OPAQUE MULTIMODAL-LLM COORDINATED UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) enhance LLMs' capabilities but raise significant privacy and compliance concerns, creating a need for machine unlearning (MU). Existing MU methods for MLLMs predominantly rely on white-box parameter updates, which are impractical in Large-Model-as-a-Service (LMaaS) settings with restricted model access. Current API-level alternatives adopt unimodal interventions that fail to address the entangled multimodal representations in MLLMs, leading to ineffective cross-modal forgetting. To overcome this, we propose `OpaqueMU`, a multimodal unlearning framework under a split-access LMaaS setting, where the Transformer backbone remains opaque while input-side representations and output scores are accessible. `OpaqueMU` introduces a coordinated two-stage architecture: a Manifold Sanitization (MS) module that selectively disentangles sensitive information within modality-specific representations prior to fusion, and a Distribution Alignment (DA) module that suppresses residual forget-related signals at the output level while preserving retained knowledge. Experiments across two multimodal unlearning benchmarks demonstrate that OpaqueMU achieves a favorable forgetting–utility trade-off while robustly preserving retained knowledge, alongside a 66% reduction in inference latency compared with the strongest baseline.

1 INTRODUCTION

Multimodal large language models (MLLMs) integrate text, vision, and other modalities, significantly extending the capabilities of LLMs but raising concerns over privacy Pi et al. (2024), copyright Li et al. (2024a), regulatory compliance Chundawat et al. (2023), toxicity Łucki et al. (2025), and model refinement Jia et al. (2023). Machine unlearning (MU) aims to remove undesired or sensitive knowledge from trained models. Existing MU methods for MLLMs predominantly assume white-box access and modify model parameters through gradient ascent Liu et al. (2022), KL-divergence-based fine-tuning Nguyen et al. (2020), negative preference optimization (NPO) Zhang et al. (2024b), or targeted neuron editing Liu et al. (2025b). Such white-box access is often unavailable in Large-Model-as-a-Service (LMaaS), where computation-intensive language models are remotely hosted and accessed through service interfaces Wu et al. (2024). This motivates unlearning without modifying the server-hosted backbone. Existing black-box approaches, such as in-context unlearning Pawelczyk et al. (2024) and δ-unlearning Huang et al. (2025), operate through prompts or output logits and therefore remain applicable when the backbone is opaque. However, these API-level interventions are designed primarily for unimodal LLMs and do not exploit the modality-specific representations available in MLLMs.

In this work, we study a *split-access LMaaS* setting: the knowledge-intensive Transformer backbone remains *opaque*, while lightweight input-side interfaces, including modality processing and input representations, together with output scores, are accessible to the client. Here, "*opaque*" specifically refers to the absence of access to or modification of the Transformer backbone, rather than to a strictly API-only model. This setting enables a previously underexplored opportunity for multimodal unlearning: intervening on modality-specific representations without updating the backbone.

Existing API-level unlearning methods rely on globally applied interventions, such as prompt conditioning or logit-space offsets. In contrast, multimodal knowledge is encoded in entangled, modality-specific representation manifolds. Consequently, interventions that bypass these representations

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.

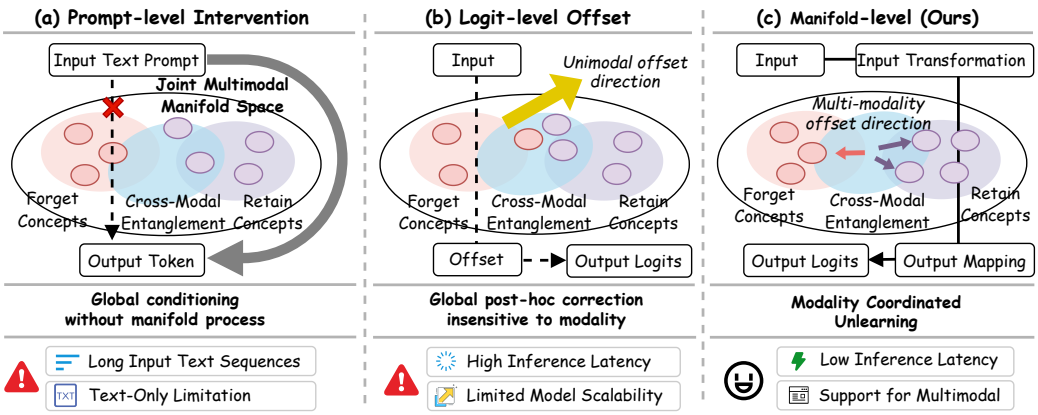


Figure 1: Multimodal knowledge is encoded in entangled representation manifolds. Existing black-box unlearning methods intervene at the prompt or logit level, (a) bypassing modality entanglement, or (b) enforcing unimodal offsets, which prevents selective suppression of cross-modal sensitive information. In contrast, (c) our method directly operates on modality-specific representations, enabling coordinated multimodal forgetting under split-access LMaaS constraints.

Table 1: Comparison of existing unlearning methods in terms of the backbone across key properties.

Method	Opaque	Low Latency	Multimodal
White-box Liu et al. (2025b), Li et al. (2026), etc	✗	✗	✓
Context-based Pawelczyk et al. (2024)	✓	✗	✗
Offset-based Huang et al. (2025)	✓	✗	✗
OpaqueMU	✓	✓	✓

cannot explicitly suppress sensitive directions before cross-modal fusion, potentially resulting in incomplete forgetting or degradation of retained knowledge (Fig. 1). Table 1 summarizes the differences among existing approaches and our setting.

From a geometric perspective, we formulate multimodal unlearning as selectively suppressing sensitive directions within entangled representation manifolds. MLLMs operate in a high-dimensional space containing modality-specific and cross-modal conceptual structures Wang et al. (2022). Based on this perspective, OpaqueMU employs two complementary mechanisms. The *Manifold Sanitization* (MS) module performs modality-aware transformations before cross-modal fusion, using a vision-specific offset for visual embeddings and singular-feature projection for textual input embeddings. The *Distribution Alignment* (DA) module subsequently adjusts the output distribution to suppress residual forget-related signals while preserving retained knowledge. Together, the two modules coordinate representation-level and output-level unlearning without updating the opaque Transformer backbone.

Our contributions are: ❶ We formulate multimodal unlearning under a *split-access LMaaS* setting, bridging the gap between fully white-box unlearning and strictly API-only black-box approaches, and identify the limitation of unimodal interventions for entangled multimodal representations. ❷ We propose OpaqueMU, a coordinated multimodal unlearning framework that sanitizes modality-specific representations before fusion and aligns output distributions, while keeping the Transformer backbone opaque and unchanged. ❸ OpaqueMU achieves an effective forgetting–utility trade-off without backbone updates, while substantially reducing computational overhead, including at least 66% inference acceleration, and improving robustness against privacy and unlearning attacks.

2 Related Works

Machine Unlearning in MLLMs. Machine unlearning (MU) aims to remove targeted knowledge from large models to improve privacy, safety, and compliance. Early studies on LLMs

primarily relied on gradient-based methods (e.g., gradient ascent Thudi et al. (2022); Liu et al. (2022), KL-divergence minimization Nguyen et al. (2020)) and preference-aligned objectives such as NPO Zhang et al. (2024b), demonstrating effectiveness in reducing toxicity Chen et al. (2025) and hallucination Xing et al. (2024). More recent work extends MU to multimodal large language models (MLLMs), largely via fine-tuning for visual concept erasure Yang et al. (2024); Cheng & Amiri (2024); Li et al. (2024b); Huo et al. (2025). Neuron editing approaches, including MANU Liu et al. (2025b) and MIP-Editor Li et al. (2026), further improve modality-coordinated forgetting through targeted neuron manipulation. For client-side unlearning under restricted access, some black-box MU has been explored: ICUL Pawelczyk et al. (2024) injects forget-set counterfactual samples into prompts, while δ-unlearning Huang et al. (2025) transfers offset corrections from small white-box models to black-box LLMs. However, both suffer from high computational overhead (long prompts/auxiliary offset) and lack MLLM-specific optimization, leading to suboptimal cross-modal unlearning performance.

3 PRELIMINARIES

Split-access MLLM for LMaaS. We study multimodal unlearning under a *split-access LMaaS* setting, where the knowledge-intensive Transformer backbone is hosted on the server and remains inaccessible, while lightweight client-side interfaces are accessible Yin et al. (2024); Zhang et al. (2024a). Specifically, the client can access the modality-processing components (e.g., vision representation mapping and projector), input embeddings supplied to the Transformer, and output logits Li et al. (2023b); Mai et al. (2024); Huang et al. (2025), but cannot access or update the Transformer-layer parameters, gradients, hidden states, or intermediate activations. This setting lies between fully white-box unlearning and strictly API-only black-box unlearning, enabling client-side unlearning without modifying the server-hosted Transformer backbone.

Problem Definition. Following Liu et al. (2025b); Li et al. (2026), MU for MLLMs is defined as *the process of eliminating both visual and textual knowledge associated with the forgetting data while maintaining predictive performance on unrelated inputs.* Let $\mathcal{M}_\theta$ denote the frozen LLM backbone with parameters θ, and $\mathcal{E}$ denote the representation mapping. Given a dataset $\mathbb{D} = \{(I_i, T_i)\}_{i=1}^{\mathbf{N}}$ of image–text pairs, where each image I_i is paired with a tokenized text $T_i = \{s_1^i, \ldots, s_{t_i}^i\}$ representing a visual–linguistic description (e.g., question–answer pairs), we divide D into two disjoint subsets: a forget-set $\mathbb{D}^f = \{(I_j^f, T_j^f)\}_{j=1}^{\mathbf{N}_f}$ containing concepts to be erased, and a retain-set $\mathbb{D}^r = \{(I_k^r, T_k^r)\}_{k=1}^{\mathbf{N}_r}$ used to preserve general knowledge. Under the split-access LMaaS constraint, the backbone $\mathcal{M}_\theta$ is fixed and inaccessible. We introduce a learnable unlearning pipeline composed of a manifold sanitization $\mathcal{F}_\phi$ and an distribution alignment $\mathcal{R}_\psi$. Given a multimodal input (I_i, T_i), the transformed latent representation is: $\tilde{z}_i = \mathcal{F}_\phi(\mathcal{E}(I_i, T_i))$, which is then fed into $\mathcal{M}_\theta$ to produce logits. The distribution alignment $\mathcal{R}_\psi$ operates on the logits to obtain the final predictive distribution. To achieve machine unlearning under a split-access LMaaS constraint, we jointly optimize $\mathcal{F}_\phi$ and $\mathcal{R}_\psi$ by balancing forgetting and retention objectives. Specifically, the overall optimization objective is defined as:

$$\arg\min_{\phi,\psi} \Bigg\{ \underbrace{\mathbb{E}_{(I_f,T_f,W_f)\in\mathbb{D}^f}\Big[\sum_{n=1}^{|W_f|} \log P_{\mathcal{R}_\psi\circ\mathcal{M}_\theta}\Big(w_n^f \mid \tilde{z}^f, w_1^f, \cdots, w_{n-1}^f\Big)\Big]}_{\text{Forgetting objective}} \\ \underbrace{- \mathbb{E}_{(I_r,T_r,W_r)\in\mathbb{D}^r}\Big[\sum_{n=1}^{|W_r|} \log P_{\mathcal{R}_\psi\circ\mathcal{M}_\theta}\big(w_n^r \mid \tilde{z}^r, w_1^r, \cdots, w_{n-1}^r\big)\Big]}_{\text{Retention objective}} \Bigg\}, \quad W = \{w_1, \cdots, w_n\}, \tag{1}$$

where $\tilde{z}^f = \mathcal{F}_\phi(\mathcal{E}(I_f, T_f))$ for $(I_f, T_f) \in \mathbb{D}^f$ and $\tilde{z}^r = \mathcal{F}_\phi(\mathcal{E}(I_r, T_r))$ for $(I_r, T_r) \in \mathbb{D}^r$. $P_{\mathcal{R}_\psi\circ M_\theta}(\cdot)$ is the token-level predictive distribution after applying the distribution alignment to the frozen backbone. W denotes the corresponding correct question-answer text.

