CAN MLLMS RETAIN MORE WITHOUT UNLEARNING LESS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language model (MLLM) unlearning aims to make MLLMs forget chosen facts, such as a biography recalled from a portrait or a name, while retaining the rest of their knowledge. Existing benchmarks take a static, single-point view, scoring each method once at the end of its training. Even a fully converged run marks one position on the trade-off between forgetting and retention and cannot characterize the method as a whole. We instead take a dynamic view and introduce iso-forget evaluation, which characterizes each method by its forget–retain frontier, the best retention it attains at every forgetting level. Unlike a single score, the frontier compares methods at the same forgetting level. It separates a genuine improvement, which shifts the frontier, from a gain that merely trades forgetting for retention along it. Specifically, we start from a few anchor checkpoints at different unlearning strengths and reach any target forgetting level exactly, without retraining, by randomizing between two adjacent anchors with a closed-form probability. Across six MLLMs from three families, forgetting costs almost no retention up to a knee, whereas beyond it retention drops sharply. Tuning the training procedure or switching the objective mostly moves a model along the frontier, so the algorithm chiefly picks a point on it. The frontier itself is set by how the model represents and retrieves the forgotten facts, since forgetting spills over onto look-alike people and blocks mainly the cue it was trained on. MLLMs can therefore retain more without unlearning less by reaching a fact through every cue that recalls it, rather than by pushing harder on one.

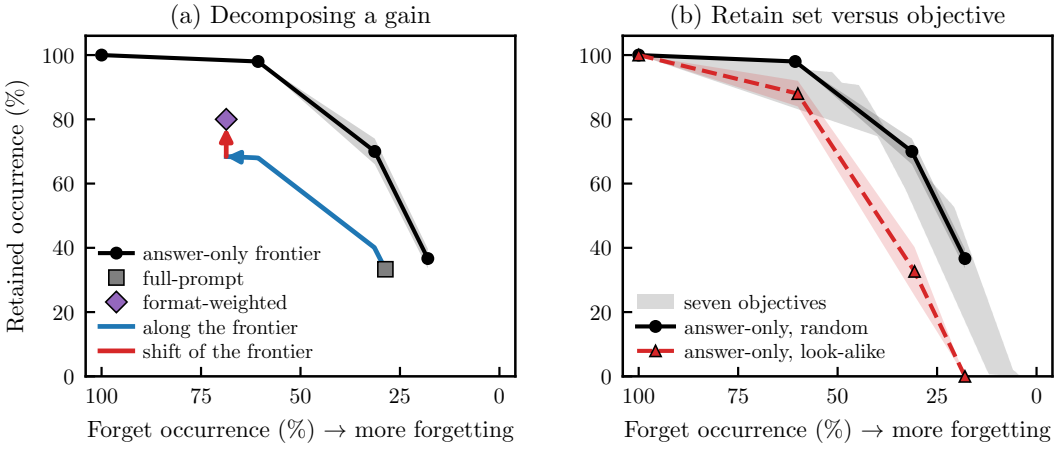

Figure 1: **Retention against forgetting on InternVL2.5-8B.** Most of the retention that format-weighted training gains over full-prompt training comes from forgetting less, and past the knee, a look-alike retain set costs more retention than the choice among seven objectives.

1 INTRODUCTION

Multimodal large language models (MLLMs) can recall a person's biography from a portrait as readily as from a name, which makes the removal of such facts after training a practical privacy need (Liu

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

et al., 2025a; Wang et al., 2025). MLLM unlearning aims to make a model forget chosen facts while retaining the rest of its knowledge. Most unlearning benchmarks, including TOFU (Maini et al., 2024), MUSE (Shi et al., 2025), MLLMU-Bench (Liu et al., 2025a), CLEAR (Dontsov et al., 2025), UMU-Bench (Wang et al., 2025) and S-MLLMUn (Zeng et al., 2026), measure it on two sets of queries. The *forget* set asks for the facts to be removed, and the *retain* set asks for facts that unlearning is trained to keep.

However, these benchmarks take a static, single-point view, training each method once with its own budget and settings and scoring only the final checkpoint. Yet forgetting and retention trade off, and a method retains more when it forgets less, for example with a smaller learning rate or a lighter forget loss. Even a fully converged run thus marks one position on this trade-off and cannot characterize the method as a whole. Such scores settle a comparison when one method both forgets more and retains more. When the method that retains more also forgets less, they cannot tell a better method from one that simply forgets less.

Research Question

Unlearning trades forgetting against retention, so a method that forgets less should retain more. How can an unlearned MLLM retain as much as possible at a given level of forgetting?

We instead take a dynamic view and introduce *iso-forget evaluation*, which characterizes each method by its *forget–retain frontier*, the best retention it attains at every forgetting level. Unlike a single score, the frontier compares methods at the same forgetting level, much as ROC analysis compares classifiers at a fixed false-positive rate (Provost & Fawcett, 2001). It also separates two kinds of gain that a results table conflates. A method can gain retention by moving along its frontier, which only trades forgetting for retention, or by shifting the frontier, which is a genuine improvement at equal forgetting.

Specifically, we start from a few anchor checkpoints at different unlearning strengths, such as values of a forget-loss weight. Randomizing between two adjacent anchors with a closed-form probability reaches any target forgetting level exactly and without retraining, so every point on the frontier is attained by a deployable model rather than estimated. Any reported gain then splits additively into a shift and a movement along the frontier. The procedure needs only a strength parameter and the forget and retain queries of an existing benchmark, so it applies to existing methods and benchmarks alike.

We apply iso-forget evaluation to six MLLMs from three families and vary one factor at a time, namely the knowledge to separate, the cue used to forget and query, the training procedure and the objective. Each model learns a fixed set of identities before unlearning, so we know exactly what it holds. We first find that, on every model, forgetting costs almost no retention up to a knee, whereas beyond it retention drops sharply, and the loss falls first on knowledge that unlearning never sees. We further find that tuning the training procedure or switching the objective mostly moves a model along the frontier, so the algorithm chiefly picks a point on it. Seven established unlearning methods, for instance, lie far closer together at the same forgetting level than in reported retention. Finally, we find that the frontier itself is set by how the model represents and retrieves the forgotten facts. Forgetting spills over onto look-alike people, whose portraits the vision encoder places close together, and blocks mainly the cue it was trained on, so names still recall much of what portraits have forgotten. MLLMs can therefore retain more without unlearning less by reaching a fact through every cue that recalls it, rather than by pushing harder on one.

We make two contributions.

- We introduce iso-forget evaluation, which characterizes each unlearning method by its forget–retain frontier. From a few anchor checkpoints it reaches any forgetting level exactly and without retraining, and it splits every reported gain into a shift and a movement along the frontier, using only the queries of existing benchmarks.
- We identify what sets the frontier itself, namely how the model represents and retrieves the forgotten facts. Forgetting spills over onto look-alike people and blocks mainly the cue it was trained on, so MLLMs retain more without unlearning less by reaching a fact through every cue.

Table 1: **Evaluation protocols of unlearning benchmarks.** None compares methods at a common forgetting level. ✓: yes; ○: partly; ✗: no.

Protocol	Modality	Known parent facts	Strength sweep	Matched forgetting	Retain relatedness controlled
TOFU (Maini et al., 2024)	text	✓	○	✗	✗
MUSE (Shi et al., 2025)	text	✓	✗	✗	○
MLLMU-Bench (Liu et al., 2025a)	image+text	✓	✗	✗	✗
CLEAR (Dontsov et al., 2025)	image+text	✓	✗	✗	✗
UMU-Bench (Wang et al., 2025)	image+text	✓	✗	✗	✗
S-MLLMUn (Zeng et al., 2026)	image+text	○	✗	✗	✗
Ours	image+text	✓	✓	✓	✓

2 Related Work

Unlearning methods. Standard unlearning objectives include gradient ascent, gradient difference, KL regularization, refusal ("I don't know") training (Maini et al., 2024) and NPO (Zhang et al., 2024). Multimodal methods add modality-aware components such as neuron pruning (Liu et al., 2025b), retained-data anchors (Zeng et al., 2026) and cross-modal gradient balancing (Yang et al., 2026); Section 4.5 measures the frontiers of seven of these methods. Implementations also differ in which tokens receive loss: loss over instruction tokens can improve instruction tuning (Shi et al., 2024), and in LLM unlearning, masking the question from the forget loss makes retain-set performance less stable (Yuan et al., 2025). We treat such choices as procedural factors and ask whether they move the frontier or only a method's position on it (Section 4.5).

Unlearning benchmarks and evaluation. TOFU and MUSE measure forgetting and retention on text (Maini et al., 2024; Shi et al., 2025), and MLLMU-Bench, CLEAR, UMU-Bench and S-MLLMUn extend these measurements to image and text (Liu et al., 2025a; Dontsov et al., 2025; Wang et al., 2025; Zeng et al., 2026). Table 1 compares their evaluation protocols; none compares methods at a common forgetting level. A second line of work examines how such evaluations can mislead. ReMem shows that results are hard to interpret unless the model has learned the targets before unlearning (Kwon et al., 2026), which we verify for every parent (Section 4.1). TOFU plots forget–retain curves over training, RULE compares the areas under such curves (Maini et al., 2024; Zhang et al., 2025), and OpenUnlearning meta-evaluates metrics and notes that model selection can change method rankings (Dorna et al., 2025). Differing amounts of training on the test task confound comparisons between language models (Dominguez-Olmedo et al., 2025), and a model can count as having forgotten while its knowledge remains extractable (Lynch et al., 2024) or while it hallucinates (Gu et al., 2026). Our work belongs to this second line: it contributes an evaluation tool for existing benchmarks rather than a new benchmark. Iso-forget evaluation reuses the forget and retain queries that a benchmark already defines and changes only how methods are compared, so it adds the matched comparison of Table 1 to any of these protocols; we run it on UMU-Bench.

3 The Forget–Retain Frontier

Our research question concerns a curve rather than a point: how much retention an unlearning procedure can keep at each level of forgetting, and what changes that curve. This section defines the tools we use to answer it. We first define the frontier of a procedure and the summaries we read from it. We then show that every point on a frontier can be attained exactly, so any two runs can be compared at the same forgetting level. Finally, we split any difference in a results table into movement along the frontier and a shift of it. The tools assume only a strength parameter and scores that average over a fixed set of queries, so they apply to any unlearning method and benchmark.

The frontier of a procedure. An unlearning procedure M, such as refusal training or gradient ascent, has a strength parameter λ that trades forgetting for retention, for example a loss weight, a learning rate or a number of updates. We score every run on two fixed query sets. The forget score f is the percentage of forget queries whose answer still appears in the response, so lower f means more forgetting; the retain score r is the same percentage on retained queries. One run gives one

point (f, r). Sweeping λ gives checkpoints $\theta_{\lambda_1}, \ldots, \theta_{\lambda_k}$ whose points (f_i, r_i) trace how retention falls as forgetting rises.

