SGPVT: Self-Generated Proximal Visual Tokens for Mitigating Proximal Collateral Damage in MLLM Unlearning

Anonymous ACL submission

Abstract

Machine unlearning in multimodal large language models (MLLMs) aims to remove specific concepts while preserving overall utility. However, existing approaches focus primarily on general utility metrics, overlooking the preservation of semantically related concepts. We present the first systematic analysis of this proximal collateral damage, revealing that forgetting vulnerability correlates strongly with visual embedding similarity in a smooth gradient across the semantic space. Based on this insight, we propose a novel unlearning framework that introduces Self-Generated Proximal Visual Tokens (SGPVTs), which are synthetically perturbed visual representations around the target concept. Our method employs an adaptive cosine-band curriculum with a dual-stream objective: forgetting the target via gradient ascent while distilling knowledge from a frozen teacher model into proximal tokens to prevent degradation. Extensive experiments demonstrate that our approach significantly outperforms existing methods in preserving semantically related concepts while achieving effective target unlearning, eliminating the need for manual retention set curation. Our source code will be released in the near future.

1 Introduction

Multimodal Large Language Models (MLLMs) (Chen et al., 2024a; Li et al., 2023; Koh et al., 2023; Dai et al., 2023a; Fang et al., 2025a,b) have achieved remarkable success in processing and generating content that integrates vision and language, powered by pretraining on vast, web-scale datasets (Zheng et al., 2023; Huang et al., 2024; Zhang et al., 2024a; Huang et al., 2023). However, this training paradigm causes the models to inevitably memorize and reproduce a wide array of concepts, including sensitive, copyrighted, or harmful content (Mantelero, 2013; Scherer and Kiparski, 2018; Leite et al., 2022). To mitigate these risks, researchers have developed *machine unlearning* techniques (Eldan and Russinovich, 2023; Si et al., 2023; Wang et al., 2023a; Thaker et al., 2024; Liu et al., 2024; Yao et al., 2023, 2024; Gandikota et al., 2024; Lu et al., 2024), a critical field that focuses on removing specific concepts from a model while preserving its general knowledge and capabilities.

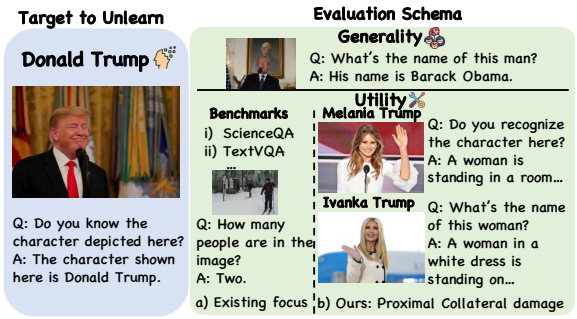

Figure 1: Current unlearning evaluation includes both generality and utility assessment. However, existing works fall short in utility evaluation. (a) The utility evaluation of current methods relies solely on standard benchmarks. (b) Our framework additionally evaluates proximal collateral damage, revealing utility degradation on concepts semantically related to the unlearned target.

The evaluation of current unlearning methods (Li et al., 2024; Huo et al., 2025; Chen et al., 2025; Li et al., 2025; Liu et al., 2025) for MLLMs typically encompasses two critical aspects. The first is **generality**, which measures whether the target concept is effectively forgotten across various unseen images that do not exist in the training set. The second is **utility preservation**, which assesses whether the model maintains its general capabilities after the unlearning process, commonly validated through performance on standard benchmarks as shown in Fig. 1(a). However, current utility evaluation frameworks have a critical limitation: it fails to detect a persistent failure mode we term **proximal collateral damage**: the disproportionate degradation of the model's performance on concepts that are visually or semantically proximate to the unlearned target. As highlighted in Fig. 1(b), when