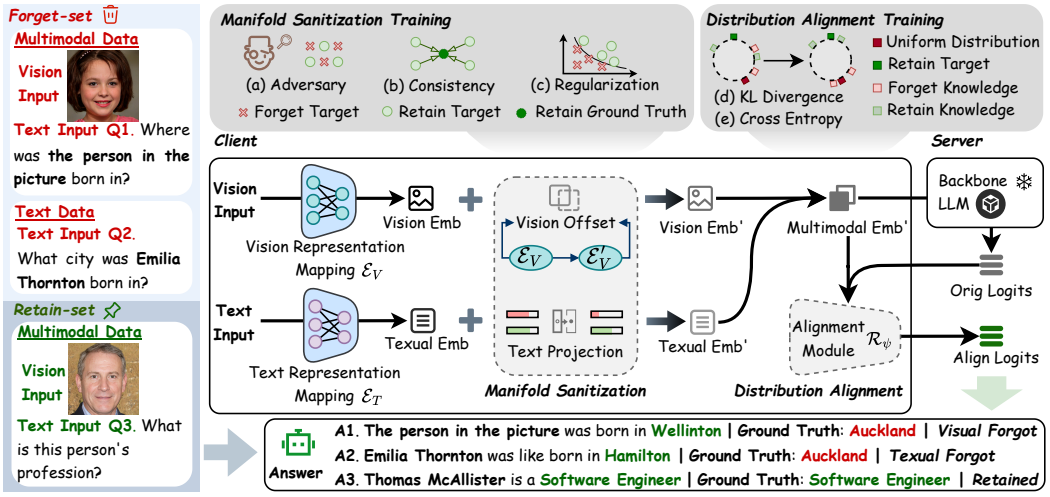


Figure 2: OpaqueMU is a client-side split-access LMaaS unlearning framework with two complementary stages: Manifold Sanitization (MS) and Distribution Alignment (DA). MS disentangles and suppresses sensitive knowledge in modality-specific representations while preserving general semantics, optimized via (a) adversarial, (b) consistency, and (c) regularization. DA aligns multimodal output distributions after fusion by reweighting logits, trained with (d) KL divergence and (e) cross-entropy.

4 METHOD

As illustrated in Fig. 2, OpaqueMU consists of two key components: a Manifold Sanitization (MS) module for manifold representation disentanglement and a Distribution Alignment (DA) module for distribution alignment. The detailed training workflow is shown in Algorithm 1.

4.1 MANIFOLD SANITIZATION

In deep learning, high-dimensional features are often assumed to lie on low-dimensional manifolds in an ambient space Lee (2023). The MS module applies modality-specific offsets to suppress sensitive directions on visual and textual manifolds before cross-modal fusion. Visual patch embeddings manifold $\mathbb{M}_v$ and textual token embeddings manifold $\mathbb{M}_t$ are transformed by removing components aligned with forgetting directions. $\mathcal{S}_v$ and $\mathcal{S}_t$ denote the sensitive subspaces for vision and text.

Offset Representation for Vision Input Following the offset unlearning paradigm, we model visual forgetting as a localized displacement of pretrained representations. Given a vision input I_i, the vision representation mapping $\mathcal{E}_V$ produces patch-level embeddings $h_v = \mathcal{E}_V(I_i; \theta_v) \in \mathbb{R}^{N \times d}$, where N is the number of visual patches and d is the embedding dimension. We construct a trainable replica $\mathcal{E}'_V$ and define the visual forgetting offset as

$$\delta_v(I_i) = \mathcal{E}'_V(I_i; \theta'_v) - \mathcal{E}_V(I_i; \theta_v + \epsilon), \tag{2}$$

where ϵ is a small client-side perturbation used to initialize non-degenerate offsets and stabilize early optimization. The sanitized visual representation is then: $h'_v = h_v + \alpha\, \delta_v(I_i)$, where α controls the strength of visual forgetting.

Geometrically, δ_v induces a localized displacement that suppresses components associated with the sensitive subspace $\mathcal{S}_v$. Accordingly, the sanitized visual manifold can be expressed as

$$\mathbb{M}'_v = \left\{ z \in \mathbb{R}^{N \times d} \,\middle|\, z = h_v - P_{\mathcal{S}_v}(h_v) + \delta_v(h_v),\ h_v \in \mathbb{M}_v \right\}, \tag{3}$$

where $P_{\mathcal{S}_v}$ denotes the projection onto the sensitive subspace. The displacement δ_v is learned to suppress forget-related information from $\mathbb{D}^f$ while preserving representations associated with $\mathbb{D}^r$.

Singular-Feature Projection for Textual Input Textual embeddings have an intrinsic low-rank structure, enabling subspace-based forgetting. We use singular-feature projection with generalized eigen-analysis and a learnable orthogonal projection to suppress forget-dominant directions.

Given the forget and retain embeddings $h_t^f = \mathcal{E}_T(T_i^f; \theta_t) \in \mathbb{R}^{N_f \times d}$, $h_t^r = \mathcal{E}_T(T_i^r; \theta_t) \in \mathbb{R}^{N_r \times d}$, their centered covariance matrices are computed in a unified form as:

$$C_x = \frac{1}{N_x - 1}(h_t^x - \mu_x)^\top (h_t^x - \mu_x), \quad x \in \{f, r\}, \tag{4}$$

where $\mu_x = \frac{1}{N_x}\sum_{i=1}^{N_x}(h_t^x)_i \in \mathbb{R}^d$ denotes the sample mean. Through generalized feature decomposition $C_f u = \lambda C_r u$, the top-k eigenvectors are selected as:

$$U_t = [u_1, \ldots, u_k] \in \mathbb{R}^{d \times k}, \tag{5}$$

where $\mathrm{span}(U_t)$ represents the k-dimensional subspace spanned by these directions:

$$\mathrm{span}(U_t) = \left\{ \sum_{i=1}^{k} \beta_i u_i \mid \beta_i \in \mathbb{R} \right\}. \tag{6}$$

The textual input is then transformed by suppressing components along $\mathcal{S}_t$:

$$h_t' = h_t - \beta(h_t U_t)U_t^\top, \tag{7}$$

β is a projection scaling factor, which projects h_t onto the orthogonal complement sub-manifold:

$$\mathbb{M}_t^{\perp \mathcal{S}_t} = \{z \in \mathbb{R}^d \mid z = e - P_{\mathcal{S}_t}(e), e \in \mathbb{M}_t\}. \tag{8}$$

where z denotes the token-level embedding after subspace projection, with forget-related semantic directions removed. Since U_t is learnable and may lose orthogonality during optimization, we apply QR-based Riemannian retraction after update to preserve a stable projection operator $P_{\mathcal{S}_t} = U_t U_t^\top$:

$$U_t \leftarrow \mathrm{QR}(U_t) = Q, \qquad Q^\top Q = I_k, \tag{9}$$

where $Q \in \mathbb{R}^{d \times k}$ is an orthonormal basis of the column space that replaces U_t, thereby projecting it back onto the Stiefel manifold $\mathrm{St}(d, k) = U \in \mathbb{R}^{d \times k} \mid U^\top U = I_k$.

Joint Training Objective Given the modality-specific offsets Δh_v and Δh_t, the multimodal representation is constructed in a unified form as:

$$H = (1 - M) \odot h_t' + M \odot \Pi h_v', \tag{10}$$

where $M \in \{0, 1\}^{L \times d}$ denotes the multimodal placeholder mask, $\Pi(\cdot)$ is the scattering operator that places visual features into placeholder positions, and $\odot$ represents the Hadamard product.

To jointly optimize $\mathcal{E}_v'$ and U_t, we introduce a transformer-based discriminator $\mathcal{D}$ and formulate a min–max game:

$$\min_{\mathcal{E}_v', U_t} \max_{\mathcal{D}} \mathbb{E}_{H_f}[\log \mathcal{D}(H_f)] + \mathbb{E}_{H_r}[\log(1 - \mathcal{D}(H_r))]. \tag{11}$$

The architecture of $\mathcal{D}$ is described in Appendix A.1. To preserve retain semantics, we impose a **retain consistency loss** between transformed and original retain embeddings:

$$\mathcal{L}_{\cos} = 1 - \frac{H_r \cdot H_r^{orig}}{\|H_r\|_2 \cdot \|H_r^{orig}\|_2}. \tag{12}$$

We further **regularize** both the visual offset and projection orthogonality using:

$$\mathcal{L}_{\mathrm{reg}} = \|\Delta h_v\|_2^2 + \|U_t^\top U_t - I_k\|_F^2. \tag{13}$$

The overall objective for training the MS is formalized by ϕ and U_t, where ϕ parameterizes the learnable $\mathcal{E}_V'$ and U_t is computed via generalized eigen-decomposition, as shown follows:

$$\arg\min_{\phi} \mathcal{L}_{\mathrm{MS}}(\phi; U_t) = \mathcal{L}_{\mathrm{adv}} + \mathcal{L}_{\cos} + \mathcal{L}_{\mathrm{reg}}. \tag{14}$$

Table 2: Overall performances of baseline methods and OpaqueMU on machine unlearning tasks with 5% forget ratio. F: Forget-set; R: Retain-set; VQA: Vision Question Answer; QA: Question Answer; VGEN: Vision Generation; GEN: Generation. The same abbreviations are used for subsequent tables.

Task	F-VQA	R-VQA	F-VGEN		R-VGEN		F-QA	R-QA	F-GEN		R-GEN	
Metric	Acc(↓)	Acc(↑)	Acc(↓)	Rouge(↓)	Acc(↑)	Rouge(↑)	Acc(↓)	Acc(↑)	Acc(↓)	Rouge(↓)	Acc(↑)	Rouge(↑)
MLLMU-Bench												
Vanilla	36.80%	40.46%	24.00%	0.5200	24.74%	0.4989	45.60%	42.11%	14.00%	0.5965	17.05%	0.6201
Grad_Diff	44.80%	40.08%	24.00%	0.2400	20.42%	0.2042	46.40%	42.19%	16.00%	0.1600	14.00%	0.1400
MANU	36.00%	39.62%	22.00%	0.5163	26.21%	0.4988	41.60%	40.13%	41.60%	0.5963	40.13%	0.6210
ICUL	46.40%	48.46%	18.00%	0.3939	19.68%	0.3765	20.80%	22.31%	12.80%	0.3824	14.00%	0.4043
δ-unlearning	36.80%	40.46%	16.00%	0.4618	20.76%	0.4766	44.00%	41.98%	12.80%	0.5863	15.20%	0.6136
OpaqueMU	32.80%	41.60%	18.00%	0.4261	24.42%	0.4960	31.20%	36.97%	6.00%	0.5372	12.95%	0.5770
CLEAR												
Vanilla	95.21%	98.23%	-	0.3505	-	0.3283	90.43%	91.14%	-	0.3774	-	0.3803
Grad_Diff	92.02%	93.92%	-	0.2282	-	0.2301	73.40%	66.58%	-	0.3655	-	0.3676
MANU	95.74%	97.72%	-	0.3421	-	0.3268	88.83%	93.42%	-	0.3676	-	0.3732
ICUL	90.43%	90.96%	-	0.1910	-	0.1907	96.81%	93.92%	-	0.5072	-	0.4556
δ-unlearning	95.21%	96.96%	-	0.3401	-	0.3277	90.43%	91.14%	-	0.3768	-	0.3798
OpaqueMU	92.55%	95.74%	-	0.2057	-	0.2374	89.89%	94.64%	-	0.3711	-	0.4036

4.2 Distribution Alignment

While MS disentangles modality-specific representations, DA performs lightweight logit-level alignment to correct cross-modal behavior. Inspired by post-processing techniques Yang & Klein (2021); Hiranandani et al. (2025), DA operates directly on output logits to reweight probabilities and adapt to distributional shifts. Importantly, DA is implemented entirely on the client side to ensure compatibility with MLLM APIs.

