ROUTEMASK: UNLEARNING IDENTITIES ALONG ACCESS ROUTES IN VLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) learn from large collections of web images and text and are now widely used in assistants and search. Some of what they learn concerns real people: a single photograph can be enough for a model to reveal a person's birthplace, occupation, or education. Machine unlearning aims to remove such information after training. Existing methods for VLMs fine-tune the model with gradient ascent, preference optimization, or refusal targets, but they mostly change what the model says, not what it knows. After refusal training, the model still picks the true answer in multiple choice, and forgetting trained with names does not carry over to photographs. We address this in two steps. First, we locate where identity is accessed: injecting name states into frozen models while blocking attention to the image reveals an entity state that faces and names share in middle layers, and a visual bypass through which late layers read the image again. Second, we propose ROUTEMASK, which updates only the MLP gate rows in these layers. At each step, a row mask compares the attribution of the current forget answers with fixed retain and perception profiles and scales the parameter update, and every forget term is paired with a retain term under both name and image cues. Experiments on MLLMU-Bench and S-MLLMUn with LLaVA-1.5 and Qwen2.5-VL show that ROUTEMASK gives the lowest forget factuality on MLLMU-Bench, improves on SMFA in every score on S-MLLMUn, and forgets more than uniform or shuffled updates of the same step size.

1 INTRODUCTION

Vision–language models (VLMs) such as LLaVA and Qwen2.5-VL are now widely used in assistants, search, and accessibility tools, where they answer open questions about images by combining visual evidence with knowledge acquired in training (Liu et al., 2024; Bai et al., 2025). Part of that knowledge concerns individual people. A model that recognizes a face can state the person's birthplace, occupation, or education, and privacy requests may require such information to be removed after a model has been trained and released. Retraining without the affected data is rarely practical, which motivates machine unlearning: editing an existing model so that it no longer provides the targeted information while keeping everything else.

VLMs make this harder than in text-only models. The same fact can be requested with the person's name or with their photograph. A common recipe is *refusal training*: the model is fine-tuned to answer "I don't know." to questions about forget identities while it keeps its answers on retain identities (Maini et al., 2024). Refusal training changes what the model says, but a content probe can still show that the true occupation outranks the alternatives (Sun et al., 2025; Ye et al., 2026); Figure 1a shows the same gap on MLLMU-Bench. Training harder also suppresses facts about other people and refuses questions about visible clothing. An unlearning edit must therefore be judged on both cues and on the information it should preserve.

Multimodal benchmarks separate several of these requirements. MLLMU-Bench reports generation and classification, and S-MLLMUn evaluates personal memory alongside visual understanding (Liu et al., 2025a; Zeng et al., 2026). We treat classification and factuality in open answers as equally important outcomes, and protect perception questions about the same images.

Where should an edit act? Studies of language models describe factual recall as the formation of a subject representation followed by attribute extraction in later layers (Meng et al., 2022; Geva et al.,

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.

2023). An image cue adds a question that has no text-only counterpart: how does the identity in the image reach that extraction? Section 4 answers it with frozen interventions. Face and name representations align in middle layers. A name state injected there can replace the photographed person across three attributes, but only after later layers are blocked from reading the image. The image itself is a second, persistent source of identity. We call the middle layers where the shared state forms the *entity band*, and the late layers that read identity from the image again the *re-entry window*; the path from the image to that window is a visual bypass. An edit confined to the entity band leaves this bypass open.

We use this localization to decide which parameters to update. In a gated MLP, the gate projection decides, for each input, how strongly each channel is used. Scaling a channel's activation by a fixed factor would change every input that uses the channel; we instead scale the *update* of the gate row, so the edited gate still responds to the input. A row mask, recomputed for each forget example, selects which gate rows in the entity band and the re-entry window may move, and by how much. The mask is needed only during training. A single mask estimated in advance would be simpler, but forget attributions averaged over identities are unstable and dominated by a few examples (Section 5).

Our contributions are:

- a causal description of identity access in a VLM, separating a shared entity state from a visual bypass;
- an update restricted to native gate rows at those locations and weighted by a row mask, recomputed for each forget example, that compares the attribution of its true answer with fixed retain and perception profiles;
- an objective in which every forget term has a retain mirror on the same behavior, with ablations of the supervised cues, the objective terms, and the update location, and a comparison against controls that match the layers and step size of the mask.

Editing gate projections, restricting updates with saliency masks, and attributing weights with forget and retain data each have precedents (Sondej et al., 2025; Fan et al., 2024; Jia et al., 2024). What we add is where to apply them, and we test directly whether the mask beats uniform steps of the same size at the same locations.

2 Related work

Machine unlearning. Unlearning was first studied as exact removal by retraining on data shards (Bourtoule et al., 2021). For large language models, approximate methods fine-tune the trained model instead: gradient ascent on the forget set (Jang et al., 2023), preference-style objectives such as NPO (Zhang et al., 2024), refusal targets (Maini et al., 2024), or changes to internal representations (Li et al., 2024b). TOFU evaluates these methods on fictitious authors (Maini et al., 2024). Several studies find that the removed information often remains recoverable: it returns after a few steps of fine-tuning or after quantization (Łucki et al., 2024; Zhang et al., 2025), and it can reappear when the question format changes (Sun et al., 2025; Ye et al., 2026). DF-MCQ therefore flattens the answer distribution over choices (Sun et al., 2025); we flatten answer contents rather than option letters, and we evaluate both open answers and multiple choice.

Unlearning in VLMs. FIUBench, MLLMU-Bench, and S-MLLMUn evaluate forgetting of personal information in VLMs, together with retain knowledge, generation, and visual understanding (Ma et al., 2025; Liu et al., 2025a; Zeng et al., 2026). Single Image Unlearning fine-tunes on one image per concept (Li et al., 2024a), MMUnlearner restricts updates with a saliency map that separates visual from textual knowledge (Huo et al., 2025), and MANU prunes neurons selected by the ratio of their forget and retain importance (Liu et al., 2025b). SMFA masks updates with guidance from retain anchors on S-MLLMUn (Zeng et al., 2026). AIM forgets whole identities through the vision encoder, and MLLMEraser steers activations at inference time (Lee et al., 2026; Ding et al., 2026). For safety alignment, unlearning in the text domain alone can reduce harmful outputs for image inputs (Chakraborty et al., 2024); for personal facts, we find that forgetting learned under one cue transfers poorly to the other.