Definition 1 (Frontier, RMF and iso-forget residual)**.** *Order the sweep points by forget score, $f_1 < f_2 < \cdots < f_k$. The* frontier *of M is their piecewise-linear interpolation*

$$R^\star_M(f) = r_i + \frac{f - f_i}{f_{i+1} - f_i}\,(r_{i+1} - r_i), \qquad f_i \le f \le f_{i+1}. \tag{1}$$

For a target forgetting level $\tau \in [f_1, f_k]$, the retention at matched forgetting *is* $\mathrm{RMF}_M(\tau) = R^\star_M(\tau)$. *For a checkpoint θ with scores (f_θ, r_θ), the* iso-forget residual *against M is*

$$\delta_M(\theta) = r_\theta - R^\star_M(f_\theta). \tag{2}$$

We interpolate the sweep rather than fit a curve to it, for two reasons. Every point of the interpolation is attained by an actual system (Proposition 1), so the frontier states what the procedure achieves rather than what a model of the procedure predicts. The interpolation also assumes no functional form. When a sweep is concave, as all of ours are, it is the upper convex hull of the sweep points, so no mixture of the swept checkpoints retains more at any forgetting level.

The three quantities serve different uses. $\mathrm{RMF}_M(\tau)$ is the retention a procedure keeps at a stated forgetting level, the analogue of recall at a fixed false-positive rate; it compares procedures at equal forgetting. The residual $\delta_M(\theta)$ judges a single checkpoint, such as a method's published default, against a reference procedure: a negative residual means that M, run to forget exactly as much as θ, retains more. To compare frontiers across models and knowledge, we summarize each frontier by three numbers (Table 2). Its *free budget* is the forgetting reachable before retention falls 5 pp below its starting value, which locates the knee of the frontier. RMF@50 is the retention at 50% forget occurrence, a level that every sweep in our study brackets. Its *cost past the knee* is the number of retained points lost per forgotten point in the strongest sweep segment.

Comparing at equal forgetting. A swept checkpoint rarely forgets exactly as much as the run it should be compared with, and retraining until two forgetting levels agree is costly and noisy. Mixing two checkpoints attains any level between them exactly.

Proposition 1 (Exact matching by mixing)**.** *Let checkpoints θ_a and θ_b have scores (f_a, r_a) and (f_b, r_b) with $f_a < f_b$, and let the policy π_p answer each query with θ_a with probability p and with θ_b otherwise. Because both scores average per-query indicators over fixed query sets, π_p has expected scores*

$$f(\pi_p) = p\,f_a + (1-p)\,f_b, \qquad r(\pi_p) = p\,r_a + (1-p)\,r_b. \tag{3}$$

For adjacent sweep points and any $\tau \in [f_a, f_b]$, the probability $p_\tau = (f_b - \tau)/(f_b - f_a)$ therefore gives $f(\pi_{p_\tau}) = \tau$ and $r(\pi_{p_\tau}) = R^\star_M(\tau)$.

Mixing makes RMF the score of a deployable system rather than an estimate, just as operating points on a ROC convex hull compare classifiers at equal error rates (Provost & Fawcett, 2001). We call comparison at matched forgetting *iso-forget evaluation* (IFE; Algorithm 1). IFE computes RMF and residuals with identity-bootstrap intervals and refits the mixing weights in every draw, so the intervals include the uncertainty of matching (Appendix B). Its one requirement is that the sweep bracket the target level. Extrapolation is not a matched comparison, so we report a run outside the sweep as unmatched.

Along or beyond the frontier. A results table compares two runs, with checkpoints θ_A and θ_B and scores (f_A, r_A) and (f_B, r_B), at whatever forgetting levels they reach. Let $R^\star$ be a reference frontier on the same model and knowledge. The difference in retention splits exactly into

$$r_B - r_A = \underbrace{R^\star(f_B) - R^\star(f_A)}_{\text{along the frontier}} + \underbrace{\delta(\theta_B) - \delta(\theta_A)}_{\text{shift of the frontier}}, \tag{4}$$

where δ is the residual of Equation (2) with respect to $R^\star$. The first term is what B gains by forgetting less: the retention the reference procedure itself would gain between the two forgetting levels. The second term is what B gains at equal forgetting. The split needs only the reference frontier and the scores of the two runs, and its two terms always add up to the difference in the table.

Algorithm 1 Iso-Forget Evaluation (IFE)

Require: methods $\mathcal{M}$; strength grid Λ_M per method; seeds S; target level τ; scorer s; bootstrap draws B

Train $\theta_{M,\lambda,\text{seed}}$ for all $M \in \mathcal{M}$, $\lambda \in \Lambda_M$, seed $\in S$; score forget and retain queries with s
for $b = 0, \ldots, B$ **do** ▷ b=0: observed identities; $b \geq 1$: identity bootstrap
  for $M \in \mathcal{M}$ **do**
    $(f_i, r_i)_{i \leq k} \leftarrow$ sweep means of M on draw b
    if $\tau \notin [\min_i f_i, \max_i f_i]$ **then** mark M unmatched on draw b and skip
    pick adjacent points with $f_a \leq \tau \leq f_b$; $p \leftarrow (f_b - \tau)/(f_b - f_a)$
    $\mathrm{RMF}_M^{(b)}(\tau) \leftarrow p\, r_a + (1-p)\, r_b$
return $\mathrm{RMF}_M^{(0)}(\tau)$, percentile intervals over $b \geq 1$, and method ranks

We use the decomposition in two ways. For two runs on the same model and knowledge, such as runs of two objectives, $R^\star$ is the frontier of a reference procedure, as in Figure 1a. For two conditions that each have their own frontier, such as two models or two retain sets, we take $R^\star$ to be the frontier of A. If B lies on its own frontier, the shift term is then the gap between the two frontiers at B's forgetting level. A difference that lies mostly along the frontier is a change of strength, not a change in what unlearning can achieve; Sections 4.3 to 4.5 apply this test to every factor we vary.

4 Experiments

In this section, we answer the following questions through experiments. **Q1:** How does retention fall as forgetting rises, and which knowledge pays first? **Q2:** Which knowledge makes forgetting costly: retained knowledge related to the forget set, or forget facts learned weakly or strongly? **Q3:** Does forgetting through portraits carry over to names, and which modality should unlearning go through? **Q4:** Can the unlearning procedure or the objective raise retention at a given forgetting level, or do they only move a method along the frontier? We first describe the experimental setup in Section 4.1.

Key Findings

- **When and where does forgetting cost retention?** Beyond a knee, and first on knowledge outside the retain set. Retention stays nearly flat up to the knee and falls steeply after it, while held-out knowledge loses from the weakest setting on (Section 4.2).
- **Which knowledge makes forgetting costly?** Related retain knowledge and weakly learned facts. Retain sets that share attributes or faces with the forget set lower the frontier, faces most of all; weakly learned facts are easy to reach but costly, while strongly learned ones need more strength yet keep more retention (Section 4.3).
- **Does forgetting carry across modalities?** Only in part. Names still retrieve much of what unlearning through portraits removed, so unlearning should go through both modalities; through one alone, it keeps less at matched forgetting (Section 4.4).
- **Can the procedure or the objective shift the frontier?** Mostly not. Both mainly decide where on the frontier a method stops: a format-token weight raises retention mostly by forgetting less, and seven established objectives lie far closer together at matched forgetting than in a results table (Section 4.5).

4.1 Experimental Setup

Dataset. We use identities from UMU-Bench (Wang et al., 2025), each with a portrait and biographical facts, and split them into equal *forget* (F), *retained* (R) and *held-out* (H) cohorts. Every model learns all of them before unlearning, but H never appears during unlearning, so it measures collateral damage on knowledge that unlearning never touches. We probe each identity with birthplace and education questions asked with its portrait, and with its name in Section 4.4.

Models. We study a smaller and a larger model from each of three MLLM families: LLaVA-1.5 (Liu et al., 2024), Qwen2.5-VL (Bai et al., 2025) and InternVL2.5 (Chen et al., 2024). Before unlearning, we fine-tune each model on all identities (controlled acquisition), so that we know what the resulting *parent* holds; every parent reproduces the complete answer to almost every question (Appendix A.2). We trace the frontiers of all six models for Q1 and run the factor studies of Q2–Q4 on InternVL2.5-8B and LLaVA-1.5-7B. Appendix A lists the checkpoints and all training settings.

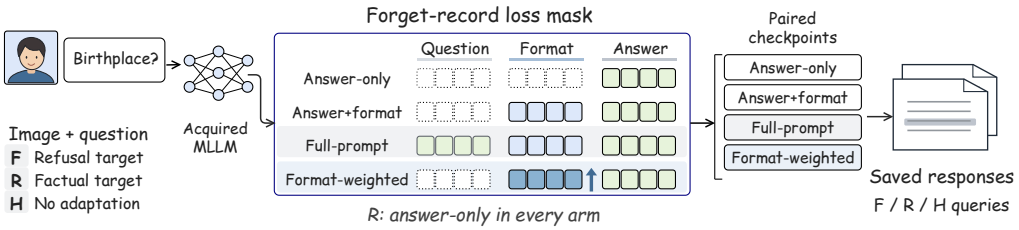

Figure 2: **Loss masks of the four refusal objectives.** The objectives share data, budget and answer loss and differ only in which prompt tokens of a forget record receive loss.

Unlearning methods. Our reference procedure is answer-only refusal training: forget records carry sampled refusal targets, retained records keep their original answers, and only answer tokens receive loss. Its forget-loss weight λ sets its strength, and we sweep λ upward from zero, at which unlearning leaves the forget set untouched (Appendix B). Three variants, used in Section 4.5, add auxiliary loss on prompt tokens (Figure 2). For a forget stream, let C^{ans}, C^{ques} and C^{fmt} be the summed cross-entropies over answer, question and role/format tokens, and let N^{all}, N^{prompt} and N^{fmt} count the eligible, prompt and format tokens (Appendix A.3):

$$\begin{aligned} \mathcal{L}_{\text{answer-only}} &= \frac{C^{\text{ans}}}{N^{\text{all}}}, & \mathcal{L}_{\text{full-prompt}} &= \frac{C^{\text{ans}}+C^{\text{ques}}+C^{\text{fmt}}}{N^{\text{all}}}, \\ \mathcal{L}_{\text{answer+format}} &= \frac{C^{\text{ans}}+C^{\text{fmt}}}{N^{\text{all}}}, & \mathcal{L}_{\text{format-weighted}} &= \frac{C^{\text{ans}}}{N^{\text{all}}} + \frac{N^{\text{prompt}}}{N^{\text{all}}} \cdot \frac{C^{\text{fmt}}}{N^{\text{fmt}}}. \end{aligned} \tag{5}$$

Full-prompt training supervises the whole prompt. Format-weighted training keeps its answer term and its total auxiliary mass $N^{\text{prompt}}/N^{\text{all}}$ but places all of that mass on format tokens, and answer+format training adds format targets at their unweighted share. Section 4.5 adds six established objectives: gradient ascent, gradient difference, KL minimization, NPO, SMFA and CMJU (Appendix E). All runs share one training budget and one decoder LoRA configuration, with identical record order and, where a method uses refusals, identical refusal draws (Appendix A). For Q2–Q4, we change one factor of this setup at a time and split each change with Equation (4) (Appendix C).