Let $\mathcal{M}'$ denote the multimodal manifold after sanitization, and $\Delta^{|\mathcal{V}|}$ the vocabulary simplex. The frozen model induces a reference output manifold $\mathcal{M}_{\text{out}} \subset \Delta^{|\mathcal{V}|}$, while the sanitized inputs yield a shifted manifold $\mathcal{M}'_{\text{out}}$. DA learns a corrective operator $\mathcal{R}_\psi : \mathcal{M}'_{\text{out}} \to \mathcal{M}_{\text{out}}$ to realign output distributions on retain data via a lightweight plugin $\mathcal{T}_\psi$.

Training Given retain and forget sets $\mathbb{D}^r, \mathbb{D}^f$, each sample (I, T) is processed by the server to generate an output distribution $\mathbf{b} = M_\theta(\mathcal{E}(I, T))$. On the client, we compute $\tilde{z} = \mathcal{F}_\phi(\mathcal{E}(I, T))$ and predict a scaling vector $\mathbf{r} = \mathcal{T}_\psi\big(M_\theta(\tilde{z})\big)$. The aligned distribution is:

$$p_i = \frac{b_i \odot r_i}{\|b_i \odot r_i\|_1}, \qquad i = 1, \ldots, m. \tag{15}$$

We optimize retention with cross-entropy and enforce forgetting via KL divergence:

$$\arg\min_{\psi} \mathcal{L}_{\text{DA}} = -\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{|\mathcal{V}|} \log(\mathbf{p}_i^r) \odot \mathbf{y}_j^r, + \tau \frac{\tau}{m}\sum_{i=1}^{m}\sum_{j=1}^{|\mathcal{V}|} u_j \log \frac{u_j}{p_{i,j}^f}, \tag{16}$$

where $u_j = \frac{1}{|\mathcal{V}|}, \quad j = 1, \ldots, |\mathcal{V}|$, which is uniform over $\mathcal{V}$, $\mathbf{y}_j \in \{0, 1\}^{|\mathcal{V}|}$ denotes the one-hot encoding of the target token. Only ψ is updated while M_θ is frozen.

Inference At inference time, the unlearned MLLM produces the base distribution $\mathbf{b}$, and $\mathcal{T}_\psi$ outputs the correction $\hat{\mathbf{r}}$. Token generation proceeds autoregressively as:

$$x_t = \arg\max_{v \in \mathcal{V}} b_t(v)\, \hat{r}_t(v), \tag{17}$$

thereby aligning the output distribution with the reference output manifold on retained knowledge.

5 Experiments

To assess the effectiveness of OpaqueMU, we evaluate it on Qwen3-VL-8B-Instruct Yang et al. (2025); performance on other models is shown in Appendix C.3. Experiments are conducted

on two dedicated multimodal unlearning benchmarks, MLLMU-Bench Liu et al. (2025a) and CLEAR Dontsov et al. (2024), which provide structured *forget* and *retain* splits across diverse multimodal tasks, including visual question answering (VQA) and text-based QA, covering both generative and discriminative settings. We compare `OpaqueMU` with four strong baselines: Grad_Diff Liu et al. (2022) (GD, a representative MU method), MANU Liu et al. (2025b) (an MLLM-specific unlearning approach), two black-box unlearning methods ICUL Pawelczyk et al. (2024), and δ-unlearning Huang et al. (2025) (δ-unl.). Vanilla denotes the original model without unlearning. All methods are trained under identical configurations for fair comparison. Following standard protocol, MLLMU-Bench employs forget-set ratios of 5%, 10%, and 15%, while CLEAR uses 1%, 5%, and 10%. Hyperparameter configurations are provided in Appendix C.1.

5.1 Main Results

Table 2 reports unlearning results under a 5% forget ratio on MLLMU-Bench and CLEAR, evaluating both forgetting effectiveness and utility preservation across classification and generation tasks. For MLLMU-Bench generation, we report Acc and Rouge, while CLEAR is evaluated using Rouge due to its open-ended QA setting. On MLLMU-Bench, `OpaqueMU` achieves the best balance between forgetting and retention. In multimodal VQA, it reduces forget-set accuracy to 32.8%, outperforming Vanilla (36.8%) and MANU (36.0%), while maintaining competitive retain accuracy (41.6%). For vision generation, `OpaqueMU` lowers F-VGEN Rouge to 0.4261 with only minor impact on R-VGEN, indicating effective forgetting without utility collapse. In contrast, white-box methods either fail to forget (e.g., MANU on F-VGEN) or degrade retain performance (e.g., Grad_Diff on R-VGEN). Text-based black-box methods are unstable in multimodal settings: ICUL achieves strong forgetting but causes severe retain degradation, especially in QA and GEN. On CLEAR, `OpaqueMU` continues to suppress forget-set performance (e.g., F-VQA 92.55% vs. 95.21% for Vanilla) while preserving retain utility (R-VQA 95.74%, R-GEN 0.4036), avoiding the retain collapse seen in ICUL and Grad_Diff. Additional results are provided in Appendix C.3.

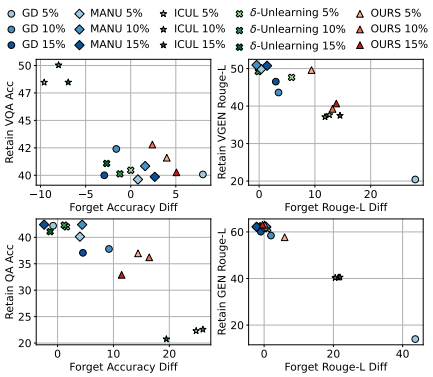


Figure 3: The overall trade-off between unlearning effectiveness and model utility across four dimensions under varying forget ratios on MLLMU-Bench.

Utility Trade-off To assess the trade-off between targeted forgetting and utility preservation, we visualize forget-set performance drop versus retain-set accuracy on MLLMU-Bench across four tasks (VQA, VGEN, QA, and GEN), where higher values on both axes indicate stronger forgetting and better retention, respectively. As shown in Fig. 3, `OpaqueMU` consistently occupies the favorable upper-right region across tasks and forget ratios, indicating effective forgetting with limited utility loss. This advantage is most evident in multimodal tasks (VQA and VGEN): while baselines such as ICUL achieve strong forgetting, they incur substantial retain-set degradation, reflecting over-aggressive unlearning. In contrast, `OpaqueMU` maintains stable retain performance while steadily improving forgetting strength. In the GEN task, the trade-off gap appears less pronounced. This is mainly due to the limited sensitivity of Rouge-L, which focuses on surface-level semantic overlap and may not fully capture forget-set-related changes, resulting in compressed performance differences across methods. More results of the utility trade-off are reported in Appendix C.4.

5.2 Ablation Studies

Table 3 presents ablation results of `OpaqueMU` on MLLMU-Bench with a 5% forget ratio, highlighting the complementary roles of MS and DA. Removing DA (Ours (*MS*)) leads to weaker forgetting and unstable retain performance, particularly in generation tasks, indicating that MS alone cannot realign the output distribution after modal perturbations and leaves forget and retain samples entangled in the downstream representation space. In contrast, using DA alone (Ours (*DA*)) yields lim-

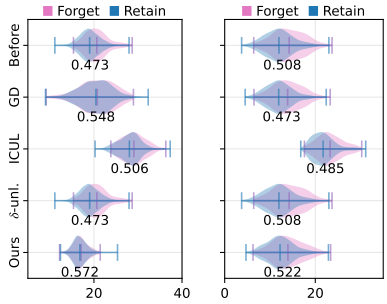
Table 3: Ablation studies of OpaqueMU on MLLMU-Bench with 5% forget ratio by Qwen3-VL.

Task	F-VQA	R-VQA	F-VGEN		R-VGEN		F-QA	R-QA	F-GEN		R-GEN	
Metric	Acc(↓)	Acc(↑)	Acc(↓)	Rouge(↓)	Acc(↑)	Rouge(↑)	Acc(↓)	Acc(↑)	Acc(↓)	Rouge(↓)	Acc(↑)	Rouge(↑)
Ours	32.80%	41.60%	18.00%	0.4261	24.42%	0.4960	31.20%	36.97%	6.00%	0.5372	12.95%	0.5770
Ours (*MS*)	34.40%	39.07%	20.00%	0.5163	24.95%	0.4924	41.60%	44.21%	14.00%	0.6122	18.11%	0.6137
Ours (*DA*)	35.12%	36.04%	19.47%	0.3534	20.00%	0.3836	31.20%	33.60%	6.00%	0.4994	9.37%	0.5294
Ours (*Offset*)	24.00%	28.00%	17.26%	0.3814	18.00%	0.3515	6.40%	5.77%	0.00%	0.2104	0.42%	0.2066
Ours (*Proj*)	30.40%	31.12%	16.00%	0.3882	15.79%	0.3527	43.20%	40.84%	6.00%	0.5500	11.96%	0.5775
0-Shot (*DA*)	21.60%	19.87%	0.00%	0.1938	1.37%	0.2030	35.20%	33.26%	10.00%	0.3356	4.32%	0.2794

ited forgetting gains and noticeably degraded retain accuracy, especially on R-VQA, suggesting that output-level manipulation is ineffective when forget-specific signals remain embedded in unchanged multimodal representations. Applying uniform offsets at the MS (Ours (*Offset*)) causes severe utility collapse across retain tasks, as indiscriminate perturbations disrupt instruction semantics, confirming that naive offset-based MS lacks the selectivity required for safe unlearning. Replacing MS with explicit visual projection moderately (Ours (*Proj*)) improves forgetting in some cases but significantly harms retain performance, particularly for VGEN and GEN, which can be attributed to distortion of visual spatial structure when the projector is sensitive to patch order or positional encoding. The zero-shot DA variant (0-Shot (*DA*)) performs worst overall, exhibiting trivial forgetting and near-random retain accuracy, demonstrating that DA critically relies on learned, representation-aligned transformations to induce effective and stable unlearning.

5.3 Manifold Distribution

We analyze the geometric effect of unlearning by examining how forget and retain samples distribute along a discriminative direction in the multimodal representation. Detailed settings are provided in Appendix C.2. This direction is constructed from the pre-unlearning model to capture the principal axis separating forget and retain data, and is fixed for all methods to ensure a fair comparison. We project representations onto this axis and visualize the resulting distributions using violin plots. Effective unlearning is reflected by strong overlap between forget and retain distributions, indicating the removal of sensitive information while preserving retain semantics. As shown in Fig. 4, OpaqueMU achieves the highest distribution overlap, with overlap ratios of 0.572 on MLLMU-Bench and 0.522 on CLEAR.



(a) MLLMU-Bench (b) CLEAR

Figure 4: Distribution shift of forget manifold toward retain manifold before and after unlearning across different baselines.

5.4 Cost

We evaluate the time and memory cost of OpaqueMU and baselines on MLLMU-Bench for both training and inference in Fig 5. Training costs for all trainable methods are negligible (nearly 0.2 s per epoch) with similar memory usage, while inference reveals notable differences: OpaqueMU achieves 0.61 s latency and under 19 GB memory, compared to 1.82 s and nearly 20 GB for ICUL and over 2 s for other baselines, providing roughly a 66% speedup over the strongest baseline. In real scenarios, where models are repeatedly queried after forgetting, inference efficiency is more critical than training cost, and OpaqueMU offers a favorable trade-off between runtime and memory usage. The communication overhead both in training and inference phase are reported in Table 4, including multimodal embeddings uploaded during manifold sanitization and logits transmitted during distribution alignment. Under Qwen3-VL-8B, total transmitted data per sample for the inference is below 10 MB for both MLLMU-Bench and CLEAR, demonstrating that OpaqueMU remains practical under standard network bandwidth constraints.

