Appendix : RAZOR: Ratio-Aware Layer Editing for Targeted Unlearning

Pseudo-code : RAZOR (Ratio-Aware Layer Editing for Targeted Unlearning)

Algorithm 1: RAZOR: Ratio-Aware Layer Editing for Targeted Unlearning

Input: Pretrained model f_θ with parameters θ and components $L = \{1, \ldots, |L|\}$
Forget set D_f, retain set D_r, validation split D_{val}
Ratio hyperparameter ρ; coefficients λ_f, λ_m
Orthogonality exponent α; stability constant ϵ
Initial saliency threshold τ_{init}; max iterations $T_{\max}$
Metric constraints Target (e.g., thresholds on M1–M5)
Output: Edited parameters θ^*

Stage 0: Baseline statistics for mismatch loss
$\theta^0 \leftarrow \theta$ `// Frozen copy of initial model`
Compute baseline signals (e.g., embeddings, logits) for L_{mismatch} using f_{θ^0} on D_f

Stage 1: One-shot gradients & ratio-aware saliency
Compute $g_f^\ell = \nabla_{\theta_\ell} L_{\text{forget}}(\theta; D_f)$ for all $\ell \in L$
Compute $g_r^\ell = \nabla_{\theta_\ell} L_{\text{retain}}(\theta; D_r)$ for all $\ell \in L$
foreach $\ell \in L$ **do**
  num $\leftarrow \|g_f^\ell\|_2^2$, denom $\leftarrow \|\theta_\ell\|_2^2 + \epsilon$
  $\phi(\ell) \leftarrow \left(\frac{\text{num}}{\text{denom}}\right) \left(1 - \cos(g_f^\ell, g_r^\ell)\right)^\alpha$
$\mathcal{K} \leftarrow \{\, \ell : \phi(\ell) > \tau_{\text{init}} \,\}$
if $\mathcal{K} = \emptyset$ **then**
  $\mathcal{K} \leftarrow \{\arg\max_\ell \phi(\ell)\}$ `// Ensure at least one component`

Stage 2: Update initially selected components
foreach $\ell \in \mathcal{K}$ **do**
  Compute $g_m^\ell = \nabla_{\theta_\ell} L_{\text{mismatch}}(\theta; D_f, \theta^0)$
  $g_{\text{RAZOR}}^\ell \leftarrow -\lambda_f \rho\, g_f^\ell + g_r^\ell + \lambda_m g_m^\ell$
  $\lambda_\ell \leftarrow \texttt{BinarySearchStep}(\theta, \ell, g_{\text{RAZOR}}^\ell, D_{\text{val}}, \text{Target})$
  $\theta_\ell \leftarrow \theta_\ell - \lambda_\ell g_{\text{RAZOR}}^\ell$

Stage 3: Iterative refinement of the active set
$t \leftarrow 1$
while $t \le T_{\max}$ **do**
  metrics $\leftarrow \texttt{EvaluateMetrics}(f_\theta, D_{\text{val}})$
  if *metrics satisfy Target* **then**
    break `// Desired forgetting/utility achieved`
  Recompute g_f^ℓ and g_r^ℓ for all $\ell \in L$
  foreach $\ell \in L$ **do**
    num $\leftarrow \|g_f^\ell\|_2^2$, denom $\leftarrow \|\theta_\ell\|_2^2 + \epsilon$
    $\phi(\ell) \leftarrow \left(\frac{\text{num}}{\text{denom}}\right) \left(1 - \cos(g_f^\ell, g_r^\ell)\right)^\alpha$
  $\ell^* \leftarrow \arg\max_{\ell \notin K} \phi(\ell)$
  if $\phi(\ell^*) \le 0$ **then**
    break `// No useful additional component to edit`
  $\mathcal{K} \leftarrow \mathcal{K} \cup \{\ell^*\}$
  Compute $g_m^{\ell^*}$ and $g_{\text{RAZOR}}^{\ell^*}$ as above
  $\lambda_{\ell^*} \leftarrow \texttt{BinarySearchStep}(\theta, \ell^*, g_{\text{RAZOR}}^{\ell^*}, D_{\text{val}}, \text{Target})$
  $\theta_{\ell^*} \leftarrow \theta_{\ell^*} - \lambda_{\ell^*} g_{\text{RAZOR}}^{\ell^*}$
  $t \leftarrow t + 1$
return $\theta^* \leftarrow \theta$

Binary search for per-layer step size λ_l

Input: Current parameters θ; selected component index l; blended gradient g^l_{RAZOR}
Validation set D_{val}; target metric constraints TARGET
Output: Layer-local step size λ_l

$\lambda_{\min} \leftarrow 0, \quad \lambda_{\max} \leftarrow \lambda_{\text{init}}$
$\lambda_{\text{best}} \leftarrow 0, \quad s_{\text{best}} \leftarrow -\infty$
while $\lambda_{\max} - \lambda_{\min} > \delta$ **do**
  $\lambda_{\text{mid}} \leftarrow (\lambda_{\min} + \lambda_{\max})/2$
  `// Propose a temporary RAZOR update on component` l
  $\theta^{\text{temp}} \leftarrow \theta$
  $\theta^{\text{temp}}_l \leftarrow \theta_l - \lambda_{\text{mid}}\, g^l_{\text{RAZOR}}$
  `// Evaluate forgetting/retention trade-off under the proposal`
  metrics $\leftarrow$ `EvaluateMetrics`$(f_{\theta^{\text{temp}}}, D_{\text{val}})$
  $s \leftarrow$ `ScoreMetrics`(metrics, TARGET)
  if *model is stable and meets basic constraints under* θ^{temp} **then**
    if $s > s_{best}$ **then**
      $s_{\text{best}} \leftarrow s$
      $\lambda_{\text{best}} \leftarrow \lambda_{\text{mid}}$
    $\lambda_{\min} \leftarrow \lambda_{\text{mid}}$ `// Safe to try a larger step`
  else
    $\lambda_{\max} \leftarrow \lambda_{\text{mid}}$ `// Step too large; shrink the interval`
return λ_{best}

A. Detailed Related Work

This section is an expanded version of the Related Work in Section 2, offering deeper technical context, additional citations, and broader coverage of prior unlearning methods across CLIP, diffusion, and VLM architectures.

