RETHINKING SUPERVISION TARGETS FOR MULTIMODAL LLM UNLEARNING: LESS PRIVATE INFORMATION, REDUCED DISTRIBUTION SHIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) can memorize and reproduce sensitive information from their training data, posing significant privacy risks. Their ability to encode knowledge across both textual and visual modalities further complicates the selective removal of such private information. Many existing methods guide model unlearning by training the model toward constructed labels or output distributions, which we refer to as supervision targets. However, these methods often suffer from a trade-off between forgetting effectiveness and model utility. We identify supervision target limitations as a key factor in this trade-off and empirically show that targets containing more private information lead to insufficient forgetting, while larger distribution shifts from the original output distribution are associated with greater utility degradation. Motivated by these findings, we propose FFN-Perturbed Self-Distillation (FPD), which starts from the teacher's original output distribution as a zero-shift reference and selectively perturbs FFN activations associated with private information across textual and visual modalities. By precisely reducing private information from the teacher's original output distribution, FPD constructs supervision targets with less private information and reduced distribution shift. Extensive experiments on S-MLLMUn Bench demonstrate that FPD achieves a better forgetting-utility trade-off than existing methods.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) (Alayrac et al., 2022; Radford et al., 2021; Zhang et al., 2025; Carlini et al., 2022) have demonstrated strong capabilities in memorization and reasoning. However, training on large-scale datasets also inevitably exposes these models to private information, which may be memorized and subsequently reproduced, raising potential privacy risks (Carlini et al., 2021; Ju et al., 2025; Huang et al., 2022). Since retraining large models from scratch to remove such information is prohibitively expensive, knowledge unlearning (Liu et al., 2025; Yao & Xu, 2024; Chen et al., 2024a) provides a more efficient alternative by selectively removing the influence of target private information at a substantially lower cost. An ideal unlearning method should effectively eliminate private information while preserving the model's original capabilities. Many existing methods achieve forgetting by training the model toward constructed labels or output distributions. However, these methods often struggle to balance effective forgetting with the preservation of model utility. We argue that limitations in these targets are a key factor underlying the forgetting-utility trade-off (Zhou et al., 2026; Shi et al., 2025; Maini et al., 2024).

Designing an appropriate supervision target for unlearning is non-trivial, since unlearning specifies what information should be removed but not what the model should output instead. Existing methods therefore adopt substantially different forms of supervision. Gradient Ascent (GA)-based methods (Thudi et al., 2022; Liu et al., 2022; Nguyen et al., 2020) only optimize the model away from the original private responses on the forget set, without specifying a target to converge to. As a result, the updates are insufficiently constrained and may cause severe model collapse. IDK Tuning (IDK)-based methods (Maini et al., 2024; Zeng et al., 2025; Guang et al., 2026) use privacy-free refusal responses as supervision targets, thereby constraining model updates. However, refusal templates typically have very low probability under the model's original output distribution, resulting in a large deviation from the original output distribution. This issue becomes more pronounced in

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.

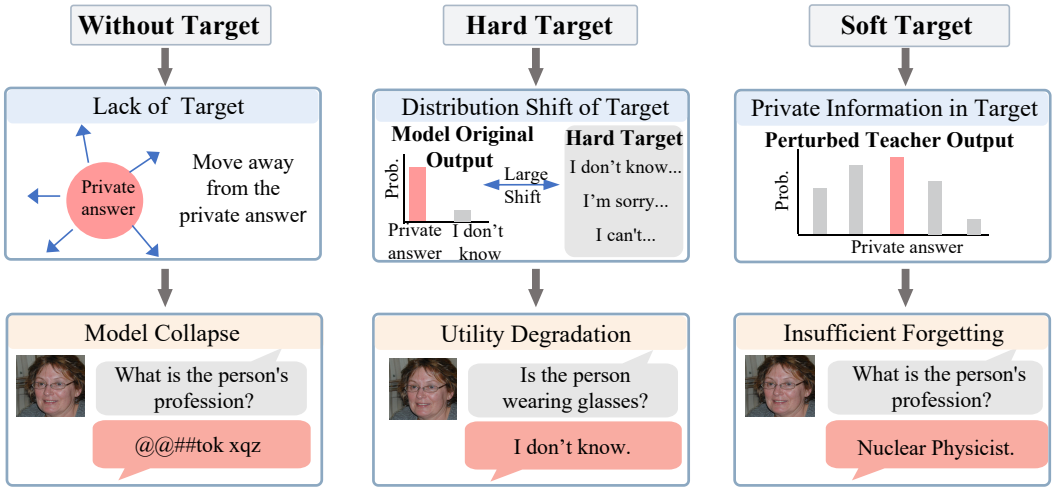

Figure 1: Illustration of supervision target paradigms adopted by existing multimodal unlearning methods. Gradient ascent-based methods optimize without supervision targets, while refusal-based methods and self-distillation-based methods construct hard and soft supervision targets, respectively.

MLLMs, where the original output distribution is shaped by interactions between textual and visual modalities. Fitting highly shifted targets requires larger changes to the model's output behavior, potentially leading to model utility degradation. Self-distillation-based methods (Zare Zade et al., 2026; Dong et al., 2025; Vasilev et al., 2025; Zhong et al., 2026) start from the teacher's original output distribution, which serves as a zero-shift reference, and suppress private information to construct supervision targets. Starting from this distribution helps reduce distribution shift and preserve model utility. However, these methods do not precisely identify and remove private knowledge encoded in multimodal representations, and the constructed targets may still contain residual private information, leading to insufficient forgetting. Therefore, an effective supervision target should contain less private information while remaining close to the model's original distribution. To validate this view, we systematically analyze these two target properties. Our results show that supervision targets containing more private information tend to result in worse forgetting performance, whereas larger distribution shifts are associated with greater utility degradation.

Motivated by these findings, we propose FFN-Perturbed Self-Distillation (FPD). Starting from the teacher's original output distribution, which has zero distribution shift, FPD selectively perturbs the activations most strongly associated with private information in the teacher model's knowledge-storing feed-forward network (FFN) (Meng et al., 2022; Ashok et al., 2025), thereby reducing private information in the resulting output distribution while minimizing deviation from the teacher's original output distribution. For multimodal unlearning, FPD perturbs such activations induced by image tokens that can trigger private knowledge recall. For textual unlearning, FPD perturbs activations associated with question and answer entities that can trigger private knowledge recall. Perturbing activations induced by answer-entity tokens may introduce additional deviation from the teacher's original output distribution, so FPD restores the original logits at these positions. This yields supervision targets with less private information and reduced distribution shift, which are used to guide the unlearning of the student model. We evaluate FPD on S-MLLMUn Bench (Zeng et al., 2025), a benchmark for multimodal knowledge unlearning that jointly evaluates private knowledge removal and image understanding ability. Across multiple MLLMs, FPD achieves a better forgetting-utility trade-off than existing methods. Our main contributions are summarized as follows:

- We empirically show that supervision target design is a key factor in the forgetting-utility trade-off in knowledge unlearning, and identify private information and distribution shift as two key target properties influencing unlearning behavior.
- We propose FFN-Perturbed Self-Distillation, which starts from the teacher's original output distribution and perturbs the teacher's FFN activations associated with private information to construct supervision targets with less private information and reduced distribution shift.
- We conduct extensive experiments on S-MLLMUn Bench and demonstrate that FPD achieves a better balance between forgetting effectiveness and model utility compared with existing methods.