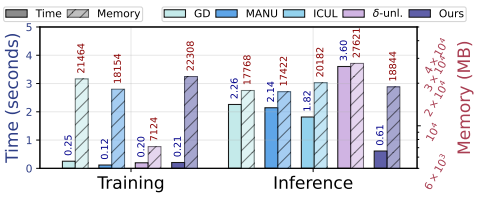


Figure 5: Comparison of time and memory overhead between OpaqueMU and baseline methods on MLLMU-Bench.

Dataset	MLLMU (tokens≈97)		CLEAR (tokens≈121)	
Train	MS (Embs) 776KB	DA (Logits) 28784.75KB	MS (Embs) 968KB	DA (Logits) 35906.75KB
Inference	MS (Embs) 702.16KB	DA (Logits) 2804.29KB	MS (Embs) 657.76KB	DA (Logits) 9225.96KB

Table 4: The communication overhead of OpaqueMU on MLLMU-Bench and CLEAR datasets.

Table 5: Success rates of unlearning attacks against different baselines on MLLMU-Bench.

Attack	Rewrite			Induce		
Ratio	5%	10%	15%	5%	10%	15%
Vanilla	27.00%	19.50%	24.33%	34.00%	33.00%	35.67%
GD	16.00%	14.00%	20.00%	24.00%	27.50%	30.00%
MANU	23.00%	19.50%	23.67%	30.00%	34.50%	30.33%
δ-unl.	15.00%	11.0%	18.67%	21.00%	21.00%	29.67%
ICUL	16.00%	18.50%	20.67%	26.00%	36.50%	34.33%
Ours	**7.00%**	**9.00%**	**12.67%**	**19.00%**	**14.50%**	**10.67%**

Table 6: Membership inference performance on MLLMU-Bench. Lower AUC/ACC indicates weaker membership signals and better privacy protection.

Method	Vanilla	GD	MANU	ICUL	δ-unl.	Ours
ACC	71.65%	49.51%	64.01%	78.15%	59.87%	**41.73%**
AUC	78.36%	56.87%	77.30%	85.61%	55.74%	**50.15%**
Forget-ACC	84.20%	88.74%	52.92%	94.40%	87.77%	**33.98%**
Retain-ACC	82.83%	85.70%	54.31%	81.92%	88.67%	**34.09%**
Non Member	54.46%	91.40%	78.97%	68.69%	66.91%	**54.46%**

5.5 Robustness

Unlearning Attack. We evaluate the robustness of OpaqueMU and baseline methods against unlearning attacks on MLLMU-Bench. Specifically, we consider two attack strategies: Rewrite Xu et al. (2024) and Induce Li et al. (2023a), which attempt to recover forgotten information by rewriting queries or inducing targeted outputs from the unlearned model. Since CLEAR consists of open-ended QA and does not support these attack formulations, evaluations are conducted on MLLMU-Bench under different forget ratios. As shown in Table 5, OpaqueMU consistently achieves the lowest attack success rates across all forget ratios and attack types. Under Rewrite attacks, the success rate of OpaqueMU remains below 13% even at a 15% forget ratio, substantially outperforming all baselines. A similar trend is observed for Induce attacks, where OpaqueMU reduces attack success to 10.67% at 15%, while other methods remain significantly more vulnerable. These results demonstrate that OpaqueMU provides stronger resistance to unlearning attacks, effectively preventing the recovery of forgotten information.

Membership Inference. We evaluate privacy protection using a black-box membership inference attack, where the attacker observes only output logits. Forget and retain members are drawn from the train set, while non-members are held-out samples. A fixed attack classifier is trained on the original model logits and evaluated after applying unlearning. As shown in Table 6, MANU exhibits substantial leakage, with high attack accuracy (64.01%) and AUC (77.30%), indicating training samples are easily distinguishable. Grad_Diff and δ-unlearning reduce overall AUC to 56.87% and 55.74%, but forget-member attack success remains high (88.74% and 87.77%), suggesting non-selective removal of membership signals. In contrast, OpaqueMU achieves the lowest MIA performance, with overall AUC near random (50.15%) and forget-member attack success dropping to 33.98%, demonstrating effective suppression of sensitive membership signals under a black-box threat model. Additional results are provided in Appendix C.6 to validate the robustness of OpaqueMU under stronger MIAs.

6 Conclusion

We study multimodal unlearning under a split-access LMaaS setting and propose OpaqueMU, which coordinates representation-level sanitization with output-level alignment while keeping the Transformer backbone opaque and unchanged. Extensive experiments demonstrate effective targeted forgetting with preserved utility, providing a practical approach to multimodal unlearning under restricted model access.

AI USE STATEMENT

In this work, we used generative AI tools solely for language editing and polishing, including improving grammar, clarity, and readability of the manuscript. We did not use generative AI tools to generate research ideas, formulate scientific claims, conduct experiments, analyze data, or produce experimental results. All AI-assisted edits were reviewed and verified by the authors to ensure that they did not alter the intended technical meaning or scientific claims. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

LIMITATIONS

`OpaqueMU` is designed on the premise that the modality representation mapping is client-side and trainable. This premise acts as a key foundation for inserting the Manifold Sanitization module before cross-modal fusion, yet it is not satisfied in commercial multimodal services where representation mappings are coupled with the backbone model and only HTTP interfaces are available. In addition, our current evaluation focuses on standard multimodal benchmarks and predefined unlearning splits. Important future research directions include exploring lightweight adaptive versions of OpaqueMU for low-access commercial multimodal service scenarios to expand its practical application scope.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013. ISSN 0162-8828.

Junkai Chen, Zhijie Deng, Kening Zheng, Yibo Yan, Shuliang Liu, PeiJun Wu, Peijie Jiang, Jia Liu, and Xuming Hu. SafeEraser: Enhancing safety in multimodal large language models through multimodal machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14194–14224. Association for Computational Linguistics, 2025.

Jiali Cheng and Hadi Amiri. Multidelete for multimodal machine unlearning. In *European Conference on Computer Vision*, pp. 165–184. Springer, 2024.

Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18:2345–2354, 2023.

Alexey Dontsov, Dmitrii Korzh, Alexey Zhavoronkin, Boris Mikheev, Denis Bobkov, Aibek Alanov, Oleg Y Rogov, Ivan Oseledets, and Elena Tutubalina. Clear: Character unlearning in textual and visual modalities. *arXiv preprint arXiv:2410.18057*, 2024.

Gaurush Hiranandani, Haolun Wu, Subhojyoti Mukherjee, and Sanmi Koyejo. Logits are all we need to adapt closed models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 23261–23289. PMLR, 2025.

James Y Huang, Wenxuan Zhou, Fei Wang, Fred Morstatter, Sheng Zhang, Hoifung Poon, and Muhao Chen. Offset unlearning for large language models. *Transactions on Machine Learning Research*, 2025.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. Mmunlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

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.

Minhyeok Lee. The geometry of feature space in deep learning models: A holistic perspective and comprehensive review. *Mathematics*, 11(10):2375, 2023.

Haodong Li, Gelei Deng, Yi Liu, Kailong Wang, Yuekang Li, Tianwei Zhang, Yang Liu, Guoai Xu, Guosheng Xu, and Haoyu Wang. Digger: Detecting copyright content mis-usage in large language model training. *CoRR*, 2024a.

Haoran Li, Dadi Guo, Wei Fan, Mingshi Xu, Jie Huang, Fanpu Meng, and Yangqiu Song. Multi-step jailbreaking privacy attacks on chatgpt. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4138–4153, 2023a.

Jiaqi Li, Qianshan Wei, Chuanyi Zhang, Guilin Qi, Miaozeng Du, Yongrui Chen, Sheng Bi, and Fan Liu. Single image unlearning: Efficient machine unlearning in multimodal large language models. *Advances in Neural Information Processing Systems*, 37:35414–35453, 2024b.

Kunhao Li, Wenhao Li, Di Wu, Lei Yang, Jun Bai, Ju Jia, and Jason Xue. Cross-modal unlearning via influential neuron path editing in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35589–35597, 2026.

Yansong Li, Zhixing Tan, and Yang Liu. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212*, 2023b.

Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pp. 243–254. PMLR, 2022.

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, 2025a.

Zheyuan Liu, Guangyao Dou, Xiangchi Yuan, Chunhui Zhang, Zhaoxuan Tan, and Meng Jiang. Modality-aware neuron pruning for unlearning in multimodal large language models. *Proceedings of the 63nd Annual Meeting of the Association for Computational Linguistics*, 2025b.

Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. An adversarial perspective on machine unlearning for ai safety. *Transactions on Machine Learning Research*, 2025.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. Split-and-denoise: protect large language model inference with local differential privacy. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 34281–34302, 2024.

Lingchen Meng, Jianwei Yang, Rui Tian, Xiyang Dai, Zuxuan Wu, Jianfeng Gao, and Yu-Gang Jiang. Deepstack: Deeply stacking visual tokens is surprisingly simple and effective for lmms. *Advances in Neural Information Processing Systems*, 37:23464–23487, 2024.

Quoc Phong Nguyen, Bryan Kian, Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 16025–16036, 2020.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: language models as few-shot unlearners. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 40034–40050, 2024.

Renjie Pi, Tianyang Han, Jianshu Zhang, Yueqi Xie, Rui Pan, Qing Lian, Hanze Dong, Jipeng Zhang, and Tong Zhang. Mllm-protector: Ensuring mllm's safety without hurting performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16012–16027, 2024.

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)*, pp. 303–319. IEEE, 2022.

Zijie Wang, Aichun Zhu, Jingyi Xue, Xili Wan, Chao Liu, Tian Wang, and Yifeng Li. Look before you leap: Improving text-based person retrieval by learning a consistent cross-modal common manifold. In *Proceedings of the 30th ACM international conference on multimedia*, pp. 1984–1992, 2022.

Panlong Wu, Qi Liu, Yanjie Dong, Zhaorui Wang, and Fangxin Wang. Lmaas: Exploring pricing strategy of large model as a service for communication. *IEEE Transactions on Mobile Computing*, 23(12):12748–12760, 2024.

Shangyu Xing, Fei Zhao, Zhen Wu, Tuo An, Weihao Chen, Chunhui Li, Jianbing Zhang, and Xinyu Dai. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1167–1181, 2024.

Xilie Xu, Keyi Kong, Ning Liu, Lizhen Cui, Di Wang, Jingfeng Zhang, and Mohan Kankanhalli. An llm can fool itself: A prompt-based adversarial attack. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Kevin Yang and Dan Klein. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3511–3535. Association for Computational Linguistics, 2021.

Tianyu Yang, Lisen Dai, Zheyuan Liu, Xiangqi Wang, Meng Jiang, Yapeng Tian, and Xiangliang Zhang. Cliperase: Efficient unlearning of visual-textual associations in clip. *arXiv preprint arXiv:2410.23330*, 2024.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. Mm-llms: Recent advances in multimodal large language models. In *ACL (Findings)*, 2024a.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling*, 2024b.

A Algorithm Details

A.1 Details of Discriminator

The discriminator $\mathcal{D}(\cdot)$ is a lightweight binary classifier used to distinguish forget and retain representations in the joint embedding space. It supports both token-level inputs (B, L, d) and pooled embeddings (B, d). For sequence inputs, a shallow Transformer encoder is applied followed by mean pooling, while pooled inputs bypass the encoder directly.

Table 7: Architecture of the Adversarial Discriminator $\mathcal{D}(\cdot)$.

Component	**Specification**	**Output Shape**
Input representation	Token sequence or pooled embedding	(B, L, d) or (B, d)
Sequence encoder:		
Transformer Encoder	1 layer, 1 head, FFN dim 512, GELU, dropout 0.1	(B, L, d)
Mean Pooling	Adaptive average pooling over sequence length ($L \to 1$)	(B, d)
Classifier head:		
Spectral Linear	Input/Output dim d (with spectral normalization)	(B, d)
LayerNorm + GELU	Feature normalization and nonlinearity	(B, d)
Linear	Output dim 1	$(B, 1)$

A.2 Details of distribution alignment module

The distribution alignment module $\mathcal{R}_\psi$ contains a lightweight plugin language model $\mathcal{T}_\psi$ that operates on transformed multimodal embeddings. It consists of a dimensionality reduction layer, a shallow frozen language-model block, and a trainable output projection head. Only the projection layers and a single transformer layer are updated during training, while the backbone parameters remain frozen.