Machine unlearning [2, 4, 13, 46, 58] addresses the selective removal of data influence from trained models, a critical need driven by privacy concerns and regulatory requirements [5]. Existing approaches mainly focus on a single task, like image classification [21, 28, 32, 57, 62, 68], image generation [17, 32, 41, 65, 67, 68, 73, 75], and LLMs text generation [70]. In this work, we propose a generic approach that is applicable to a wide range of multimodal models, including CLIP [50] for zero-shot image classification, stable diffusion models [54] for text-to-image generation, and vision language models [35] for visual question answering.

Exact vs. approximate unlearning. The gold standard, exact removal (retraining on retain set), offers provable guarantees [25], but is computationally infeasible for large-scale models [8, 24]. Research has thus shifted to *approximate unlearning* methods that modify parameters [4, 20]. SISA [2], a middle ground, trains on data shards for efficient partial retraining but requires architectural foresight and incurs storage overhead. We focus on *post hoc unlearning*: editing pre trained models without access to the original training pipeline [47, 64, 74]. Post hoc unlearning must identify which parameters to modify, as naïve full model updates risk catastrophic forgetting [64]. While early work used imprecise uniform perturbations [22], gradient-based saliency has emerged as the dominant localization paradigm. SALUN [11] ranks parameters and fine tunes only high saliency weights. SCISSORHANDS [65] improves this by computing gradients at initialization to reinitialize connections before recovery. Higher order methods, like Fisher-based Selective Synaptic Dampening [14], use curvature to preserve important weights, but computing the Fisher matrix scales poorly [42]. As a compromise, SLUG [3] trades completeness for efficiency by updating only a *single* high-impact layer. Additionally, G-Drift MIA [52] introduces a single-step unlearning mechanism for inference within large language models (LLMs). Similarly, the CatRAG framework [51, 53] refines the model's internal representations to achieve effective debiasing. These approaches highlight the core tension: localized edits are efficient but brittle, while exhaustive updates are robust but prohibitive.

Given *which parameters*, the challenge becomes *how to update them*, navigating the conflicting gradients from retain set and forget set [71]. Direct gradient ascent on the forget set [61] often destroys generalization. Gradient projection techniques [26, 48] mitigate this by constraining updates to subspaces orthogonal to the retained loss, preserving retain set performance, similar to methods in continual learning [12]. Representation space methods offer an alternative. Contrastive Unlearning

repels forget set embeddings from their class centroids [33]. Bad Teaching distills knowledge away from the forget set [35]. However, these parameter-level methods require full backpropagation and are unstable when the forget and retain sets have overlapping support. LoTUS smooths forget set predictions toward a uniform distribution [60]. This handles concept-level unlearning but struggles with instance-level deletion and often requires forget set labels [56].

Unlearning in Generative Models. These discriminative model approaches are ill-suited for generative models, where concepts are distributed compositionally [54], rendering classifier based localization ineffective [17]. Inference time interventions like Erased Stable Diffusion (ESD) [17] apply negative guidance, but this only *suppresses* expression, leaving knowledge recoverable via adversarial prompts [75]. This limitation motivates activation level surgery, such as using Sparse Autoencoders (SAEs) to find and ablate interpretable features (SAEUron [9]) or dynamically masking gradients (Fan et al. [11]). LLM unlearning faces analogous issues [1, 37, 38, 44], as creative prompting can often recover "forgotten" information [40].

Existing approaches reveal fundamental limitations that constrain practical unlearning. First, localization retention trade-offs remain unresolved: methods like SalUn [11] and SCISSORHANDS [65] identify salient parameters through gradient magnitude but lack explicit mechanisms to balance forgetting pressure against retention requirements, often leading to overly aggressive edits that degrade utility [29, 49]. Conversely, ultra conservative approaches, like SLUG [3], which confine edits to a single layer, achieve strong retention but struggle when knowledge is distributed across multiple architectural components. Second, gradient conflict management remains ad hoc: while projection methods [48] and contrastive techniques [23, 33] mitigate opposing gradients, they do so through post-hoc constraints rather than jointly optimizing for the forget-retain trade-off during parameter selection. This reactive approach cannot prevent conflicts that arise from poor localization choices. Third, architectural generalization is limited: most methods are designed for specific classifier model families, diffusion models, or language models, and require substantial re-engineering to transfer across domains. Methods effective for discriminative models often fail for generative architectures due to their fundamentally different knowledge encoding mechanisms [17, 69].

