. SDFT (Shenfeld et al., 2026) is especially related, as it uses self-distillation to stabilize continual updates and reduce forgetting.

Online distillation is attractive for knowledge injection because it avoids directly fitting the full authoritative answer. However, uniform reference-conditioned rollout distillation still provides coarse supervision: it can spend gradients on generic or unsupported wording, under-emphasize reference-supported rollout tokens, and supervise omitted authoritative facts only indirectly. **RoCo-ACE** improves this granularity by combining same-rollout likelihood contrast for reference-supported rollout tokens with sparse reference-side correction for authoritative anchors omitted from or weakly matched by the rollout.

3 Method

3.1 Overview

We consider a knowledge-injection dataset $\mathcal{D} = \{(x_i, a_i, I_i)\}_{i=1}^{N}$, where x_i is a user prompt, a_i is an authoritative reference answer that specifies the target factual update, and I_i is an optional image. Our goal is to adapt a base MLLM p_{θ_0} into a student model p_θ that incorporates the injected facts into its responses while limiting drift in non-updated behavior. Figure 2 gives an overview of the **RoCo-ACE** training objective. For each example, the student samples an on-policy rollout $y_i = (y_{i,1}, \ldots, y_{i,T_i}) \sim p_\theta(\cdot \mid x_i, I_i)$. This rollout exposes a supervision split for factual updating: it may contain reference-supported factual spans already reached on-policy, generic or unsupported wording and non-target visual details, and authoritative anchors that are omitted from or weakly matched by the rollout. **RoCo** handles the rollout-side signal by scoring the same rollout prefixes with reference-free and reference-conditioned teacher views. It uses their likelihood contrast to reallocate additional distillation weight to reference-supported rollout tokens, while generic wording and non-target visual details receive mainly the floor distillation signal. **ACE** handles the omission case with sparse anchored cross-entropy over extracted authoritative facts, assigning larger weights to anchors omitted from or weakly matched by the rollout. Together, **RoCo** and **ACE** inject sparse factual content without treating the full reference answer as a token-level imitation target.

3.2 RoCo: Rollout-Conditioned Contrastive Online Distillation

Online distillation (Shenfeld et al., 2026; Zhao et al., 2026) is well suited to retention-oriented knowledge injection because it trains on the student's own rollout instead of fitting the full authoritative answer. This keeps the update close to the current student policy, but uniform reference-conditioned distillation still assigns loss to every valid rollout token, even though only a small subset carries target facts. A student rollout often mixes reference-supported factual content with generic or unsupported wording. In Figure 2, for example, the rollout contains reference-supported spans such as "Princess Anne", "hospitalized", and "Gatcombe Park", but also includes non-target visual details such as "dark green hat" and "formal coat". **RoCo** addresses this supervision-granularity problem by comparing reference-free and reference-conditioned teacher likelihoods under the same rollout prefix. It reallocates additional distillation weight to tokens whose likelihood increases when the authoritative reference is available, while generic wording, non-target visual details, and answer style receive mainly the floor distillation signal. To make this selective reallocation explicit, **RoCo** measures how much the authoritative reference changes the likelihood assigned to each on-policy token under the same rollout prefix. Let q_ϕ denote the teacher model, and let $q_\phi(\cdot \mid \cdot)$ denote its next-token probability distribution over the vocabulary. For each rollout prefix $y_{i,<t}$, we evaluate two matched teacher views:

$$\begin{aligned} q^{i,t}_{\text{base}}(v) &= q_\phi(v \mid x_i, I_i, y_{i,<t}), \\ q^{i,t}_{\text{ref}}(v) &= q_\phi(v \mid x_i, I_i, a_i, y_{i,<t}). \end{aligned} \tag{1}$$

The subscript base denotes the reference-free view, and ref denotes the reference-conditioned view. Both views score the same student-generated prefix and share the same architecture and parameters.

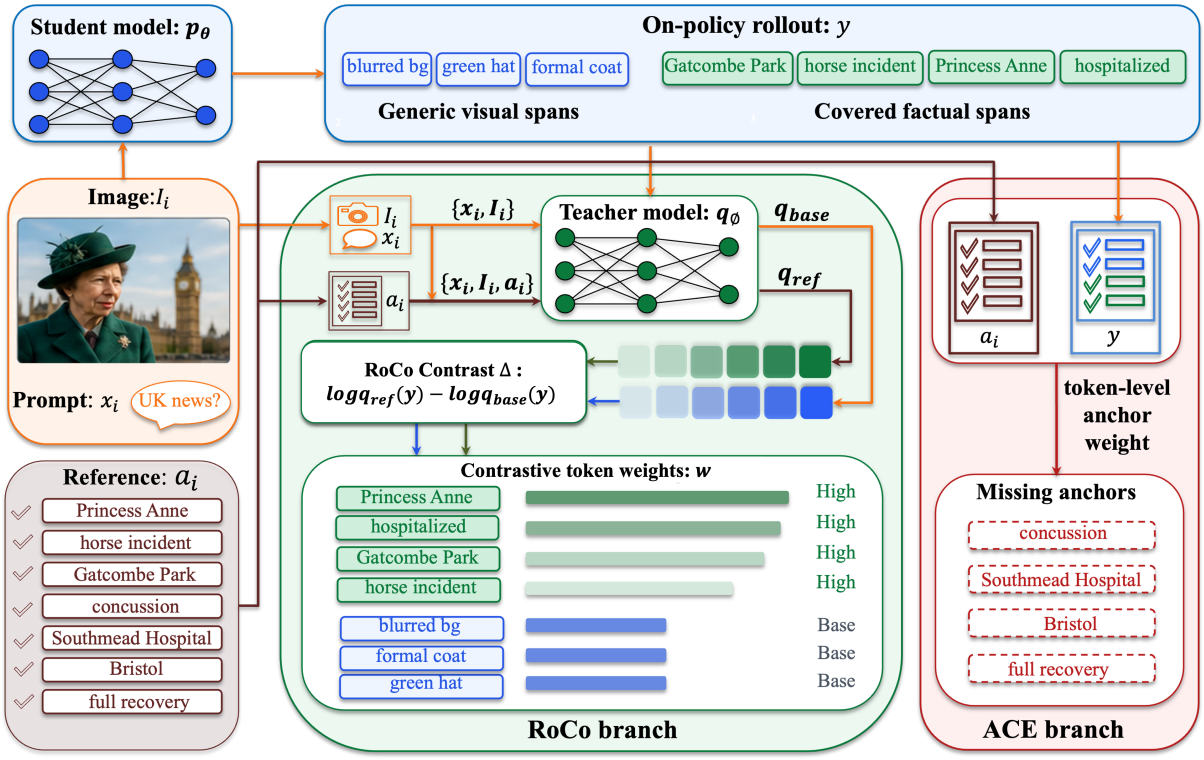

Figure 2: Overview of **RoCo-ACE**. The student samples an on-policy rollout that mixes generic or non-target visual spans, reference-supported factual spans, and omitted authoritative anchors. **RoCo** scores the same rollout under reference-free and reference-conditioned teacher views, then uses their likelihood contrast to assign larger distillation weights to reference-supported rollout tokens. **ACE** complements this rollout-side signal with sparse anchored cross-entropy for authoritative anchors omitted from or weakly matched by the rollout.

They differ only in whether the authoritative reference a_i is included in the prompt context. In our online setting, q_ϕ is an exponential-moving-average teacher synchronized from the student, keeping teacher scoring close to the current policy while avoiding a degenerate self-target at the current optimization step. For each generated token $y_{i,t}$, the RoCo contrast is defined as the log-likelihood ratio between the two views:

$$\Delta_{i,t} = \log q_{\text{ref}}^{i,t}(y_{i,t}) - \log q_{\text{base}}^{i,t}(y_{i,t}). \quad (2)$$

A high value of $q_{\text{ref}}^{i,t}(y_{i,t})$ alone may simply reflect a generic token that is already likely under the original prompt, whereas a positive $\Delta_{i,t}$ suggests additional support induced by the authoritative reference. We therefore use the contrast to estimate reference-induced support for rollout tokens: the token is generated on-policy, and the reference-conditioned view makes it more likely under the same teacher and prefix. Let $\mathcal{V}_i$ denote valid completion-token positions after removing padding and ignored tokens. The contrast is converted into dense token weights:

$$w_{i,t} = w_0 + \beta_\Delta \max(0, \Delta_{i,t} - \tau), \quad t \in \mathcal{V}_i, \quad (3)$$

where w_0 is a floor weight, β_Δ scales the contrast signal, and τ suppresses small log-likelihood differences. The floor term keeps a stable online-distillation signal over the rollout, while the contrast term reallocates additional gradient mass to tokens whose likelihood becomes sufficiently larger under the reference-conditioned view. When a detected factual expression is split into multiple subword tokens, we aggregate the token weights within the expression and broadcast the resulting span score back to its tokens. For simplicity, we still denote the resulting weights by $w_{i,t}$. This span-aware smoothing reduces subword-level noise for names, dates, numbers, and multilingual expressions.

The resulting distillation objective is

$$\mathcal{L}_{\text{RoCo}}^{(i)} = \frac{\sum_{t \in \mathcal{V}_i} w_{i,t} d_{i,t}}{\sum_{t \in \mathcal{V}_i} w_{i,t} + \epsilon}, \quad d_{i,t} = D_{\text{KL}}\left(q_{\text{ref}}^{i,t}(\cdot) \| p_\theta(\cdot \mid x_i, I_i, y_{i,<t})\right). \quad (4)$$

The denominator makes the objective a weighted average over valid rollout tokens, so RoCo changes the relative distribution of distillation weight rather than simply changing the overall loss scale. The reference-conditioned teacher provides the distri-

butional target, while the reference-versus-base contrast determines where additional distillation weight should concentrate. Thus, **RoCo** strengthens reference-supported content already reached on-policy, while keeping training anchored to the student's own completion and down-weighting generic wording, non-target visual details, and answer style.