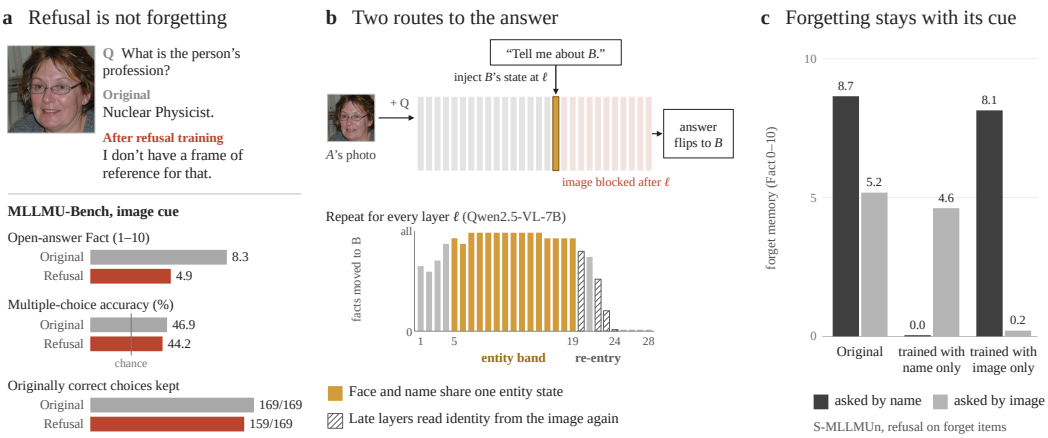

Figure 1: **Task and motivation.** (a) After refusal training, the model declines an open question about a forget identity but still picks the true answer in multiple choice. (b) How the two routes are found. The model sees A's photograph and a question. At layer ℓ we replace the hidden state at the last question token with B's name state, taken from "Tell me about B.", and, in every layer after ℓ, mask the attention from text tokens to the image tokens, so the photograph can no longer be read and the injected name is the only source of identity. If the answer then switches to B's facts, layer ℓ holds an identity state that faces and names share; repeating this at every layer marks these layers in yellow, the *entity band*. Hatched layers are those where allowing attention to the image again restores A's answer: the *re-entry window*, where the photograph supplies identity a second time. (c) Refusal trained on one cue leaves memory under the other cue (Table 8).

Locating recall and restricting updates. Feed-forward layers act as key–value memories (Geva et al., 2021), and individual neurons have been linked to specific facts (Dai et al., 2022). ROME and MEMIT tie factual editing to feed-forward computation in middle layers, and attribute extraction has been traced to later attention (Meng et al., 2022; 2023; Geva et al., 2023). The layer where causal tracing places a fact does not always predict where an edit works best (Hase et al., 2023), so we select layers by intervention rather than by tracing alone. In VLMs, questions about images draw on earlier MLP and attention layers than in language models (Basu et al., 2024), and object information moves from image tokens to the final token in middle to late layers (Neo et al., 2025), consistent with the re-entry window we find for identity. SalUn restricts unlearning with gradient saliency, WAGLE attributes weights through forget and retain objectives, and MUDMAN updates gate projections with an elementwise mask of gradient signs that changes at every step (Fan et al., 2024; Jia et al., 2024; Sondej et al., 2025). Our mask acts on whole rows, is recomputed for each forget example from the attribution of its true answer against fixed retain and perception profiles, and scales the parameter update.

3 Setting and Terminology

Let θ_0 be the benchmark's Original checkpoint, which contains the personal facts. Forget identities $\mathcal{F}$ and retain identities $\mathcal{R}$ are disjoint, and a fact is an identity, a relation, and its value. A *cue* is how the question identifies the person: a name (T) or an image (V). Perception questions $\mathcal{P}$ ask about visible properties, including those of forget images, and must remain answerable. For a language MLP at layer ℓ,

$$f_\ell(h) = W_{d,\ell}\,[\mathrm{SiLU}(W_{g,\ell}h) \odot (W_{u,\ell}h)] \tag{1}$$

(Shazeer, 2020; Bai et al., 2025). A *gate row* is one row of $W_{g,\ell}$; it computes the coefficient of one MLP channel from the residual state. We edit gate rows only and keep everything else fixed. Layers are numbered from 1.

Table 1: Frozen interventions on A's photograph (S-MLLMUn, Qwen2.5-VL-7B; 36 pairs, 108 facts). Injection at the output of layer 17; blocking removes attention to the visual span from layer 18 onward. "Leads" compares the two complete attribute values. Generation covers occupation only.

Condition	A leads /108	B leads /108	Gen. A /36	Gen. B /36
A's photograph	108	0	36	0
B's photograph (ceiling)	0	108	0	36
Block only	67	41	4	3
Inject B's name, no block	76	32	26	9
Block, inject A's name	100	8	34	1
Block, inject B's name	7	101	1	33
Block, noise of equal norm	66	42	5	3
Replace visual span with B's (no block)	5	103	2	33

4 How identity reaches attribute extraction

All experiments in this section use the frozen Original models. We describe them on the S-MLLMUn Original of Qwen2.5-VL-7B (28 layers) and repeat the layer screen on the MLLMU-Bench Original of LLaVA-1.5 (32 layers). Studies of factual recall in language models locate the assembly of subject information in middle layers (Meng et al., 2022; Geva et al., 2023); we therefore use a middle layer, layer 17 of 28, as the running example, and then repeat every intervention at all layers.

Face and name representations align in middle layers. For each identity, we take the residual state at the last name token of a neutral prompt ("Tell me about ⟨name⟩.") and the state following the person's photograph. Thirty-six identities are used for development and thirty-six others for validation. At layer 17, the photograph retrieves its own name among 36 candidates, and vice versa, with mean top-1 accuracy 54.2%, against 6.9% for the base model that never saw these identities (chance 2.8%). Restricting candidates to the same recorded gender gives 55.6% versus 9.7% (chance 5.6%); adding the same 20-year birth cohort gives 52.8% versus 13.9% (chance 11.1%). Alignment alone does not show that the model uses one state for both cues; the next experiment tests this.

The image keeps supplying identity. We select 36 pairs of identities (A, B) whose occupation, birthplace, and education both identities answer correctly under both cues, giving 108 facts. The input is always A's photograph. We inject B's neutral name state at the question end, at the output of layer 17. Optionally, all later positions are blocked from attending to the visual span from layer 18 onward. We score the attribute value after a shared answer prefix (for example, "This person was born in"), excluding the prefix; free generation uses no prefix.

Table 1 shows that the injected name is specific and sufficient once the image route is closed. Blocking alone leaves a reference level of 41/108, close to noise of equal norm (42/108). With blocking, B's name moves 101/108 facts to B, while A's own name keeps 100/108 with A. The switch holds for occupation, birthplace, and education (35/36, 34/36, 32/36), although the injected state was extracted without any attribute question; it therefore carries the person rather than one relation's answer. Free generation of occupation changes to B's value in 33/36 pairs. B's true value also gains 4.07 nat per token, so the effect is not only a loss for A. Without blocking, the same injection moves only 32/108 facts: later layers read A's identity from the image again.

A wide entity band and a late re-entry window. On the 36 pairs, blocked injection of B's name still moves 99/108 facts at layer 19 but only 29/108 at layer 23, and blocking image reading from layer 21 onward lowers A's true value by 1.47 nat per token, against 0.18 from layer 25.