We introduce **RAZOR**, a ratio-aware framework that unifies parameter localization and update computation. RAZOR scores layers and heads by a forget-to-retain gradient ratio, quantifying the forget-pressure vs. retention-alignment trade-off. This enables principled multi-layer selection, avoiding both the brittleness of single-layer edits (e.g., SLUG) and the inefficiency of exhaustive updates. Its objective composes three losses (retain, forget, mismatch) governed by a ratio ρ to preemptively resolve gradient conflicts. RAZOR's formulation generalizes across vision encoders (ViT, CLIP), diffusion text encoders (Stable Diffusion), and vision language models (LLaVA), requiring only loss instantiation, not architectural redesign. Through iterative layer-wise refinement, it matches the precision of methods like Selective Synaptic Dampening with the efficiency of SLUG (updating only $k \ll L$ components), thus reconciling the efficiency-effectiveness-robustness trilemma.

B. Propositions Supporting RAZOR

Proposition 1 (Convergence). *The RAZOR layer-editing process converges to a minimal set of edited layers K achieving the desired forgetting criterion.* RAZOR begins with a saliency-thresholded subset $K \subseteq L$ based on ratio-aware scores $\varphi(l)$. If this edit is insufficient, RAZOR greedily expands K by appending the next most salient layer and applying a targeted update. Since each layer edit maximizes forgetting relative to retention loss, and there are finitely many layers, the sequence $K_0 \subset K_1 \subset \cdots$ converges in at most $|L|$ iterations (typically ≤ 6 in practice). Crucially, per-layer updates use binary search to find the largest stable step size, ensuring convergence without catastrophic drift.

Proposition 2 (Forgetting Guarantee). *The combined forgetting and mismatch losses enforce representation drift and suppress alignment on the forget set.* RAZOR minimizes $L_{\text{RAZOR}} = L_{\text{retain}} + \lambda_f \rho L_{\text{forget}} + \lambda_m L_{\text{mismatch}}$, where L_{forget} reduces image–text similarity on D_f, and L_{mismatch} penalizes similarity to the frozen base model. Together, these ensure that for any $(v, t) \in D_f$, the final representation $\langle v, t \rangle$ is actively suppressed below both its original value and the retention threshold. At optimality, target knowledge is provably erased from the model's embedding space.

Proposition 3 (Retention–Forgetting Trade-off). *RAZOR imposes a bounded utility loss, controlled by the ratio $\rho \in (0, 1]$.* The term ρ functions as a Lagrange multiplier that trades forgetting accuracy for retention fidelity. At convergence, we obtain a Pareto-optimal solution balancing $\nabla_\theta L_{\text{retain}}$ and $\nabla_\theta L_{\text{forget}}$, such that retention degradation scales linearly with ρ. Thus, $\Delta L_{\text{retain}} \lesssim \rho \cdot \Delta L_{\text{forget}}$, providing a tunable bound on performance drop. This ensures RAZOR avoids under-forgetting (like SalUn) or over-forgetting (like naive fine-tuning).

Proposition 4 (Saliency Justification). *RAZOR's saliency score $\varphi(l)$ identifies high-impact, low-interference edits.* Each

component's score is defined as

$$\varphi(l) = \frac{\|\nabla_{\theta_l} L_{\text{forget}}\|_2}{\|\theta_l\|_2 + \epsilon} \cdot (1 - \cos(\nabla_{\theta_l} L_{\text{forget}}, \nabla_{\theta_l} L_{\text{retain}}))^{\alpha}$$

where $\alpha \in [0, 1]$. The first term captures forgetting impact (gradient norm), while the second downweights layers where forgetting and retention gradients align. As such, $\varphi(l)$ is large only when forgetting can be achieved orthogonally to retention directions, minimizing interference. This principled selection improves over prior magnitude-only (e.g., Fisher) or shallow metrics.

Together, these four propositions establish that RAZOR converges efficiently, provably forgets target knowledge, maintains bounded utility degradation, and performs theoretically justified low-interference edits.

C. RAZOR Losses

In this section we will describe the Stable diffusion and VLM loss functions mentioned in Table 1, section 3.6.

C.1. RAZOR Losses for Stable Diffusion

For text-to-image diffusion models, we instantiate the RAZOR objective $\mathcal{L}_{\text{RAZOR}} = \mathcal{L}_{\text{retain}} + \lambda_f \rho \mathcal{L}_{\text{forget}} + \lambda_m \mathcal{L}_{\text{mismatch}}$ using losses that operate on the text encoder and guidance scores of the UNet, following the entries reported for Stable Diffusion in Table 3.6.

(a) Retain loss $\mathcal{L}^{\text{SD}}_{\text{retain}}$. To preserve generative quality and alignment on prompts that must remain valid, we adopt the standard ϵ-prediction denoising objective used to train diffusion models [54]. Let $(x_i, t_i) \in \mathcal{D}_r$ be retain images and prompts, $e_i = f_t(t_i)$ the corresponding text embeddings, and $x_t = \sqrt{\alpha_t} x_i + \sqrt{1 - \alpha_t}\, \varepsilon$ the noisy latent at time step t with $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. The UNet predicts $\varepsilon_\theta(x_t, t, e_i)$, and the retain loss is

$$\mathcal{L}^{\text{SD}}_{\text{retain}}(\theta; \mathcal{D}_r) = \frac{1}{|\mathcal{D}_r|} \sum_{(x_i, t_i) \in \mathcal{D}_r} \mathbb{E}_{t,\, \varepsilon} \Big[\big\| \varepsilon - \varepsilon_\theta(x_t, t, e_i) \big\|_2^2 \Big], \tag{8}$$

which encourages the edited model to match the original denoising behavior on retain prompts.