3.3 ACE: Anchored Cross-Entropy

The on-policy design also defines a coverage boundary for **RoCo**: its loss is evaluated on the valid rollout positions $\mathcal{V}_i$, so its strongest supervision comes from content that the student already reaches in its own completion. When a target fact is absent from the rollout, reference-conditioned distillation may still influence nearby continuation distributions and help future rollouts move toward that fact, but the current rollout provides only indirect supervision for the omitted anchor. **ACE** addresses this remaining coverage gap by applying explicit supervised correction to extracted authoritative anchors, with larger weights for anchors that are missing or weakly covered in the rollout.

Let $a_i = (a_{i,1}, \ldots, a_{i,L_i})$ denote the tokenized authoritative answer. We segment a_i into factual spans $\mathcal{S}_i$. In our implementation, spans are constructed from entity-like strings, dates, numbers, quoted titles, and compact factual phrases after text normalization. The same normalization is used when comparing reference spans with rollout text. For each span $s \in \mathcal{S}_i$, let $\rho_i(s, y_i) \in [0, 1]$ be its coverage score in the student rollout, where 1 denotes full coverage and 0 denotes absence. We assign each span an anchor weight

$$u_{i,s} = u_0 + \beta_{\text{miss}}\big(1 - \rho_i(s, y_i)\big), \tag{5}$$

where u_0 is a base weight for extracted factual anchors and β_{miss} controls the strength of missing- or weak-coverage correction. Let $\mathcal{I}_i(s) \subseteq \{1, \ldots, L_i\}$ denote the reference-token positions covered by span s, and let $\mathcal{A}_i = \bigcup_{s \in \mathcal{S}_i} \mathcal{I}_i(s)$ be the set of anchor-token positions in the reference answer. We broadcast each span weight to its covered reference tokens. For overlapping spans, the token-level anchor weight is

$$u_{i,k} = \max_{s:\, k \in \mathcal{I}_i(s)} u_{i,s}, \quad k \in \mathcal{A}_i. \tag{6}$$

The ACE loss for example i is then

$$\mathcal{L}_{\text{ACE}}^{(i)} = -\frac{\sum_{k \in \mathcal{A}_i} u_{i,k} \log p_\theta(a_{i,k} \mid x_i, I_i, a_{i,<k})}{\sum_{k \in \mathcal{A}_i} u_{i,k} + \epsilon}. \tag{7}$$

Only tokens in $\mathcal{A}_i$ receive supervised loss, and non-anchor reference tokens serve only as teacher-forced context. Thus, **ACE** is sparser than ordinary SFT even when $u_0 > 0$ for extracted anchors. It complements **RoCo** by adding reference-side correction for target facts that are absent or weakly covered in the rollout, while **RoCo** reallocates distillation over student-reachable rollout positions.

3.4 Overall Objective

For a mini-batch $\mathcal{B}$, the final training objective combines rollout-side distillation and reference-side anchor correction:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(\lambda_{\text{RoCo}} \mathcal{L}_{\text{RoCo}}^{(i)} + \lambda_{\text{ACE}} \mathcal{L}_{\text{ACE}}^{(i)} \right), \tag{8}$$

where $\lambda_{\text{RoCo}} \geq 0$ and $\lambda_{\text{ACE}} \geq 0$ are loss coefficients. Here, $\mathcal{L}_{\text{RoCo}}^{(i)}$ is evaluated on the valid student-rollout positions $\mathcal{V}_i$, and $\mathcal{L}_{\text{ACE}}^{(i)}$ is evaluated only on the reference anchor-token set $\mathcal{A}_i$. Together, these two losses provide complementary supervision: **RoCo** reallocates distillation over reference-aligned reachable rollout positions, while **ACE** adds sparse reference-side correction for missing or weakly covered anchors, enabling targeted factual injection without full-reference imitation.

4 Experiments

4.1 Settings

Benchmarks. We evaluate knowledge injection along two axes: injected-knowledge acquisition and evaluated retention. For acquisition, we use EVOKE (Jiang et al., 2025a), which contains image-grounded entity and news updates with authoritative reference answers, and VP/Sci from MLLM-CL (Zhao et al., 2025), which test broader visual-perception and science-domain update settings. Appendix H provides detailed benchmark descriptions. For retention, we evaluate every trained model on TreeBench (Wang et al., 2026), VStar (Wu and Xie, 2024), MathVision (Wang et al., 2024), MMStar (Chen et al., 2024), BabyVision (Chen et al., 2026), and MM-SafetyBench (Liu et al., 2024b), abbreviated as MM-Safety in tables. These benchmarks cover visual grounding, fine-grained perception, multimodal mathematical rea-

soning, general multimodal understanding, core visual primitives, and multimodal safety behavior; we report their arithmetic mean as Ret. Avg.

Evaluation metrics. We report accuracy for injected-knowledge acquisition and evaluated retention. Open-ended injection responses are judged against the authoritative answer by an LLM judge, which focuses on factual consistency-including entity identity, event identity, date, number, location, and relation. Benchmark-specific tasks follow their standard answer formats and normalization rules. For multiple-choice tasks, we extract the predicted option from the model response. Appendix I provides details of the judge model, answer extraction, and normalization protocol.

Models and baselines. The main experiments use **Qwen3-VL-8B** as the base model; Qwen3-VL-30B results are reported in Appendix D. We compare direct fitting baselines (LoRA (Hu et al., 2022) and SFT (Ouyang et al., 2022)), reference-conditioned online distillation (SDFT with authoritative reference (Shenfeld et al., 2026)), constrained-update baselines (KORE (Jiang et al., 2025b), KeepLoRA (Luo et al., 2026), MoELoRA (Luo et al., 2024), and SEFE (Chen et al., 2025)), and our variants, **RoCo** and **RoCo-ACE**. Unless otherwise stated, all methods use the same injected examples, training budget, image preprocessing, and evaluation protocol; only the update rule or supervision objective changes. KORE is the only exception, as it additionally uses auxiliary general data to estimate its retention constraint. Implementation details and hyperparameters are provided in Appendix K.

4.2 Baseline Comparison

Figure 3 visualizes the EVOKE injection–retention trade-off, and Table 1 summarizes Inj. Acc. and Ret. Avg. across EVOKE, VP, and Sci. Full per-benchmark retention values and VP/Sci radar plots are provided in Appendix B. Representative EVOKE responses are shown in Appendix C.1. Overall, Table 1 shows that existing baselines tend to favor either injection or retention, while **RoCo-ACE** improves injected-knowledge accuracy with limited drift in evaluated retention.

Fitting-oriented baselines prioritize injection. LoRA and SFT directly optimize authoritative answers, making them natural baselines for injected-knowledge acquisition. However, full-answer supervision also fits generic wording, answer style, and non-target behavior rather than only target factual anchors. In Table 1, SFT improves EVOKE Inj. Acc. from 16.4 to 20.1, but Ret. Avg. drops from 57.2 to 31.8. LoRA has less severe drift, reaching 18.6 Inj. Acc. and 45.2 Ret. Avg. The same trend appears on VP and Sci, where both direct-fitting baselines improve Inj. Acc. over the base model but reduce Ret. Avg. These results show that full-answer fitting provides a strong injection signal, but is not ideal for retention-oriented knowledge injection.

Method	EVOKE		VP		Sci	
	Inj. Acc.	Ret. Avg.	Inj. Acc.	Ret. Avg.	Inj. Acc.	Ret. Avg.
Qwen3-VL-8B	16.4	**57.2**	69.9	**57.2**	73.1	**57.2**
LoRA	18.6	45.2	74.5	54.8	79.6	52.6
SFT	20.1	31.8	72.7	51.0	83.7	51.0
SDFT (w/ ref.)	20.9	55.1	74.1	56.4	81.5	***56.7***
KORE	***25.2***	47.0	***76.4***	52.5	***86.7***	48.3
KeepLoRA	17.3	***56.6***	74.0	52.0	74.6	54.2
MoELoRA	17.4	48.7	71.2	55.7	74.0	55.7
SEFE	19.8	44.8	74.8	54.3	77.8	52.1
RoCo	23.7	***56.6***	76.2	***56.5***	85.6	56.4
RoCo-ACE	**27.6**	56.5	**77.8**	56.2	**87.5**	56.3

Table 1: Baseline comparison across three knowledge-injection settings. Inj. Acc. reports injected-knowledge accuracy, and Ret. Avg. averages the six held-out retention benchmarks. Higher is better. Bold and bold-italic mark the best and second-best scores.

Constrained methods can favor stronger injection. KORE is the strongest non-ours method for Inj. Acc. in all three settings, reaching 25.2 on EVOKE, 76.4 on VP, and 86.7 on Sci in Table 1. This suggests that knowledge-injection-specific augmentation and constraints can strengthen target acquisition. The cost is evaluated retention: KORE obtains Ret. Avg. scores of 47.0, 52.5, and 48.3 on EVOKE, VP, and Sci, respectively, lower than the base model and online-distillation baselines. KORE therefore represents the injection-dominant side of the trade-off.