We then repeated the interventions at every decoder layer of both models (18 pairs, 54 facts each; Appendix B), with the same three representative relations (a job, a place, an institution). We define the *entity band* as the span of layers where blocked injection moves at least 90% of the model's maximum number of facts, and the *re-entry window* as the layers where reopening the image recovers the true value by at least 0.20 nat per token. This gives layers 5–19 and 20, 22–24 for Qwen2.5-VL, and layers 5–16 and 11, 15–17 for LLaVA-1.5, whose window overlaps the end of its band.

Consequences for editing. Identity reaches attribute extraction twice: once through the entity state and again through the visual bypass in the re-entry window. An edit restricted to the entity band leaves the bypass open for image cues, and an edit of expression alone, such as refusal training in late layers, leaves both routes intact. Section 7 shows the behavioral counterpart: supervision on one cue barely transfers to the other. We therefore allow updates to gate rows in both regions. The bypass itself is read by attention, but not by a small set of heads: blocking the top heads of the re-entry window together reproduces only part of the effect of blocking the whole window (Appendix C). We therefore edit the gate rows that process what attention retrieves, not the heads.

5 ROUTEMASK: GATE UPDATES ALONG IDENTITY ROUTES

5.1 ELIGIBLE LAYERS AND ROW SCORES

The eligible layer set $\mathcal{L}$ contains the entity band and the re-entry window, determined for each model by the protocol of Section 4 before any mask is compared (Table 7).

Row scores are taken at the MLP intermediate activation $z_\ell = \mathrm{SiLU}(W_{g,\ell}h) \odot (W_{u,\ell}h)$ of Equation 1, whose channel c is switched on or off by gate row c. Although the image re-enters through attention, the channels read here are those of the MLP in the same layers, which processes what attention retrieves; Appendix C shows that the trained edit acts there. For an example x with reference answer y and answer NLL $\ell(x, y)$ averaged over tokens, the attribution of channel c is

$$a_{c,\ell}(x) = \Big[- \sum_{i \in \mathrm{prompt}(x)} z_{\ell,i,c} \, \frac{\partial \ell(x,y)}{\partial z_{\ell,i,c}} \Big]_+, \tag{2}$$

the first-order increase of the answer NLL if the channel were removed at the prompt positions, kept only where the channel supports the answer. For F and R, y is the true attribute value; for P it is Original's perception answer. The forget target "I don't know." is never used: its gradient marks rows that would write a refusal, not rows that use forget knowledge.

Fixed profiles and a forget score for each example. The retain and perception profiles $r_\ell = \mathbb{E}_{x \in \mathcal{C}_R} a_\ell(x)$ and $p_\ell = \mathbb{E}_{x \in \mathcal{C}_P} a_\ell(x)$ are computed once at Original, balanced over identities, relations, and cues, and kept fixed. The forget score is not averaged. When we estimated a single forget profile in advance from 50 forget examples, masks built from disjoint halves of the identities agreed poorly (Spearman 0.13–0.20 within a layer; 6–9% of their top-5% rows shared, against 5% by chance), and in Qwen the five largest examples carried 87–93% of the gradient energy. An average over identities therefore describes a few people rather than the one being forgotten. We instead attribute each forget example x of the current step at the current weights. After normalizing each profile by its mean over the rows of a layer,

$$\widetilde{v}_{c,\ell} = \frac{v_{c,\ell}}{d_{\mathrm{ff}}^{-1} \sum_{c'} v_{c',\ell}}, \qquad m_{c,\ell}(x) = \frac{\widetilde{a}_{c,\ell}(x)}{\widetilde{a}_{c,\ell}(x) + \widetilde{r}_{c,\ell} + \widetilde{p}_{c,\ell}}. \tag{3}$$

For the forget examples B_t of step t, we average $m_\ell(x)$ over $x \in B_t$ and divide by its mean over the rows of the layer, so that the average multiplier is one; the row mask is $M_\ell(B_t) = \bar{m}_\ell(B_t)\mathbf{1}^\top$ on eligible layers and zero elsewhere. The mask is continuous and not clipped. The attribution pass uses its own hooks, and its gradients are discarded before the training loss is computed.

5.2 MATCHING FORGET AND RETAIN TERMS ON EACH BEHAVIOR

For a complete answer y, $\ell_\theta(x, y)$ is its NLL averaged over tokens, including EOS, with prompt and image tokens excluded. Content scoring uses complete attribute values without option letters. For the candidate set of relation r, $A_r = \{a_1, \ldots, a_K\}$,

$$s_\theta(x, a) = \frac{1}{|a|} \log p_\theta(a \mid x), \quad q_\theta(j \mid x) = \frac{\exp(s_\theta(x, a_j))}{\sum_{k=1}^{K} \exp(s_\theta(x, a_k))}, \tag{4}$$

where a excludes EOS. q_θ is normalized over the candidate set only. Candidates, prompts, and option orders are fixed before training and use only training data.

Table 2: Objective terms by behavior. y_{IDK} = `I don't know.`; U_K is uniform over the K candidates; q_{θ_0} is Original's candidate distribution.

Behavior	Forget (F)	Retain mirror (R)
Say	$\ell_{F,\text{say}}(x) = \ell_\theta(x, y_{\text{IDK}})$	$L_{R,\text{say}} = \mathbb{E}_{D_R} \ell_\theta(x, y)$
Recognize	$\ell_{F,\text{rec}}(x) = \text{KL}(U_K \| q_\theta(\cdot \mid x))$	$L_{R,\text{rec}} = \mathbb{E}_{D_{R,\text{rec}}} \text{KL}(q_{\theta_0} \| q_\theta)$
Answer format	–	$L_{R,\text{fmt}}$: Original's distribution over answers made of a letter and EOS, for rotated retain options
Perceive	L_P: token KL to Original on fixed perception answers, for F and R images	

Table 3: Internal controls (15% split, image cue; mean over three seeds). MLLMU: Cls in %, Fact 1–10. S-MLLMUn: Fact 0–10.

	MLLMU-Bench				S-MLLMUn		
Control	F Cls↓	R Cls↑	F Fact↓	R Fact↑	F mem.↓	R mem.↑	Visual↑
All-layer gate	40.5	35.4	2.92	4.43	4.46	5.74	6.77
Location-only	42.3	**47.6**	3.21	5.18	4.25	6.45	7.05
Step-matched uniform	44.2	46.2	3.52	5.31	4.92	**6.47**	7.07
Shuffled	44.6	45.8	3.71	5.24	5.06	6.43	7.03
RouteMask	**37.4**	45.1	**1.16**	**5.62**	**3.90**	6.38	**7.16**

Each forget term is paired with a retain mirror on the same behavior (Table 2); Section 7 tests each component on the unmasked all-layer baseline. A forget term alone can be satisfied by a shortcut that also applies to retain inputs, such as refusing every portrait question or flattening every candidate list; the mirror penalizes that shortcut.