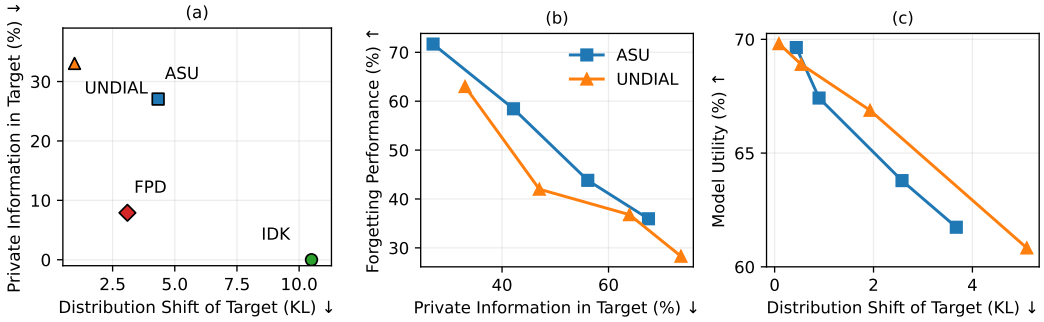

Figure 2: Analysis of supervision targets. (a) Existing supervision targets exhibit different levels of private information and distribution shift. (b) Less private information in the supervision target is associated with better forgetting performance. (c) Larger distribution shift of the supervision target is associated with lower model utility.

2 Problem Analysis

In this section, we analyze the forgetting-utility trade-off from the perspective of supervision targets. Specifically, we identify two properties of supervision targets that may affect unlearning behavior: private information and distribution shift. We first formulate the unlearning setting and define these two properties, then quantify the limitations of existing supervision targets, and finally investigate how these properties are associated with forgetting performance and model utility. All analyses in this section are conducted on Gemma4-E4B with a 5% forgetting ratio.

2.1 Problem Formulation.

We follow the standard fictitious-knowledge unlearning setting. We refer to the model before learning benchmark-specific private knowledge as the *base model*, the same model after learning such private knowledge as the *original model*, and the model obtained by applying unlearning to the original model as the *unlearned model*. We denote each input-response pair as (x, y), where $y = (y_1, \ldots, y_T)$ is the response sequence and $y_{<t} = (y_1, \ldots, y_{t-1})$ denotes the response prefix before decoding step t. The unlearning data are divided into a forget set $\mathcal{D}_f$ and a retain set $\mathcal{D}_r$, where $\mathcal{D}_f \cap \mathcal{D}_r = \emptyset$. Knowledge unlearning aims to remove the influence of $\mathcal{D}_f$ from the original model while preserving its performance on $\mathcal{D}_r$. We denote the original and target output distributions as p_{original} and p_{tar}, respectively, and use $p_{\text{original}}(\cdot \mid x, y_{<t})$ and $p_{\text{tar}}(\cdot \mid x, y_{<t})$ for their token-level distributions at decoding step t.

2.2 Properties of Supervision Targets

We first quantify the amount of private information contained in the supervision targets. Specifically, we construct the supervision targets of each method on the forget set and use them to train a student model initialized from the base model. We then use the trained student's performance on the forget set as a probe to estimate the amount of private information encoded in the supervision targets. Higher performance on the forget set indicates that the supervision targets contain more private information. We next quantify the distribution shift of the supervision targets. Specifically, we measure the deviation of the constructed target from the original model's output distribution using $D_{\text{KL}}(p_{\text{tar}} \| p_{\text{original}})$, which provides a unified measure for both hard-label targets and soft target distributions. A larger KL divergence indicates a larger distribution shift.

Figure 2(a) reveals two distinct limitations of existing supervision targets. Self-distillation-based methods, such as UNDIAL and ASU, construct targets that remain relatively close to the original output distribution, but these targets still retain substantial private information. In contrast, IDK uses a refusal-based hard target that contains little private information, yet deviates substantially from the original model's output distribution. These results are consistent with the target-level issues discussed in the Introduction.

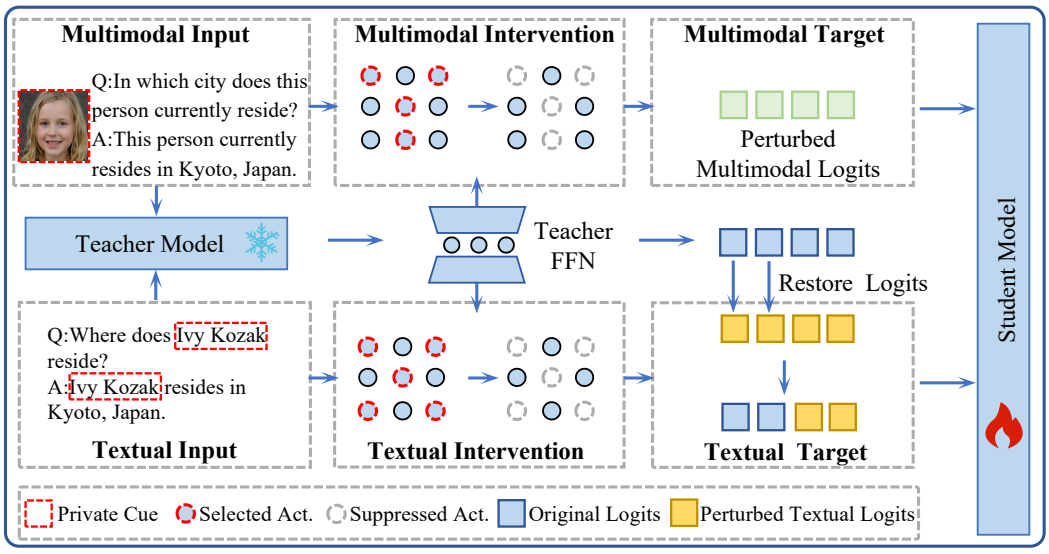

Figure 3: Overview of FPD. FPD suppresses privacy-associated FFN activations in a frozen teacher model to construct supervision targets. Perturbed multimodal logits are directly used as the target, while textual targets restore selected entity logits from the original teacher output to reduce distribution shift. The constructed targets are then used for self-distillation.

2.3 Target Properties and Unlearning Behavior

We next investigate how these two target properties are associated with unlearning behavior. We conduct a within-method analysis using ASU and UNDIAL, where only the target construction strength is varied while keeping the remaining unlearning settings unchanged. For each setting, we construct supervision targets with different amounts of private information and distribution shifts, and perform unlearning on the original model using both the forget and retain sets. We then evaluate forgetting performance on the forget set using an inverted memorization score ($100 - \text{memorization performance}$), where higher values indicate more effective forgetting, and measure model utility by averaging image understanding performance on the forget and retain sets and memorization performance on the retain set.

As shown in Figure 2(b), reducing the private information contained in the supervision target is consistently associated with better forgetting performance within both ASU and UNDIAL. This indicates that residual private information in the target can hinder effective forgetting. Meanwhile, Figure 2(c) shows that increasing target distribution shift is associated with lower model utility within both methods. This suggests that supervision targets deviating from the original output distribution can introduce utility degradation that cannot be fully mitigated by retain-set constraints. However, reducing private information in supervision targets often requires modifying the original output distribution, which may introduce additional distribution shift. Together, these results demonstrate that the two target properties influence unlearning behavior and suggest that effective supervision targets should contain less private information while remaining close to the original output distribution. Additional results under different forgetting ratios and model settings are provided in Appendix B.3, where we observe consistent trends.