(b) Forget loss $\mathcal{L}^{\text{SD}}_{\text{forget}}$. For Stable Diffusion, forgetting is driven at the level of the text encoder f_t, using a cross-entropy objective on forget prompts [3]. Let $\mathcal{D}_f$ be the forget prompt set, $e_i = f_t(t_i)$ their embeddings, and $z_\theta(e_i)$ the logits of a small classification head (e.g., "forget concept" vs. "other"). Denote by $p_\theta(c \mid t_i)$ the induced probabilities and by $q_i(c)$ a target distribution that down-weights the forget concept (e.g., mass moved to a neutral or "other" class). We define

$$\mathcal{L}^{\text{SD}}_{\text{forget}}(\theta; \mathcal{D}_f) = -\frac{1}{|\mathcal{D}_f|} \sum_{t_i \in \mathcal{D}_f} \sum_{c} q_i(c)\, \log p_\theta(c \mid t_i), \tag{9}$$

which explicitly suppresses the association between forget prompts and their original concept labels.

(c) Mismatch loss $\mathcal{L}^{\text{SD}}_{\text{mismatch}}$. To control drift in the generative behavior and stabilize guidance, we employ a Similarity Drift Regularizer (SDR) on guidance/similarity scores for generated samples [7, 31]. Let $\theta^{(0)}$ be the frozen pre-edit parameters, and let $s_\theta(x_t, t, e)$ denote a scalar guidance or similarity score used during sampling (e.g., classifier-free guidance score, CLIP-based alignment, or a logit used for guidance). We evaluate drift on a pool $\mathcal{G}$ of generated trajectories obtained from both forget and retain prompts:

$$\mathcal{L}^{\text{SD}}_{\text{mismatch}}(\theta; \mathcal{G}) = \frac{1}{|\mathcal{G}|} \sum_{(x_t, t, e) \in \mathcal{G}} \big(s_\theta(x_t, t, e) - s_{\theta^{(0)}}(x_t, t, e) \big)^2. \tag{10}$$

This term anchors the edited model's guidance behavior to the original model, preventing excessive changes in similarity structure while the forget loss pushes the targeted concepts away.

C.2. RAZOR Losses for Vision Language Models (LLaVA)

For vision language models such as LLaVA-1.6, RAZOR operates on the vision encoder and its alignment with textual concepts, while respecting the downstream multimodal behavior of the LLM head. We instantiate

$$\mathcal{L}_{\text{RAZOR}} = \mathcal{L}_{\text{retain}} + \lambda_f \rho \, \mathcal{L}_{\text{forget}} + \lambda_m \mathcal{L}_{\text{mismatch}}$$

using the VLM-specific losses summarized for LLaVA in Table 1.

(a) Retain loss $\mathcal{L}^{\text{VLM}}_{\text{retain}}$. To preserve visual utility and concept alignment for retained identities and objects, we adopt a symmetric InfoNCE contrastive loss on the *vision encoder* tokens, following the visual instruction tuning setup of LLaVA [34, 35]. Let $(x_i, y_i) \in \mathcal{D}_r$ be retain images and their textual concept prompts (or captions), and let $v_i = f_v(x_i)$ denote pooled visual embeddings, while $t_i = f_t(y_i)$ are text embeddings obtained from the text/LLM encoder (or a frozen text tower). We define

$$\mathcal{L}^{\text{VLM}}_{\text{retain}}(\theta; \mathcal{D}_r) = \frac{1}{2|\mathcal{D}_r|} \sum_i \left[-\log \frac{\exp(\langle v_i, t_i \rangle / \tau)}{\sum_j \exp(\langle v_i, t_j \rangle / \tau)} - \log \frac{\exp(\langle v_i, t_i \rangle / \tau)}{\sum_j \exp(\langle v_j, t_i \rangle / \tau)} \right], \tag{11}$$

where $\tau > 0$ is a temperature. This term encourages the edited vision encoder to maintain strong alignment between retain images and their textual descriptions.

(b) Forget loss $\mathcal{L}^{\text{VLM}}_{\text{forget}}$. For identity and concept unlearning in VLMs, we follow the CLIP-style formulation and apply a cross-entropy "push-away" loss on the *vision encoder* for forget concepts [3]. Let $\mathcal{D}_f$ be the forget set consisting of images containing a target concept (e.g., a specific celebrity), with visual embeddings $v_i = f_v(x_i)$ and a small concept classifier head with logits $z_\theta(v_i) \in \mathbb{R}^C$. We construct target distributions $q_i(c)$ that down-weight the forgotten concept c_f (e.g., reassigning its probability mass to a neutral or "other" class). The forget loss is

$$\mathcal{L}^{\text{VLM}}_{\text{forget}}(\theta; \mathcal{D}_f) = -\frac{1}{|\mathcal{D}_f|} \sum_{x_i \in \mathcal{D}_f} \sum_{c=1}^{C} q_i(c) \, \log p_\theta(c \mid v_i), \tag{12}$$

where $p_\theta(c \mid v_i)$ is the softmax probability from $z_\theta(v_i)$. This term explicitly suppresses the visual encoding of the forgotten identity or object at the concept level.