$L_{R,\text{fmt}}$ uses option letters only to protect the multiple-choice answer format; letters are never a forget signal. L_P uses the top-k tokens and one tail bucket (Appendix A). For a forget example, let $z_x = 1$ if at least two candidates are eligible. Its forget loss is

$$L_F = \mathbb{E}_{x \in D_F} \left[(1 - z_x/2)\, \ell_{F,\text{say}}(x) + (z_x/2)\, \ell_{F,\text{rec}}(x) \right], \tag{5}$$

so an example without eligible candidates receives the say term with weight one. The total loss is

$$L = L_F + \lambda_R \left(L_{R,\text{say}} + L_{R,\text{rec}} + L_{R,\text{fmt}} \right) + \lambda_P L_P, \tag{6}$$

where λ_R weights the retain mirrors and λ_P the perception term (values in Table 7). All arms of a comparison share losses, examples, teacher trajectories, and coefficients. All retain references come from Original.

5.3 Masking the accepted step

At step t, AdamW proposes a step $u_{t,\ell}$ after clipping and moment normalization. We accept

$$W_{g,\ell}^{t+1} = W_{g,\ell}^{t} + M_\ell(B_t) \odot u_{t,\ell}. \tag{7}$$

The optimizer moments keep their unmasked values. Masking gradients instead would be largely undone by Adam's normalization, and optimizing a free parameter inside $W_g^0 + M \odot \Delta W_g$ would change the moment dynamics. Weight decay is zero, and all other parameters stay fixed.

The deployed gate is the trained gate, $W_{g,\ell}^{*} = W_{g,\ell}^{0} + \sum_t M_\ell(B_t) \odot u_{t,\ell}$. The mask is used only during training; the deployed model has the original architecture with edited gate weights.

6 Experiments

6.1 Setup

We pair S-MLLMUn with its Qwen2.5-VL Original and MLLMU-Bench with its LLaVA-1.5 Original (Zeng et al., 2026; Liu et al., 2025a; Bai et al., 2025; Liu et al., 2024); every arm starts from the

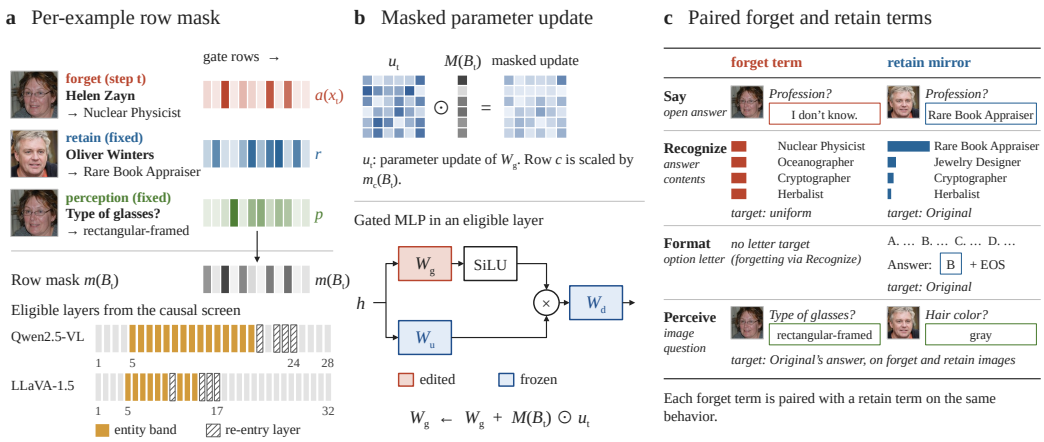

Figure 2: **RouteMask.** (a) At every step, the true answers of the step's forget examples are attributed to gate rows and compared with fixed retain and perception profiles (shown: S-MLLMUn samples); the forget shares of the examples are averaged into one row mask per layer, used only in the eligible layers, the entity band plus the re-entry layers found by the all-layer screen of each model (Section 4). Cell shades are illustrative. (b) The mask scales the parameter update of W_g; W_u and W_d stay frozen, and the deployed gate still depends on the input. (c) Each forget term has a retain mirror on the same behavior (Table 2); candidate answers are professions from other S-MLLMUn profiles, and letters serve only the retain format term.

Table 4: MLLMU-Bench, 15% forget split, LLaVA-1.5-7B. Cls in %, Fact 1–10; mean over three seeds. Best non-Original value in bold.

	Forget ↓				Retain ↑			
	Cls		Fact		Cls		Fact	
Method	T	V	T	V	T	V	T	V
Original	41.4	46.9	2.72	8.34	40.6	47.2	3.14	7.17
GA-Diff	36.0	39.6	1.26	4.12	34.7	41.6	1.68	4.31
NPO + retain	36.9	41.4	1.40	4.57	36.3	41.2	1.75	4.18
SMFA	37.8	42.3	1.38	3.12	**39.3**	45.2	2.71	5.48
MUDMAN	36.2	43.9	1.72	5.17	38.7	**46.8**	2.62	**5.67**
MANU	32.4	39.0	1.34	4.33	37.2	41.7	1.99	4.53
All-layer gate	35.4	40.5	1.66	2.92	35.5	35.4	2.54	4.43
RouteMask	**32.1**	**37.4**	**1.02**	**1.16**	39.2	45.1	**2.86**	5.62

Table 5: S-MLLMUn, 15% forget split, Qwen2.5-VL-7B; full evaluation set; Fact 0–10; mean over three seeds. *GA-Diff collapses to degenerate outputs and is not ranked.

	F memory ↓		R memory ↑		Visual ↑	
Method	T	V	T	V	F	R
Original	8.53	6.54	8.73	6.68	7.24	7.52
GA-Diff*	0.03	0.02	0.02	0.02	0.03	0.03
NPO + retain	4.68	**3.32**	5.29	4.32	6.51	6.89
SMFA	3.02	5.85	6.81	5.48	6.02	6.56
MUDMAN	4.96	3.60	7.13	5.15	6.78	7.26
MANU	6.43	4.91	3.84	3.55	6.47	6.06
All-layer gate	3.58	4.46	7.35	5.74	6.57	6.97
RouteMask	**2.97**	3.90	**7.99**	**6.38**	**6.87**	**7.44**

corresponding Original. On MLLMU we report classification accuracy with the official prompt and parser and the paper's 1–10 Fact rubric with Qwen-Plus as a substitute judge, separately for F/R and T/V. On S-MLLMUn we report the benchmark's scores for memory and visual understanding on the full evaluation set.

Training uses AdamW with learning rate 2×10^{-5}, no weight decay, and gradient clipping at one. Each step contains four forget questions, whose masks are computed at that step, together with retain questions, retain candidate sets, and perception questions; the schedule covers every forget identity and both cues, and the trained weights are deployed without interpolation. All experiments use the 15% forget split; Appendix H repeats the main comparison on the 5% and 10% splits. Main comparisons are averaged over three training seeds; ablations use one seed. Appendix A lists the remaining settings.