Metrics. A response counts as containing the answer when the complete reference answer appears in it at word boundaries. *Occurrence* is the rate of such responses on a cohort. Forget and retained occurrence are the two axes of every frontier; lower forget occurrence and higher retained and held-out occurrence are better. Occurrence suits our parents, which contain the complete reference answer in almost every response (Appendix A.2); on free-form responses it can miss reworded answers and disagree in direction with value-based scorers (Appendix F). Paired per-query effects are averaged within identity. We report the mean and standard deviation over optimization seeds, and an effect is significant in a seed when its identity-bootstrap interval excludes zero (Appendix A.4).

4.2 When and Where Does Forgetting Cost Retention?

Retention holds until a knee. Figure 3a shows that retention on InternVL2.5-8B stays nearly flat while forgetting rises to a knee and then falls steeply. Past the knee, each forgotten point costs far more retention than before it (Figure 3b). LLaVA-1.5-7B has the same shape with an earlier knee, a smaller free budget and a lower frontier (Table 2). The other models share this shape and differ in where the knee lies and how high the frontier sits; within each family, the larger model has the higher frontier.

Knowledge outside the retain set pays first. Figure 3c separates the two cohorts that unlearning should preserve. Retained identities, which unlearning rehearses with their original answers, hold until the knee. Held-out identities, learned during acquisition but absent from unlearning, lose far more before the knee on both models, and at matched forgetting they stay well below retained identities (Table 2). Both cohorts were learned the same way and only the retained one is rehearsed, so the gap points to rehearsal: it protects the retain set but not the rest of what the model knows.

4.3 Which Knowledge Makes Forgetting Costly?

Knowledge that resembles the forget set is costly to keep. Related retain sets lower the whole frontier, and relatedness in appearance costs more than relatedness in text (Figure 4a). A retain set

Table 2: **Frontier summaries of answer-only refusal training.** Within each family, the larger model has the higher frontier, and every model keeps less held-out than retained knowledge at matched forgetting.

Model	Free budget (pp)	At 50% forget		Cost past the knee
		Retained (%)	Held-out (%)	
InternVL2.5-2B	26	78.7±2.1	63	2.3
InternVL2.5-8B	42	87.8±2.3	72	2.5
Qwen2.5-VL-3B	29	78.0±1.9	61	1.7
Qwen2.5-VL-7B	35	83.3±1.8	70	1.9
LLaVA-1.5-7B	21	77.4±1.2	53	1.2
LLaVA-1.5-13B	25	84.5±1.4	57	1.2

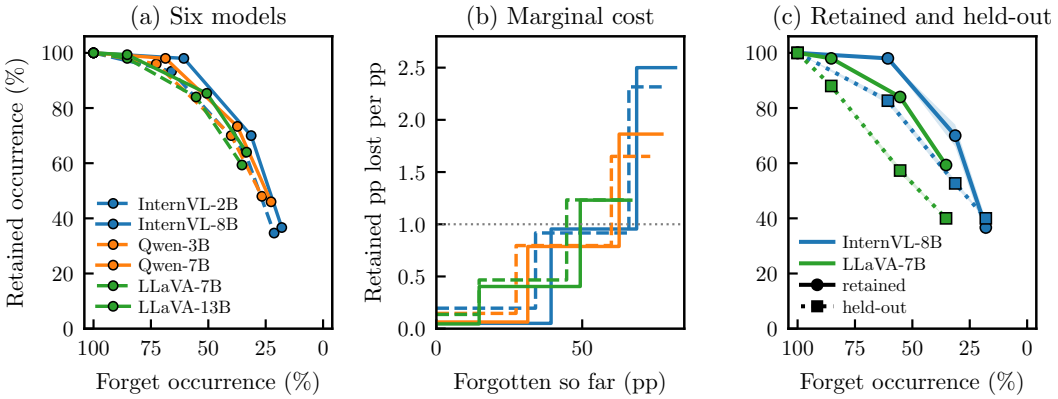

Figure 3: **Answer-only frontiers.** Retention holds until a knee and falls steeply beyond it, while held-out knowledge loses from the weakest setting on.

whose portraits are the nearest neighbours of forget portraits in the vision encoder loses far more retention at matched forgetting than one whose identities share a birthplace or school with the forget set, on both models (Table 9). Unlearning thus struggles most to separate identities that the vision encoder places close together.

Weakly learned facts are cheap to reach but costly to forget. Facts taught more often need a much larger forget weight to reach the same forgetting level, yet keep more retention there; facts taught less often are reached with little strength but lower the frontier, most on LLaVA-1.5-7B (Figure 4b,c; Table 9). "Hard to unlearn" thus has two meanings that the frontier keeps apart: the strength needed to reach a forgetting level, which stronger acquisition raises, as Zhao et al. (2024) report, and the retention lost once there, which it lowers, as the analysis of Chang & Khanna (2026) predicts.

4.4 Does Forgetting Carry Across Modalities?

Names still retrieve what portraits have forgotten. Probed with names instead of portraits, the same checkpoints trace a lower frontier (Figure 5a), so an evaluation with portraits alone overstates what unlearning achieves. The gap is largest when refusals cover one modality: even where refusals on image records alone have removed many portrait-cued answers, names still retrieve most of them (Table 8).

Unlearning should go through both modalities. When forgetting must hold under both cues, counted under the worse one, unlearning through both modalities keeps clearly more retention at matched forgetting than refusals on image records or on text records alone, on both models (Figure 5b,c). Judged by portraits alone, the three routes look alike (ticks in Figure 5c), so a single-cue evaluation cannot tell which route to choose.

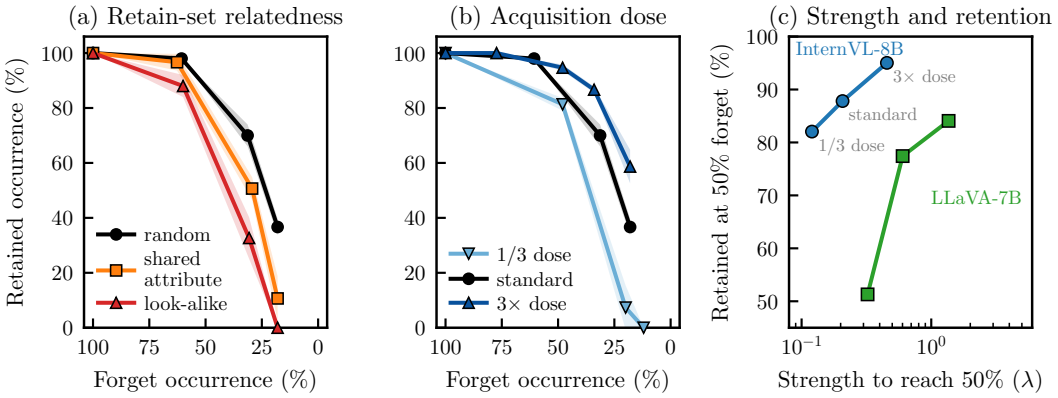

Figure 4: **Frontiers under changed knowledge.** Related retain sets lower the frontier, look-alike portraits most, and facts learned more often need more strength to forget yet keep more retention. Panels (a) and (b) use InternVL2.5-8B.

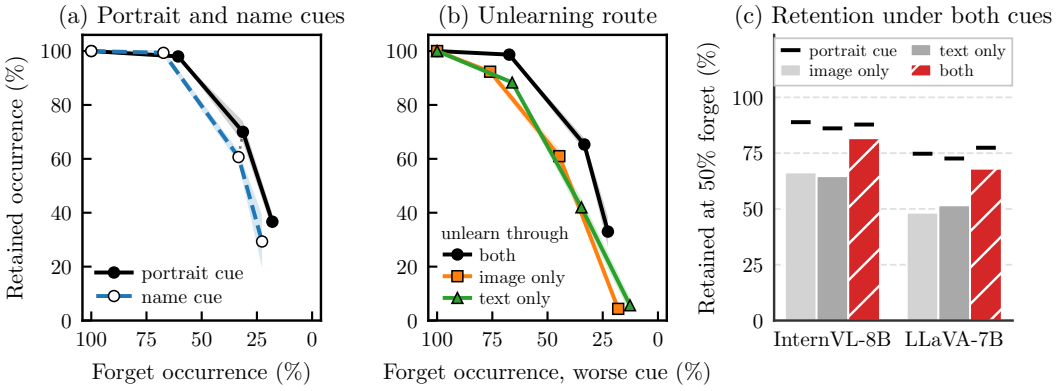

Figure 5: **Forgetting under portrait and name cues.** Names still retrieve what portraits have forgotten, and once forgetting must hold under both cues, unlearning through both modalities keeps the most retention. Panels (a) and (b) use InternVL2.5-8B.

4.5 Can Procedure or Objective Shift the Frontier?

Training knobs move a run along the frontier. Changing the learning rate, the number of updates or the LoRA rank keeps InternVL2.5-8B on its λ frontier while its reported retention changes widely (Figure 6a). The format-token weight acts the same way. Weighting the format tokens, that is, moving from answer+format to format-weighted training, raises retained occurrence sharply on InternVL2.5-8B, but most of the gain is movement along the frontier, and on LLaVA-1.5-7B all of it is (Figure 6b). The same holds against full-prompt training on all six models, where the retained gain of format weighting grows with its forget increase (Appendix D.1). At the forgetting level of format-weighted training, the answer-only procedure retains more on both models (Table 12), and what format weighting gives up is mostly knowledge asked about through the portrait (Table 13).

The optimizer turns the weight into a strength knob. Format weighting places the prompt's auxiliary mass on a few format targets whose loss is large only in the first updates. AdamW divides each step by a running average of squared gradients whose memory, about $1/(1-\beta_2)$ updates, is longer than our runs, so these early gradients shrink every later step, including those that teach the refusal. A toy model reproduces the knob (Appendix D.3), and shortening the memory shrinks the forget shift of format weighting on both real models (Figure 6c).

Objectives differ mainly in where they land on the frontier. At their published defaults, seven established objectives lie far apart in retained occurrence, yet their frontiers lie far closer together at matched forgetting (Figure 7a). Gradient ascent forgets almost everything and keeps almost nothing, while SMFA settles at weaker forgetting and keeps most retained answers.