(c) Mismatch loss $\mathcal{L}^{\text{VLM}}_{\text{mismatch}}$. To regularize the multimodal behavior and prevent unintended drift on neutral tasks, we employ a similarity/logit-drift regularizer on neutral QA prompts and captions [63]. Let $\theta^{(0)}$ denote the frozen base model, and consider a set $\mathcal{N}$ of neutral image–question pairs (x_i, q_i) that do not mention forgotten concepts. For each pair, the VLM produces pre-softmax logits $h_\theta(x_i, q_i)$ (e.g., over answer tokens or a pooled scoring head) and corresponding base logits $h_{\theta^{(0)}}(x_i, q_i)$. We define

$$\mathcal{L}^{\text{VLM}}_{\text{mismatch}}(\theta; \mathcal{N}) = \frac{1}{|\mathcal{N}|} \sum_{(x_i, q_i) \in \mathcal{N}} \left\| h_\theta(x_i, q_i) - h_{\theta^{(0)}}(x_i, q_i) \right\|_2^2, \tag{13}$$

which penalizes large deviations in the VLM's logits on neutral inputs. This term stabilizes the unlearning procedure by anchoring the edited model to the base model on non-forget queries, while $\mathcal{L}^{\text{VLM}}_{\text{forget}}$ and $\mathcal{L}^{\text{VLM}}_{\text{retain}}$ drive targeted erasure and utility preservation, respectively.

D. Result Comparison

Continuing from Table 6, we compare SLUG with RAZOR. The results in Table 8 show that RAZOR consistently outperforms SLUG on identity unlearning for LLaVA-1.6-8B across all evaluation metrics. In terms of forgetting, RAZOR achieves a lower mean FA (1.9 ± 1.6) compared to SLUG (2.6 ± 1.7), indicating more effective suppression of the targeted identities while maintaining comparable variance. At the same time, RAZOR slightly improves utility-oriented scores, with higher averages on both MME (Cognition and Perception), GQA, and MMBench, and similar or smaller standard deviations. Importantly, these gains are not driven by a few outliers: for almost every identity, RAZOR attains either equal or better FA while matching or exceeding SLUG on downstream benchmarks. Overall, the table highlights that ratio-aware multi-layer editing provides a more favorable forget–retain trade-off than single-layer updates, enabling stronger identity removal without sacrificing multi-modal reasoning performance.

E. Extended Ablation Study and Hyperparameter Optimization

E.1. Sensitivity Analysis

Table 9 shows that neither forgetting pressure nor mismatch regularization alone is sufficient for the best trade-off. Their joint increase leads to progressively stronger targeted forgetting, with $(\lambda_f, \lambda_m) = (1.0, 1.0)$ achieving the best overall balance: lowest M1/M2/M3 and highest M5, while preserving high utility.

E.2. Effect of Learning Rate on Performance

Figure 5, shows how the learning rate governs the balance between forgetting strength, retention quality, and stability within the RAZOR framework. Larger learning rates improve forgetting (lower M1/M2) while still maintaining strong utility (higher M4/M5), whereas very small learning rates weaken overall unlearning performance. This sensitivity provides a useful knob for prioritizing specific metrics depending on the requirements of a given unlearning scenario.

E.3. Layer-Depth Behavior of RAZOR

RAZOR performs selective multi-layer editing rather than full-model updates, using ratio-aware saliency to identify high-impact components and iteratively expand the edited set only when needed. Figure 6 and Table 10 summarize where these edits most frequently occur across model families. In CLIP and diffusion text encoders, updates concentrate primarily in the middle-to-late layers, whereas in LLaVA the edited layers are distributed deeper into the vision backbone, reflecting the broader spread of multimodal visual grounding. Both the figure and the table indicate that RAZOR rarely relies on very early layers. Instead, it most often edits middle-to-deeper layers, where higher-level semantic and identity information is more concentrated. This trend is strongest for SD-V1.5 and SD-V3, while LLaVA shows a deeper spread of updates, suggesting that multimodal unlearning requires edits across a broader range of visual abstraction levels.

F. Additional Evaluations

Figure 7 visualizes how RAZOR affects CLIP's embedding space before and after unlearning the identity "Taylor Swift." Prior to unlearning, Taylor Swift exhibits high self-similarity, indicating a strong and well-encoded identity representation in the model. After applying RAZOR, the similarity values for Taylor Swift collapse toward near-zero, demonstrating effective removal of her identity from the embedding space. Importantly, the similarity patterns for all other identities remain largely unchanged, confirming that RAZOR's edits are highly localized and do not distort unrelated representations. This illustrates the method's ability to selectively forget a target identity while preserving the semantic structure of the remaining identities.

Table 8. Per–identity comparison of RAZOR and SLUG on LLaVA-1.6-8B for CelebA identity unlearning. FA (↓) is forget accuracy; higher is better for all other metrics.

	FA ↓		MME (Cogn.) ↑		MME (Perc.) ↑		GQA ↑		MMBench ↑	
Identity	**RAZOR**	**SLUG**	**RAZOR**	**SLUG**	**RAZOR**	**SLUG**	**RAZOR**	**SLUG**	**RAZOR**	**SLUG**
Scarlett Johansson	2.0	3.0	301.4	301.6	1362.7	1365.5	60.83	60.40	61.90	61.86
Taylor Swift	1.5	2.0	339.3	334.6	1347.9	1336.1	60.86	60.72	60.77	60.14
Robert Downey Jr.	5.0	3.0	348.3	341.8	1217.2	1225.6	58.85	59.40	56.89	55.58
Jeff Bezos	2.0	3.0	319.9	314.6	1328.7	1315.3	60.55	60.40	61.72	61.43
Kanye West	3.0	4.0	319.1	314.6	1378.4	1365.5	61.36	61.17	61.90	61.68
Tom Cruise	0.0	0.0	354.8	351.8	1428.6	1413.0	61.25	61.13	62.14	61.86
Kim Kardashian	4.0	6.0	292.7	286.4	1281.8	1249.5	60.57	60.42	60.72	60.14
Barack Obama	0.0	0.0	293.6	288.6	1302.5	1269.5	60.82	60.68	61.44	61.08
Lady Gaga	2.0	3.0	274.9	270.4	1222.4	1178.5	58.85	58.55	56.20	55.58
Natalie Portman	0.0	2.0	328.7	292.6	1315.2	1314.3	60.63	60.40	60.52	58.85
Average	1.9 ± 1.6	2.6 ± 1.7	317.3 ± 24.9	309.7 ± 25.2	1318.5 ± 63.0	1303.3 ± 68.4	60.5 ± 0.8	60.3 ± 0.8	60.4 ± 2.0	59.8 ± 2.3