6.2 BASELINES AND CONTROLS

External baselines are GA with retain learning (GA-Diff), NPO with retain learning, SMFA, MUDMAN, and MANU, run with official code where it supports the benchmark (Zhang et al., 2024; Zeng et al., 2026; Sondej et al., 2025; Liu et al., 2025b). Internal controls keep the objective, data, and schedule of ROUTEMASK and change only which rows move. **All-layer gate** updates every gate row in every layer without a mask; **location-only** updates every gate row in the eligible layers without a mask; **step-matched uniform** gives every row of an eligible layer the same multiple of its step, so that the norm of each layer's update equals that of ROUTEMASK; **shuffled** permutes the mask of each step within each layer and rescales it to the same norm. Location-only isolates the layer choice; step-matched uniform and shuffled test row selection at equal step size.

6.3 MAIN RESULTS

Tables 4 and 5 compare ROUTEMASK with the external baselines. We see three patterns. First, GA-Diff collapses S-MLLMUn to near-zero scores on every split, and NPO and MANU lower retain classification and factuality almost as much as they lower forget scores. Second, SMFA forgets well under the name cue on S-MLLMUn and lowers factuality on MLLMU-Bench, but it leaves much of the memory under the image cue on S-MLLMUn and changes classification only moderately; supervision that mostly reaches the entity route leaves the visual bypass open. Third, ROUTEMASK gives the lowest forget factuality on MLLMU-Bench under both cues, with classification forgetting competitive with MANU at a smaller retain cost. On S-MLLMUn it improves on SMFA in all six scores: it forgets more under both cues while keeping more retain memory and visual understanding. NPO and MUDMAN forget more under the image cue on S-MLLMUn, but at a much larger cost to retain memory, and on MLLMU-Bench MUDMAN keeps more retain factuality while forgetting less.

6.4 DOES ROW SELECTION MATTER?

Table 3 compares ROUTEMASK with the internal controls on the image cue. Restricting updates to the eligible layers already keeps much more retain knowledge than the all-layer gate at similar forgetting. At the same update norm, step-matched uniform and shuffled updates forget much less than ROUTEMASK with similar retain scores, so a smaller step does not explain the difference. Shuffled keeps the distribution of multipliers in each layer but assigns them to the wrong rows, and it behaves like uniform: what matters is which rows receive the large multipliers, not how peaked the mask is.

7 ABLATIONS

We test the three design choices of ROUTEMASK one at a time: supervising both cues, pairing each forget term with a retain mirror, and restricting which parameters move. To see each choice on its own, the ablations use the unmasked all-layer gate and a single seed. All experiments use the 15% forget split. On MLLMU-Bench we score 2,040 questions (360 classification and 150 generation items for each of F, R, and each cue) with the official classification prompt and the benchmark's Fact rubric, judged by Qwen-Plus. We leave out Fact under the name cue on MLLMU-Bench, where Original already scores low. "Kept" counts the image questions that Original answered correctly and that are still answered correctly.

7.1 WHICH CUES MUST BE SUPERVISED?

If the two cues reach attribute extraction through partly separate routes (Section 4), forgetting learned under one cue should not carry over to the other. We train the refusal objective on S-MLLMUn with forget questions posed by name only, by image only, or by both, and keep everything else fixed (Figure 1c; Appendix E).

Forgetting stays with the cue it was trained on. Training on names removes memory under the name cue but leaves most of the memory under the image cue, and training on images does the reverse.

Table 6: Objective terms (MLLMU-Bench, LLaVA-1.5; all 32 gate matrices, unmasked). Cls in %; Fact 1–10 on generation under the image cue; kept out of 169 (F) and 170 (R). Pair 1: a short schedule with forget items on name and birth date, identical retain mirrors and perception term; only the form of the recognize term differs. Pair 2: a longer schedule; say only trains refusal on forget items with image cues with a retain answer loss, and the full objective adds the recognize term and all retain mirrors. Removals: one term set to zero in the full objective; Pair 1 already removes the say term. α interpolates the weight change toward Original. Further runs are in Appendix G.

		F Cls ↓		R Cls ↑		F Fact	R Fact	Kept	
	Objective	T	V	T	V	↓	↑	F ↓	R ↑
	Original	41.4	46.9	40.6	47.2	8.34	7.17	169	170
Pair 1	Recognize on option letters	34.7	30.3	33.3	26.7	4.73	4.60	85	62
	Recognize on answer contents	39.7	40.8	41.1	44.2	4.97	5.19	131	125
Pair 2	Say only, α=1	40.3	46.9	44.2	50.6	1.20	4.31	149	168
	Full objective, α=1	36.4	40.3	37.2	34.4	3.05	4.28	117	96
	Say only, α=0.5	40.6	44.2	45.0	50.3	4.92	6.01	159	164
	Full objective, α=0.5	35.6	43.6	36.4	43.1	4.82	5.27	137	137
Removal	− R recognize mirror	36.7	38.9	32.8	30.6	2.93	4.19	108	79
	− perception term	37.5	40.0	35.3	33.9	2.84	3.90	115	90

We therefore apply every forget and retain term under both cues. Training on images alone also hurts visual understanding the most, which is why the objective includes a perception term.

7.2 Which objective terms are needed?

The objective combines a say term, a recognize term, and retain mirrors. Table 6 reports two paired comparisons and the removal of single terms from the full objective. Pair 2 differs in more than one term, so it compares the full objective with plain refusal rather than isolating one factor.

Recognize answers, not letters. When the recognize term flattens the distribution over option letters (Pair 1), the model finds a shortcut: it gives the same letter to every question asked with the photograph, which costs more retain choices than forget choices. Flattening the answer contents removes this shortcut. A ranking test without letters shows that it pushes the true answer down much more for forget facts than for retain facts, mostly on the relations seen in training.

Refusal alone leaves classification intact. In Pair 2, refusal brings Fact under the image cue close to its floor, yet almost every originally correct forget choice is still chosen. Adding the recognize term and the mirrors finally moves forget classification, but retain classification drops with it, since the unmasked update also changes rows that retain identities use. The objective affects both open answers and multiple choice, but it does not decide by itself which parameters should change; that is the role of the mask. The removals confirm that each mirror protects its own behavior: without the retain recognize mirror, retain classification and kept choices drop while forgetting barely changes, and without the perception term retain factuality falls.

Update location. With the same objective and parameter count, down projections give a similar trade-off to gate projections on MLLMU-Bench but damage visual understanding far more on S-MLLMUn (Appendix F).

8 Conclusion

Identity reaches attribute extraction in a VLM twice: through an entity state shared by names and faces, and through a visual bypass that re-reads the image in late layers. RouteMask restricts unlearning updates to the gate rows at these locations, scales them with a row mask recomputed for each forget example, and pairs every forget term with a retain mirror. At matched step size, the mask recomputed for each example forgets more than uniform or shuffled allocations with a similar retain cost. Appendix I discusses limitations.

AI USE STATEMENT

We used generative AI tools (Claude, Anthropic) for the following tasks: implementing and debugging experiment code, including the training, evaluation, and intervention scripts; drafting and editing the text of the paper; searching for, summarizing, and formatting related literature; and giving feedback on experimental design and on the interpretation of results. The research questions, the method, and the final experimental choices were decided by the authors. We have reviewed all AI-assisted work: every number in the paper comes from our own runs and was checked against the logged outputs, the code was run and tested by the authors, and every citation was checked against the original publication.