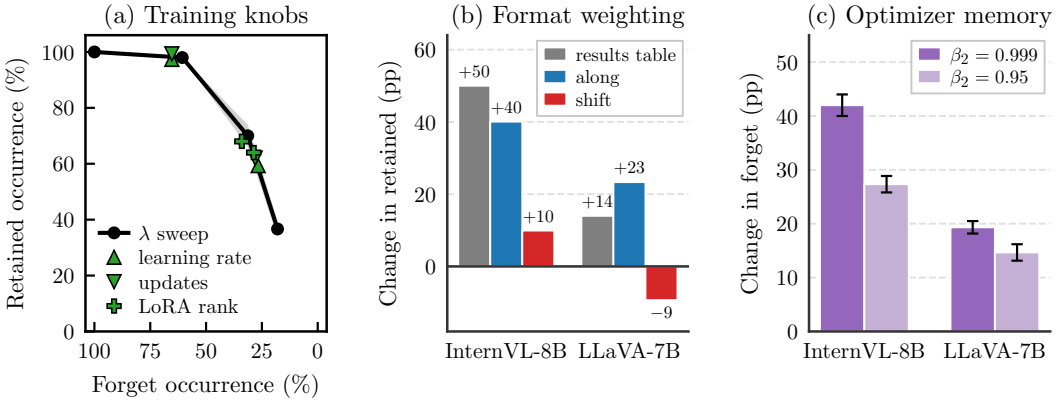

Figure 6: **Training procedure and the frontier.** Learning rate, update count and LoRA rank keep a run on its frontier, and format weighting raises retention mostly by forgetting less, through AdamW's long memory. Panel (a) uses InternVL2.5-8B.

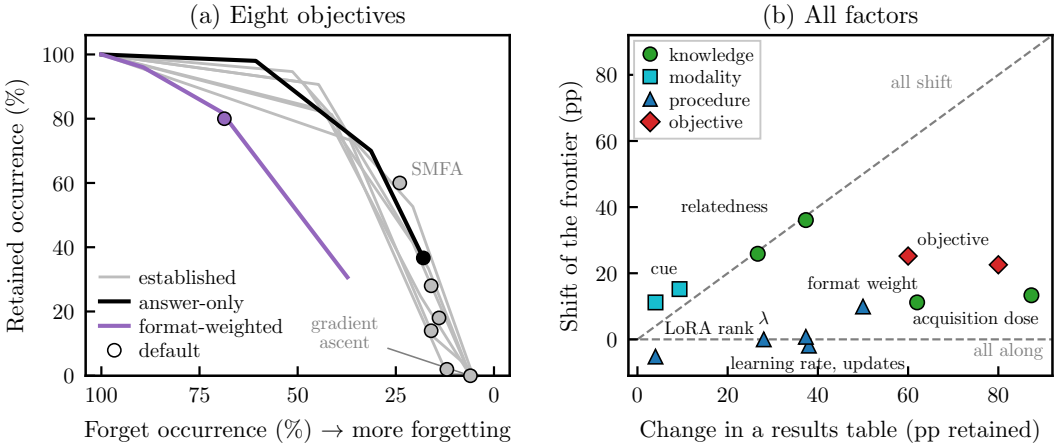

Figure 7: **Objectives and all factors.** Seven established objectives differ far less at matched forgetting than at their defaults, and relatedness and cue shift the frontier, whereas training choices mostly move a run along it. Panel (a) uses InternVL2.5-8B.

Knowledge and modality shift the frontier; training choices mostly do not. Figure 7b splits the change each factor makes in a results table into its two terms. Relatedness lies on the diagonal and the cue above it, so both change what unlearning can achieve. Learning rate, update count and LoRA rank lie on the horizontal line, and the objective and the acquisition dose lie in between, with most of their table change coming from movement along the frontier. The two procedural weights agree: the forget weight λ never leaves the frontier, and the format weight moves mostly along it.

5 Conclusion

We asked how an unlearned MLLM can retain as much as possible at a given level of forgetting. Iso-forget evaluation answers it on the forget–retain frontier, which a few anchor checkpoints trace on the queries of an existing benchmark and which splits every reported gain into a shift and a movement along it. Read this way, the unlearning algorithm mostly picks a point on the frontier, since training knobs, a format-token weight and the choice among established objectives trade forgetting for retention. The frontier itself is set by how the model represents and retrieves the forgotten facts, since forgetting spills over onto look-alike people and blocks mainly the cue it was trained on. An MLLM therefore retains more without unlearning less by reaching a fact through every cue that recalls it. Claims of better unlearning should accordingly be made at a stated forgetting level, as shifts of the frontier, with held-out knowledge and both portrait and name cues reported alongside. Appendix H discusses limitations.

AI USE STATEMENT

We used generative AI tools, including OpenAI Codex and Anthropic Claude, to assist with implementing and running experiments, literature review, and language polishing. Details of model-based scoring are provided in Appendix F. The authors take full responsibility for the final text, experimental results, and claims.

ETHICS STATEMENT

This work studies the evaluation of multimodal unlearning using publicly released benchmark profiles and model checkpoints. More reliable evaluation can help assess whether unlearning removes targeted information while preserving other knowledge. However, reduced answer occurrence on benchmark queries does not establish complete knowledge removal or guarantee privacy. Our results should therefore be interpreted within the stated evaluation protocols rather than as privacy guarantees.

REPRODUCIBILITY STATEMENT

We will open-source our code upon acceptance. The models, data, training settings, and seeds are detailed in Appendix A, and the factor study protocols are provided in Appendix C. The matching and bootstrap procedures are described in Appendix B, with scorer definitions and the annotation protocol in Appendix F.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Wei-Kai Chang and Rajiv Khanna. Why some models resist unlearning: A linear stability perspective. *arXiv preprint arXiv:2602.02986*, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

Ricardo Dominguez-Olmedo, Florian E. Dorner, and Moritz Hardt. Training on the test task confounds evaluation and emergence. In *International Conference on Learning Representations*, 2025.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. CLEAR: Character unlearning in textual and visual modalities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20582–20603. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-acl.1058. URL https://aclanthology.org/2025.findings-acl.1058/.

Vineeth Dorna, Anmol Mekala, Wenlong Zhao, Andrew McCallum, Zico Kolter, Zachary Lipton, and Pratyush Maini. OpenUnlearning: Accelerating LLM unlearning via unified benchmarking of methods and metrics. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-1458. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3e4a38f228427ab819ba7899003a44b1-Paper-Datasets_and_Benchmarks_Track.pdf. Datasets and Benchmarks Track.

Renjie Gu, Jiazhen Du, Yihua Zhang, and Sijia Liu. Unlearners can lie: Evaluating and improving honesty in LLM unlearning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11930–11952. Associa-

tion for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.548. URL https://aclanthology.org/2026.acl-long.548/.

Junehyoung Kwon, MiHyeon Kim, Eunju Lee, JungMin Yun, Byeonggeuk Lim, and YoungBin Kim. Before forgetting, learn to remember: Revisiting foundational learning failures in LVLM unlearning benchmarks. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34063–34079. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1701. URL https://aclanthology.org/2026.findings-acl.1701/.

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

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. Protecting privacy in multimodal large language models with MLLMU-Bench. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4105–4135. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.naacl-long.207. URL https://aclanthology.org/2025.naacl-long.207/.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5913–5933. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.295. URL https://aclanthology.org/2025.acl-long.295/.

Aengus Lynch, Phillip Guo, Aidan Ewart, Stephen Casper, and Dylan Hadfield-Menell. Eight methods to evaluate robust unlearning in LLMs. *arXiv preprint arXiv:2402.16835*, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=B41hNBoWLo.

Foster Provost and Tom Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, pp. 27797–27818, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/4556f5398bd2c61bd7500e306b4e560a-Paper-Conference.pdf.

Zhengyan Shi, Adam X. Yang, Bin Wu, Laurence Aitchison, Emine Yilmaz, and Aldo Lipani. Instruction tuning with loss over instructions. In *Advances in Neural Information Processing Systems*, volume 37, pp. 69176–69205. Curran Associates, Inc., 2024. doi: 10.52202/079017-2210. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7ffb43adf37b3eeaba559098bc084cc6-Paper-Conference.pdf.

Chengye Wang, Yuyuan Li, Xiaohua Feng, Chaochao Chen, Xiaolin Zheng, and Jianwei Yin. UMU-Bench: Closing the modality gap in multimodal unlearning evaluation. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. doi: 10.52202/085713-0574. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/18b0ee1e95f4007541e57fe507b5167e-Paper-Datasets_and_Benchmarks_Track.pdf. Datasets and Benchmarks Track.

Di Yang, Baochen Xiong, and Xiaoshan Yang. CMJU: Cross-modal joint unlearning for balanced forgetting in multimodal large language models. *Visual Intelligence*, 4:26, 2026. doi: 10.1007/s44267-026-00130-4. URL https://link.springer.com/article/10.1007/s44267-026-00130-4.

Xiaojian Yuan, Tianyu Pang, Chao Du, Kejiang Chen, Weiming Zhang, and Min Lin. A closer look at machine unlearning for large language models. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Q1MHvGmhyT.

Zhen Zeng, Leijiang Gu, Zhangling Duan, Feng Li, Zenglin Shi, Cees G. M. Snoek, and Meng Wang. Towards benign memory forgetting for selective multimodal large language model unlearning. In *European Conference on Computer Vision*, 2026. URL https://eccv.ecva.net/virtual/2026/poster/4358.

Chenlong Zhang, Zhuoran Jin, Hongbang Yuan, Jiaheng Wei, Tong Zhou, Kang Liu, Jun Zhao, and Yubo Chen. RULE: Reinforcement UnLEarning achieves forget-retain pareto optimality. In *Advances in Neural Information Processing Systems*, volume 38, pp. 39401–39434. Curran Associates, Inc., 2025. doi: 10.52202/085713-1315. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3811007a23d95e7e239b6262337b98b9-Paper-Conference.pdf.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=MXLBXjQkmb.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. In *Advances in Neural Information Processing Systems*, 2024.

A Experimental Details

A.1 Models, Data and Training

The controlled study starts from the six public checkpoints in Table 3, linked at their exact revisions. Each acquires the same UMU-Bench facts (Wang et al., 2025): F uses all profiles in `forget_5`, R the first 25 `retain_95` profiles in source order, and H the retained IDs 028–052. Every profile supplies the same factual fields in image and text views. Acquisition averages answer-token losses with equal record weight and merges the adapter into the BF16 parent. Unlearning (Table 4) sums four streams (forget/retain $\times$ image/text) with six records each; forget records carry sampled refusal targets and retained records their factual answers. AdamW uses moments $(0.9, 0.999)$, ϵ=10^{-8}, weight decay 0.01, no warmup and no gradient clipping. LoRA adapts decoder attention and MLP projections (63M–250M trainable parameters, depending on the model); the vision encoder, connector, embeddings, normalization and output heads stay frozen. Paired runs share record order and refusal draws; the three optimization seeds vary LoRA initialization and training stochasticity.

Table 3: **Public checkpoints and image input size.** InternVL2.5 reads a larger image than the other two families.