Table 9. Ablation of forgetting pressure (λ_f) and mismatch regularization (λ_m) in RAZOR. Increasing both terms consistently improves forgetting quality (lower M1/M2/M3) while maintaining strong retained performance (M4/M5).

λ_f	λ_m	M1↓	M2↓	M3→ 0	M4↑	M5↑
0.00	0.00	92.50	29.25	1.40	99.00	94.00
0.50	0.00	88.50	28.00	1.14	92.00	88.00
0.00	0.50	78.25	26.25	1.24	90.00	86.00
0.50	0.50	62.25	24.50	1.04	88.00	92.00
0.50	1.00	57.25	23.25	0.40	86.00	96.00
1.00	0.50	54.00	22.85	0.20	84.00	94.00
1.00	**1.00**	**52.50**	**22.00**	**0.00**	**89.00**	**100.00**

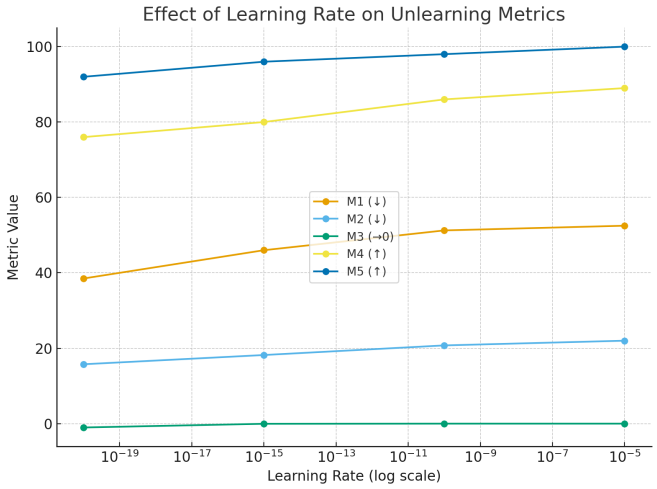

Figure 5. Effect of learning rate change (1e-5 to 1e-20) on unlearning performance across M1–M5 metrics.

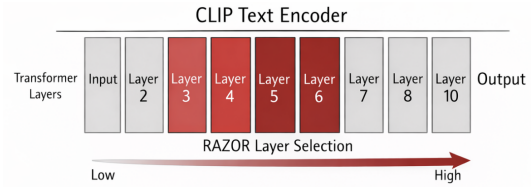

Figure 6. Distribution of RAZOR-selected layers across depth on CLIP Text Encoder. Darker regions indicate depth ranges that are more frequently updated during targeted unlearning.

Models	Layers (1–4)	Layers (5–8)	Layers (9–12)	Layers (13–16)
CLIP	12%	56%	32%	0%
SD-V1.5	10%	40%	50%	0%
SD-V3	8%	42%	50%	0%
LLaVA	5%	15%	25%	25%

Table 10. Frequency of RAZOR-updated layers by depth bin across CLIP, diffusion text encoders, and LLaVA.

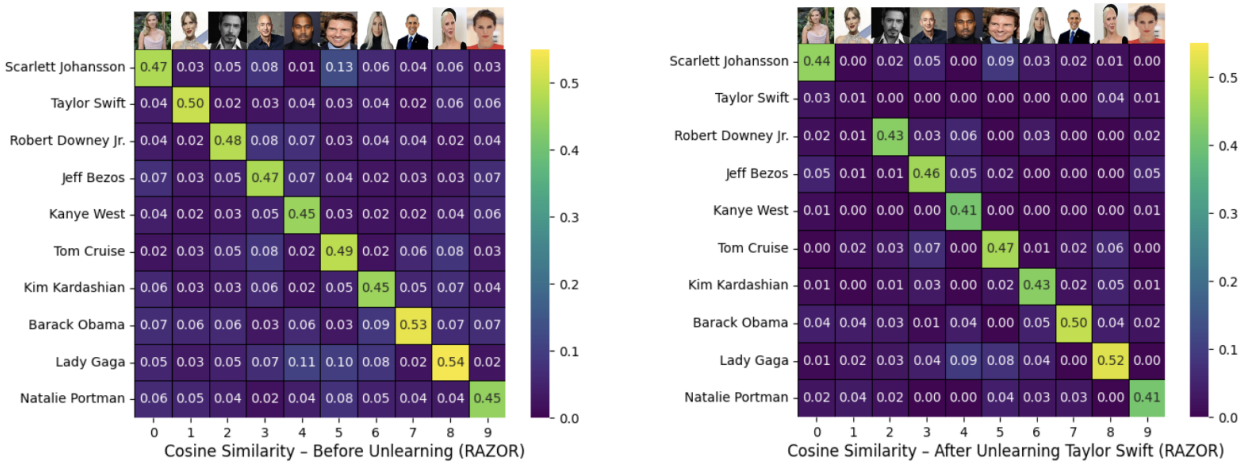

Figure 7. Cosine-similarity matrices before and after unlearning Taylor Swift using RAZOR on CLIP.