Table 8: Architecture of the plugin language model $\mathcal{T}_\psi$ in $\mathcal{R}_\psi$.

Component	**Specification**	**Output Shape**
Input embeddings	Multimodal hidden states from manifold sanitization	(B, L, d_{base})
Down Projection	Linear: $d_{\text{base}} \to d$	(B, L, d)
Shallow language model (partially trainable):		
Transformer Layers	N layers, hidden size d, h heads, FFN dim f	(B, L, d)
Output head:		
LM Head	Linear: $d \to \lvert\mathcal{V}\rvert$ (no bias)	$(B, L, \lvert\mathcal{V}\rvert)$

A.3 Workflow of training OpaqueMU

Algorithm 1: Workflow of `OpaqueMU`

Input: Frozen MLLM M_θ, multimodal representation mapping $\mathcal{E} = (\mathcal{E}_V, \mathcal{E}_T)$;
forget set $\mathbb{D}_f$, retain set $\mathbb{D}_r$; hyperparameters k
Output: Sanitized representation mapping $\mathcal{E}'_V$, subspace U_t, DA module $\mathcal{T}_\psi$

`/* Phase 1: Manifold Sanitization */`

for *each batch from* $\mathbb{D}_f \cup \mathbb{D}_r$ **do**

$h_v \leftarrow \mathcal{E}_V(I), \quad h_t \leftarrow \mathcal{E}_T(T)$

$h'_v \leftarrow h_v + \alpha(\mathcal{E}'_V(I) - \mathcal{E}_V(I))$

$U_t \leftarrow \text{TopKEig}(h_t), \quad h'_t \leftarrow h_t - \beta(h_t U_t)U_t^\top$

$U_t \leftarrow \text{QR}(U_t)$

$H \leftarrow (1 - M) \odot h'_t + M \odot \Pi h'_v$

$\mathcal{L}_{\text{MS}} \leftarrow \mathcal{L}_{\text{adv}} + \mathcal{L}_{\text{retain}} + \mathcal{L}_{\text{reg}}$

Update $\theta_{\mathcal{E}'}, U_t$

`/* Phase 2: Distribution Alignment */`

for *each* $(I_f, T_f) \sim \mathbb{D}_f,\ (I_r, T_r) \sim \mathbb{D}_r$ **do**

$p_f \leftarrow \mathcal{P}_\psi(I_f, T_f), \quad p_r \leftarrow \mathcal{P}_\psi(I_r, T_r)$

$\mathcal{L}_{\text{DA}} \leftarrow \mathcal{L}^r_{\text{CE}}(p_r) + \tau\mathcal{L}^f_{\text{KL}}(p_f)$

Update ψ

return $\mathcal{E}'_V, U_t, \mathcal{T}_\psi$

We provide the specific process for training `OpaqueMU`, including Manifold Sanitization and Distribution Alignment, as shown in Algorithm 1

A.4 Cost Analysis

Time Cost For the visual modality, a lightweight visual offset representation mapping $\mathcal{E}'_V$ computes the visual representation with complexity $\mathcal{O}(B\,N_p\,d_v)$, where N_p is the number of image patches and d_v the visual token dimension. For the textual modality, the generalized eigendecomposition used to initialize the forget subspace incurs a one-time preprocessing cost of $\mathcal{O}\big((N_f + N_r)d_t^2 + d_t^3\big)$, which is negligible in practice.

During training, the textual projection module operates with complexity $\mathcal{O}(B\,d_t\,k)$, where $k \ll d_t$. We can roughly align textual and visual embeddings into a shared space ($d = d_t = d_v$), the overall training complexity of the manifold sanitization module is $\mathcal{O}(B\,d)$.

The distribution alignment module is a lightweight adapter (e.g., a shallow Transformer or MLP) applied to logits, with training complexity $\mathcal{O}(B\,L_{\text{out}}\,r)$, r denotes the adaptation dimension where $r \ll \mathcal{V}$ is the adapter rank and $\mathcal{V}$ the vocabulary size.

Thus, the incremental training cost introduced by `OpaqueMU` is

$$\mathcal{O}(B\,d + B\,L_{\text{out}}\,r).$$

At inference time, the per-sample overhead reduces to $\mathcal{O}(d)$ for manifold sanitization and $\mathcal{O}(L_{\text{out}}\,r)$ for distribution alignment, yielding an overall incremental inference complexity of

$$\mathcal{O}(d + L_{\text{out}}\,r),$$

which is negligible compared to the frozen MLLM backbone.

Space Cost During training, client-side memory consists of trainable parameters, optimizer states, and intermediate activations. The dominant trainable components include: (i) the textual projection matrix $U \in \mathbb{R}^{d_t \times k}$, (ii) the visual offset representation mapping with parameter complexity $P_v = \mathcal{O}(L_v d_v^2)$, where L_v is the number of layers, and (iii) the distribution alignment module of size $\mathcal{O}(rd_t)$.

Accounting for AdamW optimizer states (which store two momentum buffers per parameter), the total parameter-related memory scales as $\mathcal{O}(d_t k + L_v d_v^2 + rd_t)$.

Intermediate activations and gradient buffers introduce an additional memory cost of

$$\mathcal{O}\big(B(d + L_{\text{out}} r)\big),$$

where B is the batch size and $d = \max(d_t, d_v)$.

Consequently, the overall client-side space complexity during training is

$$\mathcal{O}\big(d_t k + L_v d_v^2 + rd_t + B(d + L_{\text{out}} r)\big).$$

At inference time, optimizer states and gradient buffers are discarded, leaving only model parameters and transient activations. The resulting memory footprint reduces to

$$\mathcal{O}\big(d_t k + L_v d_v^2 + rd_t\big),$$

with activation costs being negligible.

Communication Cost In the split-access LMaaS setting, `OpaqueMU` communicates with the server in two directions per query. After manifold sanitization, the client uploads the multimodal embedding sequence $H \in \mathbb{R}^{L \times d}$, where L is the total input token length (text + visual placeholders) and d the unified embedding dimension. The server returns the full logits $\mathbf{z} \in \mathbb{R}^{L_{\text{out}} \times \mathcal{V}}$.

Assuming 16-bit floating-point representation, the uplink and downlink communication costs are $\mathcal{O}(Ld)$ and $\mathcal{O}(L_{\text{out}}\mathcal{V})$ bytes, respectively, yielding a total per-query communication complexity of

$$\mathcal{O}(Ld + L_{\text{out}}\mathcal{V}).$$

B Proof of Effectiveness of OpaqueMU

B.1 Effectiveness of Manifold Sanitization

Assumptions. Let $\mathcal{M}$ be the latent representation manifold with disjoint forget $\mathcal{M}^f$ and retain $\mathcal{M}^r$ submanifolds. $\mathcal{F}_\phi$ denotes the Manifold Sanitization module. Assume:

1. **Adversarial loss:** $\mathcal{L}_{\text{adv}}$ is minimized to ϵ_{adv}.
2. **Retain consistency loss:** $\mathcal{L}_{\cos}$ penalizes angular deviation via cosine similarity.
3. **Projection stability:** U_t satisfies $\|U_t^\top U_t - I_k\|_F^2 \le \eta$.

Notation. Let $H_f = \mathcal{F}_\phi(H_f)$, $H_r = \mathcal{F}_\phi(H_r)$, $d_{\mathcal{M}}(\cdot,\cdot)$ denote the geodesic distance, $v \in T_x\mathcal{M}$ tangent vectors, and Φ_1, Φ_2, Ψ are monotone functions that map adversarial, retain, and projection errors to corresponding manifold-distance bounds, with $\Phi_1(0) = \Phi_2(0) = 0$ and $\Psi(\eta) \to 0$ as $\eta \to 0$.

Lemma B.1 (Adversarial Separation)**.**

$$d_{\mathcal{M}}(H_f, \mathcal{M}^f) \geq \Phi_1(\mathcal{L}_{\text{adv}}), \quad \forall x \in \mathcal{M}^f.$$

Proof. Let $\mathcal{D}$ be the adversary. By definition,

$$\mathcal{L}_{\text{adv}} = -\mathbb{E}_{x\sim\mathcal{M}^f} \log \mathcal{D}(f_\theta(x)) - \mathbb{E}_{x\sim\mathcal{M}^r} \log(1 - \mathcal{D}(f_\theta(x))).$$

Since $\mathcal{D}$ is optimal for $\mathcal{F}_\phi$:

$$\mathcal{D}^*(\mathcal{F}_\phi(x)) = \frac{p_f(\mathcal{F}_\phi(x))}{p_f(\mathcal{F}_\phi(x)) + p_r(\mathcal{F}_\phi(x))}.$$

where p_f, p_r are the induced densities on $\mathcal{M}$. Then

$$\mathcal{L}_{\text{adv}} = \text{JS}\big(p_f(\mathcal{F}_\phi(x)) \| p_r(\mathcal{F}_\phi(x))\big),$$

where JS is Jensen-Shannon divergence. By standard manifold assumptions Bengio et al. (2013):

$$d_{\mathcal{M}}(H_f, \mathcal{M}^f) \geq \Phi_1(\text{JS}(p_f \| p_r)) = \Phi_1(\mathcal{L}_{\text{adv}}).$$

□

Lemma B.2 (Retain Preservation)**.**

$$d_{\mathcal{M}}(H_r, H_r^{orig}) \leq \Phi_2(\mathcal{L}_{\cos}).$$

Proof. Cosine-based retain loss:

$$\mathcal{L}_{\cos} = \frac{1}{m} \sum_i \big(1 - \cos \angle(H_r(x_i), H_r^{\text{orig}}(x_i))\big).$$

Let $\theta_i = \angle(H_r(x_i), H_r^{\text{orig}}(x_i))$, then

$$\cos\theta_i = 1 - \delta_i, \quad \delta_i \leq \mathcal{L}_{\cos}.$$

By law of cosines in tangent space:

$$d_{\mathcal{M}}(H_r(x_i), H_r^{\text{orig}}(x_i))^2 \leq 2\|H_r^{\text{orig}}(x_i)\|^2(1 - \cos\theta_i) \leq 2\|H_r^{\text{orig}}(x_i)\|^2 \mathcal{L}_{\cos}.$$

Taking the maximum over i gives the bound:

$$d_{\mathcal{M}}(H_r, H_r^{\text{orig}}) \leq \Phi_2(\mathcal{L}_{\cos}).$$

□

Lemma B.3 (Local Stability)**.**

$$(1 - \Psi(\eta))\|v\| \leq \|U_t v\| \leq (1 + \Psi(\eta))\|v\|, \quad v \in T_x\mathcal{M}.$$

Proof. From the Frobenius bound:

$$\|U_t^\top U_t - I_k\|_F^2 \leq \eta \implies \|v^\top (U_t^\top U_t - I_k)v\| \leq \eta\|v\|^2.$$

Then

$$v^\top U_t^\top U_t v = \|U_t v\|^2 = \|v\|^2 + v^\top (U_t^\top U_t - I_k)v \in [(1-\eta)\|v\|^2, (1+\eta)\|v\|^2].$$

Taking square root yields

$$(1 - \Psi(\eta))\|v\| \leq \|U_t v\| \leq (1 + \Psi(\eta))\|v\|,$$

where $\Psi(\eta) = \sqrt{1+\eta} - 1$. □

B.2 Effectiveness of Distribution Alignment

Assumptions. Let the distribution alignment $\mathcal{R}_\psi$ satisfy:

1. **Smoothness:** $\mathcal{R}_\psi$ is Lipschitz continuous.
2. **Retain generalization:** On $\mathbb{D}^r$, the mapping achieves error ϵ_{gen}.
3. **KL-based forgetting:** Forgetting is implemented via $\mathcal{L}_{\text{DA}}^f = \frac{\tau}{m}\sum_i \text{KL}(u\|p_i^f)$ with uniform prior u.