Retention-oriented baselines limit drift but under-inject. SDFT (w/ ref.) and KeepLoRA better maintain Ret. Avg. but are weaker on Inj. Acc. On EVOKE, SDFT keeps Ret. Avg. at 55.1, close to the base model's 57.2, but reaches only 20.9 Inj. Acc. KeepLoRA reaches 56.6 Ret. Avg. but only 17.3 Inj. Acc. MoELoRA and SEFE show intermediate behavior across the three settings. These results are consistent with the limitation of conser-

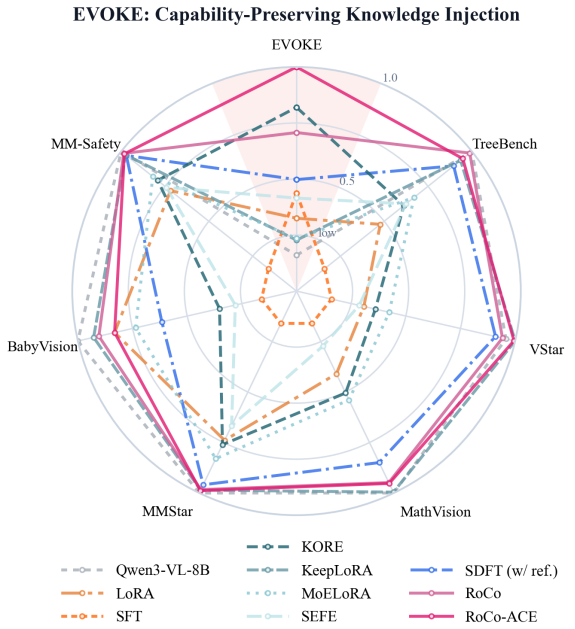

Figure 3: Radar comparison on EVOKE. The red axis reports injected-knowledge accuracy, and the other axes report evaluated retention on six held-out multimodal and safety benchmarks. Scores are normalized for visualization; exact values are reported in Appendix B. **RoCo-ACE** improves injected-knowledge accuracy while keeping evaluated retention close to online-distillation and retention-oriented baselines.

vative updates or uniform rollout distillation: they can limit evaluated retention drift, but leave some sparse target facts under-injected.

RoCo-ACE improves the trade-off. **RoCo** improves over the SDFT reference point by replacing uniform rollout distillation with same-rollout likelihood contrast. On EVOKE, it increases Inj. Acc. from 20.9 to 23.7 and Ret. Avg. from 55.1 to 56.6. Adding **ACE** further supplies sparse correction for authoritative anchors omitted from or weakly matched by the rollout. **RoCo-ACE** achieves the best EVOKE Inj. Acc. among compared methods, 27.6, while keeping Ret. Avg. at 56.5, only 0.7 below the base model. Compared with KORE, it improves EVOKE Inj. Acc. by 2.4 points and Ret. Avg. by 9.5 points. Compared with SFT, it improves them by 7.5 and 24.7 points. The same pattern holds on VP and Sci in Table 1. **RoCo-ACE** obtains the best Inj. Acc. on both settings, 77.8 and 87.5, while keeping Ret. Avg. at 56.2 and 56.3, close to SDFT (w/ ref.) and higher than KORE. Overall, the results support the central design: **RoCo** reweights reference-supported rollout tokens through same-rollout likelihood contrast, while **ACE** adds sparse reference-side correction without full-answer imitation.

4.3 Ablation Study

Table 2 isolates the main design choices. SDFT (w/ ref.) and **RoCo** use the same rollout, reference-conditioned teacher target, optimizer, and training data. The key difference is whether rollout tokens receive uniform weights or weights from same-rollout likelihood contrast. **RoCo** improves EVOKE Inj. Acc. from 20.9 to 23.7 and Ret. Avg. from 55.1 to 56.6, showing that reweighting reference-supported rollout tokens improves over uniform reference-conditioned distillation.

We further include teacher-side controls. First, a fixed-teacher variant disables EMA teacher updates while keeping the same **RoCo-ACE** objective. Its EVOKE Inj. Acc. drops from 27.6 to 25.6, while Ret. Avg. remains close to the EMA-teacher setting. This suggests that a fixed teacher can still keep evaluated retention stable, but provides a weaker injection signal than an EMA teacher that tracks the current student during online distillation.

Second, a mismatched-reference control feeds the reference-conditioned teacher a shuffled authoritative answer from another training example, while keeping the student rollout, original prompt, teacher, optimizer, and matched ACE anchors unchanged. This preserves the reference pathway and context length but removes semantic alignment between the input and the teacher-side reference. EVOKE Inj. Acc. drops to 18.5, indicating that matched reference content is important for the RoCo likelihood contrast. Ret. Avg. remains high, and injection does not collapse to the base level because ACE still provides sparse correction for omitted or weakly matched authoritative anchors, while the rollout-distillation path remains anchored to the original prompt.

ACE addresses the complementary omission case. SDFT+ACE reaches 25.9 EVOKE Inj. Acc., showing that sparse anchored correction helps even without same-rollout likelihood contrast. ACE only reaches 25.1 EVOKE Inj. Acc. but lowers Ret. Avg. to 53.6, suggesting that anchor correction alone improves injection but lacks the rollout-distillation signal needed to limit drift. The full **RoCo-ACE** objective combines rollout-side reweighting with reference-side anchor correction, reaching the best EVOKE Inj. Acc. of 27.6 while keeping Ret. Avg. at 56.5.

The full-reference CE controls show why anchor-level correction matters. Here, SFT applies uni-

Variant	EVOKE	Retention Benchmarks						
		TreeBench	VStar	MathVision	MMStar	BabyVision	MM-Safety	Ret. Avg.
Base model	16.4	**43.3**	83.2	**49.3**	**70.6**	***14.2***	***82.6***	**57.2**
SFT	20.1	15.8	68.1	29.8	35.5	10.7	30.9	31.8
SDFT (w/ ref.)	20.9	39.9	82.3	45.8	68.9	12.6	80.8	55.1
SDFT + SFT	21.8	38.1	80.6	43.2	66.1	11.9	77.9	53.0
SDFT + ACE	***25.9***	40.0	**85.1**	45.2	68.8	12.5	80.3	55.3
RoCo	23.7	42.9	82.9	48.1	69.9	13.8	81.7	56.6
RoCo-ACE (mismatched ref.)	18.5	39.5	83.1	46.3	68.9	13.5	82.0	56.8
RoCo-ACE (fixed teacher)	25.6	***43.0***	83.1	***48.9***	***70.1***	**14.6**	**82.8**	***57.0***
RoCo + SFT	24.8	41.1	82.1	45.2	65.2	12.4	80.1	54.4
ACE only	25.1	40.6	79.5	46.8	64.6	11.9	78.5	53.6
RoCo-ACE	**27.6**	41.6	***83.9***	48.2	***70.1***	13.5	81.5	56.5

Table 2: Ablation results for Qwen3-VL-8B with full retention metrics. EVOKE measures injected-knowledge accuracy. Retention columns report held-out general multimodal and safety accuracy, and Ret. Avg. averages TreeBench, VStar, MathVision, MMStar, BabyVision, and MM-Safety. Higher values are better. Bold and bold-italic mark the best and second-best scores in each column.

form CE over the complete authoritative answer, while SDFT+SFT and **RoCo**+SFT add the same full-reference CE term to the corresponding online-distillation objectives. SDFT+SFT and **RoCo**+SFT improve injection less effectively than their ACE counterparts and lower Ret. Avg. to 53.0 and 54.4, respectively. This shows that the gains do not come merely from adding ordinary supervised CE. Anchor-level correction is more favorable than full-answer token-level imitation because it restricts supervised loss to extracted factual spans rather than generic reference wording or answer style.

4.4 Qualitative Example

Figure 4 shows an EVOKE example from the training logs. Additional cases are provided in Appendix C. The rollout reaches the right visual and topical neighborhood, mentioning Amsterdam, NDSM, street art, and graffiti, but omits the target museum identity and several concrete facts. **RoCo** assigns larger weights to these reference-supported rollout tokens, while **ACE** adds sparse anchored correction for omitted authoritative anchors such as "STRAAT Museum", "October 9, 2020", and "86,000 sq ft". This example illustrates the intended split: rollout-side reweighting for content already reached on-policy, and reference-side correction for omitted facts.

Prompt
Could you explain the details of the museum featured in the image?

Student rollout
The model identifies a different Amsterdam museum and mixes unsupported history about a warehouse and cooperative organizations. It nevertheless stays close to the visual category by mentioning Amsterdam, NDSM, street art, graffiti, and a cultural venue.

Authoritative reference
The STRAAT Museum is an Amsterdam NDSM street-art and graffiti museum opened on October 9, 2020. Its 86,000 sq ft space contains over 180 works from more than 170 artists, includes a mezzanine gallery, and features exterior murals such as Eduardo Kobra's 2016 Anne Frank depiction.

RoCo signal
High-weight rollout spans include "museum in Amsterdam", "NDSM", "street art", and "graffiti". These partial tokens preserve useful visual and topical context.

ACE correction
Missing anchors include "STRAAT Museum", "October 9, 2020", "86,000 sq ft", "over 180 works", "more than 170 artists", "mezzanine gallery", and "Eduardo Kobra".

Figure 4: Representative EVOKE example illustrating the complementarity between **RoCo** reweighting and **ACE** anchored correction. **RoCo** emphasizes reference-supported rollout tokens already reached on-policy, while **ACE** supplies sparse correction for authoritative anchors omitted from the rollout.