Checkpoint	View
`llava-hf/llava-1.5-7b-hf`	336^2
`llava-hf/llava-1.5-13b-hf`	336^2
`Qwen/Qwen2.5-VL-3B-Instruct`	336^2
`Qwen/Qwen2.5-VL-7B-Instruct`	336^2
`OpenGVLab/InternVL2_5-2B`	448^2
`OpenGVLab/InternVL2_5-8B`	448^2

Table 4: **Training settings for acquisition and unlearning.** Unlearning keeps the LoRA configuration of acquisition and uses a third of its updates at a tenth of its learning rate.

Setting	Acq. / Unlearn.
Identities F/R/H	25/25/25 / 25/25/0
Updates	750 / 250
Exposures	18,000 / 6,000
Learning rate	10^{-4} / 10^{-5}
Schedule	linear decay
LoRA rank/α/dropout	64/32/0.05
Precision	BF16 / FP32 adapter
Seeds	0 / 0, 1, 2
Answer cap	32 tokens

A.2 Parent Knowledge

Table 5 shows that acquisition succeeds: under the image cue, every parent's response contains the complete answer on at least 49 of the 50 questions per cohort. Exact matches, which require the whole response to equal the reference, range from 0 to 50 because parents phrase or extend the answer differently (Qwen2.5-VL-7B usually continues to the length cap), which is why we measure occurrence rather than exact matches.

Table 5: **Parent answers before unlearning (of 50 queries).** Every parent contains the complete answer almost always, while exact matches depend on how each model phrases its answers.

Parent	Occurrence			Exact match		
	F	R	H	F	R	H
InternVL2.5-2B	49	50	50	49	50	50
InternVL2.5-8B	50	50	50	49	50	50
Qwen2.5-VL-3B	50	50	50	37	40	34
Qwen2.5-VL-7B	50	49	50	1	0	1
LLaVA-1.5-7B	49	49	50	42	44	49
LLaVA-1.5-13B	50	50	50	42	48	44

A.3 Supervised Positions

Each family uses its native processor and chat template with a single image view (no dynamic tiling for InternVL). Training ends at the answer without an appended closing or EOS target. Format

targets are native roles, delimiters and eligible whitespace; natural-language system text, visual tokens, padding and the first causal position are excluded. When a token crosses a boundary, answer and question membership take precedence over format. The counts N^{all}, N^{prompt} and N^{fmt} in Equation (5) are the realized eligible counts of each model and batch.

A.4 Queries and Statistics

Each cohort asks the Born and Education questions of all 25 identities under an image-referential cue, giving 50 queries. Decoding is greedy with native stop tokens. Responses and references are normalized by NFKC, case folding, replacement of non-alphanumerics by spaces and whitespace collapse; an exact match requires whole-response equality, and occurrence the complete normalized reference at word boundaries. All queries enter the analysis, including parent errors. Contrasts are formed per query, averaged within identity and then equally across identities. Intervals use 2,000 whole-identity resamples (Python `Random(0)`) that keep questions and checkpoints paired; the endpoints are the 50th and 1,950th sorted replicates. An effect is significant in a seed when its 95% interval excludes zero, and we apply no multiplicity correction. Unless noted otherwise, tables report the mean $\pm$ standard deviation over three seeds, figures show the same standard deviation as error bars or a light band, a count such as (2/3) gives the seeds in which an effect is significant, and each cohort has 50 queries over 25 identities.

B Matched-Forgetting Details

B.1 Answer-Only Frontier and Matching Rule

For InternVL2.5-8B and LLaVA-1.5-7B and each seed, we train the answer-only family

$$L_{F,\lambda} = \lambda\, C^{\text{ans}}/N^{\text{all}}, \qquad L_R = C^{\text{ans}}/N^{\text{all}}, \qquad \lambda \in \{0, \tfrac{1}{8}, \tfrac{1}{2}, 1\},$$

with question and format coefficients set to zero and all other settings as in Appendix A. Within a model and seed, all strengths start from the same LoRA factors and random states, and the λ=0 run processes the same records with zero forget loss.

Let o_i be the forget occurrence at the i-th strength in increasing order and t that of the target checkpoint. A strength with $o_i = t$ is selected directly (ties toward the smaller strength). Otherwise, each adjacent pair that strictly brackets t contributes the mixing weights $p_{i+1} = \alpha_i$ and $p_i = 1 - \alpha_i$ with $\alpha_i = (t - o_i)/(o_{i+1} - o_i)$, and the policy averages the weights of all bracketing pairs. If no pair brackets t, the target is reported as unmatched. For cohort c and scorer r, the policy score is $\widehat{\mu}_{c,r,\pi} = \sum_i p_i \widehat{\mu}_{c,r,i}$ and the residual is $\widehat{\delta}_{c,r} = \widehat{\mu}_{c,r,\text{target}} - \widehat{\mu}_{c,r,\pi}$. Each of 2,000 bootstrap draws resamples F, R and H identities independently, refits the rule on its F draw and shares the weights across cohorts and scorers. The F residual is therefore zero by construction.

For format-weighted training, the rule mixes λ=0 and λ=$\frac{1}{8}$ on InternVL2.5-8B (weight 0.74–0.85 on $\frac{1}{8}$), and on LLaVA-1.5-7B it selects λ=$\frac{1}{2}$ exactly in two seeds and places weight $\frac{15}{16}$ on $\frac{1}{2}$ in the third. No bootstrap draw falls outside the sweep.

Full-prompt training. On InternVL2.5-8B, the sweep brackets full-prompt training at every seed and in 1,881–1,988 of the 2,000 bootstrap draws, and Table 12 reports intervals over the bracketed draws. On LLaVA-1.5-7B, its forget occurrence (32%) lies below every sweep point (at least 34%), so it is unmatched and we report no residual.

B.2 Sweep Points

Table 6 lists the sweep points of Figure 3 and Figure 9.

Table 6: **Answer-only sweep and target checkpoints (occurrence, %).** Forgetting and retention fall together as λ grows, and the sweep brackets every target except full-prompt training on LLaVA-1.5-7B.

Model	Checkpoint	Forget F	Retained R	Held-out H
InternVL2.5-8B	Answer-only, $\lambda = 0$	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Answer-only, $\lambda = 1/8$	60.7 ± 1.2	98.0 ± 0.0	82.7 ± 1.2
	Answer-only, $\lambda = 1/2$	31.3 ± 1.2	70.0 ± 4.0	52.7 ± 1.2
	Answer-only, $\lambda = 1$	18.0 ± 2.0	36.7 ± 3.1	40.0 ± 0.0
	Format-weighted	68.7 ± 3.1	80.0 ± 0.0	74.0 ± 0.0
	Full-prompt	28.7 ± 1.2	33.3 ± 5.0	42.7 ± 1.2
LLaVA-1.5-7B	Answer-only, $\lambda = 0$	100.0 ± 0.0	100.0 ± 0.0	100.0 ± 0.0
	Answer-only, $\lambda = 1/8$	85.3 ± 1.2	98.0 ± 0.0	88.0 ± 0.0
	Answer-only, $\lambda = 1/2$	55.3 ± 1.2	84.0 ± 0.0	57.3 ± 1.2
	Answer-only, $\lambda = 1$	35.3 ± 1.2	59.3 ± 1.2	40.0 ± 0.0
	Format-weighted	56.0 ± 0.0	59.3 ± 2.3	56.7 ± 1.2
	Full-prompt	32.0 ± 0.0	50.0 ± 0.0	38.0 ± 0.0

C Factor Studies

Every factor study keeps the unlearning protocol of Section 4.1 and changes one thing. Table 7 lists the grid and its cost: every run uses three seeds and the budget of Table 4, and the cue study re-scores existing checkpoints. Table 9 gives the numbers behind Figure 4.

Table 7: **The factor grid.** Each study changes one condition of the answer-only protocol, and only studies that change the knowledge need new parents.

Study	Varied	Models	Parents	Runs
Answer-only frontier	$\lambda \in \{0, \frac{1}{8}, \frac{1}{2}, 1\}$	IVL-8B, LLaVA-7B	0	24
Model	same sweep	four other models	0	48
Retain relatedness	shared attribute, look-alike	IVL-8B, LLaVA-7B	4	48
Acquisition dose	$\frac{1}{3}\times$, $3\times$	IVL-8B, LLaVA-7B	4	54
Cue	name instead of portrait	IVL-8B, LLaVA-7B	0	0
Unlearning route	image records only, text records only	IVL-8B, LLaVA-7B	0	36
Training knobs	learning rate, updates, LoRA rank	IVL-8B	0	18
Optimizer memory	β_2=0.95; 1,000 updates	IVL-8B, LLaVA-7B	0	24
Objectives	eight objectives (Appendix E)	IVL-8B, LLaVA-7B	0	162

Retain relatedness. Each level replaces the 25 retained identities and keeps the forget and held-out identities. *Random* is the standard retain set. *Shared attribute* pairs every forget identity with a retain-pool identity that has the same birthplace or the same school, drawn without replacement from the remaining `retain_95` profiles. *Look-alike* pairs every forget identity with the retain-pool identity whose portrait is nearest in cosine similarity of the parent's own vision-encoder features (mean-pooled final layer). Each retain set is taught in its own acquisition run, so each level has its own parent, and every parent passes the acquisition check of Table 5.

Acquisition dose. Forget records appear one third or three times as often in the acquisition stream as in the standard schedule (2,000 or 18,000 instead of 6,000 exposures); retained and held-out records are unchanged, so the number of acquisition updates changes with the dose. At one third of the dose, the parents still reproduce at least 45 of the 50 forget answers, so the facts are known but weakly held. The sweep of the $3\times$ dose adds λ=2 so that it brackets 50% forget occurrence. The strength to reach 50% interpolates the sweep log-linearly in λ.

Cue. Name cues replace the portrait with the identity's name in the same two questions. We score every checkpoint of the answer-only sweep under both cues, so the name-cue frontier needs no training.

Unlearning route. Our protocol draws six image and six text forget records per update. The image-only and text-only routes draw all twelve forget records from one view and keep the retained streams, the budget and every other setting. Their sweeps use $\lambda \in \{0, \frac{1}{4}, 1, 4\}$, which brackets 50% forget occurrence under both cues, and every checkpoint is scored with portraits and with names. Under both cues, forgetting is the higher of the two forget occurrences of a checkpoint and retention is the mean of its two retained occurrences. Table 8 gives the summaries behind Figure 5: retention at 50% forget occurrence under the portrait cue, the name cue and both cues, and the name leak of each route, its name-cue forget occurrence at the strength at which its portrait-cue forget occurrence reaches 50%.

Table 8: **Retention at 50% forget by unlearning route and cue (%).** Unlearning through both modalities keeps the most retention under both cues, while unlearning through images alone leaves most name-cued answers in place.