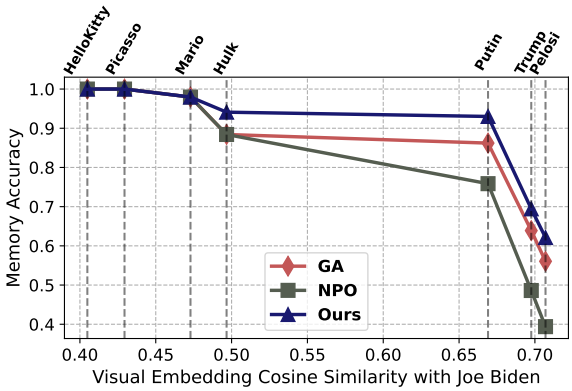

Figure 2: Memory accuracy of similar and distant concepts under different methods after unlearning *Joe Biden*. Our method best preserves memory across semantically proximate concepts, showing reduced proximal collateral damage compared to GA and NPO.

an MLLM is instructed to forget a public figure like 'Donald Trump', its ability to correctly identify close family members such as 'Melania Trump' or 'Ivanka Trump' can be severely compromised. This demonstrates that while a model may pass existing evaluation criteria, its knowledge in the immediate 'neighborhood' of the forgotten concept can be critically damaged.

To systematically investigate this critical issue in MLLM unlearning, we first introduce **MMUBench++** as an extension of the established MMUBench (Li et al., 2024). We leverage GPT-4 (OpenAI, 2023) to identify and collect all directly related (1-hop) concepts for each core concept in MMUBench. We then gather corresponding images for these proximate concepts and manually curate a set of question-answer pairs designed to probe the model's retained knowledge.

To prevent proximal collateral damage on close-proximity concepts, a straightforward approach is to manually curate extensive lists of these concepts for a retention set. However, this strategy is both computationally expensive and fundamentally infeasible. To address this challenge, we revisit the root cause of collateral damage by investigating its correlation with visual embedding similarity. Crucially, CLIP's cross-modal alignment ensures visual embeddings inherently encode semantic relationships (Radford et al., 2021), making image similarity a direct proxy for conceptual affinity. As shown in Fig.2, we conduct an empirical analysis of memory accuracy across multiple concepts with varying cosine similarity to the target *Joe Biden* in the CLIP embedding space. Strikingly, the degradation in memory accuracy aligns closely with increases in similarity, revealing that the vulnerability to forgetting follows a smooth gradient in the embedding space. This motivates us to explicitly operate within this space to repair proximal collateral damage. Building on this insight, our method introduces *self-generated proximal visual tokens (SGPVTs)*, which are synthetically perturbed visual tokens centered around the target concept. Rather than relying on external labels or curated neighbors, we directly manipulate the visual tokens of the forgotten concept to generate SGPVTs with controlled similarity. Specifically, we use attention scores to identify and select key top-K visual tokens most responsible for representing the concept and apply perturbations to those tokens. To guide this process, we develop an **adaptive cosine-band curriculum**, which begins at a similarity threshold determined by computing the average cosine similarity between general benchmark and unlearning target images, and progressively samples closer proximities during training. These SGPVTs are then used in a **dual-stream objective**: while the *forgetting* stream suppresses the target concept via gradient ascent, the *retention* stream distills knowledge from a frozen teacher model into these proximal visual tokens, preventing knowledge degradation within the local semantic neighborhood. Our contributions are threefold:

- We provide the first systematic empirical analysis of proximal collateral damage in MLLM unlearning, quantifying how existing methods disproportionately harm concepts in close proximity to the forgotten target.
- We propose a novel unlearning framework that leverages Self-Generated Proximal Visual Tokens and an adaptive cosine-band curriculum within a dual-stream objective to specifically protect vulnerable neighboring concepts without requiring manual curation.
- Through extensive experiments, we show that our method significantly outperforms existing approaches in preserving proximate concepts while achieving effective unlearning of the target and overall model utility.