5 Conclusion

We presented **RoCo-ACE**, a rollout-conditioned contrastive online distillation objective for knowledge injection. **RoCo** uses same-rollout reference-free/reference-conditioned likelihood contrast to reweight reference-supported rollout tokens, while **ACE** adds sparse anchored correction for authoritative anchors omitted from or weakly matched by the rollout. Together, the two objectives inject sparse factual content without fitting the full reference answer as a token-level imitation target. Experiments across three knowledge-injection settings and multiple retention benchmarks show improved injected-knowledge accuracy while keeping evaluated retention close to the base model.

Limitations

RoCo-ACE is designed for knowledge-injection settings where each update is accompanied by an authoritative natural-language reference. Its supervision is organized around student rollout tokens and extracted factual anchors, making it most directly suited to entity, event, domain, and visual-world updates whose target facts can be expressed as compact text spans. In label-only settings, very short answers, or cases where the relevant evidence is mostly implicit in the image, the available anchor signal becomes weaker.

Although images are included in both student and teacher contexts, the current objective does not use region-level, OCR-level, or object-level evidence annotations to explicitly ground each corrected anchor. A natural extension is to tie anchored factual correction more directly to visual evidence through image-ablation contrast, OCR/entity grounding, or region-level supervision. Finally, retention is evaluated empirically on held-out multimodal and safety benchmarks; the method is not intended to provide formal guarantees over all non-updated capabilities or deployment distributions.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. 2023. On-policy distillation of language models: Learning from self-generated mistakes. *arXiv preprint arXiv:2306.13649*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Chang Che, Ziqi Wang, Pengwan Yang, Qi Wang, Hui Ma, and Zenglin Shi. 2025. Lora in lora: Towards parameter-efficient architecture expansion for continual visual instruction tuning. *arXiv preprint arXiv:2508.06202*.

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

Liang Chen, Weichu Xie, Yiyan Liang, Hongfeng He, Hans Zhao, Zhibo Yang, Zhiqi Huang, Haoning Wu, Haoyu Lu, Y. charles, Yiping Bao, Yuantao Fan, Guopeng Li, Haiyang Shen, Xuanzhong Chen, Wendong Xu, Shuzheng Si, Zefan Cai, Wenhao Chai, and 10 others. 2026. BabyVision: Visual reasoning beyond language. *Preprint*, arXiv:2601.06521.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven C. H. Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2023. Minillm: Knowledge distillation of large language models. *arXiv preprint arXiv:2306.08543*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Kailin Jiang, Yuntao Du, Yukai Ding, Yuchen Ren, Ning Jiang, Zhi Gao, Zilong Zheng, Lei Liu, Bin Li, and Qing Li. 2025a. When large multimodal models confront evolving knowledge: Challenges and pathways. *arXiv preprint arXiv:2505.24449*.

Kailin Jiang, Hongbo Jiang, Ning Jiang, Zhi Gao, Jinhe Bi, Yuchen Ren, Bin Li, Yuntao Du, Lei Liu, and Qing Li. 2025b. Kore: Enhancing knowledge injection for large multimodal models via knowledge-oriented augmentations and constraints. *arXiv preprint arXiv:2510.19316*.

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. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Gengsheng Li, Tianyu Yang, Junfeng Fang, Mingyang Song, Mao Zheng, Haiyun Guo, Dan Zhang, Jinqiao Wang, and Tat-Seng Chua. 2026. Unifying group-relative and self-distillation policy optimization via sample routing. *arXiv preprint arXiv:2604.02288*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730–19742.

Zhizhong Li and Derek Hoiem. 2016. Learning without forgetting. In *Proceedings of the European Conference on Computer Vision*, pages 614–629. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. 2024a. MIBench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22417–22428, Miami, Florida, USA. Association for Computational Linguistics.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. 2024b. MM-SafetyBench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pages 386–403. Springer.

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

Tongxu Luo, Jiahe Lei, Fangyu Lei, Weihao Liu, Shizhu He, Jun Zhao, and Kang Liu. 2024. Moelora: Contrastive learning guided mixture of experts on parameter-efficient fine-tuning for large language models. *arXiv preprint arXiv:2402.12851*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. 2024. MLLM-protector: Ensuring MLLM's safety without hurting performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16012–16027, Miami, Florida, USA. Association for Computational Linguistics.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Hejian Sang, Yuanda Xu, Zhengze Zhou, Ran He, Zhipeng Wang, and Jiachen Sun. 2026. Crisp: Compressed reasoning via iterative self-policy distillation. *arXiv preprint arXiv:2603.05433*.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. *arXiv preprint arXiv:2601.19897*.

Mingyang Song and Mao Zheng. 2026. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*.

Pengwei Tang, Xiaolin Hu, Yong Liu, Lizhong Ding, Dongjie Zhang, Xing Wu, and Debing Zhang. 2025. Put the space of lora initialization to the extreme to preserve pre-trained knowledge. *arXiv preprint arXiv:2503.02659*.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. 2026. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *The Fourteenth International Conference on Learning Representations*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with the math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Shipeng Wang, Xiaorong Li, Jian Sun, and Zongben Xu. 2021. Training networks in null space of feature covariance for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 184–193.

Penghao Wu and Saining Xie. 2024. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13084–13094.

Weixiang Wu, Dongxu Wang, Renze Xu, Xinnan Ma, Zhen Zeng, Xiaoshuai Guo, Gao Huang, and Yinghui Wang. 2025. Automating steering for safe multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

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

Tianzhu Ye, Li Dong, Xun Wu, Shaohan Huang, and Furu Wei. 2026. On-policy context distillation for language models. *arXiv preprint arXiv:2602.12275*.

Hongbo Zhao, Fei Zhu, Haiyang Guo, Meng Wang, Rundong Wang, Gaofeng Meng, and Zhaoxiang Zhang. 2025. Mllm-cl: Continual learning for multimodal large language models. *arXiv preprint arXiv:2506.05453*.

Siyan Zhao, Zhihui Xie, Mengchen Liu, Jing Huang, Guan Pang, Feiyu Chen, and Aditya Grover. 2026. Self-distilled reasoner: On-policy self-distillation for large language models. *arXiv preprint arXiv:2601.18734*.

A Anchor Extraction and Supervision-Weight Diagnostics

Token category	Share	Component	Diagnostic rule
Reference-supported rollout tokens	30.8%	RoCo enhanced KD	Valid rollout tokens satisfying $\Delta_{i,t} > \tau$. These positions receive additional RoCo contrast weight.
Generic or weakly supported rollout tokens	57.3%	RoCo KD floor	Valid rollout tokens without RoCo contrast activation; these positions are mainly retained by the online-distillation floor.
Missing target-anchor tokens	11.9%	ACE correction	Extracted authoritative-anchor tokens absent from the rollout; these positions receive ACE missing-span correction.

Table 3: Token-allocation diagnostics on EVOKE training rollouts. The diagnostic pool consists of valid student-rollout tokens plus missing reference-anchor tokens. The three categories are mutually exclusive and sum to one, showing how RoCo-ACE allocates RoCo and ACE supervision across reference-supported rollout content, generic rollout content, and missing target anchors. Percentages are computed from logged RoCo-ACE training intervals under the paper's default anchor weighting.

This section describes the lightweight span extraction used by **ACE** and RoCo span smoothing, and reports a supervision-weight diagnostic that connects these extracted units to the claims in the main paper. The diagnostic is not intended to estimate final gradient norms or normalized loss mass. Instead, it shows how training units are assigned to the RoCo contrast term, the RoCo floor term, and reference-side ACE correction.

Span extraction. Span construction is a tokenizer-level procedure and does not rely on external NER or parsing tools. We decode valid response tokens, remove special tokens, punctuation-only tokens, and common function words, and group contiguous content tokens containing CJK characters, Latin letters, or digits into spans. A span is kept if it contains at least two visible characters or any digit, and we cap each span by a maximum token length. This simple extractor is designed to capture compact factual expressions such as entity names, dates, numbers, titles, locations, and short noun phrases.

Anchor matching for ACE. For ACE, the same extractor is applied to the tokenized authoritative answer. Each reference span is normalized by lowercasing and retaining only alphanumeric and CJK characters; the student rollout is normalized in the same way. For each extracted reference span s, we compute an anchor match score $m_i(s, y_i) \in [0, 1]$ against the rollout. In our default implementation, $m_i(s, y_i) = 1$ when the normalized span string appears in the normalized rollout and 0 otherwise; the same notation also supports partial-match variants. Thus, matched anchors receive the base anchor weight, while omitted or weakly matched anchors receive larger weights. The ACE token set $\mathcal{A}_i$ is the union of reference-token positions covered by the extracted anchors. Tokens outside $\mathcal{A}_i$ serve only as teacher-forced context and do not receive supervised CE loss, making ACE narrower than full-reference SFT.

Span smoothing for RoCo. The same extractor is also used for RoCo span-aware smoothing on the rollout side. After preliminary RoCo weights are computed from the same-rollout likelihood contrast, we aggregate token weights within each extracted rollout span and broadcast the span score back to all tokens in that span. For notation, the resulting smoothed weights are still denoted by $w_{i,t}$. This reduces subword-level noise for names, dates, numbers, multilingual expressions, and other compact factual units without requiring external entity taggers. The diagnostic in Table 3 classifies rollout tokens by their pre-smoothing contrast activation, so it reflects the source of the RoCo weight signal before span-level broadcasting.