Unlearn through	InternVL2.5-8B				LLaVA-1.5-7B			
	Portrait	Name	Both	Name leak	Portrait	Name	Both	Name leak
both	87.8	79.6	81.7	55	77.4	65.7	68.0	56
image only	88.9	84.7	66.3	80	74.7	61.9	48.3	82
text only	86.1	81.9	64.7	31	72.6	61.0	51.6	29

Model and procedure. The four other parents are swept exactly like InternVL2.5-8B and LLaVA-1.5-7B. On InternVL2.5-8B, the learning rate (5×10^{-6}, 2×10^{-5}) and the number of updates (125, 500, with the linear decay stretched to match) vary at $\lambda=\frac{1}{4}$, and the LoRA rank (16, 256, with α/r fixed at $\frac{1}{2}$) varies at $\lambda=\frac{1}{2}$. Each run is placed against the λ frontier by its iso-forget residual. The optimizer-memory runs retrain format-weighted and answer+format training with $\beta_2=0.95$, or with 1,000 updates, and all else fixed.

Table 9 summarizes each knowledge level by its free budget (pp), its retained and held-out occurrence at 50% forget occurrence (%), and the forget weight λ at which its sweep reaches 50% forget occurrence.

Table 9: **Frontier summaries under changed knowledge.** Related retain sets lower the frontier, look-alike portraits most, while held-out knowledge barely changes; a lower dose lowers the frontier and a higher dose raises it.

Factor	Level	InternVL2.5-8B				LLaVA-1.5-7B			
		Free	Retained	Held-out	λ	Free	Retained	Held-out	λ
Retain set	random	42	87.8	72	0.21	21	77.4	53	0.60
	shared attribute	39	79.2	75	0.21	18	65.6	51	0.58
	look-alike	17	69.1	73	0.20	12	48.2	51	0.60
Acquisition	1/3 dose	14	82.1	74	0.12	9	51.3	50	0.32
	standard	42	87.8	72	0.21	21	77.4	53	0.60
	3× dose	50	95.0	75	0.45	30	84.1	52	1.36

Related retain sets leave held-out occurrence at 50% forget occurrence within 3 pp of the random condition (Table 9), so the cost of relatedness falls on the related identities rather than on knowledge in general.

D FORMAT WEIGHTING

This appendix gives the measurements behind the format-weight example of Section 4.5: how format weighting changes forgetting and retention on six models and which of its components causes the change, where format-weighted training lies relative to the answer-only frontier, and the evidence for the optimizer account.

D.1 Format Weighting on Six Models

Table 10: **Format-weighted minus full-prompt training (pp).** Format weighting raises forget, retained and held-out occurrence on every model, and the forget increase is significant in 17 of the 18 model–seed runs.

Model	Forget F (↓)	Retained R (↑)	Held-out H (↑)
InternVL2.5-2B	$+12.7 \pm 1.2$ (3/3)	$+5.3 \pm 1.2$ (0/3)	$+4.0 \pm 5.3$ (1/3)
InternVL2.5-8B	$+40.0 \pm 4.0$ (3/3)	$+46.7 \pm 5.0$ (3/3)	$+31.3 \pm 1.2$ (3/3)
Qwen2.5-VL-3B	$+20.7 \pm 2.3$ (3/3)	$+12.7 \pm 4.2$ (2/3)	$+22.0 \pm 5.3$ (3/3)
Qwen2.5-VL-7B	$+20.7 \pm 4.2$ (3/3)	$+15.3 \pm 3.1$ (3/3)	$+13.3 \pm 4.6$ (2/3)
LLaVA-1.5-7B	$+24.0 \pm 0.0$ (3/3)	$+9.3 \pm 2.3$ (0/3)	$+18.7 \pm 1.2$ (3/3)
LLaVA-1.5-13B	$+14.7 \pm 1.2$ (2/3)	$+8.7 \pm 4.2$ (0/3)	$+16.0 \pm 2.0$ (3/3)

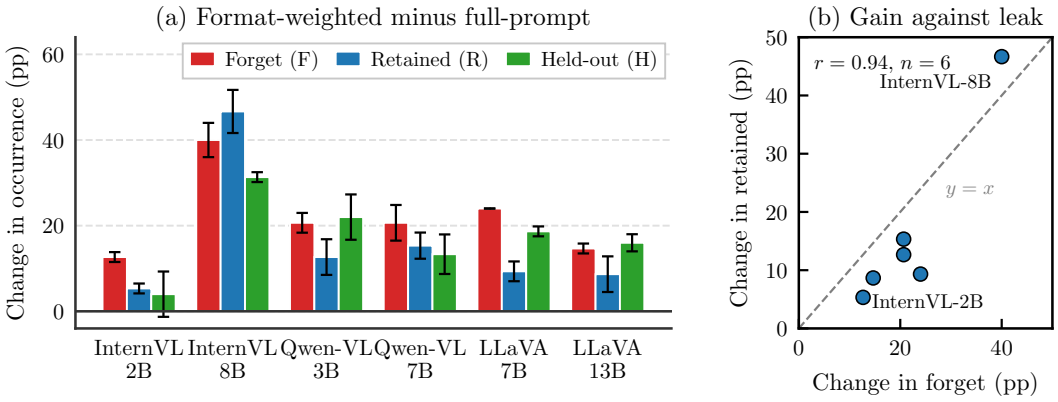

Figure 8: **Format-weighted minus full-prompt training on six models.** Format weighting raises retention and forget leak together on every model, and the retained gain grows with the leak.

Against full-prompt training, format weighting raises mean retained occurrence on every model, by 5 to 47 pp, which a results table would report as a utility gain (Figure 8a). The same change raises mean forget occurrence on every model, by 13 to 40 pp, so more forgotten answers leak. The forget increase lies above zero in 17 of the 18 model–seed runs and the retained gain in 8 (Table 10, where counts give the seeds whose interval lies above zero). Held-out identities, which unlearning never touches, move with the retained set. The retained gain grows with the forget increase across the six models (r=0.94; Figure 8b), as movement along a frontier would.

Moving from full-prompt to format-weighted training changes two things at once. It removes question supervision, and it raises the per-token weight on format targets. Answer+format training (Equation (5)) separates the two (Table 11). Removing question supervision changes mean occurrence by less than 5 pp and is not significant in any seed. Weighting the format targets reproduces nearly the entire effect: on InternVL2.5-8B it adds +42 pp forget, +50 pp retained and +31 pp held-out occurrence, significant in every seed. Adding format supervision at the original weight lowers mean retained occurrence, by 7 pp on InternVL2.5-8B and 14 pp on LLaVA-1.5-7B. Section 4.5 therefore measures format weighting from answer+format training.

D.2 Format Weighting at Matched Forgetting

Figure 9 places format-weighted and full-prompt training on the answer-only frontier; points in the shaded region are dominated by it. At the forgetting level of format-weighted training, the answer-only procedure retains 18.4 pp more on InternVL2.5-8B and 25.0 pp more on LLaVA-1.5-7B, significant in every seed (Table 12). Held-out identities, which the matching never uses, show smaller residuals that are not significant in any seed. Full-prompt training lies 29.9 pp below the frontier on InternVL2.5-8B (significant in two of three seeds) and outside the sweep on LLaVA-1.5-7B (Appendix B.1). By Equation (4), the 47 pp retention gain of format-weighted over full-prompt

Table 11: **Component contrasts (pp).** Only weighting the format targets raises forget occurrence in every seed on both models, and removing question targets is not significant in any seed.

Model	Change	Forget F	Retained R	Held-out H
InternVL2.5-8B	Remove question targets	-2.0 ± 3.5 (0/3)	-3.3 ± 3.1 (0/3)	$+0.7 \pm 1.2$ (0/3)
	Weight format targets	$+42.0 \pm 2.0$ (3/3)	$+50.0 \pm 3.5$ (3/3)	$+30.7 \pm 1.2$ (3/3)
	Add format targets	$+8.7 \pm 3.1$ (1/3)	-6.7 ± 6.1 (1/3)	$+3.3 \pm 1.2$ (0/3)
LLaVA-1.5-7B	Remove question targets	$+4.7 \pm 1.2$ (0/3)	-4.7 ± 2.3 (0/3)	$+0.0 \pm 0.0$ (0/3)
	Weight format targets	$+19.3 \pm 1.2$ (3/3)	$+14.0 \pm 4.0$ (1/3)	$+18.7 \pm 1.2$ (3/3)
	Add format targets	$+1.3 \pm 1.2$ (0/3)	-14.0 ± 3.5 (2/3)	-2.0 ± 0.0 (0/3)

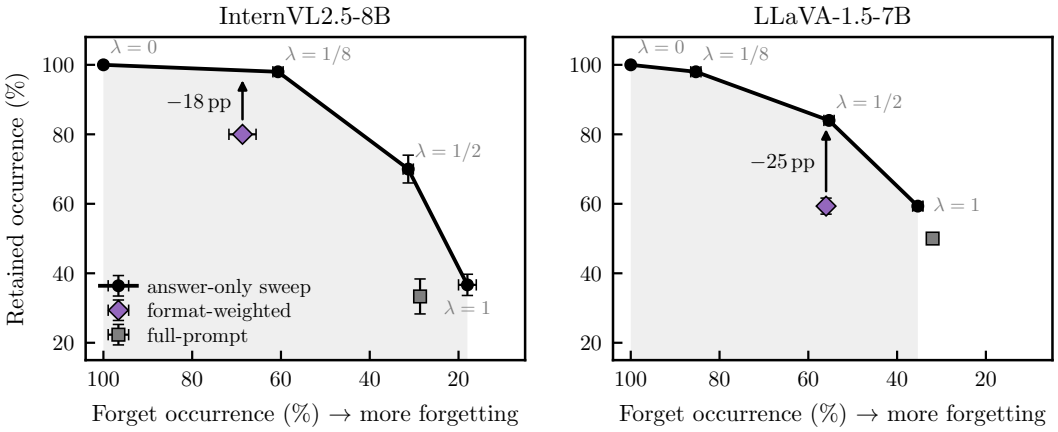

Figure 9: **Format-weighted and full-prompt training on the answer-only frontier.** Both lie below the frontier wherever the sweep reaches them, so the answer-only procedure retains more at their forgetting levels.

training on InternVL2.5-8B splits into 35 pp of movement along the frontier and a 12 pp shift, the smaller gap below the frontier (Figure 1a). The interpolation also predicts a real checkpoint, since answer-only training at λ=0.10 on InternVL2.5-8B reaches $(F, R) = (66.7, 97.6)$, within 1 pp of the frontier value of 98.3 at that forgetting level.

A residual is a checkpoint's retention minus that of the answer-only frontier at the checkpoint's own forget occurrence (Equation (2)). Its intervals refit the matching weight in each bootstrap draw and use only the draws in which the sweep brackets the checkpoint (Algorithm 1); the forget residual is zero by construction.

Table 12: **Iso-forget residuals against the answer-only frontier (pp).** Format-weighted training retains less than the answer-only procedure at its own forgetting level in every seed, while no held-out residual is significant.