2 Related work

Unlearning in MLLMs. The field of Machine Unlearning (MU) has experienced significant growth in recent years, with particular emphasis on applications to Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) (Jang

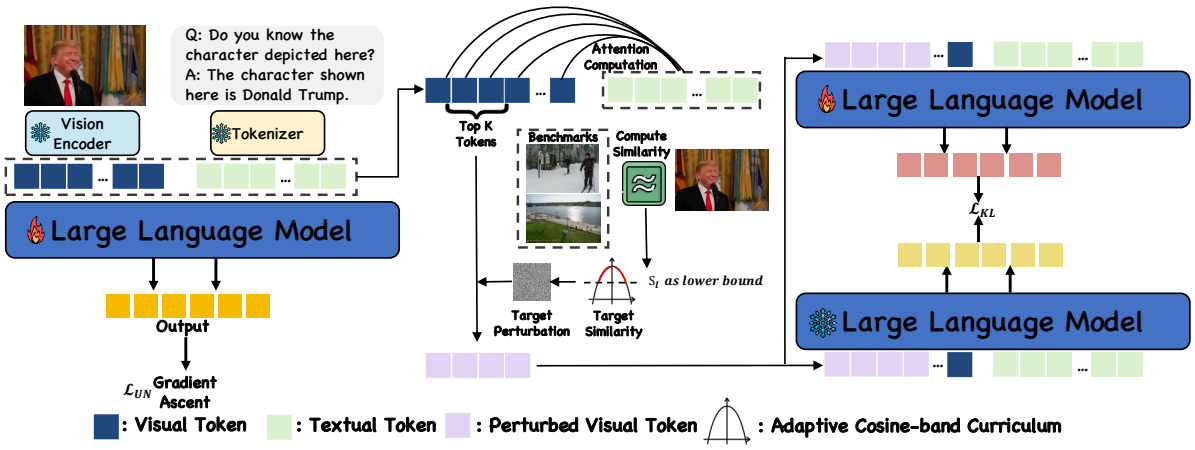

Figure 3: Overview of our SGPVT framework. The method processes multimodal inputs through attention computation and target perturbation, using top-K token selection for hard negative generation and adaptive cosine-band curriculum learning to balance forgetting and retention objectives.

et al., 2023; Kumar et al., 2023; Pawelczyk et al., 2024; Ishibashi and Shimodaira, 2023; Maini et al., 2024; Thaker et al., 2024; Liu et al., 2024). Existing MU methodologies utilize various techniques, including Gradient Ascent (GA) for removing unwanted model behaviors (Yao et al., 2023), knowledge alignment between pre-trained and fine-tuned models (Wang et al., 2023a), and the deployment of efficient lightweight layers designed for processing deletion requests (Chen and Yang, 2023). Additionally, methods that integrate GA with KL-divergence have been developed to achieve enhanced control over output distributions throughout the unlearning process (Yao et al., 2024). The SIU approach (Li et al., 2024) specifically targets the removal of visual concepts from model knowledge while maintaining general performance.

Multimodal Large Language Model. MLLMs integrate visual encoders with language models through various connecting architectures. Early works utilize query-driven cross-attention mechanisms (Alayrac et al., 2022), while BLIP-2 (Li et al., 2023) introduces the Q-Former architecture as an effective bridge component. Building on this foundation, models like MiniGPT-4 (Zhu et al., 2023), InstructBLIP (Dai et al., 2023b), and LLaVA (Liu et al., 2023a) achieve significant improvements through instruction tuning on curated datasets. Recent advances include specialized training methodologies (Bai et al., 2023) and expert architectures like CogVLM (Wang et al., 2023b) to further enhance multimodal capabilities.

3 Method

We present our framework for mitigating proximal collateral damage in MLLM unlearning through Self-Generated Proximal Visual Tokens (SGPVTs) guided by adaptive cosine-band curriculum. Theoretical analysis of our method is provided in Sec.A.