Supervision-weight diagnostic. Table 3 reports a token-level diagnostic on EVOKE training rollouts. The diagnostic pool contains valid rollout tokens and omitted authoritative-anchor tokens. These units correspond to the three supervision cases used by **RoCo-ACE**: rollout tokens receiving the RoCo contrast term, rollout tokens receiving mainly the RoCo floor term, and reference-side anchors receiving ACE omission correction.

For rollout tokens, Table 3 uses the RoCo likelihood contrast as the diagnostic rule. Tokens with $\Delta_{i,t} > \tau$ are counted as reference-supported rollout tokens; they account for 30.8% of the diagnostic pool and receive the RoCo floor plus the additional contrast term. Tokens with $\Delta_{i,t} \leq \tau$ are counted as generic or non-activated rollout tokens; they account for 57.3% and mainly receive the online-distillation floor. For reference-side units, omitted authoritative-anchor tokens account for 11.9% and receive larger ACE weights through the omission term $1 - \rho_i(s, y_i)$.

This distribution supports the intended supervision split. Uniform reference-conditioned roll-

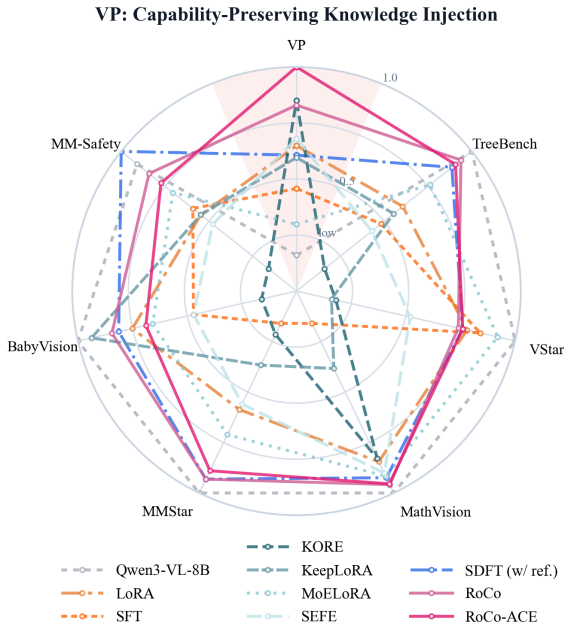

Figure 5: Radar comparison on VP, showing target knowledge injection and general capability retention across six multimodal benchmarks.

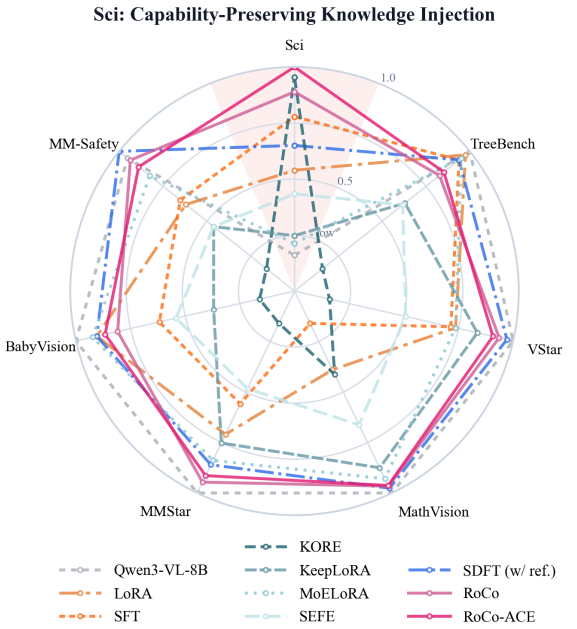

Figure 6: Radar comparison on Sci, showing target knowledge injection and general capability retention across six multimodal benchmarks.

out distillation would place its rollout-side loss over all valid rollout tokens, most of which are not contrast-activated in this diagnostic. RoCo instead reallocates additional distillation weight to reference-supported rollout tokens, while ACE supplies sparse correction for omitted authoritative anchors that are not available as generated rollout tokens in the current step. Because Table 3 reports token-count categories rather than normalized loss mass, it should be interpreted as a weight-assignment diagnostic rather than an exact gradient-budget decomposition.

B Full Baseline Comparison Results

Table 4 reports the full baseline comparison on EVOKE using Qwen3-VL-8B as the base model. In addition to the injected-knowledge accuracy, we include all individual retention benchmarks to show how each method affects general multimodal capabilities after knowledge injection. This table complements the main-text discussion by making the injection–retention trade-off explicit across direct fine-tuning, constrained-update baselines, online distillation, and our RoCo-ACE variants.

Table 5 further evaluates the same set of methods on the VP and Sci knowledge-injection benchmarks from MLLM-CL. VP focuses on ability-oriented injection, while Sci focuses on domain-oriented knowledge injection. We keep the retention benchmark suite unchanged across EVOKE, VP, and Sci, which allows a consistent comparison of whether each method can acquire new knowledge while preserving general capabilities. Across these additional settings, RoCo-ACE maintains a favorable balance between injected-knowledge accuracy and retention performance.

Figures 5 and 6 provide the corresponding radar visualizations for VP and Sci. We move these plots to the appendix because the main text summarizes the three datasets in Table 1 using injected-knowledge accuracy and Ret. Avg.

C Qualitative Analysis

Table 6 presents three non-duplicate EVOKE cases selected from the joint **RoCo-ACE** training logs. The color scheme follows the method overview: orange boxes summarize the student rollout, green boxes show the strongest **RoCo** rollout signals, and red boxes show facts corrected by Anchored Cross-Entropy. Together, these examples show how **RoCo** emphasizes useful on-policy content while ACE supplies missing or contradicted reference facts.

The baseline-family examples in Appendix C.1 further illustrate how different method families behave on sparse authoritative references. Direct fitting usually captures only a few salient facts, constrained-update methods cover a moderate sub-

Method	Knowledge Injection	Retention Benchmark						
	EVOKE	TreeBench	VStar	MathVision	MMStar	BabyVision	MM-Safety	Ret. Avg.
Qwen3-VL-8B	16.4	**43.3**	83.2	**49.3**	**70.6**	**14.2**	**82.6**	**57.2**
LoRA	18.6	26.2	70.9	35.6	59.7	13.5	65.3	45.2
SFT	20.1	15.8	68.1	29.8	35.5	10.7	30.9	31.8
SDFT (w/ ref.)	20.9	39.9	82.3	45.8	68.9	12.6	80.8	55.1
KORE	***25.2***	30.6	71.9	37.8	60.6	11.5	69.8	47.0
KeepLoRA	17.3	41.2	**84.0**	***49.2***	70.0	***13.9***	81.1	***56.6***
MoELoRA	17.4	32.5	73.1	38.6	63.5	13.1	71.5	48.7
SEFE	19.8	31.0	70.5	32.4	56.7	11.2	66.7	44.8
RoCo	23.7	***42.9***	82.9	48.1	69.9	13.8	***81.7***	***56.6***
RoCo-ACE	**27.6**	41.6	***83.9***	48.2	***70.1***	13.5	81.5	56.5

Table 4: Main baseline comparison on Qwen3-VL-8B. The knowledge-injection column reports LLM-judge accuracy on EVOKE. Retention benchmark columns report accuracy on held-out general multimodal and safety benchmarks, and Ret. Avg. averages TreeBench, VStar, MathVision, MMStar, BabyVision, and MM-Safety. Higher values are better; bold and bold-italic mark the best and second-best scores in each column.

Dataset	Method	Injection	Retention Benchmarks						
		Acc.	TreeBench	VStar	MathVision	MMStar	BabyVision	MM-Safety	Ret. Avg.
VP	Qwen3-VL-8B	69.9	**43.3**	**83.2**	**49.3**	**70.6**	**14.2**	***82.6***	**57.2**
	LoRA	74.5	39.5	82.1	45.2	67.6	13.4	81.0	54.8
	SFT	72.7	38.3	82.4	27.1	64.5	12.5	81.2	51.0
	SDFT (w/ ref.)	74.1	42.2	82.0	47.3	***70.1***	13.6	**83.0**	56.4
	KORE	***76.4***	35.2	79.1	44.8	64.9	11.5	79.3	52.5
	KeepLoRA	74.0	39.0	79.0	33.0	66.0	***14.0***	81.0	52.0
	MoELoRA	71.2	41.0	***82.8***	47.1	68.5	13.1	81.7	55.7
	SEFE	74.8	37.8	80.8	46.7	67.4	12.5	80.7	54.3
	RoCo	76.2	***42.7***	81.9	***48.2***	***70.1***	13.7	82.3	***56.5***
	RoCo-ACE	**77.8**	42.4	82.0	48.1	69.8	13.2	82.0	56.2
Sci	Qwen3-VL-8B	73.1	**43.3**	**83.2**	**49.3**	**70.6**	**14.2**	***82.6***	**57.2**
	LoRA	79.6	***43.1***	79.5	31.2	67.9	13.7	80.5	52.6
	SFT	83.7	42.8	79.2	24.4	66.5	12.2	80.7	51.0
	SDFT (w/ ref.)	81.5	42.8	***82.8***	***48.6***	69.3	13.7	**82.9**	***56.7***
	KORE	***86.7***	36.5	71.4	31.9	62.8	9.8	77.6	48.3
	KeepLoRA	74.6	40.3	80.9	45.6	68.3	10.9	79.5	54.2
	MoELoRA	74.0	42.9	79.5	47.2	69.1	***13.8***	81.8	55.7
	SEFE	77.8	40.2	76.3	39.3	65.8	11.8	79.5	52.1
	RoCo	85.6	41.9	82.3	48.3	***70.1***	13.2	82.5	56.4
	RoCo-ACE	**87.5**	42.1	81.9	48.2	69.8	13.5	82.2	56.3

Table 5: Baseline comparison on the MLLM-CL VP and Sci knowledge-injection benchmarks (Zhao et al., 2025) with full retention metrics. Acc. reports injected-knowledge accuracy for the corresponding dataset. Retention columns report accuracy after training on that dataset, and Ret. Avg. averages TreeBench, VStar, MathVision, MMStar, BabyVision, and MM-Safety. Higher values are better; bold and bold-italic mark the best and second-best scores in each column.

set, online distillation preserves natural response style but under-injects, and **RoCo-ACE** recovers more anchors without copying the full reference.