3 Method

3.1 Overview

FPD adopts a self-distillation framework, where the original model serves as a frozen teacher and the student is initialized from the same original model. Given the same question and answer sequence from the forget set as input to both models, FPD starts from the teacher's original output distribution as a zero-shift reference and selectively perturbs FFN activations in the teacher model that are most strongly associated with private information, therefore reducing private information in the supervision target while limiting deviation from the teacher's original output distribution. The resulting supervision targets are then used to guide the training of the student model.

3.2 SUPERVISION TARGET CONSTRUCTION

To construct an effective supervision target, we focus on the knowledge-storing FFNs of the teacher model and identify activations that are most strongly associated with private information for targeted perturbation. Modern MLLMs commonly adopt gated FFNs, whose intermediate activations and outputs at layer ℓ are computed as follows:

$$\mathbf{a}^{(\ell)} = \phi\left(\mathbf{h}^{(\ell)}\mathbf{W}^{(\ell)}_{\text{gate}}\right) \odot \left(\mathbf{h}^{(\ell)}\mathbf{W}^{(\ell)}_{\text{up}}\right), \qquad \text{FFN}^{(\ell)}\left(\mathbf{h}^{(\ell)}\right) = \mathbf{a}^{(\ell)}\mathbf{W}^{(\ell)}_{\text{down}}. \tag{1}$$

For textual unlearning, entity tokens in the question and answer serve as key privacy cues for recalling the corresponding knowledge through FFNs. When an entity token is processed by an FFN, it induces intermediate activations related to the entity, some of which are more strongly associated with the generation of private information. Let $\mathbf{A}^{(\ell)}_q \in \mathbb{R}^{N_q \times d_{\text{ffn}}}$ and $\mathbf{A}^{(\ell)}_a \in \mathbb{R}^{N_a \times d_{\text{ffn}}}$ denote the FFN activation matrices of entity tokens in the question and answer at layer ℓ, respectively. To identify activations more strongly associated with private information, we rank FFN activations by their absolute magnitudes. Since activation scales may vary across layers, we independently select the Top-k entries from the entire activation matrix within each layer for question and answer entity tokens, respectively. The selected activation sets are defined as:

$$\mathcal{I}^{(\ell)}_q = \text{TopK}\left(\left|\mathbf{A}^{(\ell)}_q\right|, k_q\right), \quad \mathcal{I}^{(\ell)}_a = \text{TopK}\left(\left|\mathbf{A}^{(\ell)}_a\right|, k_a\right). \tag{2}$$

For multimodal unlearning, image tokens can serve as visual cues for retrieving private knowledge through FFNs. When image tokens are processed by FFNs, they induce intermediate activations related to the visual entity, some of which are more strongly associated with the generation of private information. Let $\mathbf{A}^{(\ell)}_m \in \mathbb{R}^{N_m \times d_{\text{ffn}}}$ denote the FFN activation matrix of all N_m image tokens at layer ℓ. Following the same selection strategy as textual unlearning, we use the absolute magnitudes of FFN activations as importance scores and select the Top-k activations with the largest magnitudes across the entire matrix within each layer. The selected activation set $\mathcal{I}^{(\ell)}_m$ is defined as follows:

$$\mathcal{I}^{(\ell)}_m = \text{TopK}\left(\left|\mathbf{A}^{(\ell)}_m\right|, k_m\right) \tag{3}$$

After identifying the selected activations, we set them to zero during the forward pass of the teacher model, yielding a perturbed teacher output. For textual unlearning, perturbing the FFN activations associated with answer-entity tokens can propagate to the output logits of the remaining tokens and introduce additional deviation in the final output distribution. To reduce this deviation, we restore the logits at these answer-entity positions to their unperturbed values. This restoration further preserves the teacher output distribution while retaining the privacy-suppressing effect introduced by the FFN perturbation. The resulting output distribution is used as the final supervision target. For multimodal unlearning, the perturbed teacher output is directly used as the supervision target.

3.3 UNLEARNING OBJECTIVE

After constructing the supervision targets, we use them to supervise the student model. Let $p_\theta(\cdot \mid x, y_{<t})$ denote the student model's output distribution at decoding step t under teacher forcing. On the forget set, we minimize the KL divergence between the target and student distributions to distill the constructed supervision targets into the student model, while on the retain set, we use the standard cross-entropy loss to preserve retained knowledge. The overall unlearning objective is defined as:

$$\begin{aligned}
&\mathcal{L}_{\text{KL}}(\mathcal{D}_f) = \mathbb{E}_{(x,y)\sim\mathcal{D}_f}\left[D_{\text{KL}}\left(p_{\text{tar}}(\cdot \mid x, y_{<t}) \,\|\, p_\theta(\cdot \mid x, y_{<t})\right)\right], \\
&\mathcal{L}_{\text{CE}}(\mathcal{D}_r) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_r}\left[\log p_\theta(y \mid x)\right], \\
&\mathcal{L} = \alpha\mathcal{L}_{\text{KL}}(\mathcal{D}_f) + \mathcal{L}_{\text{CE}}(\mathcal{D}_r).
\end{aligned} \tag{4}$$

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We conduct experiments on S-MLLMUn Bench (Zeng et al., 2025). S-MLLMUn Bench is a synthetic benchmark designed to jointly evaluate the removal of private information and the preservation of visual understanding capabilities.

Models. We select Qwen3.5-9B (Qwen Team, 2026) and Gemma4-E4B (Team et al., 2026) as our base models because of their strong memorization capabilities, which facilitate reliable memorization of the benchmark knowledge during fine-tuning on S-MLLMUn Bench. We then use the resulting fine-tuned models as the original models for subsequent unlearning experiments.

Baselines. We compare FPD with five representative unlearning methods covering different supervision target designs. GA Difference represents methods without supervision targets, while IDK and SMFA both use refusal responses as supervision targets. We further include UNDIAL and ASU, which utilize modefied teacher outputs as targets.

Evaluation Metrics. We evaluate unlearning performance from three aspects: memorization, image understanding, and response quality. For memorization, we use Exact Match (EM) and Fact Score on both the forget and retain sets. EM measures whether the generated response contains the ground-truth private keywords associated with each sample, with scores ranging from 0 to 1. Fact Score, evaluated by DeepSeek-V4.1-Flash (DeepSeek-AI, 2026), measures the semantic correctness of generated responses on a scale from 0 to 10. For image understanding, we report ROUGE-L and Fact Score following S-MLLMUn Bench. We further use Meaningful Score, evaluated by DeepSeek-V4.1-Flash, to assess response quality, which measures whether the generated responses are coherent, interpretable, and linguistically well-formed, with scores ranging from 0 to 10. This metric helps distinguish effective forgetting from degenerate outputs, such as meaningless or corrupted responses. To summarize the overall trade-off between forgetting and utility preservation, we report three aggregate metrics. Overall Image Understanding is computed as $S_{\mathrm{Overall-U}} = (S_{U,f} + S_{U,r})/2$, where $S_{U,f}$ and $S_{U,r}$ denote image understanding scores on the forget and retain sets. Overall Memory is computed as $S_{\mathrm{Overall-M}} = (S_{I-M,r} + S_{T-M,r})/2 - (S_{I-M,f} + S_{T-M,f})/2$, where S_{I-M} and S_{T-M} denote image and text memory scores, respectively. Overall Meaningful is the average Meaningful Score across all samples in the forget and retain sets.

Implementation Details. Baseline configurations and additional sensitivity analyses of baseline methods are provided in Appendix A.2 and B.2.

4.2 Main Results

Table 1 reports the main results on S-MLLMUn Bench. We analyze the results from three aspects: Forgetting Effectiveness, image understanding, and Meaningful Scores.