ETHICS STATEMENT

This work aims to remove personal information from trained VLMs. All experiments use public unlearning benchmarks (MLLMU-Bench and S-MLLMUn) whose profiles describe fictitious people constructed for unlearning research; no real person's data is used, we collect no new personal data, and no human subjects are involved. The localization in Section 4 describes how identity reaches a model's answers and could also help an attacker target these routes; we report it because the same routes must be known to remove the information. Unlearning as evaluated here is behavioral: it does not guarantee that the information cannot be recovered, for example by relearning or quantization (Appendix I), and it does not replace careful data selection before training.

REPRODUCIBILITY STATEMENT

Section 5 and Appendix A give the objective, the row mask, the eligible layers, and all training settings (Table 7). Appendix B describes the all-layer screen used to select the layers, and Section 6 lists the benchmarks, splits, evaluation prompts, and judge. Main results are averaged over three training seeds.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923v1. Version 1.

Samyadeep Basu et al. Understanding information storage and transfer in multi-modal large language models. In *Advances in Neural Information Processing Systems*, 2024.

Lucas Bourtoule et al. Machine unlearning. In *IEEE Symposium on Security and Privacy*, 2021.

Trishna Chakraborty et al. Can textual unlearning solve cross-modality safety alignment? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Damai Dai et al. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Chenlu Ding, Jiancan Wu, Leheng Sheng, Fan Zhang, Yancheng Yuan, Xiang Wang, and Xiangnan He. MLLMEraser: Achieving test-time unlearning in multimodal large language models through activation steering, 2026. URL https://arxiv.org/abs/2510.04217v3. Version 3, 2 February 2026; first posted in 2025.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. SalUn: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.12508v5. Spotlight; preprint first posted in 2023.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. In *Advances in Neural Information Processing Systems*, 2023.

Jiahao Huo et al. MMUnlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Joel Jang et al. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Jinghan Jia, Jiancheng Liu, Yihua Zhang, Parikshit Ram, Nathalie Baracaldo, and Sijia Liu. WAGLE: Strategic weight attribution for effective and modular unlearning in large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1767. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/649ad92e7067b3553a0f15acac68806d-Abstract-Conference.html.

Wonjun Lee, Jaehyuk Jang, Kangwook Ko, Hee-Seon Kim, and Changick Kim. AIM: Anchor identity features, then match for multimodal large language model unlearning, 2026. URL https://arxiv.org/abs/2608.28312v1. Version 1; accepted to Findings of EMNLP 2026.

Jiaqi Li et al. Single image unlearning: Efficient machine unlearning in multimodal large language models. In *Advances in Neural Information Processing Systems*, 2024a.

Nathaniel Li et al. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning*, 2024b.

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. URL https://arxiv.org/abs/2310.03744v2. Highlight; preprint: arXiv:2310.03744v2.

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/.

Jakub Łucki et al. An adversarial perspective on machine unlearning for AI safety, 2024. URL https://arxiv.org/abs/2409.18025.

Yingzi Ma et al. Benchmarking vision language model unlearning via fictitious facial identity dataset. In *International Conference on Learning Representations*, 2025.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. TOFU: A task of fictitious unlearning for LLMs, 2024. URL https://arxiv.org/abs/2401.06121v1. Version 1.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2202.05262v5.

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.07229v2. Preprint first posted in 2022.

Clement Neo et al. Towards interpreting visual information processing in vision-language models. In *International Conference on Learning Representations*, 2025.

Noam Shazeer. GLU variants improve transformer, 2020. URL https://arxiv.org/abs/2002.05202v1. Version 1.

Filip Sondej, Yushi Yang, Mikołaj Kniejski, and Marcel Windys. Robust LLM unlearning with MUDMAN: Meta-unlearning with disruption masking and normalization, 2025. URL https://arxiv.org/abs/2506.12484v5. Version 5.

Guangzhi Sun, Potsawee Manakul, Xiao Zhan, and Mark Gales. Unlearning vs. obfuscation: Are we truly removing knowledge? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11457–11467. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.577. URL https://aclanthology.org/2025.emnlp-main.577/.

Xiaotian Ye, Mengqi Zhang, and Shu Wu. LLM unlearning should be form-independent. In *IEEE Symposium on Security and Privacy*, 2026. URL https://arxiv.org/abs/2506.07795v1. Preprint: arXiv:2506.07795, version 1, 2025.

Zhen Zeng, Leijiang Gu, Zhangling Duan, Feng Li, Cees G. M. Snoek, Meng Wang, and Zenglin Shi. Towards benign memory forgetting for selective multimodal large language model unlearning, 2026. URL https://arxiv.org/abs/2511.20196v2. Version 2, 26 June 2026; first posted in 2025; accepted to ECCV 2026.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning, 2024. URL https://arxiv.org/abs/2404.05868v2. Version 2.

Zhiwei Zhang et al. Catastrophic failure of LLM unlearning via quantization. In *International Conference on Learning Representations*, 2025.

A Implementation details

Table 7: Training settings.

	S-MLLMUn / Qwen2.5-VL-7B	MLLMU-Bench / LLaVA-1.5-7B
Forget split	15% (official)	15% (official)
Eligible layers (1-based)	5–20, 22–24	5–17
Learning rate	2×10^{-5}	2×10^{-5}
AdamW $(\beta_1, \beta_2, \epsilon)$; weight decay; clipping	$(0.9, 0.999, 10^{-8})$; 0; 1	$(0.9, 0.999, 10^{-8})$; 0; 1
Loss weights (λ_R, λ_P)	$(0.8,\ 0.2)$	$(0.8,\ 0.2)$
Forget questions per step	4 (2 name, 2 image)	4 (2 name, 2 image)
Training seeds	42, 43, 44	42, 43, 44
Judge	Qwen-Plus, benchmark prompt	Qwen-Plus, rubric of Liu et al. (2025a)

Layer selection. The eligible layers are the entity band and the re-entry layers of the all-layer screen in Section 4. The screen used occupation, birthplace, and education questions for both models, chosen as representatives of three answer types (a job, a place, and an institution).

Profiles and mask. The retain and perception profiles are computed once at Original with dropout disabled: for each example we backpropagate the mean token NLL of its reference answer (the true value for R, Original's answer for P) and accumulate Eq. 2 in FP32, averaging within relation and identity and then across cues. At each training step, every forget question of the step is attributed in the same way at the current weights; its attribution gradients are cleared before the training losses are computed, so they do not enter the update.

Masked update. At each step we take an ordinary AdamW step, compute the proposed change u_t of every eligible gate matrix, and replace it by $M(B_t) \odot u_t$; the optimizer moments keep their unmasked values. Because the averaged mask is normalized to mean one in each layer, rows with a high forget share can accept more than their proposed step.