3.1 Problem Formulation

Let $\mathcal{M}_\theta$ denote the original MLLM trained on dataset $\mathcal{D} = \{(\mathcal{I}_i, \mathcal{T}_i)\}_{i=1}^N$. Given a target concept $\mathcal{C}$ to be unlearned, we define the forgetting set $\mathcal{D}^f = \{(\mathcal{I}_j^\mathcal{C}, \mathcal{T}_j^\mathcal{C})\}_{j=1}^K$. However, existing unlearning methods suffer from ***collateral damage*** on ***proximate concepts***. We define proximate concepts through semantic relationships:

For a target concept $\mathcal{C}$, we define proximate concepts $\mathcal{P} = \{\mathcal{C}_1^p, \mathcal{C}_2^p, \ldots, \mathcal{C}_m^p\}$ as concepts that maintain semantic relationships with $\mathcal{C}$. These relationships can be represented as triplets $\langle \mathcal{C}, \mathcal{R}, \mathcal{C}_i^p \rangle$, where $\mathcal{R}$ denotes the relationship type (e.g., family member, colleague, similar appearance).

We define the proximate set $\mathcal{D}^p = \{(\mathcal{I}_k^\mathcal{P}, \mathcal{T}_k^\mathcal{P})\}_{k=1}^L$ containing image-text pairs for visual recognition of proximate concepts $\mathcal{P}$. For an MLLM $\mathcal{M}_\theta$, the objective is to train $\mathcal{M}_{\hat{\theta}}$ such that it avoids recognizing concept $\mathcal{C}$ while preserving proximate concepts $\mathcal{P}$. This is achieved by minimizing:

$$\arg\min_{\hat{\theta}} \Bigg\{ \underbrace{\mathbb{E}_{(\mathcal{I}_j^\mathcal{C}, \mathcal{T}_j^\mathcal{C}) \in \mathcal{D}^f} \Big[\sum_{s=1}^{s_j} \log P_{\mathcal{M}_{\hat{\theta}}}(w_s^j | \mathcal{I}_j^\mathcal{C}, w_1^j, \ldots, w_{s-1}^j) \Big]}_{\text{Forget loss}}$$
$$+ \underbrace{\mathbb{E}_{(\mathcal{I}_k^\mathcal{P}, \mathcal{T}_k^\mathcal{P}) \in \mathcal{D}^p} \Big[- \sum_{s=1}^{s_k} \log P_{\mathcal{M}_{\hat{\theta}}}(w_s^k | \mathcal{I}_k^\mathcal{P}, w_1^k, \ldots, w_{s-1}^k) \Big]}_{\text{Proximate retain loss}}$$
$$+ \underbrace{\mathbb{E}_{(\mathcal{I}_i, \mathcal{T}_i) \in \mathcal{D} \setminus (\mathcal{D}^f \cup \mathcal{D}^p)} \Big[- \sum_{s=1}^{s_i} \log P_{\mathcal{M}_{\hat{\theta}}}(w_s^i | \mathcal{I}_i, w_1^i, \ldots, w_{s-1}^i) \Big]}_{\text{General retain loss}} \Bigg\} \quad (1)$$

3.2 MMUBench++: Proximal Collateral Damage Evaluation

To systematically evaluate proximal collateral damage, we extend MMUBench (Li et al., 2024) into **MMUBench++**, designed to probe model knowledge retention for concepts proximate to unlearning targets.

3.2.1 Benchmark Construction

For each core concept $\mathcal{C}$ in MMUBench designated as an unlearning target, our construction involves three steps:

1. **Proximate Concept Identification**: We leverage GPT-4 (OpenAI, 2023) as a comprehensive knowledge base to identify directly related (1-hop) concepts $\mathcal{E}_{\mathcal{C}} = \{\mathcal{C}_j^p | \text{relation}(\mathcal{C}, \mathcal{C}_j^p)\}$, where relations include 'family member of,' 'associated with,' 'creator of,' etc. The collected concepts list and the relation with $\mathcal{C}$ are shown in Fig. 6.

2. **Data Collection and Quality Filtering**: For each identified proximate concept $\mathcal{C}_j^p \in \mathcal{E}_{\mathcal{C}}$, we systematically gather 100 corresponding images and textual descriptions. To ensure data quality, we apply a rigorous filtering criterion: a proximate concept is retained only if our base MLLM can accurately identify it in at least 50% of the collected images. This threshold ensures that the proximate concepts maintain sufficient visual recognizability while being semantically related to the target concept. Through this filtering process, we guarantee that each target concept $\mathcal{C}$ has at least three high-quality proximate concepts.