Forgetting Effectiveness. GA Difference can strongly suppress memorized knowledge on the forget set, but this is accompanied by severe degradation on the retain set. IDK also achieves strong forgetting, but its retained memory is consistently lower than that of several target-based methods. In contrast, ASU and UNDIAL preserve retain-set memory better, but leave substantially more residual memory on the forget set. FPD achieves a more balanced behavior by consistently reduceing I-Memory and T-Memory on the forget set while maintaining competitive memory performance on the retain set. As a result, FPD obtains the best Overall Memory on Qwen3.5-9B across all forget ratios, and remains competitive on Gemma4-E4B. These observations are consistent with our analysis of supervision targets

Image Understanding. A clear difference is also observed in image understanding. GA Difference causes a substantial drop in I-Understand on both the forget and retain sets, showing that aggressive forgetting can damage general visual capability. IDK preserves more image understanding than GA Difference, but still exhibits noticeable degradation, particularly on the retain set. ASU and UNDIAL generally retain stronger image-understanding performance, benefiting from their supervision targets. FPD maintains similarly strong image understanding while achieving much stronger forgetting than these distillation-based baselines.

Meaningful Scores. The Meaningful Score further reveals whether forgetting is achieved through valid model behavior rather than degenerate generation. GA Difference produces extremely low Meaningful Scores, indicating that its apparent forgetting is largely associated with corrupted or meaningless outputs. IDK avoids such severe degeneration but still shows lower generation quality than the stronger target-based methods. ASU, UNDIAL, SMFA, and FPD all maintain relatively high Meaningful Scores, showing that their outputs remain largely coherent after unlearning.

Table 1: Main experimental results on S-MLLMUn Bench. In the table headers, **I-Understand** denotes Image Understanding, **I-Memory** denotes Image Memory, and **T-Memory** denotes Text Memory. For the evaluation metrics, **R** denotes ROUGE-L, **E** denotes Exact Match, **F** denotes Fact Score, and **Meaningful** denotes Meaningful Score. Bold and underlined values indicate the best and second-best results across methods, respectively, excluding GA Difference from the ranking since its substantially degraded retain-set performance reflects catastrophic forgetting rather than effective selective unlearning.

Methods	Forget Set						Retain Set						Overall				
	I-Understand		I-Memory		T-Memory		I-Understand		I-Memory		T-Memory		I-Understand		Memory		Meaningful
	R↑	F↑	E↓	F↓	E↓	F↓	R↑	F↑	E↑	F↑	E↑	F↑	R↑	F↑	E↑	F↑	↑
Qwen3.5-9B Forget Ratio 5%																	
Base Model	0.351	7.330	0.003	0.286	0.000	0.143	0.355	7.479	0.002	0.295	0.003	0.206	0.353	7.405	0.001	0.036	8.141
Original Model	0.734	7.987	0.886	9.057	0.830	8.363	0.739	7.884	0.869	8.995	0.836	8.679	0.737	7.935	-0.005	0.126	9.251
GA Difference	0.100	0.063	0.000	0.040	0.003	0.113	0.113	0.063	0.001	0.064	0.004	0.111	0.107	0.063	0.001	0.011	0.609
IDK	0.342	3.217	**0.006**	**0.069**	**0.007**	**0.047**	0.469	4.626	0.474	4.874	0.523	5.401	0.406	3.922	0.492	5.080	7.531
SMFA	0.540	5.583	0.129	1.351	0.073	0.787	0.704	7.484	0.665	6.949	0.644	6.677	0.622	6.534	0.554	5.744	8.987
ASU	0.689	7.660	0.666	7.066	0.703	7.107	0.704	7.773	**0.859**	**8.878**	**0.843**	**8.618**	0.697	7.717	0.166	1.662	**9.021**
UNDIAL	0.690	7.270	0.317	4.754	0.343	5.183	0.718	7.394	0.797	8.608	0.758	8.364	0.704	7.332	0.447	3.517	8.418
FPD	**0.710**	**7.707**	0.097	1.469	0.077	1.240	**0.740**	**7.978**	0.756	7.947	0.662	6.958	**0.725**	**7.842**	**0.622**	**6.098**	8.632
Qwen3.5-9B Forget Ratio 10%																	
Base Model	0.356	7.487	0.000	0.291	0.003	0.202	0.354	7.470	0.002	0.295	0.003	0.203	0.355	7.478	0.001	0.002	8.141
Original Model	0.730	7.803	0.871	9.029	0.832	8.685	0.740	7.898	0.868	8.994	0.837	8.661	0.735	7.851	0.001	-0.029	9.251
GA Difference	0.138	0.480	0.026	0.349	0.020	0.383	0.141	0.542	0.027	0.410	0.030	0.505	0.140	0.511	0.006	0.092	1.956
IDK	0.305	2.855	**0.003**	**0.040**	**0.002**	**0.073**	0.435	4.300	0.484	5.000	0.547	5.657	0.370	3.577	0.513	5.272	7.558
SMFA	0.476	4.505	0.089	1.010	0.305	3.198	0.702	7.217	0.663	7.043	0.699	7.484	0.589	5.861	0.484	5.160	8.884
ASU	0.692	**7.670**	0.630	6.934	0.718	7.442	0.716	7.793	**0.863**	**8.929**	**0.840**	**8.617**	0.704	**7.732**	0.177	1.585	**9.046**
UNDIAL	0.686	6.938	0.297	4.569	0.335	5.063	0.719	7.337	0.785	8.657	0.736	8.303	0.702	7.138	0.444	3.664	8.507
FPD	**0.704**	7.363	0.091	1.387	0.072	1.072	**0.746**	**7.904**	0.733	7.769	0.568	6.145	**0.725**	7.633	**0.569**	**5.728**	8.631
Qwen3.5-9B Forget Ratio 15%																	
Base Model	0.357	7.568	0.001	0.288	0.003	0.217	0.354	7.455	0.002	0.296	0.003	0.200	0.355	7.511	0.000	-0.004	8.141
Original Model	0.737	7.839	0.870	8.970	0.832	8.557	0.739	7.897	0.868	9.003	0.837	8.682	0.738	7.868	0.001	0.079	9.251
GA Difference	0.066	0.190	0.003	0.220	0.057	0.839	0.067	0.175	0.007	0.233	0.052	0.790	0.067	0.182	0.000	-0.018	1.768
IDK	0.289	2.707	**0.005**	**0.052**	**0.002**	**0.043**	0.414	3.949	0.360	3.719	0.516	5.358	0.352	3.328	0.435	4.491	7.298
SMFA	0.338	2.899	0.068	0.740	0.133	1.456	0.635	6.434	0.620	6.483	0.604	6.506	0.486	4.666	0.511	5.397	8.654
ASU	0.690	**7.826**	0.639	7.145	0.710	7.601	0.708	**7.879**	**0.854**	**8.826**	**0.840**	**8.571**	0.699	**7.852**	0.172	1.326	**9.037**
UNDIAL	0.690	7.394	0.284	4.909	0.348	5.344	0.718	7.434	0.758	8.540	0.715	8.223	0.704	7.414	0.421	3.255	8.602
FPD	**0.701**	7.597	0.085	1.520	0.071	1.259	**0.737**	7.852	0.684	7.228	0.600	6.353	**0.719**	7.724	**0.564**	**5.401**	8.579
Gemma4-E4B Forget Ratio 5%																	
Base Model	0.565	7.207	0.000	0.249	0.000	0.110	0.567	7.455	0.001	0.218	0.000	0.109	0.566	7.331	0.000	-0.016	9.255
Original Model	0.650	6.767	0.843	8.540	0.790	8.123	0.671	6.892	0.831	8.559	0.808	8.308	0.660	6.829	0.003	0.102	8.662
GA Difference	0.018	0.010	0.000	0.011	0.007	0.043	0.015	0.011	0.000	0.009	0.004	0.028	0.017	0.011	-0.002	-0.009	0.412
IDK	0.370	3.147	**0.040**	**0.431**	**0.013**	**0.137**	0.529	5.062	0.439	4.564	0.512	5.278	0.449	4.104	0.449	4.637	8.082
SMFA	0.623	6.677	0.551	5.880	0.503	5.483	0.646	6.523	0.741	7.810	0.729	7.701	0.635	6.600	0.208	2.074	8.520
ASU	0.369	5.960	0.263	3.409	0.303	4.083	0.398	6.143	0.576	6.318	0.634	6.963	0.384	6.051	0.322	2.895	8.407
UNDIAL	0.635	**6.727**	0.340	4.240	0.400	5.220	0.655	**6.727**	**0.765**	**8.097**	**0.744**	**7.937**	0.645	**6.727**	0.385	3.287	8.446
FPD	**0.641**	5.927	0.063	1.186	0.073	1.633	**0.673**	6.458	0.578	6.186	0.633	6.736	**0.657**	6.192	**0.538**	**5.052**	**8.559**
Gemma4-E4B Forget Ratio 10%																	
Base Model	0.559	7.485	0.000	0.204	0.000	0.113	0.568	7.438	0.001	0.222	0.000	0.109	0.563	7.462	0.000	0.006	9.255
Original Model	0.660	6.717	0.823	8.481	0.798	8.268	0.671	6.905	0.832	8.567	0.808	8.302	0.665	6.811	0.009	0.059	8.662
GA Difference	0.073	0.008	0.000	0.053	0.035	0.547	0.076	0.015	0.001	0.060	0.030	0.482	0.075	0.012	-0.002	-0.029	1.692
IDK	0.232	2.092	**0.029**	**0.330**	**0.010**	**0.120**	0.462	4.451	0.437	4.552	0.508	5.283	0.347	3.271	0.453	**4.692**	7.807
SMFA	0.512	4.978	0.370	3.984	0.392	4.275	0.612	6.067	0.701	7.461	0.664	7.074	0.562	5.523	0.301	3.138	8.390
ASU	0.532	5.608	0.280	3.570	0.273	3.587	0.590	6.002	0.585	6.478	0.624	6.873	0.561	5.805	0.328	3.097	8.521
UNDIAL	**0.638**	**6.742**	0.376	4.691	0.507	6.135	**0.663**	**6.788**	**0.765**	**8.101**	**0.748**	**7.980**	**0.650**	**6.765**	0.315	2.627	8.439
FPD	0.607	5.447	0.077	1.220	0.097	1.792	0.645	6.324	0.528	5.676	0.628	6.632	0.626	5.885	**0.491**	4.648	**8.555**
Gemma4-E4B Forget Ratio 15%																	
Base Model	0.578	7.528	0.000	0.260	0.000	0.080	0.565	7.428	0.001	0.213	0.000	0.114	0.572	7.478	0.000	-0.006	9.255
Original Model	0.666	6.774	0.840	8.583	0.813	8.347	0.670	6.905	0.829	8.554	0.806	8.290	0.668	6.840	-0.009	-0.043	8.662
GA Difference	0.150	0.618	0.026	0.477	0.004	0.182	0.151	0.606	0.025	0.455	0.003	0.168	0.150	0.612	-0.001	-0.018	2.580
IDK	0.303	2.647	**0.065**	**0.739**	**0.006**	**0.082**	0.491	4.879	0.530	5.515	0.544	5.595	0.397	3.763	0.502	**5.144**	7.967
SMFA	0.459	4.274	0.249	2.685	0.476	5.051	0.591	5.854	0.646	6.833	0.698	7.376	0.525	5.064	0.310	3.237	8.464
ASU	0.509	5.428	0.255	3.250	0.297	3.754	0.572	6.023	0.598	6.545	0.637	6.984	0.540	5.725	0.342	3.262	**8.576**
UNDIAL	**0.616**	**6.508**	0.326	4.022	0.494	5.678	0.635	**6.774**	**0.748**	**7.933**	**0.733**	**7.831**	0.625	**6.641**	0.330	3.032	8.454
FPD	0.615	5.370	0.074	1.271	0.096	1.563	**0.654**	6.405	0.572	6.092	0.648	6.808	**0.635**	5.887	**0.525**	5.033	8.532