Perception teacher. L_P is a KL divergence to Original on fixed perception answers, computed over Original's top-64 tokens plus one bucket for the remaining vocabulary. The teacher distribution is cached once and shared by all arms.

Controls. Each control follows its own trajectory and computes its own AdamW proposal $u_{t,\ell}$; its accepted norm in each layer is set to the accepted norm recorded for RouteMask at the same step and seed. Step-matched uniform scales every row of layer ℓ at step t by $\kappa_{t,\ell} = a_{t,\ell}/\|u_{t,\ell}\|_2$, where $a_{t,\ell}$ is the accepted norm of RouteMask; κ may exceed one. Shuffled applies a fixed permutation within each layer to $\bar{m}(B_t)$ and rescales to the same norm.

Uncertainty. Intervals are percentile 95% intervals from 10,000 paired bootstrap draws over identities, keeping all questions and cues of an identity together.

B All-layer causal screen

We ran the interventions of Section 4 at every decoder layer of both models, on 18 pairs and three attributes (54 facts) per model. For each layer we record the facts moved to B by blocked injection, and we call a layer a *re-entry layer* if, under a rule fixed before the curves were inspected, blocking the image at that layer alone, or reopening it in a cumulative cut, changes A's log probability of the true value by at least 0.20 nat per token in at least two thirds of the pairs. In Qwen2.5-VL, blocked injection moves at least 47/54 facts at every layer from 4 to 19 and at least 52/54 at layers 7–15, then drops to 43, 40, 28, and 11 at layers 20–23 and to 0 from layer 24. The re-entry layers are 20, 22, 23, and 24 (reopening recovers 0.25–0.47 nat per token); with the name as the source, they are 14, 20, 22, and 23. In LLaVA-1.5, blocked injection moves at most 35/54 facts, with a plateau at layers

5–16 and a drop to 17 at layer 17 and to 0 from layer 20; the re-entry layers are 11, 15, 16, and 17 (0.21–0.74 nat per token). The name route of LLaVA is not identifiable on this panel, because no pair answers both identities from the name. We define the *entity band* as the span of layers where blocked injection moves 90% of the model's maximum (Qwen 5–19, LLaVA 5–16) and the *re-entry window* as these re-entry layers. In Qwen the window lies after the band; in LLaVA it overlaps the end of the band. Eighteen pairs give a coarse estimate, the band is a range rather than a unique key, and all re-entry layers were found through the conditional recovery criterion, not because any single layer is necessary.

C ATTENTION HEADS AND VISUAL READBACK

Head screen. In the re-entry layers of both models we blocked the image contribution of single heads and ranked them by how much blocking moves the answer toward the injected identity. In Qwen2.5-VL, blocking the top 14 or 28 heads together moves 46 of 54 facts in ranking but changes only 7 of 18 occupation answers, whereas blocking all heads of the window moves 54 of 54 and 18 of 18. In LLaVA-1.5, blocking the top 14 heads leaves all 108 facts with the photographed identity. The bypass is therefore spread over many heads, which is why ROUTEMASK edits the MLP gate rows that process the retrieved information rather than a head set.

Readback in an edited model. To test whether an edited model forgets by changing what attention retrieves from the image, we took a ROUTEMASK checkpoint of Qwen2.5-VL and, at the input of the attention output projection, replaced the current contribution of image tokens by the Original contribution, either in layers 18–24 or in all layers. On 32 fixed portrait questions, this visual readback leaves the edited behavior in place: the log probability of "I don't know." on forget questions stays near that of the edited model (-0.09 versus -9.43 for Original). Restoring the Original MLP outputs at the same positions instead returns it to -8.90. The edit thus acts in the MLP computation downstream of the visual read, not by changing which image content attention retrieves. This small panel does not exclude other routes through attention.

D MASK DIAGNOSTICS

H1: for a given forget example, the eligible layers contain gate rows whose attribution is high for that example and low under the retain and perception profiles. During training we log, for each step, the mean and spread of $m(x)$, the share of rows above 0.5, and the rank correlation between $\widetilde{a}(x)$ and $\widetilde{r}$; we also compare the top rows of examples from the same identity with those of different identities. If $\widetilde{a}(x)$ tracks $\widetilde{r}$, the mask cannot be selective, and the masked arm reduces to reweighted uniform steps.

Each row's accepted step is $\bar{m}_c(B_t)\, u_c$; a layer mean of one does not fix the layer's accepted norm, so the replay controls match that norm directly. A small multiplier does not bound any person's output change: small steps accumulate, and a row with $\bar{m}_c < 1$ can still affect many inputs. Equation 7 tests whether allocating of a finite step budget improves the forget–retain trade-off. Only a comparison with uniform steps of the same actual size can answer that.

E SUPERVISION CUE

Table 8 gives the full ablation of the supervised cue summarized in Section 7 and Figure 1c. We train the refusal objective on S-MLLMUn with forget questions posed by name only, by image only, or by both, keeping the retain supervision, the schedule, and the updated parameters fixed. Supervising one cue removes memory under that cue but leaves most of it under the other, at both scales. Supervision on images only also costs the most visual understanding, on forget and retain images alike, which motivates the perception term of Section 5.

Table 8: Supervision cue (S-MLLMUn, Qwen2.5-VL; all 28 gate matrices, unmasked; refusal objective; Fact 0–10 on 300 forget, 300 retain, and 300 visual questions; 15% split). Only the cue of the forget items changes.

		F memory ↓		R memory ↑		Visual ↑	
Forget cue	α	T	V	T	V	F	R
Original	–	8.65	5.18	8.42	6.49	6.71	8.32
Both	1.0	0.00	0.24	5.38	4.27	3.63	6.99
Name only	1.0	0.00	4.62	7.12	5.59	5.73	8.35
Image only	1.0	8.14	0.21	8.37	5.00	2.39	6.82
Both	0.5	3.47	3.32	8.12	6.42	6.37	8.61
Name only	0.5	3.33	5.01	8.29	6.19	6.67	8.26
Image only	0.5	9.02	1.48	8.32	6.34	6.07	8.61

F UPDATE LOCATION

On MLLMU-Bench, Table 9 compares updating native gate projections with updating the down projections of the same layers, with identical objective, schedule, and parameter count.

Table 9: Update location (MLLMU-Bench, LLaVA-1.5; refusal objective, 104 steps, all 32 layers, same parameter count). Columns as in Table 6.

		F Cls ↓		R Cls ↑		F Fact	R Fact	Kept	
Update	α	T	V	T	V	↓	↑	F ↓	R ↑
Original	–	41.4	46.9	40.6	47.2	8.34	7.17	169	170
Gate	1.0	42.2	45.0	41.7	48.6	1.17	4.18	149	165
Down	1.0	41.1	42.8	43.6	45.6	1.70	5.32	141	155
Gate	0.5	41.7	46.4	43.6	50.0	4.77	6.02	160	163
Down	0.5	40.0	43.1	45.3	48.6	4.74	5.75	150	160

With the refusal objective, the two locations give similar trade-offs on MLLMU: down projections lower forget classification slightly more at full scale and keep slightly fewer retain choices. On S-MLLMUn, updates of the down projection damage visual understanding far more than gate updates at full scale (Table 10), which is why the method edits gate rows. Neither location separates forget from retain choices under this objective; row selection must supply that separation if it exists.