Notation. Let

$$T_\psi : b \mapsto \Pi_\Delta(b \odot \mathcal{R}_\psi(b)), \quad p_i = T_\psi(b_i), \quad d_F \text{ denotes Fisher-Rao distance.}$$

Lemma B.4 (Retain Fidelity)**.**

$$d_F(p_i, y_i) \le C_1\sqrt{\mathcal{L}_{\text{DA}}^r}, \quad \mathcal{L}_{\text{DA}}^r = \frac{1}{m}\sum_i \text{KL}(y_i\|p_i), \quad C_1 = \sqrt{2}.$$

Proof.

$$\begin{aligned} \text{KL}(y_i\|p_i) &= -\log p_i[y_i] \quad \text{(one-hot target)} \\ d_F^2(p_i, y_i) &\le 2\,\text{KL}(y_i\|p_i) \quad \text{(KL-Fisher duality)} \\ \implies d_F(p_i, y_i) &\le \sqrt{2\,\mathcal{L}_{\text{DA}}^r} = C_1\sqrt{\mathcal{L}_{\text{DA}}^r}. \end{aligned}$$

□

Lemma B.5 (Sensitive Suppression)**.** *For $j \in \mathcal{M}^f$, $k \in \mathcal{M}^r$:*

$$\frac{p_k}{p_j} \ge \frac{b_k}{b_j}e^\mu, \quad \text{dist}_F(p, \mathcal{M}_{\text{out}}^f) \ge \delta(\mu, \gamma),$$

with $\mu = \min_{j,k}\log(r_k/r_j)$ and $\gamma = \text{dist}_F(\mathcal{M}_{\text{out}}^r, \mathcal{M}_{\text{out}}^f)$.

Proof.

$$\frac{p_k}{p_j} = \frac{b_k r_k}{b_j r_j} = \frac{b_k}{b_j}\frac{r_k}{r_j} \ge \frac{b_k}{b_j}e^\mu.$$

Since d_F is continuous:

$$\text{dist}_F(p, \mathcal{M}_{\text{out}}^f) \ge \delta(\mu, \gamma).$$

□

Lemma B.6 (Forget Separation)**.**

$$d_F(p_i^f, u) \le \epsilon_{\text{forget}}(\tau), \quad \epsilon_{\text{forget}}(\tau) = \sqrt{2\,\mathcal{L}_{\text{DA}}^f}.$$

Proof.

$$d_F^2(p_i^f, u) \le 2\,\text{KL}(u\|p_i^f) \le 2\,\mathcal{L}_{\text{DA}}^f \implies d_F(p_i^f, u) \le \sqrt{2\,\mathcal{L}_{\text{DA}}^f}.$$

□

Lemma B.7 (Manifold Approximation)**.**

$$\widetilde{\mathcal{M}}_\psi = T_\psi(\mathcal{M}'_{\text{out}}) \subset \Delta^{|\mathcal{V}|-1}, \quad \text{dist}_F(\widetilde{\mathcal{M}}_\psi, \mathcal{M}_{\text{out}}^r) \le \epsilon_{\text{gen}}.$$

Proof. Lipschitz continuity of $\mathcal{R}_\psi$ and smoothness of Π_Δ imply

$$\|T_\psi(b) - T_\psi(b')\| \le L\|b - b'\|,$$

which bounds the Fisher-Rao distance between the mapped submanifold and $\mathcal{M}_{\text{out}}^r$:

$$\text{dist}_F(T_\psi(\mathcal{M}'_{\text{out}}), \mathcal{M}_{\text{out}}^r) \le \epsilon_{\text{gen}}.$$

□

B.3 Coordinated Effect of Manifold Sanitization and Distribution Alignment

Assumptions. Let the system satisfy:

1. **MLLM Lipschitzness:** M_θ is locally Lipschitz on the retained input manifold $\mathcal{M}_{\text{in}}^r$ with constant L:
$$\|M_\theta(x) - M_\theta(x')\| \le L\|x - x'\|, \quad \forall x, x' \in \mathcal{M}_{\text{in}}^r.$$
2. **Manifold Sanitization Retention:** f_θ satisfies
$$\|\Delta H_i\| = \|\mathcal{F}_\phi(H_r^i) - H_r^i\| \le \Phi_2(\mathcal{L}_{\cos}), \quad H_r^i \in \mathcal{M}^r.$$
3. **Distribution Alignment Generalization:** T_ψ is Lipschitz and
$$\text{dist}_F(T_\psi(\mathcal{M}'_{\text{out}}), \mathcal{M}_{\text{out}}^r) \le \epsilon_{\text{gen}}.$$

Notation.

$\hat{y}_i = T_\psi\big(M_\theta(\mathcal{F}_\phi(x_i))\big)$, $\quad p_i = T_\psi(b_i)$, $\quad b_i = M_\theta(\mathcal{F}_\phi(H_r^i))$, $\quad d_F$ Fisher-Rao distance.

Lemma B.8 (MS Task Deviation)**.** *For $i \in \mathbb{D}^r$,*
$$d_F(M_\theta(\mathcal{F}_\phi(H_r^i)), y_i) \le L\Phi_2(\mathcal{L}_{\cos}).$$

Proof. By Lipschitzness of M_θ:
$$\begin{aligned} d_F(M_\theta(\mathcal{F}_\phi(H_r^i)), y_i) &= d_F(M_\theta(H_r^i + \Delta H_i), M_\theta(H_r^i)) \\ &\le L\|\Delta H_i\| \le L\Phi_2(\mathcal{L}_{\cos}). \end{aligned}$$
□

Lemma B.9 (DA Task Deviation)**.** *For $i \in \mathbb{D}^r$ with untransformed input H_r^i, the distribution alignment satisfies*
$$d_F(T_\psi(M_\theta(H_r^i)), y_i) \le \epsilon_{\text{gen}}.$$

Proof. Since T_ψ is a Lipschitz retraction onto $\widetilde{\mathcal{M}}_\psi$ with $\text{dist}_F(\widetilde{\mathcal{M}}_\psi, \mathcal{M}_{\text{out}}^r) \le \epsilon_{\text{gen}}$, we have directly
$$d_F(T_\psi(M_\theta(H_r^i)), y_i) \le \epsilon_{\text{gen}}.$$
□

Lemma B.10 (Coordinated MS + DA Task Bound)**.** *For $i \in \mathbb{D}^r$, the combined mapping satisfies*
$$d_F(\hat{y}_i, y_i) \le L\Phi_2(\mathcal{L}_{\cos}) + \epsilon_{\text{gen}} - \kappa_i,$$
with cross-term
$$\kappa_i = \frac{1}{2}\big(d_F(M_\theta(f_\theta(H_r^i)), y_i) + d_F(T_\psi(M_\theta(H_r^i)), y_i) - d_F(\hat{y}_i, y_i)\big) \ge 0,$$
quantifying the "coordinated gain".

Proof. Start with triangle inequality and add/subtract $M_\theta(H_r^i)$:
$$\begin{aligned} d_F(\hat{y}_i, y_i) &= d_F(T_\psi(M_\theta(f_\theta(H_r^i))), y_i) \\ &\le d_F(T_\psi(M_\theta(f_\theta(H_r^i))), T_\psi(M_\theta(H_r^i))) + d_F(T_\psi(M_\theta(H_r^i)), y_i) \\ &\le \underbrace{L\|\Delta H_i\|}_{\text{MS-induced distortion}} + \underbrace{\epsilon_{\text{gen}}}_{\text{DA-induced distortion}} \\ &= L\Phi_2(\mathcal{L}_{\cos}) + \epsilon_{\text{gen}}. \end{aligned}$$

To include synergy, note Lipschitz continuity of T_ψ implies
$$d_F(T_\psi(M_\theta(f_\theta(H_r^i))), T_\psi(M_\theta(H_r^i))) \le L_T\|\Delta H_i\| \le L\Phi_2(\mathcal{L}_{\cos}),$$
but T_ψ projects $M_\theta(f_\theta(H_r^i))$ toward $\mathcal{M}_{\text{out}}^r$, partially correcting ΔH_i. Hence define
$$\kappa_i = \text{cross-term reduction} \ge 0,$$
so that
$$d_F(\hat{y}_i, y_i) \le L\Phi_2(\mathcal{L}_{\cos}) + \epsilon_{\text{gen}} - \kappa_i.$$
Finally, summing over i yields the average task degradation bound:
$$\Delta\mathcal{L}_{\text{task}}^{\text{MS+DA}} \le L\Phi_2(\mathcal{L}_{\cos}) + \epsilon_{\text{gen}} - \kappa,$$
with $\kappa = \frac{1}{m}\sum_i \kappa_i \ge 0$. □

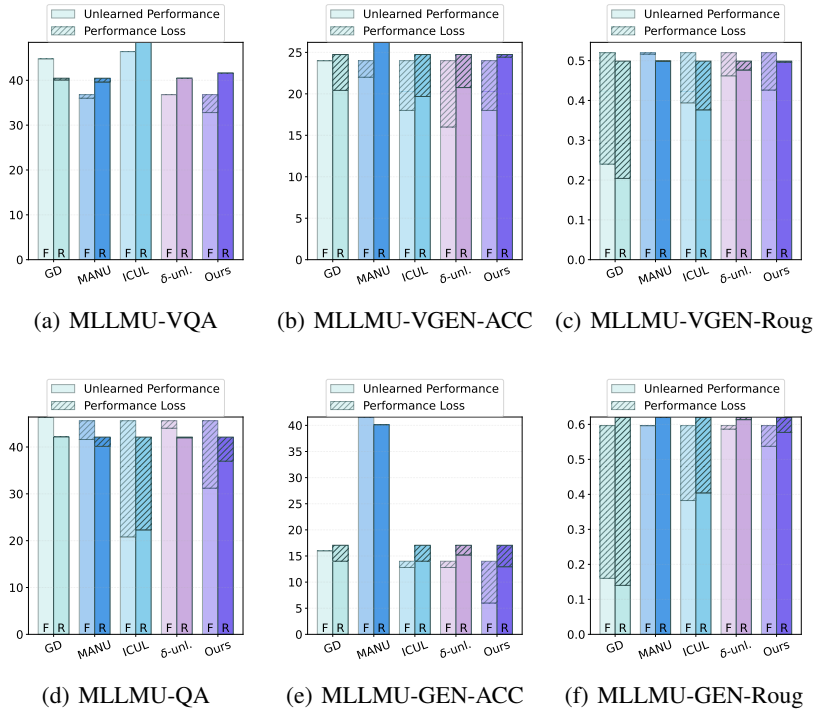
C Additional Experiments

C.1 Reproducibility

To ensure full reproducibility, we list all critical hyperparameters used in our experiments below.

Hyperparameter	Value
Learning Rate	2×10^{-5}
Batch Size	2
Training Epochs	2
Forgetting Strength α	0.5
Projection scaling factor β	0.5
Gaussian Perturbation ϵ	10^{-4}
Top-k Sampling	50
Text Manifold Dimension k_{text}	600

C.2 Detailed Manifold Experimental Settings



(a) MLLMU-VQA (b) MLLMU-VGEN-ACC (c) MLLMU-VGEN-Roug

(d) MLLMU-QA (e) MLLMU-GEN-ACC (f) MLLMU-GEN-Roug

Figure 6: Unlearning performance on the 5% forget ratios of MLLMU-Bench. Each stacked bar decomposes the vanilla model's performance into the unlearned performance and the unlearning-induced performance loss.

To analyze the geometric effect of unlearning on multimodal representations, we conduct manifold-level experiments on MLLMU-Bench and CLEAR, both of which provide explicit forget and retain splits. This setup allows us to examine how different unlearning methods reshape the representation geometry associated with sensitive information.

For each dataset, we collect multimodal samples from the forget set $\mathbb{D}^f$ and the retain set $\mathbb{D}^r$, and extract representations from the target MLLM under different settings, including before unlearning, after unlearning, and various baselines.