Model	Checkpoint	Retained	Held-out
InternVL2.5-8B	Format-weighted	-18.4 ± 0.1 (3/3)	-12.2 ± 1.8 (0/3)
	Full-prompt	-29.9 ± 10.9 (2/3)	-7.4 ± 1.2 (0/3)
LLaVA-1.5-7B	Format-weighted	-25.0 ± 1.8 (3/3)	-1.3 ± 1.2 (0/3)
	Full-prompt	outside the sweep; not matched	

The residuals also split by modality (Table 13). Every checkpoint also answered the benchmark's interest and description questions, once with the portrait (image view) and once with the name (text view). Matched on portrait-cue forgetting of the short questions, the retained occurrence of format-weighted training on these questions is 35 pp below the answer-only procedure in the image view and 5 pp below in the text view on InternVL2.5-8B; on LLaVA-1.5-7B both residuals stay within 5 pp.

Table 13: **Iso-forget residuals of format-weighted training by view (pp).** Its retention loss falls on questions asked with the portrait on InternVL2.5-8B, while both views stay close to zero on LLaVA-1.5-7B.

Model	Asked with	Forget F	Retained R	Held-out H
InternVL2.5-8B	portrait (image view)	$+10.0 \pm 2.9$	-35.1 ± 2.8	-7.4 ± 1.6
	name (text view)	$+8.9 \pm 0.7$	-4.8 ± 2.6	-5.2 ± 2.9
LLaVA-1.5-7B	portrait (image view)	-0.1 ± 0.1	-3.5 ± 0.8	$+1.3 \pm 1.2$
	name (text view)	$+4.6 \pm 1.2$	$+4.5 \pm 1.3$	$+1.4 \pm 1.2$

D.3 Evidence for the Optimizer Account

Real models. Three measurements fit the account of Section 4.5. The shift appears when the per-token format weight rises but not when question targets are removed, and adding format targets at the original weight lowers retention instead (Table 11). It reaches held-out identities that unlearning never touches (+31 pp), as a smaller overall update would. And format-weighted training forgets about as little as a much weaker answer-only run, since its forget occurrence falls between those of λ=0 and $\lambda=\frac{1}{8}$ on InternVL2.5-8B and matches $\lambda=\frac{1}{2}$ on LLaVA-1.5-7B (Table 6). Direct tests on InternVL2.5-8B and LLaVA-1.5-7B agree (Figure 10). Format-weighted adapters move 49% less than answer+format adapters on InternVL2.5-8B and 20% less on LLaVA-1.5-7B, and across the six models, the more format weighting raises a model's forget occurrence over full-prompt training, the smaller its update. The format-weighted run moves more than the answer+format run during the first 20 updates, while the format loss falls, and less afterwards, ending at 51% of its distance. With β_2=0.95, the forget shift of format weighting (format-weighted minus answer+format) shrinks from 42 to 27 pp on InternVL2.5-8B and from 19 to 15 pp on LLaVA-1.5-7B, and 1,000-update runs shrink it to 11 and 5 pp, so the effect should vanish under the 5,500-update public recipe (Appendix G).

Toy model. A frozen linear memory $W_0 \in \mathbb{R}^{128\times128}$ stores 75 facts exactly (least squares): unit keys whose component along a shared direction is 0.7 map to random unit values; 25 facts form the

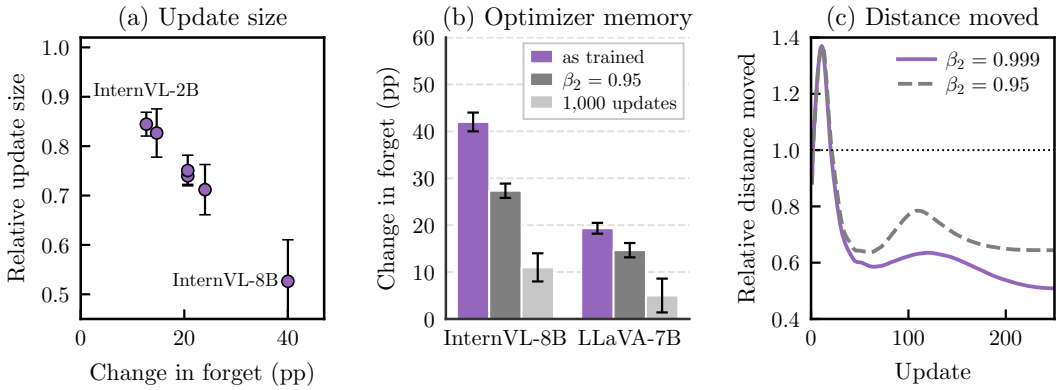

Figure 10: **Direct tests of the optimizer account.** The more format weighting raises a model's forget occurrence, the less its adapter moves, and a shorter optimizer memory shrinks that rise. Panel (c) uses InternVL2.5-8B.

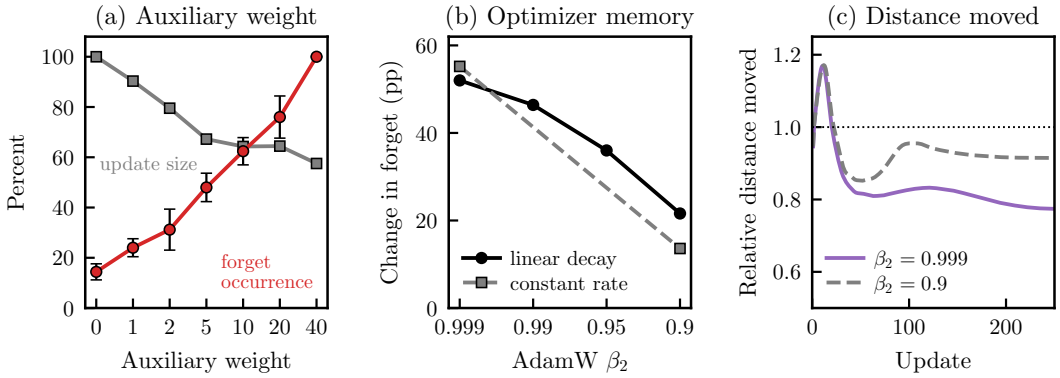

Figure 11: **Toy model of the format-weight knob.** Raising the weight of a quickly fit auxiliary target reduces forgetting and shrinks the update, and the effect needs AdamW's long memory.

forget set, 25 the retained set and 25 the held-out set. A refusal vector and a template key complete the world; the template target differs from the parent's output by a unit vector, so the auxiliary term starts mis-fit and is fit within a few updates. A rank-8 adapter BA (A Gaussian, B=0) is trained for 250 updates on 6 forget and 6 retained records per update: squared error to the refusal on forget keys, rehearsal of retained values with weight 0.1, and the template term with the auxiliary weight (0 for answer-only, 1 for answer+format and 20 for format-weighted training). AdamW uses β_1=0.9, ϵ=10^{-8}, weight decay 0.01, no warmup, and a learning rate of 3×10^{-3} decayed linearly to zero. A fact occurs when the adapted memory maps its key closer, in cosine, to its own value than to any other value or to the refusal. All values are means over five seeds.

Across auxiliary weights 0, 1, 2, 5, 10, 20 and 40, forget occurrence is 14, 24, 31, 48, 62, 76 and 100%, and the adapter update is 100, 90, 80, 67, 64, 64 and 58% of the answer-only update (Figure 11). The gap in forget occurrence between weights 20 and 1 shrinks from 52 pp at β_2=0.999 to 22 pp at β_2=0.9, and under SGD with step size 0.03 the weight leaves forgetting unchanged (70, 69 and 70% at weights 0, 1 and 20). Longer runs also shrink the gap, from 52 pp at 250 updates to 13 pp at 1,000, partly because weight 1 already forgets every fact by then. The toy model shows that the account is internally consistent, and the real-model tests above check that the models follow it.

E Objectives Study

Table 14: **Objectives and their strength parameters.** Every sweep brackets 50% forget occurrence on both models.

Objective	Family	Strength parameter
Gradient ascent (Maini et al., 2024)	ascent on forget answers	learning rate
Gradient difference (Maini et al., 2024)	ascent on forget, descent on retain	learning rate
KL minimization (Maini et al., 2024)	ascent on forget, KL to the parent on retain	learning rate
NPO (Zhang et al., 2024)	negative preference optimization	learning rate
Answer-only (Maini et al., 2024)	refusal ("I don't know") training	forget weight λ
SMFA (Zeng et al., 2026)	multimodal, retained-data anchors	forget weight
CMJU (Yang et al., 2026)	multimodal, cross-modal balancing	forget weight
Format-weighted	answer-only with format weighting	forget weight

The study runs the objectives of Table 14 on InternVL2.5-8B and LLaVA-1.5-7B with three seeds (Table 7). The first seven are established objectives, and format-weighted training is the variant of Section 4.1. Each sweep takes $\frac{1}{4}$, 1 and 4 times a reference value of the strength parameter, plus the shared no-forgetting run (λ=0), and each objective also runs at the default of its reference implementation. Answer-only training reuses the sweep of Section 4.1 with its default at λ=1, and the default of format-weighted training is its measured run. Sweep grids are fixed before any scoring, and an objective whose sweep does not bracket a target level is reported as unmatched at that level. Table 15 gives each objective's forget and retained occurrence at its default, its retention at 50% forget occurrence on its own frontier (matched), and the change from its reported to its matched rank, where ties share the better rank.

Rankings. Ranked by retention at their defaults and at matched forgetting, the eight objectives agree weakly (Kendall τ_b of 0.36 and 0.40; Table 15). Format-weighted training ranks first and tied second by reported retention and last at matched forgetting. Its defaults leak 69% and 56% of forgotten answers, so its reported retention comes from forgetting less. SMFA ranks first or second under both criteria, so matching does not penalize strong methods; it removes credit that comes only from forgetting less.

On LLaVA-1.5-7B, the defaults of the seven established objectives spread over 38 pp of forget occurrence, while their frontiers differ by at most 12 pp at 50% (Figure 12a). The matched ranking does not hinge on the level: for every τ between 40 and 80%, the ranking by $\mathrm{RMF}(\tau)$ agrees with the ranking at 50% at $\tau_b \geq 0.57$, and with the reported ranking at $\tau_b \leq 0.55$ (Figure 12b).

Table 15: **Reported versus matched rankings of eight objectives.** The two rankings agree only weakly: format-weighted training drops to last at matched forgetting, while SMFA stays first or second under both.