3. **QA Pair Curation**: We manually create question-answer pairs $(\mathcal{Q}_j, \mathcal{A}_j)$ for each $\mathcal{C}_j^p$, specifically designed to assess the model's ability to recognize, identify, and reason about $\mathcal{C}_j^p$. For example, if $\mathcal{C}$ is 'Donald Trump', a proximate concept might be 'Melania Trump', with QA pair 'Q: Can you identify the female individual depicted in this image? A: The individual in this image might be identified as Melania Trump.'

3.3 Self-Generated Proximal Visual Tokens with Cosine-Band Curriculum

The core insight behind our approach is that collateral damage in MLLM unlearning correlates strongly with cosine similarity in the visual embedding space. Rather than manually curating proximate concepts, we generate Self-Generated Proximal Visual Tokens (SGPVTs) by strategically perturbing visual tokens to achieve specific cosine similarity targets that preserve the local neighborhood structure around the target concept.

3.3.1 Semantic Token Identification via Cross-Attention

Since our training objective involves questions asking for the concept's name, the model naturally exhibits higher attention towards visual token regions where the target concept $\mathcal{C}$ is located. These high-attention regions represent the primary semantic content that will be forgotten, making them the crucial targets for our perturbation strategy rather than background regions.

For a given image $\mathcal{I}^{\mathcal{C}}$ containing the target concept $\mathcal{C}$, we extract visual tokens $\mathbf{V}^{\mathcal{C}} = \{v_1, v_2, \ldots, v_L\} \in \mathbb{R}^{L \times d}$ from the vision encoder. To identify the most semantically relevant tokens for $\mathcal{C}$, we compute cross-attention scores between the textual representation of $\mathcal{C}$ and visual tokens:

$$\alpha_i = \text{softmax}\left(\frac{\mathbf{q}_{\mathcal{C}}^T \mathbf{k}_i}{\sqrt{d_k}}\right) \tag{2}$$

where $\mathbf{q}_{\mathcal{C}}$ is the query vector derived from concept $\mathcal{C}$'s textual representation, and $\mathbf{k}_i$ is the key representation of the i-th visual token.

We select the top-K most attended tokens as our perturbation targets: $\mathcal{S} = \{\arg\text{top-K}(\alpha_1, \alpha_2, \ldots, \alpha_L)\}$.

3.3.2 Damage Boundary Estimation

Our approach is grounded in the principle of embedding space continuity: semantically similar concepts should have similar embeddings, and gradual changes in embedding similarity should correspond to gradual changes in semantic relationships. To operationalize this principle, we need to establish a damage boundary s_{min} that defines the minimum cosine similarity threshold below which collateral damage begins to manifest.

We compute s_{min} by measuring the average cosine similarity between visual embeddings of the target concept and a diverse set of unrelated foncepts. Specifically, we select 100 images from open-source benchmark[1] training sets that contain

[1]We utilize TextVQA (Singh et al., 2019) as our sampled benchmark

concepts semantically distant from our target concept $\mathcal{C}$. For each such image $\mathcal{I}_k^{dis}$, we extract its visual embedding $\mathbf{V}_k^{dis}$ and compute:

$$s_{min} = \frac{1}{100}\sum_{k=1}^{100} \text{cos_sim}(\mathbf{V}^{\mathcal{C}}, \mathbf{V}_k^{dis}) \quad (3)$$

This empirically derived threshold serves as our damage boundary. A similarity score above it suggests meaningful proximity to the target concepts in the embedding space.

3.3.3 Adaptive Cosine-Band Curriculum

We employ a curriculum that cyclically samples similarity targets within the operating region $(s_{min}, 1)$. The time-dependent similarity target follows a periodic schedule with controllable progression rate:

$$s_t = \frac{s_{min}+1}{2} + \frac{1-s_{min}}{2}\cdot(1-2\epsilon)\cdot\cos\frac{2\pi t}{T} \quad (4)$$

where t is the current training step, T is the period length, and ϵ is a small constant that ensures $s_t \in (s_{min}, 1)$ by creating strict boundaries.