On S-MLLMUn, Table 10 varies the updated matrices and layers. All rows start from S-MLLMUn's Qwen2.5-VL Original, use the refusal objective with retain answers and the same 900 development questions of the 15% split (one seed). "Anchor" adds a behavior anchor that describes portraits to the retain side. Gate and down variants update the same number of parameters. The last-14 rows are retrained from Original on layers 15–28 only.

Table 10: Ablations of the update location on S-MLLMUn (Fact 0–10).

		F memory ↓		R memory ↑		Visual ↑	
Update	Scale	T	V	T	V	F	R
Original	–	8.65	5.18	8.42	6.49	6.71	8.32
Gate, anchor	1.0	0.00	0.04	5.59	5.90	4.65	8.53
Gate, anchor	0.5	4.06	3.04	8.26	6.62	6.13	8.39
Down, anchor	1.0	0.00	0.21	5.05	3.93	2.31	4.54
Down, anchor	0.5	3.44	2.62	8.71	6.51	5.92	7.71
Gate, no anchor	1.0	0.00	0.24	5.28	4.25	3.51	7.04
Gate, no anchor	0.5	3.35	3.38	8.28	6.57	6.57	7.98
Gate, last 14 layers	1.0	0.00	0.63	6.65	4.65	2.88	7.72
Gate, last 14 layers	0.5	4.66	3.25	8.62	6.22	6.49	8.44

At full scale, updates of the down projection damage visual understanding far more than gate updates (retain visual 4.54 versus 8.53 with the anchor). The anchor protects retain memory under the image cue at full scale (5.90 versus 4.25) but not at half scale. Training only the last 14 layers gives similar forgetting under the image cue at half scale (3.25 versus 3.38) but weaker forgetting under the name cue (4.66 versus 3.35). These runs fix two choices of ROUTEMASK: it edits gate rather than down projections, and its eligible layers include the entity band rather than only the late layers. The perception term of Section 5 replaces the portrait anchor, which protects retain memory only at full scale.

G ADDITIONAL OBJECTIVE RUNS ON MLLMU-BENCH

Table 11 adds NPO under the same schedule as the refusal run with the say term only of Table 6, and interpolations of the Pair 1 run that recognizes answer contents of Pair 1.

Table 11: Additional unmasked runs on all gates on MLLMU-Bench (LLaVA-1.5, one seed). Columns as in Table 6.

	F Cls ↓		R Cls ↑		F Fact	R Fact	Kept	
Run	T	V	T	V	↓	↑	F ↓	R ↑
Original	41.4	46.9	40.6	47.2	8.34	7.17	169	170
NPO objective on gate rows, α=1	39.4	41.9	45.0	44.2	2.33	2.81	145	137
NPO objective on gate rows, α=0.5	38.6	44.4	43.6	47.2	3.88	5.26	157	168
Content recognize, α=0.75	38.9	43.3	39.4	48.9	5.70	5.34	141	151
Content recognize, α=0.5	40.6	42.5	42.8	49.2	6.83	6.50	140	154

The NPO objective, applied to the same gate rows, behaves like refusal: it lowers Fact under the image cue far more than classification, and at full scale it lowers forget and retain choices equally (145 of 169 and 137 of 170 kept).

H OTHER FORGET SPLITS

Both benchmarks define 5% and 10% forget splits in addition to the 15% split used in the main text. Each split starts from the same Original and uses the schedule of the main experiments. Across the 5% and 10% splits, ROUTEMASK keeps the lowest forget factuality on MLLMU-Bench and remains competitive with MANU in classification; on S-MLLMUn it improves on SMFA in every score on every split.

Table 12: MLLMU-Bench with 5% and 10% forget splits (image cue; Cls in %, Fact 1–10).

	5%				10%			
Method	F Cls↓	F Fact↓	R Cls↑	R Fact↑	F Cls↓	F Fact↓	R Cls↑	R Fact↑
Original	46.7	8.32	46.7	7.12	47.1	8.28	46.7	7.17
GA-Diff	39.3	3.53	41.6	4.62	39.6	3.65	41.2	4.47
NPO + retain	41.2	4.59	43.0	4.85	41.4	4.39	42.1	4.57
SMFA	41.9	3.00	**45.4**	5.68	41.3	3.01	45.0	5.67
MUDMAN	43.0	4.72	44.9	**5.94**	44.0	4.73	**45.3**	**5.83**
MANU	38.0	4.10	43.5	4.98	37.7	4.24	43.3	4.77
All-layer gate	39.8	2.92	34.2	4.31	40.3	2.97	35.1	4.39
ROUTEMASK	**36.6**	**1.13**	44.8	5.76	**36.3**	**1.13**	44.8	5.80

Table 13: S-MLLMUn with 5% and 10% forget splits (full evaluation sets; Fact 0–10; *GA-Diff collapses and is not ranked; Visual is the mean over forget and retain images).

	5%					10%				
	F mem.↓		R mem.↑		Vis.↑	F mem.↓		R mem.↑		Vis.↑
Method	T	V	T	V		T	V	T	V	
Original	8.40	6.01	8.70	6.77	7.58	8.77	6.72	8.64	6.86	7.44
GA-Diff*	0.03	0.02	0.03	0.02	0.03	0.03	0.02	0.03	0.02	0.03
NPO + retain	4.38	**2.97**	6.13	4.92	7.02	4.58	**3.43**	5.40	4.54	6.87
SMFA	2.92	4.91	7.56	5.58	6.74	3.03	5.79	7.24	5.52	6.55
MUDMAN	4.64	3.30	7.06	5.67	7.29	4.85	3.67	6.86	5.51	7.12
MANU	5.24	4.08	4.99	3.94	6.76	6.67	4.75	4.63	3.95	6.52
All-layer gate	3.51	4.19	7.01	5.76	6.75	4.02	4.70	7.29	5.68	6.69
RouteMask	**2.79**	3.41	**8.10**	**6.49**	**7.40**	**2.82**	3.76	**7.98**	**6.54**	**7.25**

I Limitations and scope

The name route of LLaVA-1.5 is not identifiable on the localization panel. Our evaluation is behavioral; beyond the readback test, we do not attack the edited models with relearning or quantization. The visual bypass is read by many attention heads rather than a small set (Appendix C); the method edits the MLP gate rows downstream of that read, and a readback test shows that its effect does not depend on changing the retrieved image content, but other attention routes are not excluded. The forget attribution is a first-order heuristic, and a high forget share does not show that a row stores a person's knowledge exclusively. Refusal training can spread to perception questions, factuality scored by a judge depends on the evaluator, and letter parsing can reward a biased label distribution. The study concerns behavioral unlearning of synthetic identities from given checkpoints; it provides no certified deletion, and real people with pretrained facts require separate evaluation.