D Supplementary Results on Qwen3-VL-30B

Table 8 keeps the benchmark columns and baseline groups consistent with the main Qwen3-VL-8B table to test whether the injection-retention trends persist at a larger model scale. The 30B results show the same qualitative pattern: direct fitting improves injected knowledge but sacrifices retention, constrained-update methods are more conservative, and **RoCo-ACE** gives the strongest injected-knowledge result while keeping retention close to or above the base model.

Table 9 reports component-level Qwen3-VL-30B ablations with every retention benchmark. The larger model follows the same trend as Qwen3-VL-

Case 1: generic under-injection in a business entity

Prompt
What details can you give about the business in the image?

Evolution

Authoritative reference
Evolution AB is a Swedish gaming technology company based in Stockholm, specializing in B2B live casino software for online operators. It was founded in 2006 by Jens von Bahr and Fredrik Österberg, became publicly listed on Nasdaq in 2015, introduced games such as Monopoly Live and Lightning Roulette, acquired Ezugi and NetEnt, and attracted investment from Kenneth Dart in 2024.

Student rollout
The rollout gives a generic description of startups, growth, innovation, vision clarity, branding, customer loyalty, and market adaptation. It never identifies Evolution AB or its concrete business facts.

RoCo signal
The strongest rollout signals are generic business phrases such as "Startups abound with growth and innovation" and "successful startups have a crystal-clear mission". This case exposes a limitation of rollout-only reweighting: without the target entity in the rollout, the contrast cannot directly teach Evolution AB.

ACE correction
Missing spans include "Evolution AB", "Swedish gaming technology company", "Stockholm", "B2B live casino software", "Jens von Bahr", "Fredrik Österberg", "Nasdaq in 2015", "Monopoly Live", "Lightning Roulette", "Ezugi", "NetEnt", and "Kenneth Dart in 2024".

Case 2: sports news numeric/event correction

Prompt
What information can you offer about the sport news presented in the image?

Authoritative reference
The Kansas City Chiefs defeated the Denver Broncos 16–14 at Arrowhead Stadium on November 10, 2024, blocking a last-second field goal to preserve their unbeaten streak. Patrick Mahomes led a late-game drive, Leo Chenal made the decisive block, and Travis Kelce set a franchise touchdown-reception record.

Student rollout
The rollout reaches the right sports-news neighborhood: it mentions Travis Kelce, the Kansas City Chiefs, the Denver Broncos, Arrowhead Stadium, and a narrow victory. However, it shifts the event to a season opener or preseason game, gives an imprecise three-point outcome, and introduces unsupported players and game details.

RoCo focus
High-weight rollout spans include "Travis Kelce and the Kansas City Chiefs", "victory against the Denver Broncos", "Arrowhead Stadium", and "brief summary of the news". These tokens preserve useful event context rather than treating the entire generated answer uniformly.

ACE correction
Missing or contradicted spans include "16–14", "last-second field goal attempt", "November 10, 2024", "Patrick Mahomes", "Leo Chenal", "franchise record", "Taylor Swift", "15th consecutive win", and "Buffalo Bills".

Case 3: India news collaborator substitution

Prompt
Can you please share some details about the india news shown in the image?

Authoritative reference
Ed Sheeran performed a duet with Punjabi star Diljit Dosanjh in Mumbai before a 50,000-strong audience, singing a Punjabi rendition of Dosanjh's hit "Lover". The surprise collaboration at Mahalaxmi Racecourse drew social-media attention during Sheeran's "Mathematics" tour.

Student rollout
The rollout remains in the broad music-news setting and mentions Ed Sheeran, Punjabi music, collaboration, crowds, and concerts, but it substitutes the collaborator with Sidhu Moose Wala and introduces unrelated claims about surveillance and cancelled tours.

RoCo focus
High-weight spans include "summary of the news", "British singer Ed Sheeran", "Punjabi", "joint performances", and "large crowds", so online distillation is concentrated on the useful music-news neighborhood.

ACE correction
Missing or contradicted spans include "Diljit Dosanjh", "Mumbai", "50,000-strong audience", "Punjabi rendition of Dosanjh's hit Lover", "Mahalaxmi Racecourse", "Mathematics tour", and "performing in Punjabi for the first time".

Table 6: Qualitative EVOKE examples from joint **RoCo-ACE** training. We retain three non-duplicate cases that cover entity under-injection, sports-event correction, and collaborator substitution. Across these cases, **RoCo** emphasizes useful content already present in the rollout, while ACE supplies authoritative facts that the rollout omits or contradicts.

8B: full-answer SFT controls improve EVOKE less efficiently and reduce retention, ACE improves target injection, and combining it with RoCo gives the best injected-knowledge accuracy among the compared variants while keeping the retention average close to the base model.

C.1 Baseline-Family Behavior Examples. Representative EVOKE cases illustrating how baseline families behave on sparse authoritative references: direct fitting captures a few salient facts, constrained updating covers a moderate subset, online distillation preserves natural response style but under-injects, and **RoCo-ACE** recovers more anchors without copying the full reference.

Case A: album entity with release and credit facts

Question
Could you introduce the album in the image?

Gold facts
Almighty So 2 is the **fifth studio album** by **Chief Keef**, released on **May 10, 2024** by **43B**. It is a **sequel to Almighty So**, includes guests such as **G Herbo, Tierra Whack, Sexyy Red, and Quavo**, has production by Keef and several collaborators, and peaked at **number 30 on the Billboard 200**.

SFT / direct fitting
Captures **Almighty So 2**, **Chief Keef**, and a 2024 album identity, but gives little detail about the label, guest artists, production, or chart performance.

KORE / constrained update
Adds the **fifth studio album**, **May 10, 2024**, and **43B** facts, but still only briefly covers guests and omits most production and chart details.

SDFT / online distillation
Keeps natural album-cover wording and visual grounding, but avoids firm claims about the exact **release label, guest list, producers, and Billboard ranking**.

RoCo-ACE / rollout-conditioned anchors
Recovers **Almighty So 2**, **fifth studio album**, **Chief Keef**, **May 10, 2024**, **43B**, several guest artists, and **number 30 on the Billboard 200**, while not copying every production detail.

Case B: film entity with series, cast, and release facts

Question
Can you share information about the film presented in the image?

Gold facts
1995 is a **2024 Canadian comedy film** written and directed by **Ricardo Trogi**. It is the **fourth film** in his semi-autobiographical series after **1981, 1987, and 1991**, stars **Jean-Carl Boucher as Trogi**, and was released on **July 31, 2024** by **Immina Films**.

SFT / direct fitting
Captures **1995**, **Canadian comedy film**, and **Ricardo Trogi**, but gives little information about the cast, release date, distributor, or story setup.

KORE / constrained update
Adds **2024**, the semi-autobiographical series relation, and **Jean-Carl Boucher**, but simplifies the premise and omits distributor details.

SDFT / online distillation
Preserves natural film-poster description, but remains cautious about the exact **series order, cast, release date, and distributor**.

RoCo-ACE / rollout-conditioned anchors
Covers **1995**, **2024 Canadian comedy**, **Ricardo Trogi**, the **fourth-entry** relation, **Jean-Carl Boucher as Trogi**, and **July 31, 2024**, while leaving out some supporting cast details.

E Supplementary Results on InternVL3.5-8B

Table 10 reports an EVOKE component comparison on InternVL3.5-8B, testing whether the component-level trend of **RoCo-ACE** transfers beyond the Qwen3-VL model family. The comparison follows the same protocol as the Qwen3-VL ablations: the base model is compared with **RoCo**, ACE-only training, and the full **RoCo-ACE** objective on EVOKE, while using the same six held-out retention benchmarks.

The results are consistent with the intended component behavior. **RoCo** raises EVOKE Inj. Acc. from 11.1 to 17.8 while keeping Ret. Avg. nearly unchanged relative to the base model. ACE-only training provides a stronger injection signal, reaching 19.2 EVOKE Inj. Acc., but lowers Ret. Avg. to 46.7, reflecting the cost of anchor correction without rollout-side distillation. The full **RoCo-ACE** objective gives the best EVOKE Inj. Acc. of 20.9 while keeping Ret. Avg. at 49.3, close to the base model's 49.5. This supports the same conclusion as the Qwen3-VL experiments: **RoCo** helps limit evaluated retention drift, while **ACE** supplies sparse correction for omitted authoritative anchors.

F Hyperparameter Sensitivity

Table 11 reports sensitivity analysis for the three hyperparameters that directly control the two components of **RoCo-ACE**. The ACE loss weight λ_{ACE} controls how strongly missing or weakly covered authoritative anchors are supervised. The missing-span multiplier β_{miss} controls how much more weight is assigned to anchors absent from the student rollout. The RoCo contrast margin τ controls how much the reference-conditioned teacher likelihood must exceed the base-teacher likelihood before a rollout token receives stronger distillation weight.