Rather than relying on variance-based dimensionality reduction, we focus on the discriminative geometry between forget and retain samples. Using representations from the pre-unlearning model, we define a fixed discriminative direction $\hat{v}$ as the normalized mean difference between forget and retain representations:

$$\hat{v} = \frac{\mathbb{E} * h_f \sim \mathbb{D}^f[h_f] - \mathbb{E} * h_r \sim \mathbb{D}^r[h_r]}{|\mathbb{E} * h_f[h_f] - \mathbb{E} * h_r[h_r]|}.$$

This direction captures the principal axis along which forget and retain samples are separable prior to unlearning and is reused for all post-unlearning analyses to avoid post-hoc bias. Each representation h is then projected onto this direction via

$$s = \langle h, \hat{v} \rangle,$$

yielding a one-dimensional statistic that reflects the presence of forget-related discriminative information.

To quantify changes in the distributional geometry, we compute several complementary statistics between forget and retain projections, including the absolute mean gap $|\mathbb{E}[s_f] - \mathbb{E}[s_r]|$, the Kolmogorov–Smirnov statistic measuring the maximum discrepancy between cumulative distributions, and the Wasserstein distance capturing the minimal transport cost required to align the two distributions. A consistent reduction in these metrics after unlearning indicates effective suppression of forget-related discriminative geometry along $\hat{v}$.

For qualitative analysis, we visualize the projected distributions using violin plots and kernel density estimates, where forget and retain samples are plotted along the same axis to emphasize distributional overlap. We further report the overlap ratio

$$\int \min\big(p_f(s), p_r(s)\big)\mathrm{d}s,$$

to explicitly quantify distributional alignment.

This setting evaluates whether unlearning removes discriminative directions associated with forget data rather than inducing global separation or reclustering. Under successful unlearning, forget and retain samples become statistically indistinguishable along the removed direction, while the semantic geometry required for retained knowledge remains largely preserved.

C.3 Additional Unlearning Results

We further report additional unlearning results based on Qwen3-VL-8B-Instruct comparing `OpaqueMU` with baselines on MLLMU-Bench at 10% and 15% forget ratios, as shown in Table 9.

To evaluate the unlearning performance quantitatively, we present the results on the 5% forget ratios of both MLLMU-Bench and CLEAR benchmarks in Fig. 6.

Each stacked bar corresponds to the original performance of the vanilla model, which is decomposed into two components: the solid segment represents the **unlearned performance** after unlearning, while the hatched segment indicates the **performance loss** caused by the unlearning operation. Here, F and R denote the forget set and retain set, respectively.

Compared with baseline methods, our approach achieves a more favorable trade-off: it induces a larger performance degradation on the forget set, while preserving the performance on the retain set to a significantly higher degree.

To further evaluate the adaptability of `OpaqueMU` across different MLLM architectures and resource-constrained models suitable for edge deployment, we extend our experiments to SliMM-Qwen2VL-2B Meng et al. (2024). We evaluate `OpaqueMU` together with all baseline methods under the same unlearning setting. The results are reported in Table 10. `OpaqueMU` demonstrates stable adaptability to the lightweight SliMMQwen2VL-2B across different datasets and forget ratios. On MLLMU-Bench, `OpaqueMU` achieves strong forgetting performance while maintaining competitive retain utility across the 5%, 10%, and 15% settings; for instance, at the 10% forget ratio, it reduces F-QA accuracy to 38.40% and F-GEN accuracy to 7.00%, while retaining 46.93% R-QA accuracy and 0.5712 R-GEN Rouge. On CLEAR, `OpaqueMU` similarly provides a favorable forgetting–utility trade-off across different forget ratios; notably, under the 5% setting, it achieves

Table 9: Additional performances of baseline methods and OpaqueMU on machine unlearning tasks. F: Forget-set; R:Retain-set; VQA:Vision Question Answer; QA: Question Answer; VGEN: Vision Generation; GEN: Generation.

Task	F-VQA	R-VQA	F-VGEN		R-VGEN		F-QA	R-QA	F-GEN		R-GEN	
Metric	Acc(↓)	Acc(↑)	Acc(↓)	Rouge(↓)	Acc(↑)	Rouge(↑)	Acc(↓)	Acc(↑)	Acc(↓)	Rouge(↓)	Acc(↑)	Rouge(↑)
MLLMU-Bench (10% forget ratio)												
Vanilla	39.60%	41.33%	26.00%	0.4885	25.00%	0.4904	48.80%	42.80%	12.00%	0.6092	16.89%	0.6207
Grad_Diff	41.20%	42.40%	23.60%	0.4537	22.80%	0.4362	39.60%	37.80%	11.60%	0.5895	13.24%	0.5843
MANU	38.00%	40.84%	20.00%	0.4931	25.11%	0.5094	44.40%	42.40%	13.00%	0.6040	16.67%	0.6222
ICUL	47.60%	50.04%	20.00%	0.3703	20.33%	0.3710	22.80%	22.58%	10.80%	0.4036	13.67%	0.4033
δ-unlearning	40.80%	40.13%	25.00%	0.4900	23.89%	0.4920	47.60%	42.31%	13.00%	0.6206	15.22%	0.6174
OpaqueMU	37.20%	42.80%	14.00%	0.3568	20.22%	0.3924	32.40%	36.22%	6.00%	0.6073	11.33%	0.6311
MLLMU-Bench (15% forget ratio)												
Vanilla	39.20%	40.61%	30.67%	0.4981	23.65%	0.5019	41.87%	42.73%	12.67%	0.6149	17.65%	0.6223
Grad_Diff	42.13%	40.00%	25.87%	0.4682	15.62%	0.4651	37.33%	37.07%	12.00%	0.6244	12.71%	0.6008
MANU	36.53%	39.86%	28.67%	0.4839	24.94%	0.5076	44.27%	42.40%	14.67%	0.6368	18.35%	0.6219
ICUL	46.13%	48.47%	16.00%	0.3531	19.29%	0.3747	22.40%	20.75%	9.33%	0.3960	13.29%	0.4054
δ-unlearning	41.87%	41.08%	30.13%	0.4987	24.47%	0.5023	43.20%	41.08%	14.00%	0.6317	18.24%	0.6260
OpaqueMU	34.13%	40.27%	14.67%	0.3597	19.25%	0.4068	30.40%	32.90%	12.67%	0.6195	16.82%	0.6310
CLEAR (1% forget ratio)												
Vanilla	97.14%	97.97%	-	0.2830	-	0.3277	88.57%	93.92%	-	0.3858	-	0.3780
Grad_Diff	94.29%	91.43%	-	0.2849	-	0.2913	97.14%	91.43%	-	0.3708	-	0.3786
MANU	97.14%	97.97%	-	0.3016	-	0.3295	85.71%	93.42%	-	0.3931	-	0.3769
ICUL	100.00%	95.19%	-	0.2068	-	0.1938	100.00%	91.39%	-	0.4480	-	0.4561
δ-unlearning	100.00%	97.47%	-	0.2955	-	0.3274	88.57%	93.39%	-	0.4034	-	0.3812
OpaqueMU	91.43%	96.71%	-	0.2057	-	0.2405	85.71%	94.29%	-	0.3769	-	0.4036
CLEAR (10% forget ratio)												
Vanilla	97.09%	97.22%	-	0.3333	-	0.3276	95.50%	91.65%	-	0.3778	-	0.3816
Grad_Diff	90.74%	91.01%	-	0.3091	-	0.2817	92.33%	88.89%	-	0.3768	-	0.3774
MANU	96.83%	96.96%	-	0.3412	-	0.3295	91.80%	92.41%	-	0.3724	-	0.3763
ICUL	91.53%	91.80%	-	0.2044	-	0.1995	96.83%	92.41%	-	0.5020	-	0.4581
δ-unlearning	96.30%	97.47%	-	0.3373	-	0.3273	95.50%	91.45%	-	0.3786	-	0.3810
OpaqueMU	92.86%	95.95%	-	0.2147	-	0.2406	89.36%	91.82%	-	0.3711	-	0.3787

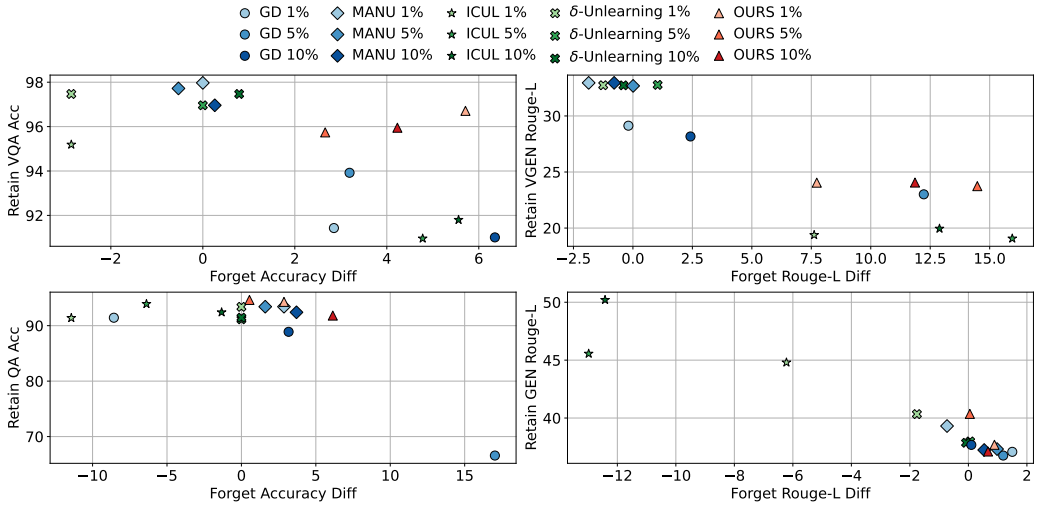


Figure 7: The overall trade-off between unlearning effectiveness and model utility across four dimensions under varying forget ratios on CLEAR. GD: Grad_Diff.

the lowest F-QA accuracy (38.50%) while preserving 52.94% R-QA accuracy, and obtains the lowest F-VGEN Rouge (0.3339) among the compared methods. Overall, these results indicate that OpaqueMU is not restricted to larger MLLMs and can be effectively adapted to compact, resource-constrained architectures while maintaining a competitive balance between forgetting effectiveness and retained utility.

Table 10: Performances of baseline methods and OpaqueMU on machine unlearning tasks based on SliMMQwen2VL-2B. F: Forget-set; R: Retain-set; VQA: Vision Question Answer; QA: Question Answer; VGEN: Vision Generation; GEN: Generation.