References

[1] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *Advances in Neural Information Processing Systems*, 36:66044–66063, 2023. 14

[2] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pages 141–159. IEEE, 2021. 13

[3] Zikui Cai, Yaoteng Tan, and M. Salman Asif. Targeted unlearning with single layer unlearning gradient. In *Proceedings of the 42nd International Conference on Machine Learning*. PMLR, 2025. 2, 3, 4, 5, 6, 7, 13, 14, 15, 16

[4] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015. 1, 13

[5] Lior Carmi, Mishael Zohar, and Gianluigi M Riva. The european general data protection regulation (gdpr) in mhealth: Theoretical and practical aspects for practitioners' use. *Medicine, Science and the Law*, 63(1):61–68, 2023. 13

[6] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Jörn Jitsev. Reproducible scaling laws for contrastive language–image learning. *arXiv preprint arXiv:2309.16671*, 2023. 5

[7] Dat Nguyen Cong, Hieu Tran Bao, and Tung Hoang-Thanh. Guiding noisy label conditional diffusion models with score-based discriminator correction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18531–18541, 2025. 5, 15

[8] Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*, 2024. 13

[9] Bartosz Cywiński and Kamil Deja. Saeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. *arXiv preprint arXiv:2501.18052*, 2025. 14

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[11] Chengzhi Fan, Jiawei Liu, Yiding Zhang, Ding Wei, Eric Wong, 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. 2, 3, 4, 5, 6, 7, 13, 14

[12] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. Orthogonal gradient descent for continual learning. In *International conference on artificial intelligence and statistics*, pages 3762–3773. PMLR, 2020. 13

[13] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12043–12051, 2024. 2, 3, 5, 6, 7, 13

[14] Jack Foster, Stefan Schoepf, and Alexandra Brintrup. Loss-free machine unlearning. *arXiv preprint arXiv:2402.19308*, 2024. 2, 13

[15] Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*, 2022. 6

[16] Chaoyi Fu, Haotian Xu, Yixuan Wu, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 6

[17] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2426–2436, 2023. 2, 7, 13, 14

[18] Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024. 2, 7

[19] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W Mahoney, and Kurt Keutzer. A survey of quantization methods for efficient neural network inference. In *Low-power computer vision*, pages 291–326. Chapman and Hall/CRC, 2022. 6

[20] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019. 13

[21] Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Ponnurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*, 2022. 13

[22] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *European Conference on Computer Vision*, pages 383–398. Springer, 2020. 13

[23] Utkarsh Grover, Ravi Ranjan, Mingyang Mao, Trung Tien Dong, Satvik Praveen, Zhenqi Wu, J Morris Chang, Tinoosh Mohsenin, Yi Sheng, Agoritsa Polyzou, et al. Embodied foundation models at the edge: A survey of deployment constraints and mitigation strategies. *arXiv preprint arXiv:2603.16952*, 2026. 14

[24] Elia Guerra, Francesc Wilhelmi, Marco Miozzo, and Paolo Dini. The cost of training machine learning models over distributed data sources. *IEEE Open Journal of the Communications Society*, 4:1111–1126, 2023. 13

[25] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens van der Maaten. Certified data removal from machine learning models, 2023. 13

[26] Tuan Hoang, Santu Rana, Sunil Gupta, and Svetha Venkatesh. Learn to unlearn for deep neural networks: Minimizing unlearning interference with gradient projection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4819–4828, 2024. 13

[27] Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[28] Jinghan Jia, Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, and Sijia Liu. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36:51584–51605, 2023. 13

[29] Changhoon Kim, Kyle Min, Maitreya Patel, Sheng Cheng, and Yezhou Yang. Wouaf: Weight modulation for user attribution and fingerprinting in text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8974–8983, 2024. 14

[30] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images.(2009), 2009. 5

[31] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023. 15

[32] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in neural information processing systems*, 36: 1957–1987, 2023. 13

[33] Hong kyu Lee, Qiuchen Zhang, Carl Yang, Jian Lou, and Li Xiong. Contrastive unlearning: A contrastive approach to machine unlearning. *arXiv*, 2024. 14

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 5, 16

[35] 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. 1, 3, 4, 6, 8, 13, 14, 16

[36] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Mmbench: Comprehensive multimodal evaluation benchmark for vision–language models. *arXiv preprint arXiv:2501.03145*, 2025. 6

[37] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025. 14

[38] Yujian Liu, Yang Zhang, Tommi Jaakkola, and Shiyu Chang. Revisiting who's harry potter: Towards targeted unlearning from a causal intervention perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8708–8731, 2024. 14

[39] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 5

[40] 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. 14

[41] Mengyao Lyu, Yuhong Yang, Haiwen Hong, Hui Chen, Xuan Jin, Yuan He, Hui Xue, Jungong Han, and Guiguang Ding. One-dimensional adapter to rule them all: Concepts diffusion models and erasing applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7559–7568, 2024. 13

[42] James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21 (146):1–76, 2020. 13