The λ_{ACE} rows show the role of anchored correction. Setting $\lambda_{\text{ACE}} = 0$ reduces the method to

Case C: person entity with sports, medical, and family facts

Question
Can you describe the human depicted in the image?

Gold facts
Lida-Maria Manthopoulou, born on **June 14, 2005**, is a **Greek para-athlete** who competed in the **2024 Summer Paralympics** and won **silver in the 100 metres T38**. She represents **Elpides Thessaloniki**, was diagnosed with **multiple sclerosis** before the Paralympics, and is connected to actor **Aias Manthopoulos** and other public family members.

SFT / direct fitting
Captures **Lida-Maria Manthopoulou**, Greek para-athlete identity, and a 2024 Paralympics association, but omits most event, medal, club, health, and family facts.

KORE / constrained update
Adds the **silver medal** and **100 metres T38** event, but only briefly mentions background and omits most medical and family details.

SDFT / online distillation
Keeps natural person-description style, but avoids firm details about the **Paralympic event, medal, diagnosis, club, and family lineage**.

RoCo-ACE / rollout-conditioned anchors
Recovers the person identity, **Greek para-athlete**, **June 14, 2005**, **2024 Summer Paralympics**, **silver in the 100 metres T38**, **Elpides Thessaloniki**, and the **Aias Manthopoulos** family link, while leaving out some extended family details.

Table 7: Baseline-family behavior examples (continued).

Group	Method	Knowledge Injection	Retention Benchmark						
		EVOKE	TreeBench	VStar	MathVision	MMStar	BabyVision	MM-Safety	Ret. Avg.
Base model	Qwen3-VL-30B	21.6	***41.3***	83.2	54.5	72.4	16.4	82.4	***58.4***
Direct fitting	LoRA	23.1	38.1	79.5	50.6	67.4	12.7	76.4	54.1
Direct fitting	SFT	25.1	36.5	77.4	49.6	62.7	11.7	71.7	51.6
Online distillation	SDFT (w/ ref.)	26.4	39.4	83.1	55.0	72.1	15.3	**83.0**	58.0
Constrained updates	KORE	***29.7***	39.6	79.5	49.3	65.1	13.8	70.5	53.0
Constrained updates	KeepLoRA	20.8	41.2	83.0	54.1	72.0	16.4	81.9	58.1
Constrained updates	MoELoRA	22.9	39.4	80.5	51.2	68.9	13.5	78.5	55.3
Constrained updates	SEFE	26.1	38.6	77.5	50.1	64.2	12.8	68.4	51.9
Ours	**RoCo**	28.9	40.8	***85.4***	**56.8**	**72.9**	***16.7***	***82.9***	**59.2**
Ours	**RoCo-ACE**	**31.2**	**42.0**	**85.6**	***55.4***	***72.8***	**16.8**	82.7	**59.2**

Table 8: Supplementary Qwen3-VL-30B comparison. EVOKE is evaluated with the same LLM-judge protocol as the main table. Retention columns report held-out general multimodal and safety accuracy, and Ret. Avg. averages TreeBench, VStar, MathVision, MMStar, BabyVision, and MM-Safety. Higher values are better; bold and bold-italic mark the best and second-best scores in each column.

RoCo-only, which keeps Ret. Avg. high but lowers EVOKE accuracy to 23.7. The default $\lambda_{\text{ACE}} = 0.2$ improves EVOKE to 27.6 with almost no retention loss, while the more aggressive setting $\lambda_{\text{ACE}} = 1.0$ slightly increases EVOKE but drops Ret. Avg. to 52.5. This suggests that ACE should remain a targeted correction term rather than dominate the online-distillation objective.

The β_{miss} and τ rows show the same balance. Moderate missing-span emphasis under-corrects omitted facts, while overly large missing-span emphasis pressures retention. For the RoCo margin, $\tau = 0.2$ gives the best injection-retention balance among the tested settings: a smaller margin is too permissive and a larger margin is too conservative. Overall, the default configuration is selected because it achieves strong injected-knowledge accuracy without turning training into full-reference imitation.

G Additional Discussion on Retention Scope and Reference Quality

RoCo-ACE does not explicitly guarantee preservation of every general capability. Its preservation mechanism is indirect: it reduces irrelevant updates during injected-data training. This distinction mat-

Variant	EVOKE	Retention Benchmarks						
		TreeBench	VStar	MathVision	MMStar	BabyVision	MM-Safety	Ret. Avg.
Base model	21.6	41.3	83.2	54.5	72.4	16.4	82.4	58.4
SDFT (w/ ref.)	26.4	39.4	83.1	55.0	72.1	15.3	**83.0**	58.0
SDFT + SFT	27.8	39.0	83.2	53.2	71.0	13.7	81.9	57.0
SDFT + ACE	28.9	***41.6***	84.5	***55.9***	72.1	16.5	82.5	***58.9***
RoCo	28.9	40.8	***85.4***	**56.8**	**72.9**	***16.7***	***82.9***	**59.2**
RoCo + SFT	29.5	40.1	82.2	55.1	71.9	15.0	81.2	57.6
ACE only	***30.7***	38.9	70.7	52.8	70.6	13.9	80.8	54.6
RoCo-ACE	**31.2**	**42.0**	**85.6**	55.4	***72.8***	**16.8**	82.7	**59.2**

Table 9: Supplementary component ablations on Qwen3-VL-30B with full retention metrics. EVOKE measures injected-knowledge accuracy. Retention columns report held-out general multimodal and safety accuracy, and Ret. Avg. averages TreeBench, VStar, MathVision, MMStar, BabyVision, and MM-Safety. Higher values are better; bold and bold-italic mark the best and second-best scores in each column.

Variant	EVOKE	Retention Benchmarks						
		TreeBench	VStar	MathVision	MMStar	BabyVision	MM-Safety	Ret. Avg.
InternVL3.5-8B	11.1	**37.4**	71.2	**34.3**	***67.1***	11.1	**75.8**	**49.5**
RoCo	17.8	36.3	**71.9**	***34.0***	67.0	**11.8**	74.5	49.3
ACE only	***19.2***	34.4	69.2	31.9	66.8	9.2	74.1	46.7
RoCo-ACE	**20.9**	***36.7***	***71.6***	34.0	**67.5**	***11.6***	***74.7***	***49.3***

Table 10: InternVL3.5-8B component comparison on EVOKE. EVOKE measures injected-knowledge accuracy. Retention columns report held-out multimodal and safety benchmark accuracy, and Ret. Avg. averages TreeBench, VStar, MathVision, MMStar, BabyVision, and MM-Safety. Higher values are better; bold and bold-italic mark the best and second-best scores in each column.

Group	Setting	EVOKE	Ret. Avg.	Expected role
λ_{ACE}	0.0	23.7	**56.6**	RoCo only; no anchored missing-fact correction
	0.2	***27.6***	***56.5***	Default setting
	1.0	**28.1**	52.5	Aggressive anchor fitting
β_{miss}	3.0	25.1	**56.8**	Moderate missing-span emphasis
	5.0	***27.6***	***56.5***	Default setting
	7.0	**28.2**	54.1	Aggressive missing-anchor emphasis
τ	0.1	23.9	***56.3***	Mild contrast margin
	0.2	**27.6**	**56.5**	Default setting
	0.4	***25.2***	55.9	Conservative contrast margin

Table 11: Hyperparameter sensitivity grid for **RoCo-ACE** on EVOKE. EVOKE measures injected-knowledge accuracy, and Ret. Avg. averages TreeBench, VStar, MathVision, MMStar, BabyVision, and MM-Safety. The RoCo-only and default rows reuse the corresponding results from Table 4; bold and bold-italic mark the best and second-best scores within each hyperparameter group.

ters because many regressions are not caused by learning a new fact itself, but by repeatedly training on long, narrow, stylistically similar answers. The method reduces that source of regression by making the update selective with respect to factual support from the reference.

The method also depends on reference quality. If the authoritative answer is noisy, overly broad, or stylistically biased, the RoCo contrast can amplify the wrong signal and ACE can reinforce incorrect spans. Span-level diagnostics are therefore important: the system should log which generated spans receive high contrast weights and which reference spans receive anchor correction. Such diagnostics make failure modes visible and help distinguish genuine factual injection from style imitation. The

open-ended injection metric also relies on an LLM judge, so judge-agreement or human-audit results should be reported when the benchmark setting allows them.

H Benchmark Details

Knowledge injection benchmarks. EVOKE (Jiang et al., 2025a) contains image-grounded entity and news updates with authoritative reference answers. This setting is deliberately different from rich-demonstration continual-learning settings studied by SDFT-style methods: the reference information is often short, sparse, and localized to a small number of factual spans, making uniform online distillation an incomplete injection signal. To test whether the method generalizes beyond time-sensitive news and entity updates, we additionally include two MLLM-CL benchmarks (Zhao et al., 2025): VP from the ability continual-learning suite and Sci from the domain continual-learning suite. MLLM-CL organizes ability data as non-IID tasks that introduce new visual capabilities, and domain data as IID tasks drawn from mainstream visual domains. In our setting, VP evaluates capability-oriented visual perception injection, while Sci evaluates science-domain knowledge injection.

We follow the original data splits released with each benchmark. EVOKE uses its provided entity/news knowledge-injection split with authoritative reference answers, and VP and Sci use the corresponding MLLM-CL training and test splits without additional resampling. All compared methods are trained and evaluated on the same split for each benchmark; no retention benchmark examples are mixed into the injection training set.