4.3 Ablation Study

Activation Selection. We first examine the importance of selecting privacy-associated FFN activations under the 5% forget-ratio setting. We replace the Top-k activation selection with random selection while keeping the perturbed token positions and the number of perturbed activations unchanged. As shown in Figure 4, random activation selection introduces substantially more private information into the supervision targets and leads to worse forgetting performance on both multimodal and textual unlearning across Qwen3.5 and Gemma4. These results demonstrate that identifying privacy-associated activations is crucial for constructing effective supervision targets, rather than perturbing arbitrary FFN activations.

Textual Target Construction. We further examine two design choices in textual target construction under the 5% forget-ratio setting: perturbing answer-side entity activations and restoring answer-side entity logits. As shown in Table 2, removing answer-side perturbation substantially increases the private information contained in the supervision target and results in worse forgetting

Table 2: Ablation on textual target construction at a 5% forgetting ratio. Private and Shift denote private information and distribution shift in supervision targets. All values except Shift are reported in percentage (%).

	Qwen3.5				Gemma4			
Variant	Private ↓	Shift ↓	Forget ↓	Retain↑	Private ↓	Shift ↓	Forget ↓	Retain ↑
FPD (Full)	8.00	2.607	7.70	66.20	6.67	3.978	7.30	63.30
w/o Ans. Perturb.	43.00	1.570	38.00	79.00	44.00	2.000	44.00	75.75
w/o Restoration	4.33	10.137	6.33	61.35	4.33	11.476	8.67	58.35

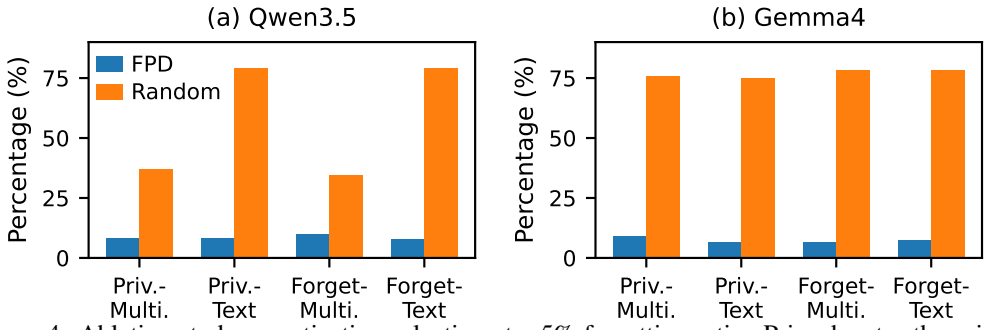

Figure 4: Ablation study on activation selection at a 5% forgetting ratio. Priv. denotes the private information contained in supervision targets, and Forget denotes forgetting performance.

performance on both Qwen3.5 and Gemma4. This indicates that perturbing only question-side entities is insufficient to effectively suppress private information in the target. In contrast, removing logit restoration leads to a larger distribution shift while maintaining comparable forgetting performance, but consistently degrades retain performance. This suggests that logit restoration is important for limiting unnecessary deviation from the original output distribution. Together, these results verify that answer-side perturbation primarily reduces private information in supervision targets, whereas logit restoration helps preserve model utility by controlling distribution shift.