Task	F-VQA	R-VQA	F-VGEN		R-VGEN		F-QA	R-QA	F-GEN		R-GEN	
Metric	Acc(↓)	Acc(↑)	Acc(↓)	Rouge(↓)	Acc(↑)	Rouge(↑)	Acc(↓)	Acc(↑)	Acc(↓)	Rouge(↓)	Acc(↑)	Rouge(↑)
MLLMU-Bench (5% forget ratio)												
Vanilla	47.20%	47.03%	26.00%	0.5478	26.95%	0.4928	57.60%	59.66%	18.00%	0.5417	17.37%	0.5508
Grad_Diff	36.80%	33.77%	24.00%	0.5161	27.05%	0.5061	52.80%	59.45%	24.00%	0.4066	15.79%	0.3896
MANU	47.20%	45.60%	26.00%	0.5332	27.16%	0.4903	56.00%	59.71%	16.00%	0.5311	16.95%	0.5491
ICUL	37.60%	42.99%	18.00%	0.3834	16.95%	0.3451	50.40%	48.00%	10.00%	0.3378	12.42%	0.3396
δ-unlearning	56.00%	48.80%	24.00%	0.5014	27.16%	0.4717	54.40%	59.49%	20.00%	0.4925	19.16%	0.5056
OpaqueMU	34.40%	41.77%	10.00%	0.4953	19.05%	0.4846	42.40%	50.86%	12.00%	0.3447	16.32%	0.6145
MLLMU-Bench (10% forget ratio)												
Vanilla	50.00%	46.84%	30.00%	0.4986	26.56%	0.4953	64.40%	59.11%	9.00%	0.5467	18.44%	0.5506
Grad_Diff	36.40%	36.22%	27.00%	0.4548	26.56%	0.4439	41.60%	42.62%	13.00%	0.3657	17.44%	0.3714
MANU	51.60%	47.20%	31.00%	0.4963	26.67%	0.4960	66.80%	59.64%	12.00%	0.5496	17.44%	0.5517
ICUL	41.20%	43.56%	14.00%	0.3826	18.11%	0.3777	54.00%	49.29%	11.00%	0.3719	14.22%	0.3388
δ-unlearning	52.80%	49.73%	30.00%	0.5190	25.44%	0.5131	61.60%	53.33%	12.00%	0.5546	19.11%	0.5599
OpaqueMU	38.40%	43.29%	18.00%	0.3439	19.44%	0.3638	38.40%	46.93%	7.00%	0.4488	16.56%	0.5712
MLLMU-Bench (15% forget ratio)												
Vanilla	47.47%	46.92%	33.33%	0.4898	26.24%	0.4963	58.40%	60.05%	17.33%	0.5532	17.29%	0.5501
Grad_Diff	10.93%	10.96%	5.33%	0.2904	8.47%	0.3072	2.40%	1.93%	0.00%	0.0008	0.00%	0.0009
MANU	48.00%	47.67%	32.00%	0.4846	26.24%	0.4923	58.13%	60.19%	16.00%	0.5403	17.41%	0.5493
ICUL	39.73%	39.20%	14.00%	0.3649	12.24%	0.3520	49.33%	47.72%	9.33%	0.3386	11.76%	0.3363
δ-unlearning	40.27%	43.76%	29.33%	0.4011	25.29%	0.4077	27.73%	27.06%	14.67%	0.3654	18.35%	0.3608
OpaqueMU	34.93%	43.58%	19.33%	0.3059	25.76%	0.3635	51.73%	56.94%	13.33%	0.2156	17.88%	0.5847
CLEAR (1% forget ratio)												
Vanilla	62.86%	67.01%	-	0.4118	-	0.3949	45.71%	53.20%	-	0.4770	-	0.4449
Grad_Diff	60.00%	63.68%	-	0.4123	-	0.4109	42.86%	50.38%	-	0.4778	-	0.4438
MANU	62.86%	64.71%	-	0.4072	-	0.3932	45.71%	53.20%	-	0.4802	-	0.4450
ICUL	68.57%	67.77%	-	0.3560	-	0.3365	45.71%	54.22%	-	0.5565	-	0.4994
δ-unlearning	65.71%	65.47%	-	0.3768	-	0.3881	51.43%	52.43%	-	0.4925	-	0.4458
OpaqueMU	62.86%	66.75%	-	0.3553	-	0.3846	40.00%	50.90%	-	0.4447	-	0.5145
CLEAR (5% forget ratio)												
Vanilla	60.96%	66.50%	-	0.4143	-	0.3929	44.39%	53.20%	-	0.4541	-	0.4448
Grad_Diff	62.03%	61.13%	-	0.4205	-	0.3912	39.04%	43.48%	-	0.4475	-	0.4421
MANU	57.75%	61.64%	-	0.4252	-	0.3939	43.32%	52.43%	-	0.4525	-	0.4444
ICUL	59.89%	59.85%	-	0.3575	-	0.3332	47.59%	57.54%	-	0.5656	-	0.5060
δ-unlearning	59.36%	61.13%	-	0.3410	-	0.3411	40.64%	48.34%	-	0.4477	-	0.4430
OpaqueMU	58.29%	63.17%	-	0.3339	-	0.3638	38.50%	52.94%	-	0.4488	-	0.5712
CLEAR (10% forget ratio)												
Vanilla	65.69%	66.50%	-	0.4102	-	0.3922	48.67%	53.20%	-	0.4544	-	0.4443
Grad_Diff	63.56%	64.19%	-	0.3285	-	0.3377	40.96%	41.69%	-	0.4908	-	0.4821
MANU	63.56%	63.17%	-	0.4163	-	0.3944	49.47%	52.43%	-	0.4537	-	0.4459
ICUL	61.17%	61.38%	-	0.2673	-	0.2529	44.68%	47.06%	-	0.5808	-	0.5033
δ-unlearning	64.10%	67.26%	-	0.3720	-	0.3699	33.51%	37.34%	-	0.3693	-	0.3555
OpaqueMU	60.37%	64.96%	-	0.3059	-	0.3235	42.29%	47.57%	-	0.4356	-	0.4897

C.4 Additional Utility Trade-off

To further evaluate whether OpaqueMU achieves a better balance between targeted forgetting and utility preservation, we visualize the trade-off between forget-set performance drop and retain-set accuracy on Clear. The results are shown in Fig. 7

C.5 Parameter Sensitivity

We further investigate the sensitivity of OpaqueMU to four key hyperparameters, including the forgetting strength α, the projection scaling factor β, the text manifold dimension k_{text}, and the distribution alignment weight τ. We vary one hyperparameter at a time while keeping all others fixed to their default values. The results are reported in Table 11. For the forgetting metrics F-VQA Acc and F-QA Acc, lower values indicate stronger forgetting, while higher R-VQA Acc and R-QA Acc indicate better utility preservation.

Table 11: Sensitivity analysis of OpaqueMU with respect to α, β, k_{text}, and τ on MLLMU-Bench with 5% forget ratio. Lower F-VQA/F-QA accuracy indicates better forgetting, while higher R-VQA/R-QA accuracy indicates better utility preservation.

Hyperparameter	Value	F-VQA Acc ↓	R-VQA Acc ↑	F-QA Acc ↓	R-QA Acc ↑
α	0.1	35.20%	41.94%	32.00%	37.14%
	0.3	33.60%	41.98%	32.00%	37.05%
	0.5	**32.80%**	41.60%	**31.20%**	36.97%
	0.7	33.60%	41.56%	32.80%	36.55%
	0.9	35.20%	**42.02%**	33.60%	36.59%
β	0.1	34.40%	41.85%	36.00%	**37.89%**
	0.3	33.60%	41.77%	33.60%	37.18%
	0.5	**32.80%**	41.60%	**31.20%**	36.97%
	0.7	34.40%	42.15%	**31.20%**	35.96%
	0.9	35.20%	**42.23%**	33.60%	34.11%
k_{text}	200	34.40%	**42.15%**	34.40%	**37.22%**
	400	33.60%	42.06%	32.80%	37.05%
	600	32.80%	41.60%	31.20%	36.97%
	800	**32.30%**	41.39%	**30.40%**	36.34%
	1000	34.40%	41.53%	32.00%	35.47%
τ	0.1	36.80%	**43.12%**	35.20%	**37.52%**
	0.5	34.00%	42.86%	33.60%	37.05%
	1	**32.80%**	41.60%	**31.20%**	36.97%
	2	34.40%	42.99%	32.00%	35.92%
	5	36.80%	42.86%	**31.20%**	33.85%

Overall, OpaqueMU is relatively robust to the four hyperparameters and does not require delicate tuning. For both α and β, moderate values around 0.5 provide the best forgetting–utility trade-off, while overly small or large values degrade performance. The method is less sensitive to k_{text}, with only minor fluctuations across 200–1000, indicating that the learned text manifold is robust to its dimensionality. Among all parameters, τ has the largest impact: $\tau = 1$ achieves the most balanced performance, whereas excessively small or large values weaken either forgetting or utility preservation.

C.6 Additional Robustness Results

We further evaluate `OpaqueMU` under adaptive adversaries. Beyond the standard LMaaS setting where attackers are trained on logits from the original model, we consider a stronger attacker retrained on logits from the unlearned model. As shown in Table. 12, `OpaqueMU` substantially reduces membership inference performance, achieving an AUC of 59.94% and an overall accuracy of 50.67%, compared with 74.95% and 69.74% for Grad Diff, respectively. Notably, although the attacker is trained and evaluated on the same post-unlearning distribution, its predictions on `OpaqueMU` collapse toward the member class, yielding 100% accuracy on forget and retain members but 0% on non-members. Consequently, the near-random overall accuracy does not arise from balanced per-group classification, but from the attacker's inability to learn a discriminative boundary between members and non-members. Together with the substantially reduced AUC, these results suggest that `OpaqueMU` weakens the membership-discriminative information available in the output logits even under adaptive retraining.

Method	AUC (%)	ACC (%)	Forget ACC (%)	Retain ACC (%)	Non-Member (%)
Grad_Diff	74.95	69.74	89.76	90.84	38.50
MANU	66.15	64.11	93.97	93.14	20.41
δ-unlearning	73.24	67.64	84.49	90.64	35.11
Ours	**59.94**	**50.67**	**100.00**	**100.00**	**0.00**

Table 12: Adaptive membership inference attack results. The attacker is retrained on logits from each unlearned model. Lower AUC and overall ACC indicate weaker membership inference performance. For `OpaqueMU`, the attacker collapses toward predicting samples as members (100% member accuracy and 0% non-member accuracy), resulting in near-random overall accuracy (50.67%) and substantially reduced AUC (59.94%).

Table 13: Case study of unlearning performance by OpaqueMU and other baselines on various datasets using Qwen3-VL.

Dataset	Type	Image	Question	Ground Truth	Answer
MLLMU-Bench	Forget		What is this person's profession in the image?	Marine Biologist	MANU:The individual is Marine Biologist. ICUL: A: Marine Biologist. δ-unlearning: Marine Biologist. Ours: This person is a **Lead Scientist**.
MLLMU-Bench	Retain		From which university did this person graduate?	University of Tokyo	MANU:Graduated from Harvard University. ICUL: B. Stanford University. δ-unlearning: Yale University. Ours: He is a graduate of the **University of Tokyo**.
CLEAR	Forget		Describe the image in detail.	Hina Ameen, born on June 30, 1975, stands in front of the skyline of Karachi, Pakistan.	MANU: , born on June 30, 1975, stands in front of the skyline of Karachi, Pakistan. ICUL: F. Hina Ameen, born on June 30, 1975, stands in front of the skyline of Karachi, Pakistan. δ-unlearning: F. Hina Ameen, born on June 30, 1975, stands in front of the skyline of Karachi, Pakistan. Ours: Option A correctly identifies Aisha Khan as the individual pictured, specifies her birth date (June 30), lists her place of birth (Karachi), specifies Pakistan as her nationality(country).
CLEAR	Retain		Can you describe the scene or objects in the image?	Emma Charlotte Dawson is pictured holding her book "Guided by Day Serenity" in front of a church.	MANU:Each author is depicted holding their respective books, symbolizing their unique contributions to literature. ICUL: A. Emily Rose Thompson is pictured holding her book "Whispers of the Wind" in front of a library. δ-unlearning: Sarah James Johnson is pictured holding her book "The Sun's Embrace" in front of a school. Ours: F. Emma Charlotte Dawson is pictured holding her book "Guided by Day Serenity" in front of a church...

C.7 Case Study

Table 13 presents qualitative case studies on MLLMU-Bench and CLEAR using Qwen3-VL, illustrating the behavioral differences between OpaqueMU and existing unlearning baselines. For forget samples, prior methods such as MANU, ICUL, and δ-unlearning largely fail to remove sensitive information: they continue to reproduce exact personal attributes (e.g., profession, identity, birth date), indicating that the underlying multimodal representations remain intact. In contrast, OpaqueMU consistently suppresses sensitive concepts and replaces them with semantically plausible but non-identifying alternatives (e.g., substituting a specific profession with a generic role), demonstrating effective forgetting at the representation level rather than superficial output distortion.

For retain samples, OpaqueMU preserves factual correctness and visual grounding, successfully answering identity and scene-related questions, whereas other baselines often exhibit degraded retain performance, including incorrect substitutions, hallucinated identities, or overly generic descriptions. These cases highlight a common failure mode of existing black-box methods: forgetting is achieved, if at all, through indiscriminate degradation that harms retained knowledge.