Retention benchmarks. TreeBench (Wang et al., 2026) evaluates traceable visual grounded reasoning and evidence localization. VStar, following V*Bench (Wu and Xie, 2024), evaluates fine-grained visual search in high-resolution and cluttered scenes. MathVision (Wang et al., 2024) measures multimodal mathematical reasoning over visual problems. MMStar (Chen et al., 2024) evaluates vision-indispensable multimodal understanding with reduced language-only shortcuts and data leakage. BabyVision (Chen et al., 2026) probes core visual abilities that should be solved with minimal reliance on language priors. MM-SafetyBench (Liu et al., 2024b) evaluates multimodal safety behavior under image-conditioned unsafe instructions. Together, these benchmarks test whether knowledge injection preserves visual grounding, fine-grained perception, mathematical reasoning, general multimodal understanding, early visual primitives, and safety behavior.

I Evaluation Protocol

We use accuracy as the primary metric. For open-ended injected-knowledge examples, we use Qwen3-30B-A3B-Instruct-2507 as the LLM judge (Qwen Team, 2025). The judge compares each model response against the authoritative answer and returns a binary correctness label. It is instructed to focus on factual consistency with the reference, including entity identity, event identity, date, number, location, relation, and other task-specific facts, rather than surface wording. We report the resulting LLM-judge accuracy over the evaluation set.

For general capability benchmarks, we follow the benchmark-specific answer format and compute accuracy after answer normalization. For multiple-choice tasks, the predicted option is extracted from the model response. When the output is ambiguous or free-form, the same LLM-judge protocol is used to decide whether it matches the ground-truth answer. We report injected-knowledge accuracy together with general benchmark accuracy, since a method is only useful if it improves injected facts without causing substantial capability degradation.

J Model and Baseline Details

The main experiments use **Qwen3-VL-8B** as the base model because it exposes realistic retention failures while keeping the full baseline suite computationally tractable. We include **Qwen3-VL-30B** experiments in the supplementary material to check whether the observed trends persist at a larger scale. Unless otherwise stated, the vision tower is frozen, text and projector-side trainable components follow the corresponding baseline configuration, and **RoCo** uses one on-policy generation per prompt with a synchronized reference teacher.

Direct fitting baselines. LoRA (Hu et al., 2022) is the parameter-efficient fitting baseline, while SFT (Ouyang et al., 2022) directly fits authoritative answers and follows the standard supervised instruction-tuning paradigm. These baselines test how much injected knowledge can be acquired through direct answer fitting and how much retention is lost under ordinary supervised updates.

Reference-conditioned self-distillation baseline. SDFT (w/ authoritative reference) (Shenfeld et al., 2026) uses the same on-policy student rollout, synchronized teacher, and reference-conditioned teacher prompt as **RoCo**, but applies a uniform distillation loss over all valid rollout tokens. This is the main online-distillation reference point because every injected example provides an authoritative answer. A no-reference vanilla KD baseline is not included in the main comparison: without the authoritative reference, the teacher cannot observe the injected facts and the experiment mainly measures policy preservation rather than knowledge acquisition.

Constrained-update baselines. KORE (Jiang et al., 2025b) represents knowledge-injection-specific constrained updating and is expected to be strong on target injection, but prior results also show that strong injection does not uniformly preserve every general capability dimension. Following the KORE protocol, we sample auxiliary general examples from LLaVA-OneVision (Li et al., 2024) to estimate the general representation space used by its retention constraint; these auxiliary examples are not used by our method or by other baselines. KeepLoRA (Luo et al., 2026), MoELoRA (Luo et al., 2024), and SEFE (Chen et al., 2025) represent parameter-space constrained adaptation methods that reduce forgetting through residual-gradient updates, expertized LoRA modules, or forgetting-aware constraints. Among them, KeepLoRA (Luo et al., 2026) is a retention-oriented baseline that can preserve general behavior more conservatively but may provide weaker injection when the new knowledge is sparse.

Our variants. The main variants are **RoCo** and **RoCo-ACE**. The former isolates rollout-conditioned contrastive online distillation over student rollouts, while the latter combines it with missing-span anchored supervision. Additional component controls, including SDFT+SFT, SDFT+ACE, **RoCo**+SFT, and ACE only, are reported in the ablation tables.

Implementation form of baselines. For Qwen3-VL-8B, SFT, SDFT, and **RoCo** variants use full language-side updating with the vision tower and multimodal aligner frozen. LoRA, KeepLoRA, MoELoRA, SEFE, and KORE are implemented as adapter-based or constrained-update baselines following their method definitions, using the same rank, target-module, and training-budget settings unless the method requires its own constraint or two-stage procedure. For Qwen3-VL-30B, the non-adapter methods likewise use full language-side updating with frozen vision and aligner modules, while adapter-defined methods remain adapter-based. This keeps the comparison aligned with the intended update mechanism of each baseline rather than converting all methods into a single adapter form.

K Training Configuration and Hyperparameters

Unless otherwise stated, all primary experiments are run for five epochs with learning rate 1×10^{-5}, warmup ratio 0.05, bfloat16 precision, FlashAttention, DeepSpeed ZeRO-3, per-device batch size 1, and gradient accumulation 8. We set the maximum text sequence length to 1536 and the maximum generated completion length to 1024. For image inputs, we set the minimum and maximum pixel budgets to $65{,}536$ and $3{,}379{,}200$, corresponding to an image-token range of 64 to 3300. The vision tower and multimodal aligner are frozen. Full-model updating is used for SFT, SDFT, and **RoCo** variants, while adapter-based baselines use LoRA-style updates with rank 16, alpha 32, dropout 0.05, and all linear layers as target modules.

For all online distillation variants, the student samples one on-policy completion per prompt (`num_generations=1`) with temperature 1.0, top-$p = 1.0$, and one return sequence. The teacher uses the same base model architecture and the same Qwen3-VL chat template as the student. The student prompt contains only the target user question and the optional image. The reference-conditioned teacher prompt replaces the user message with a reference-augmented instruction of the form: target question, authoritative reference answer, and an instruction to generate a new answer that is factually consistent with the reference. The base teacher prompt uses the original student input without the authoritative reference. The teacher never generates the training completion in our main setting; it only scores the student's rollout under these two prompt contexts. The teacher is synchronized with the student by an exponential-moving-average update after every optimization step, so the teacher tracks the current policy while remaining a stable scoring model.

Teacher prompt template. For English reference-conditioned teacher scoring, we use the following template, where `{question}` is the original user query and `{gold_response}` is the authoritative reference answer:

```
Target Question:
{question}

Response to the target question:
{gold_response}

The assistant answer above is the
authoritative reference for this training
example.
Generate a new answer that stays fully
consistent with the same entity/event
identity and factual claims.
Use the image only to add visible grounding
details and natural wording.
Do not contradict, negate, or correct the
reference answer.
Do not claim that the image is unrelated,
blank
```

For **RoCo**, we set the distillation loss weight $\lambda_{\mathrm{RoCo}} = 1.0$, the floor token weight $w_0 = 1.0$, the reference-contrast scale $\beta_\Delta = 1.0$, and the contrast margin $\tau = 0.2$. Dense token weighting, span aggregation, and weight normalization are enabled. We do not use top-m truncation or any auxiliary gating for rollout-conditioned contrastive tokens. The synchronized teacher is updated after every optimization step with exponential moving average decay 0.999, corresponding to an effective update coefficient of 0.001 from the current student.

For Anchored Cross-Entropy, we set the objective weight $\lambda_{\mathrm{ACE}} = 0.2$ unless otherwise stated. The base weight inside extracted anchor spans is $u_0 = 1.0$, and the missing-span weight is $\beta_{\mathrm{miss}} = 5.0$; reference tokens outside extracted factual anchors are not included in the ACE loss. Missing-span weighting and span aggregation are enabled for ACE variants. For SDFT (w/ ref.), we keep the same student rollout and reference-conditioned teacher-scoring pipeline as **RoCo**, but replace the rollout-conditioned contrastive weights with uniform weights and set $\lambda_{\mathrm{ACE}} = 0$. For hybrid SDFT+SFT and **RoCo**+SFT variants, the supervised loss weight is set to 1.0 over the full reference answer.

Runtime cost. Table 12 reports approximate wall-clock training and inference cost on EVOKE with Qwen3-VL-8B, three training epochs, and 16 PPU accelerators. **RoCo-ACE** is slower than direct fitting and adapter-only baselines because it performs online rollout scoring with reference-conditioned and base-teacher contexts, but its cost remains close to SDFT since both use the same online-distillation backbone. At inference time, all methods use model weights of comparable scale, so their per-query latency remains approximately the same under the same decoding setup.

Method	Training time (h)	Inference latency
LoRA	∼5.0	∼1.0 s/query
SFT	∼10.0	∼1.0 s/query
SDFT (w/ ref.)	∼20.0	∼1.0 s/query
KORE	∼9.0	∼1.0 s/query
KeepLoRA	∼6.0	∼1.0 s/query
MoELoRA	∼5.0	∼1.0 s/query
SEFE	∼9.0	∼1.0 s/query
RoCo-ACE	∼22.0	∼1.0 s/query

Table 12: Estimated runtime comparison on the EVOKE setting with Qwen3-VL-8B, three training epochs, and 16 PPU accelerators. Training time is reported as wall-clock hours from representative runs. Inference latency is similar across methods because the deployed model sizes are comparable; we observe roughly one second per query under the same inference setup.