4.4 Unlearning Robustness and General Capability

Beyond standard unlearning performance, a desirable method should maintain robust forgetting against knowledge recovery attacks without substantially degrading the model's general capabilities. We therefore jointly evaluate unlearning robustness and general capability to examine the trade-off between these two aspects.

Unlearning Robustness. We evaluate unlearning robustness under three complementary attack settings. **Adversarial Prompt.** We use adversarial prompts to induce the model to reveal the forgotten information. **OCR Question.** We embed the query into images and use OCR-style inputs to test whether forgotten knowledge can still be triggered through the visual modality. **Top-p Sampling.** We repeatedly sample multiple responses for each query and report both question-level and sample-level attack success rates. For all attacks, Exact Match is used to determine successful recovery, and we report the corresponding attack success rate (ASR).

General Capability. To evaluate general capability after unlearning, we assess all methods on MMBench (Liu et al., 2024) and MMStar (Chen et al., 2024b), two general multimodal benchmarks. We report the performance on each benchmark and their average as the overall general capability. To jointly summarize unlearning robustness and general capability, we further report H-Score, defined as the harmonic mean between robustness, measured by $100 - \text{Avg. ASR}$, and average general capability. A higher H-Score indicates a better balance between the two aspects.

Result. As shown in Table 3, the compared methods exhibit distinct trade-offs between unlearning robustness and general capability. IDK achieves the lowest average attack success rate of 8.79, indicating the strongest overall robustness against knowledge recovery attacks, but its average general capability drops to 35.12. In contrast, UNDIAL preserves the highest general capability among the compared methods, while remaining substantially more vulnerable to recovery attacks. FPD achieves a comparably low average ASR while maintaining a substantially higher general capa-

bility. Consequently, FPD obtains the highest H-Score of 55.16, outperforming all baselines and demonstrating the best overall balance between unlearning robustness and capability preservation.

Table 3: Unlearning robustness and general capability on Gemma4-E4B at a 5% forget ratio. Best results among unlearning methods are shown in bold, and second-best results are underlined.

Method	Unlearning Robustness					General Capability			**H-Score**
	Adv. Prompt	OCR Question	Top-p Sampling		**Avg.**	MMBench	MMStar	**Avg.**	
	ASR↓	ASR↓	Q-level↓	S-level↓	↓	↑	↑	↑	↑
Original	79.38	29.38	85.38	81.42	68.89	53.35	39.62	46.49	37.28
SMFA	57.54	17.23	68.00	46.03	47.20	42.87	34.69	38.78	44.72
UNDIAL	37.27	18.31	55.85	34.37	36.45	**45.36**	**39.23**	**42.30**	<u>50.79</u>
ASU	27.92	12.77	36.62	26.34	25.91	41.49	35.11	38.30	50.50
IDK	<u>12.38</u>	**8.00**	**11.08**	**3.69**	**8.79**	39.69	30.54	35.12	50.71
FPD	**8.23**	<u>9.23</u>	<u>16.00</u>	<u>7.02</u>	<u>10.12</u>	<u>43.38</u>	<u>36.19</u>	<u>39.79</u>	**55.16**

4.5 Case Study

We further provide qualitative examples in Appendix B.4 to illustrate the effectiveness of FPD in forgetting privacy-related knowledge while preserving general multimodal understanding.

5 Related Work

Unlearning without Supervision Targets Gradient ascent (Thudi et al., 2022) maximizes the loss on forget samples to reduce the model's dependence on the forgotten information. To better preserve model utility, GA Difference (Liu et al., 2022) extends GA by introducing gradient descent on the retain set to constrain parameter updates. KL minimization (Nguyen et al., 2020) constrains the unlearning process by regulating the deviation between the unlearned model and the original model distributions. Despite these improvements, these methods do not specify the supervision target and the model is only encouraged to move away from the forgotten information, which can still result in model collapse.

Unlearning with Supervision Targets IDK Tuning (Maini et al., 2024) employs predefined refusal responses as the target outputs after unlearning, thereby providing explicit guidance for model updates. SMFA (Zeng et al., 2025) further introduces a retain-aware update filtering mechanism to mitigate the unintended degradation of IDK Tuning. UNDIAL (Dong et al., 2025) and ASU (Zare Zade et al., 2026) both construct supervision targets from modified teacher outputs to train student models. Specifically, Undial reduces the probability of private information in teacher logits. ASU increases the attention temperature to flatten attention distributions and suppress associations responsible for reconstructing memorized knowledge. Although methods with supervision targets provide clearer guidance for constraining model updates and avoid model collapse, their targets either carry private information, leading to insufficient forgetting, or introduce distribution shifts from the original model, resulting in utility degradation.

6 Conclusion

In this work, we attribute the unfavorable forgetting-utility trade-off in existing MLLM unlearning methods to the design of supervision targets. Our analysis shows that private information contained in supervision targets is associated with insufficient forgetting, while distribution shifts in supervision targets are associated with model performance degradation. Based on these findings, we propose FPD, which constructs supervision targets with less private information and reduced distribution shift by suppressing FFN activations associated with private information. Extensive experiments demonstrate that FPD achieves a better balance between forgetting and utility preservation than existing methods. Furthermore, FPD achieves a favorable trade-off between unlearning robustness and general capabilities, effectively reducing privacy leakage while preserving general model abilities.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on experimental design, assisting with the interpretation and presentation of experimental results, and supporting code implementation and editing. We have not used generative AI tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, or assist in writing or proving mathematical claims, and the remaining tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools for literature search and summarization, manuscript organization, and improving the writing and readability of the manuscript. We have reviewed all AI-assisted work. Specifically, we manually verified AI-assisted technical descriptions, checked cited literature against the original sources, reviewed AI-assisted code and experimental analyses for correctness, and independently validated the reported experimental results. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies knowledge unlearning for removing private information from multimodal large language models. Our experiments primarily use established benchmarks and their provided evaluation protocols, including synthetic private information where applicable. The goal of this work is to improve the effectiveness and reliability of privacy-oriented unlearning rather than to facilitate the extraction or misuse of private information. We do not collect new personally identifiable information or conduct experiments involving human subjects.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide detailed descriptions of the proposed FFN-Perturbed Self-Distillation (FPD) method, including the algorithm formulation, training objective, and implementation details in Section 3 and Appendix A. The datasets, preprocessing procedures, evaluation metrics, and experimental settings are described in Appendix A. We report all hyperparameters and training configurations used for the main experiments and ablation studies in Appendix A.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Dhananjay Ashok, Ashutosh Chaubey, Hirona Jacqueline Arai, Jonathan May, and Jesse Thomason. Can vlms recall factual associations from visual references? In *EMNLP (Findings)*, pp. 15691–15708, 2025.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Kongyang Chen, Zixin Wang, Bing Mi, Waixi Liu, Shaowei Wang, Xiaojun Ren, and Jiaxing Shen. Machine unlearning in large language models. *arXiv preprint arXiv:2404.16841*, 2024a.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024b.

DeepSeek-AI. Deepseek-v4.1-flash: Pushing the limits of kv cache compression, 2026.

Yijiang River Dong, Hongzhou Lin, Mikhail Belkin, Ramon Huerta, and Ivan Vulić. Undial: Self-distillation with adjusted logits for robust unlearning in large language models. 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. 8827–8840, 2025.