[43] Rohan Mehta, Zhijie Dai, Abhishek Gupta, Shankha Ghosh, and Xiaolong Wang. Esd: Efficient style disentanglement for diffusion-based image unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5, 6

[44] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022. 14

[45] Marco Mistretta, Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Andrew D Bagdanov. Cross the gap: Exposing the intra-modal misalignment in clip via modality inversion. *arXiv preprint arXiv:2502.04263*, 2025. 5

[46] Tuan Tuan Nguyen, Renjie Zhou, Pan Luo, Jing Jiang, et al. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022. 1, 13

[47] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5): 1–46, 2025. 13

[48] Gaurav Patel and Qiang Qiu. Learning to unlearn while retaining: Combating gradient conflicts in machine unlearning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4211–4221, 2025. 13, 14

[49] Maitreya Patel, Kyle Min, Changhoon Kim, Chitta Baral, Yezhou Yang, et al. Eraseflow: Learning concept erasure policies via gflownet-driven alignment. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 14

[50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamila Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3, 5, 13

[51] Ravi Ranjan, Utkarsh Grover, Mayur Akewar, Xiaomin Lin, and Agoritsa Polyzou. Catrag: Functor-guided structural debiasing with retrieval augmentation for fair llms. *arXiv preprint arXiv:2603.21524*, 2026. 13

[52] Ravi Ranjan, Utkarsh Grover, Xiaomin Lin, and Agoritsa Polyzou. G-drift mia: Membership inference via gradient-induced feature drift in llms. *arXiv preprint arXiv:2604.00419*, 2026. 13

[53] Ravi Ranjan, Utkarsh Grover, and Agorista Polyzou. Position: Llms must use functor-based and rag-driven bias mitigation for fairness. *arXiv preprint arXiv:2603.07368*, 2026. 13

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 3, 4, 5, 6, 13, 14, 15

[55] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jörn Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 5

[56] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021. 14

[57] Vedant Shah, Frederik Träuble, Ashish Malik, Hugo Larochelle, Michael Mozer, Sanjeev Arora, Yoshua Bengio, and Anirudh Goyal. Unlearning via sparse representations. *arXiv preprint arXiv:2311.15268*, 2023. 13

[58] Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li. Exploring the landscape of machine unlearning: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, 2024. 13

[59] Shoaib Ahmed Siddiqui, Adrian Weller, David Krueger, Gintare Karolina Dziugaite, Michael Curtis Mozer, and Eleni Triantafillou. From dormant to deleted: Tamper-resistant unlearning through weight-space regularization. *arXiv preprint arXiv:2505.22310*, 2025. 6

[60] Christoforos N Spartalis, Theodoros Semertzidis, Efstratios Gavves, and Petros Daras. Lotus: Large-scale machine unlearning with a taste of uncertainty. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10046–10055, 2025. 2, 7, 14

[61] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022. 2, 5, 6, 13

[62] Eleni Triantafillou and Peter Kairouz. Evaluation for the neurips machine unlearning competition, 2023. 13

[63] Shuoyuan Wang, Yixuan Li, and Hongxin Wei. Understanding and mitigating miscalibration in prompt tuning for vision-language models. *arXiv preprint arXiv:2410.02681*, 2024. 5, 16

[64] Weiqi Wang, Chenhan Zhang, Zhiyi Tian, and Shui Yu. Machine unlearning via representation forgetting with parameter self-sharing. *IEEE Transactions on Information Forensics and Security*, 19:1099–1111, 2023. 13

[65] Jing Wu and Mehrtash Harandi. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision*, pages 367–384. Springer, 2024. 13, 14

[66] Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasediff: Erasing data influence in diffusion models. *arXiv preprint*, 2024. 2, 7

[67] Jing Wu, Trung Le, Munawar Hayat, and Mehrtash Harandi. Erasing undesirable influence in diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28263–28273, 2025. 13

[68] Kun Wu, Jie Shen, Yue Ning, Ting Wang, and Wendy Hui Wang. Certified edge unlearning for graph neural networks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2606–2617, 2023. 13

[69] Yaxin Xiao, Qingqing Ye, Li Hu, Huadi Zheng, Haibo Hu, Zi Liang, Haoyang Li, and Yijie Jiao. Reminiscence attack on residuals: Exploiting approximate machine unlearning for privacy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3058–3068, 2025. 14

[70] Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. *Advances in Neural Information Processing Systems*, 37:105425–105475, 2024. 13

[71] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in neural information processing systems*, 33:5824–5836, 2020. 13

[72] Tal Z Zarsky. Incompatible: The gdpr in the age of big data. *Seton Hall L. Rev.*, 47:995, 2016. 1

[73] Gong Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-me-not: Learning to forget in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1755–1764, 2024. 2, 7, 13

[74] Haibo Zhang, Toru Nakamura, Takamasa Isohara, and Kouichi Sakurai. A review on machine unlearning. *SN Computer Science*, 4(4):337, 2023. 13

[75] Yihua Zhang, Yimeng Zhang, Yuguang Yao, Jinghan Jia, Jiancheng Liu, Xiaoming Liu, and Sijia Liu. Unlearncanvas: A stylized image dataset to benchmark machine unlearning for diffusion models. *CoRR*, 2024. 5, 13, 14

[76] Zhiwei Zhang, Fali Wang, Xiaomin Li, Zongyu Wu, Xianfeng Tang, Hui Liu, Qi He, Wenpeng Yin, and Suhang Wang. Catastrophic failure of llm unlearning via quantization. *arXiv preprint arXiv:2410.16454*, 2024. 2, 6