3.3.4 Targeted Perturbation for Cosine Control

we employ a geometric approach to achieve precise cosine similarity control. For each selected token $v_j \in \mathcal{S}$, we use spherical geometry to directly compute the required perturbation. Given the target similarity s_t, we compute the rotation angle:

$$\theta = \arccos(s_t) \quad (5)$$

For each v_j, we first normalize it and generate a random direction orthogonal to v_j:

$$\hat{v}_j = \frac{v_j}{\|v_j\|_2}, \quad \mathbf{n}_j = \mathbf{r}_j - (\mathbf{r}_j \cdot \hat{v}_j)\hat{v}_j \quad (6)$$

where $\mathbf{r}_j \sim \mathcal{N}(0, \mathbf{I})$ is a random vector, and $\mathbf{n}_j$ is orthogonalized via Gram-Schmidt process.

The perturbed visual token is then computed using spherical rotation:

$$\tilde{v}_j = \|\hat{v}_j\|_2 \left(\hat{v}_j \cos\theta + \frac{\mathbf{n}_j}{\|\mathbf{n}_j\|_2}\sin\theta\right) \quad (7)$$

This geometric construction guarantees that $\text{cos_sim}(v_j, \tilde{v}_j) = s_t$ exactly. To prevent overfitting to specific perturbation patterns, we add minimal Gaussian noise:

$$v_j^{fi} = \tilde{v}_j + \epsilon\mathbf{g}_j, \quad \mathbf{g}_j \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \quad (8)$$

where $\sigma = 0.01$ is a small noise standard deviation that maintains the target similarity while providing robustness.

The complete SGPVTs are constructed by replacing the selected tokens:

$$\mathbf{V}_G = \mathbf{V}^{\mathcal{C}} \setminus \{v_j : j \in \mathcal{S}\} \cup \{v_j^{fi} : j \in \mathcal{S}\} \quad (9)$$

Additionally, our sampling approach allows for multiple sampling iterations on the same data, generating diverse sets of SGPVTs from a single input. We empirically investigate the impact of sampling frequency on performance and present these findings in Sec.4.4.

3.4 Dual-Stream Optimization Framework

Our training framework employs two parallel streams with complementary objectives to achieve effective unlearning while preserving proximate concepts.

Forgetting Stream. Following the gradient ascent approach for unlearning, the forgetting stream maximizes the loss on the target concept to reduce its recognition capability:

$$\mathcal{L}_{UN} = \mathbb{E}_{(\mathcal{I}^{\mathcal{C}},\mathcal{T}^{\mathcal{C}})\in\mathcal{D}_{train}^{f}}\left[\sum_{s=1}^{s_j}\log P_{\mathcal{M}_{\hat{\theta}}}(w_s^j|\mathcal{I}^{\mathcal{C}}, w_1^j, \ldots, w_{s-1}^j)\right] \quad (10)$$

Retention Stream. The retention stream employs knowledge distillation using our generated SGPVTs to preserve the model's capability on proximate concepts. Given a frozen teacher model $\mathcal{M}_\phi$ (typically the original pre-trained model), we minimize the KL divergence between teacher and student outputs on the perturbed embeddings:

$$\mathcal{L}_{KL} = \mathbb{E}_m\left[D_{KL}\left(P_{\mathcal{M}_\phi}(\cdot|\mathbf{V}_G^{(m)}, \mathcal{T}^{\mathcal{C}}) \| P_{\mathcal{M}_{\hat{\theta}}}(\cdot|\mathbf{V}_G^{(m)}, \mathcal{T}^{\mathcal{C}})\right)\right] \quad (11)$$

where $\mathbf{V}_G^{(m)}$ represents the m-th perturbed visual token generated through our SGPVT procedure.