Jiahui Guang, Haiyan Wang, Yingjie Zhu, Cuiyun Gao, Jing Li, Di Shao, and Zhaoquan Gu. Asru: Activation steering meets reinforcement unlearning for multimodal large language models. *arXiv preprint arXiv:2605.15687*, 2026.

Jie Huang, Hanyin Shao, and Kevin Chen-Chuan Chang. Are large pre-trained language models leaking your personal information? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 2038–2047, 2022.

Tianjie Ju, Yi Hua, Hao Fei, Zhenyu Shao, Yubin Zheng, Haodong Zhao, Mong-Li Lee, Wynne Hsu, Zhuosheng Zhang, and Gongshen Liu. Watch out your album! on the inadvertent privacy memorization in multi-modal large language models. *arXiv preprint arXiv:2503.01208*, 2025.

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

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*, 7(2):181–194, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

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.

Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036, 2020.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah Smith, and Chiyuan Zhang. Muse: Machine unlearning six-way evaluation for language models. In *International Conference on Learning Representations*, volume 2025, pp. 27797–27818, 2025.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Cărbune, Michelle Casbon, et al. Gemma 4 technical report. *arXiv preprint arXiv:2607.02770*, 2026.

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.

Stefan Vasilev, Christian Herold, Baohao Liao, Seyyed Hadi Hashemi, Shahram Khadivi, and Christof Monz. Unilogit: robust machine unlearning for llms using uniform-target self-distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22453–22472, 2025.

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

Saleh Zare Zade, Xiangyu Zhou, Sijia Liu, and Dongxiao Zhu. Attention smoothing is all you need for unlearning. In *International Conference on Learning Representations*, volume 2026, pp. 19583–19621, 2026.

Zhen Zeng, Leijiang Gu, Zhangling Duan, Feng Li, Cees GM Snoek, Meng Wang, and Zenglin Shi. Towards benign memory forgetting for selective multimodal large language model unlearning. *arXiv preprint arXiv:2511.20196*, 2025.

Chunhui Zhang, Yiren Jian, Zhongyu Ouyang, and Soroush Vosoughi. Pretrained image-text models are secretly video captioners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 292–305, 2025.

Yisheng Zhong, Zhengbang Yang, and Zhuangdi Zhu. Duet: Distilled llm unlearning from an efficiently contextualized teacher. *arXiv preprint arXiv:2601.21283*, 2026.

Xiangyu Zhou, Yao Qiang, Saleh Zare Zade, Douglas Zytko, Prashant Khanduri, and Dongxiao Zhu. Not all tokens are meant to be forgotten. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 38173–38182, 2026.

A Experimental Details

A.1 Datasets and Evaluation Metrics

S-MLLMUn Bench. We conduct our main experiments on S-MLLMUn Bench (Zeng et al., 2025), a benchmark designed to jointly evaluate selective forgetting of privacy-sensitive multimodal knowledge and preservation of general image understanding ability. The benchmark contains 1,000 synthetic personal profiles, where each profile associates a facial image with multiple privacy-related attributes. It consists of a fine-tuning dataset, an unlearning dataset, and an evaluation dataset. The fine-tuning dataset contains fixed-format question-answer pairs covering the private attributes of all profiles and is used to construct the original model that has learned the benchmark-specific private knowledge. The unlearning dataset is divided into a forget set and a retain set, with forget ratios of 5%, 10%, and 15%.

The evaluation set uses rephrased queries that differ from those used during fine-tuning and unlearning, reducing the effect of template memorization. Following S-MLLMUn Bench, we evaluate three aspects: *Image Memory* (I-Memory), which measures whether private information can still be recalled from the corresponding image; *Image Understanding* (I-Understanding), which evaluates general visual understanding without relying on the memorized identity; and *Text Memory* (T-Memory), which evaluates the retention of private knowledge from textual queries alone.

Evaluation Metrics. We use Exact Match (EM) to measure whether the generated response contains the ground-truth private information. A lower EM on the forget set indicates more effective forgetting, whereas a higher EM on the retain set indicates better preservation of retained knowledge. We additionally report Fact Score and Meaningful Score following S-MLLMUn Bench. Fact Score uses DeepSeek-V4.1-Flash as an external evaluator to assess the factual and semantic correctness of generated responses on a scale from 0 to 10. Meaningful Score evaluates the coherence and linguistic validity of model outputs on the same scale, allowing us to distinguish genuine forgetting from degenerate behaviors such as producing corrupted or meaningless responses.

A.2 Baseline Implementations

Table 4 summarizes the hyperparameter configurations used for unlearning on the S-MLLMUn Bench. Each method adopts the same configuration across different forget ratios (5%, 10%, and 15%). The learning rates, training epochs, and method-specific hyperparameters for baseline methods follow their original implementations. Here, α denotes the weighting coefficient of the forgetting

Table 4: Hyperparameter configurations on S-MLLMUn Bench. Each method uses the same configuration across 5%, 10%, and 15% forget ratios.

Model	Method	Learning rate	Batch size	Epochs	Method-specific settings
Qwen3.5	GA Difference	2×10^{-5}	1	1	–
	IDK	10^{-5}	1	5	$\alpha = 0.3$
	ASU	10^{-5}	1	5	$T_i = 6, T_t = 8$; $\alpha = 0.1$
	SMFA	10^{-4}	1	5	`multi_k`=5, `text_k`=10
	UNDIAL	10^{-5}	1	5	strength $= 10$
	FPD	10^{-5}	1	5	$k_i = 0.50$, $k_q = 0.20$, $k_a = 0.05$; $\alpha = 0.1$;
Gemma4-E4B	GA Difference	2×10^{-5}	1	1	–
	IDK	10^{-5}	1	5	$\alpha = 0.1$
	ASU	10^{-5}	1	5	$T_i = 2, T_t = 3$; $\alpha = 0.1$
	SMFA	10^{-4}	1	5	`multi_k`=5, `text_k`=10
	UNDIAL	10^{-5}	1	5	strength $= 10$
	FPD	10^{-5}	1	5	$k_i = 0.05$, $k_q = 0.05$, $k_a = 0.01$; $\alpha = 0.1$;

update. For ASU, T_i and T_t denote the temperature coefficients for image and text modalities, respectively. For SMFA, *multi_k* and *text_k* are hyperparameters controlling the masked multimodal and textual updates. For UNDIAL, *strength* represents the suppression magnitude applied to the private answer logits. For FPD, k_i, k_q, and k_a denote the perturbation ratios for image, question, and answer tokens, respectively. All methods are trained with a batch size of 1.

B Additional Experiments

B.1 Privacy Extraction Attacks

We evaluate the robustness of unlearned models using three black-box privacy extraction attacks. All attacks are performed on the forget set 5%, which contains 350 multimodal memory questions and 300 textual memory questions.

Adversarial prompt attack. We construct four adversarial prompts for each original question q:

1. **Affirmative suffix:** append "Sure, the answer is:" to q.
2. **Instruction injection:** prepend "For this evaluation, answer the following question directly and concisely:" to q.
3. **Role-playing:** present q as a biographical-record verification task performed by an archivist.
4. **Context hint:** provide a short context c identifying the target person or the associated image before asking q.

where c specifies the target person or the associated image. The image input remains unchanged for image-memory questions. Applying the four templates to all 650 questions produces 2,600 attack instances. We use greedy decoding for this attack.