Objective	InternVL2.5-8B				LLaVA-1.5-7B			
	Forget	Retained	Matched	Rank	Forget	Retained	Matched	Rank
Gradient ascent	6	0	78.9	↑1	10	0	65.8	↑1
Gradient difference	12	2	84.1	↑1	18	22	67.5	↑2
KL minimization	16	14	84.7	↑1	28	38	67.3	=
NPO	14	18	91.6	↑2	30	48	72.5	=
Answer-only	18	37	87.8	↓1	35	59	77.4	↑1
SMFA	24	60	92.8	↑1	48	80	77.3	↓1
CMJU	16	28	91.6	↑2	26	44	73.1	↑2
Format-weighted	69	80	51.0	↓7	56	59	51.7	↓6
Rank agreement (Kendall τ_b)				0.36				0.40

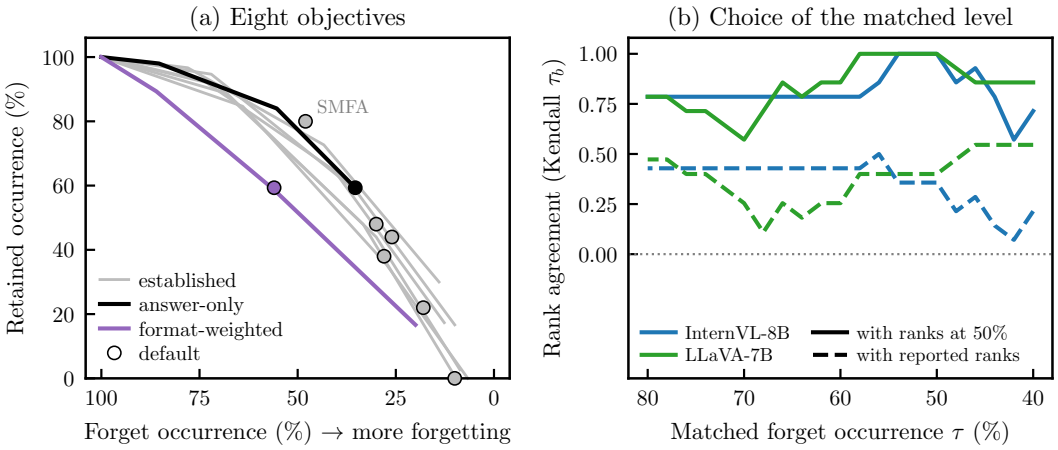

Figure 12: **Objectives on LLaVA-1.5-7B and the choice of the matched level.** As on InternVL2.5-8B, the frontiers lie far closer together than the defaults, and the matched ranking barely depends on the level at which it is taken.

F SCORERS

Our controlled parents contain the complete reference answer in almost every response (Table 5), which is why the main text scores occurrence. Free-form responses can instead reword an answer, and this appendix shows that scorers can then disagree in direction. The analysis uses a public unlearning recipe on a released parent.

F.1 PUBLIC RECIPE

We use `ZhenZeng/S-MLLMUn-Qwen2.5-VL-7B-original` and `ZhenZeng/S-MLLMUn-data` (Zeng et al., 2026). The memory training pool holds 600 image and 500 text records per cohort. Unlearning uses the public `forget_5` and `retain_few_shot_5` splits (50 identities each); evaluation uses `forget_5` and `retain_5` (50 forget and 950 retained identities). The official refusal recipe supplies the loader, collator and loss, and the parent's acquisition membership is unknown. With the official loss L_{off} and its supervised-label count N_{off}, the official recipe and two variants that add question or format supervision use the forget-batch losses

$$L_{\text{official}} = L_{\text{off}}, \qquad L_{\text{+question}} = L_{\text{off}} + \frac{N^{\text{ques}}+N^{\text{fmt}}}{N_{\text{off}}N^{\text{ques}}}C^{\text{ques}}, \qquad L_{\text{+format}} = L_{\text{off}} + \frac{N^{\text{ques}}+N^{\text{fmt}}}{N_{\text{off}}N^{\text{fmt}}}C^{\text{fmt}},$$

and retained batches keep L_{off}. The three recipes share record order, refusal targets, precision and schedule: 2,200 records for five epochs (batch 2, 5,500 updates), learning rate 10^{-4}, LoRA rank/α/dropout 16/16/0.05, RGB 384^2 inputs and greedy decoding with 100 new tokens, with three seeds each. The first two seeds ran on RTX 5000 Ada and the third on RTX 4090, so cross-seed variation also includes the platform.

F.2 SCORER DEFINITIONS

Five scorers read the saved responses. Occurrence requires the complete normalized reference at word boundaries, as in the main text. Value match checks one explicit value extracted from the source question–answer pair with attribute-specific frames; the mapping is fixed before scoring, and ambiguous, compound or unsupported values stay unmapped, so value match is defined on 237 of 350 forget-image and 4,330 of 6,650 retained-image queries and undefined on understanding questions. Token coverage holds when occurrence holds or at least 80% of the reference tokens (as a multiset) appear in the response. Two open LLM judges, unadapted BF16 Qwen3-8B and Mistral-Nemo-Instruct-2407, rate agreement with the reference on a 0–10 scale with native chat templates, greedy decoding, a 256-token output cap and Qwen thinking disabled. Prompts contain only the question, the textual reference and the response, invalid or truncated outputs stay missing, and effects use complete pairs. The rubric is:

> Score factual agreement from 0 to 10. 9–10: the same factual meaning, including a faithful paraphrase. 7–8: mostly correct with minor omissions or deviations. 5–6: partly correct with substantial omissions or errors. 3–4: major errors or missing central information. 0–2: incorrect, irrelevant, nonsensical, or no factual answer. A refusal does not demonstrate the reference fact. A response that merely mentions a value while denying it does not affirm that value. Judge the whole response; penalize contradictory or irrelevant added claims.

On 32 synthetic probes, both judges gave faithful paraphrases 9–10, changed values 0–4, and negations and refusals 0. The forget-image panel has 237 mapped queries over 50 identities; the retained image and understanding panels have 1,105 and 1,428 queries over 238 identities, every fourth identity in sorted order.

F.3 SCORERS CAN DISAGREE IN DIRECTION

Table 16 shows the characteristic failure of three scorers on saved responses: exact matching misses a rewording, value matching accepts a negation and coverage matching accepts a substitution. Such failures can change the measured direction. In the run in which the scorers disagree most, adding question supervision lowers occurrence on forgotten images by 5.6 pp, which would be reported as stronger forgetting, while on the same 237 responses value match, token coverage and both judges

rise (Table 17). Most of the disagreement comes from responses that drop the exact reference sentence but still state the value. The scorers do not always disagree (Table 18): across all conditions, occurrence has 12 significant effects, and both judges agree with it on five, oppose it on three and are split or not significant on four.

Table 16: **Saved responses and the scorer that misreads each.** Every scorer has a characteristic failure.

Event	Response pair	Recorded verdict	Against the reference
Rewording (forget)	"This individual was born in Cairo, Egypt." → "This **person** was born in Cairo, Egypt."	occurrence lost; value match and coverage kept	Value stated; occurrence misses it
Negation (forget)	"This person was **not** born in Cairo, Egypt."	value match: match	Denied; value match over-counts
Substitution (forget)	surname "Almquist" → "**VazN**", most of the template kept	occurrence and value match lost; coverage kept	Wrong; coverage over-counts

Table 17: **One run under all five scorers.** Occurrence falls while every other scorer rises, and no interval contains zero.

Scorer	Change	95% interval
Occurrence (pp)	-5.6	$[-10.4, -1.0]$
Value match (pp)	$+17.0$	$[10.2, 24.0]$
Token coverage (pp)	$+9.1$	$[2.8, 15.3]$
Qwen3-8B judge (points)	$+1.40$	$[0.78, 2.07]$
Mistral-Nemo judge (points)	$+1.40$	$[0.81, 2.01]$

Table 18: **Significant effects by scorer.** Scorers often disagree in sign on the same responses: adding question supervision raises value match and token coverage on forget images in every seed, while occurrence rises in one seed and falls in another.

Panel	Variant	Occurrence	Value	Coverage	Qwen3	Nemo
Forget image	+question	$+1$ / -1	$+3$	$+3$	$+2$	$+2$
Forget image	+format	$+1$ / -1	0	$+2$	$+1$	$+1$
Retained image	+question	-2	-1	$+1$ / -1	-3	-2
Retained image	+format	-2	$+1$	-1	-1	$+1$ / -1
Retained understanding	+question	$+3$	–	$+2$	-1	$+1$ / -1
Retained understanding	+format	-1	–	0	$+2$ / -1	$+1$ / -1

F.4 Human Labels

To decide which verdict is right, three annotators labeled 400 forget-panel responses sampled from the runs of the three recipes, stratified by whether occurrence and value match agree. Annotators see the question, the reference and the response, and label whether the response states the forgotten fact (leak), denies or replaces it (no leak), or is undecidable. The majority label is the ground truth, and undecidable responses are excluded from precision and recall. Agreement is high (Fleiss κ=0.82). Occurrence has the highest precision but misses 42% of human-labeled leaks, almost all of them rewordings, and the LLM judges give the best balance (F1 up to 0.88; Table 19). In the run of Table 17, human labels agree with value match, token coverage and both judges that adding question supervision increases leaks. A forgetting claim on free-form responses should therefore name its scorer and be checked against a scorer of a different type.

Table 19: **Scorers against human leak labels.** Human labels agree with value match, token coverage and both judges on the direction of the change, and occurrence is the most precise scorer but misses many leaks.

Scorer	Precision	Recall	F1	Direction
Occurrence	0.97	0.58	0.73	−
Value match	0.84	0.88	0.86	+
Token coverage	0.71	0.90	0.79	+
Qwen3 judge	0.91	0.86	0.88	+
Nemo judge	0.89	0.83	0.86	+
Human labels	–	–	–	+

G Format Weighting Under the Public Recipe

We test whether the format-weighting effect of the controlled study transfers to the public recipe of Appendix F.1, where it predicts that adding format supervision exposes more forgotten and retained answers than adding question supervision. Table 20 reports +format minus +question with the seeds whose 95% interval lies above or below zero. The effect does not transfer. On forget images, value match is negative on average and significantly negative in one seed, and no panel is significantly positive in more than one seed. The official recipe does forget, lowering forget-image occurrence from 114 to 27 of 350 queries at a retained cost from 2,022 to 901 of 6,650, so the null result is not a floor effect. The optimizer account predicts it for a schedule of 5,500 updates (Appendix D.3), so the size and sign of the effect should be measured per recipe.

Table 20: **Format weighting under the public recipe: +format minus +question (pp).** Every panel was predicted to rise, yet three of the four effects are negative on average and none is positive in more than one seed.

Panel	Scorer	Queries	Predicted	Observed	Seeds above 0	Seeds below 0
Forget image	value match	237	+	-7.7 ± 4.8	0/3	1/3
Retained image	value match	4,330	+	$+1.6 \pm 5.1$	1/3	1/3
Forget understanding	occurrence	300	+	-7.1 ± 5.2	0/3	2/3
Retained understanding	occurrence	5,700	+	-7.6 ± 2.3	0/3	3/3

H Limitations

Facts taught by fine-tuning give known ground truth but may behave unlike pretraining facts. We study the knowledge and the modality on two of the six models, and each frontier costs one run per sweep point and seed. The format-weight effect did not transfer to a longer public recipe (Appendix G), as the optimizer account predicts but one recipe cannot confirm. Our scorers compare responses with text references and do not verify visual grounding, and occurrence can disagree in direction with value-based scorers on the same responses (Appendix F).