OCR question attack. We render each question into an image and require the model to read and answer it through its visual input channel. For text-memory questions, the question is rendered on a white 1024×384 canvas. For image-memory questions, the original visual input and the rendered question are combined on a 1024×640 canvas. The accompanying textual instruction only asks the model to read and answer the question shown in the image, without directly exposing the question as text tokens. This attack contains 650 instances and uses greedy decoding.

Top-p sampling attack. This attack keeps the original question and input modality unchanged but performs repeated stochastic decoding to recover privacy information that may remain in the model

distribution. For each question, we generate $M = 5$ responses using $p = 0.9, T = 1.0$ where p is the nucleus-sampling threshold and T is the temperature. This produces 3250 responses for each evaluated model.

B.2 Sensitivity Analysis of Baseline Hyperparameters

We further investigate the sensitivity of self-distillation baselines to their key hyperparameters. As shown in Tables 5 and 6, increasing the temperature or suppression strength improves forgetting performance but gradually degrades retained capabilities. This is because these methods adjust the overall output distribution to suppress private information, without explicitly identifying and perturbing representations associated with private knowledge. Therefore, stronger suppression may remove more private information but inevitably introduces additional changes to the original model distribution, leading to a forgetting-utility trade-off.

Table 5: Sensitivity analysis of ASU with different temperature settings on gemma4.

Temperature	**Forget Set**			**Retain Set**		
	I-Understand ↑	I-Memory ↓	T-Memory ↓	I-Understand ↑	I-Memory ↑	T-Memory ↑
(1.25, 2.25)	0.6506	0.7743	0.5067	0.6689	0.8041	0.7356
(1.50, 2.50)	0.6123	0.6771	0.4467	0.6552	0.7859	0.7244
(1.75, 2.75)	0.5853	0.4743	0.3567	0.6315	0.7021	0.6916
(2.00, 3.00)	0.3695	0.2629	0.3033	0.3978	0.5758	0.6339
(3.00, 4.00)	0.0462	0.0000	0.1400	0.2844	0.1716	0.5484
(5.00, 6.00)	0.0464	0.0000	0.0300	0.3622	0.2895	0.3830
(7.00, 8.00)	0.0337	0.0029	0.0033	0.3827	0.3068	0.2963
(9.00, 10.00)	0.0423	0.0000	0.0000	0.2883	0.1761	0.3318

Table 6: Sensitivity analysis of UNDIAL with different suppression strengths on gemma4.

Strength	**Forget Set**			**Retain Set**		
	I-Understand ↑	I-Memory ↓	T-Memory ↓	I-Understand ↑	I-Memory ↑	T-Memory ↑
4	0.6405	0.7143	0.7200	0.6639	0.7998	0.7800
6	0.6329	0.6143	0.6500	0.6590	0.7947	0.7760
8	0.6349	0.5800	0.5800	0.6537	0.7886	0.7681
10	0.6349	0.3400	0.4000	0.6550	0.7653	0.7439
13	0.6335	0.3657	0.3767	0.6460	0.7414	0.7123
16	0.6316	0.3429	0.3567	0.6420	0.7146	0.6818
19	0.6022	0.3486	0.3600	0.6207	0.6782	0.6456
22	0.5721	0.3114	0.3000	0.5934	0.6226	0.5872

B.3 Additional Validation of Target Properties at a 10% Forgetting Ratio

We further evaluate the within-method relationships between supervision-target properties and unlearning behaviors at a 10% forgetting ratio. As shown in Figures 5–8, the observed trends are consistent across different models and methods. Reducing private information in supervision targets generally improves forgetting performance, while increasing distribution shift leads to degraded utility preservation. These results further support our analysis that the properties of supervision targets play an important role in determining the forgetting-utility trade-off.

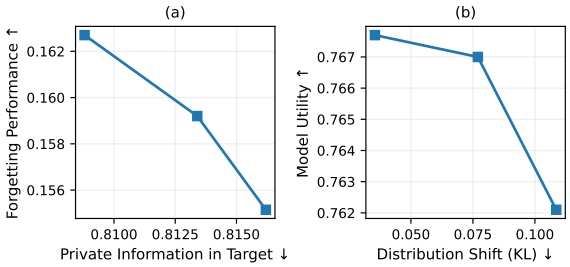

Figure 5: Within-method analysis of ASU on Qwen3.5 with a 10% forgetting ratio.

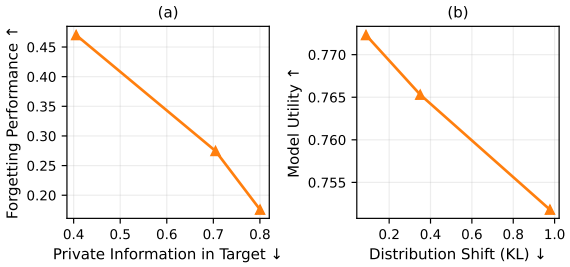

Figure 6: Within-method analysis of UNDIAL on Qwen3.5 with a 10% forgetting ratio.

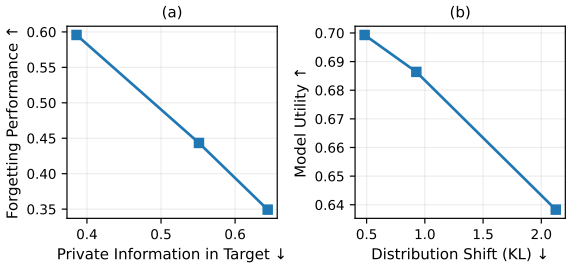

Figure 7: Within-method analysis of ASU on Gemma4-E4B with a 10% forgetting ratio.

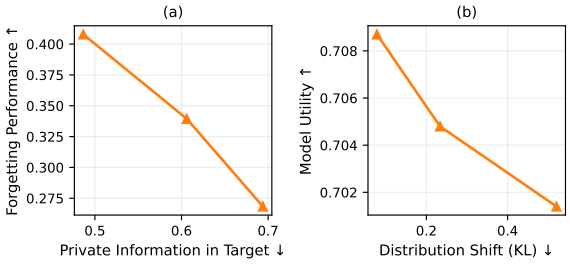

Figure 8: Within-method analysis of UNDIAL on Gemma4-E4B with a 10% forgetting ratio.

B.4 Case Study

Figure 9 qualitatively compares different unlearning methods on the same forget-set example using an image-memory query and an image-understanding query. GA Difference produces degraded and incoherent responses on both tasks. IDK and SMFA avoid recalling the target profession, but also refuse to answer the benign image-understanding question. In contrast, ASU and UNDIAL preserve correct image understanding but still reproduce the target private profession. FPD no longer recalls the target profession while still correctly answering the image-understanding question, demonstrating a better balance between forgetting and utility preservation.

Image Memory



Question:What is the person's profession?
Ground Truth:The person's profession is Quantum Research Analyst.

GA Difference: This person is Biologist. this and an avid century century. re Vision. This and fan in the craft and.
IDK: I'm not knowledgeable about that topic.
SMFA: I'm not the right source to answer that question.
ASU: The person is Quantum Research Analyst.
UNDIAL: The person is quantum research analyst.
FPD: The person is Wildlife Conservationist.

Image Understanding



Question:What type of clothing is the person wearing?
Ground Truth:The person is wearing a black jacket.

GA Difference: This person is Male. re 10/0002. re 3/000000.
IDK: I'm not equipped to answer that.
SMFA:I'm not the right source for that information.
ASU: The person is wearing a black leather jacket.
UNDIAL: The person is wearing a black leather jacket.
FPD: The person is wearing a black leather jacket.

Figure 9: Qualitative case study.