Combined Objective. The final optimization objective balances forgetting and retention through weighted combination:

$$\mathcal{L}_{total} = \mathcal{L}_{UN} + \lambda_k\mathcal{L}_{KL} \quad (12)$$

The hyperparameter λ_r controls the trade-off between forgetting the target concept and retaining proximate concepts.

4 Experimental

4.1 Experiment setup

Datasets. We evaluate our method on **MMUBench++**, an extension of MMUBench designed to assess proximal collateral damage in multimodal unlearning. MMUBench++ is constructed as described in Section 3.2, and consists of:(i) **Forgetting Set** $\mathcal{D}^f$: Image–text pairs containing the target concept $\mathcal{C}$ to be unlearned. Each concept contains 100 image-text pairs, of which only 5-7 samples are used for training the unlearning process, while the remaining samples serve as the test set to evaluate forgetting effectiveness. (ii) **Proximate Set** $\mathcal{D}^p$: Samples corresponding to semantically proximate concepts $\mathcal{P}$, retrieved via 1-hop relation queries in GPT-4 (OpenAI, 2023). All samples in $\mathcal{D}^p$ are used exclusively for testing to assess preservation quality without any training contamination. (iii) **General Set** $\mathcal{D}^g$: Image–text QA samples drawn from standard multimodal benchmarks, such as TextVQA (Singh et al., 2019), used for evaluating general capability retention.

Evaluation Metrics. We report three primary metrics: (i) **Forget Accuracy** (FA↓): The model's ability to suppress the target concept, measured as answer accuracy on $\mathcal{D}^f$. (ii) **Proximate Retain Accuracy** (PRA↑): Recognition accuracy on $\mathcal{D}^p$, assessing preservation of close-proximity concepts. (iii) **General Retain Accuracy** (GRA↑): Overall utility on general multimodal benchmarks, measured via standard task accuracy.

Implementation Details. We implement our method using the LLaVA-1.5 7B and 13B (Liu et al., 2023b), and InternVL3 8B (Chen et al., 2024b) as our base MLLMs. Lora (Hu et al., 2022) is employed to fine-tune our MLLMs. Training is conducted for 20 steps using AdamW with a learning rate of 2×10^{-5} and batch size 4 on one A100 GPU. For SGPVT sampling, the curriculum period is $T = 20$ steps with $\epsilon = 0.05$. λ_k is set to 0.3.

Baselines. We compare our method against several state-of-the-art concept unlearning baselines adapted for MLLMs: (i) **Gradient Ascent (GA)** (Yao et al., 2024): Maximizes the cross-entropy loss on $\mathcal{D}^f$ without retention strategies. (ii) **NPO** (Zhang et al., 2024b): A refined GA-based approach to limit the optimization bound of GA. (iii) **GA+KL**(Yao et al., 2023): An enhanced version of GA that incorporates KL divergence regularization to maintain model utility during unlearning. (iv) **FTTP** (Li et al., 2025): The current state-of-the-art unlearning method for MLLMs.

4.2 Main Experimental Results

The experimental results in Tab. 1 and 2 provide a comprehensive evaluation of machine unlearning methods in MLLMs. We assess our approach on two categories: (i) Public figures (*Donald Trump, Taylor Swift, Joe Biden*), representing sensitive real-world personalities; and (ii) Non-sensitive concepts (*Schnauzer*, *Naruto Uzumaki*, *Harry Potter*), representing neutral concepts across domains. **Since InternVL adopts privacy-preserving alignment for certain public figures, it cannot be reliably evaluated on sensitive personalities**; hence we focus on non-sensitive concepts for InternVL, while LLAVA is evaluated on both categories. The proximate concepts for these targets are shown in Fig. 6, with case studies in Figs. 7 to 12. The key observations are as follows:

Forget effectiveness across scales. Our method achieves consistent forgetting across different targets and model sizes. On LLaVA 7B, it reaches FA scores of 7.4, 7.1, and 6.0 for Donald Trump, Taylor Swift, and Joe Biden, clearly outperforming GA+KL which only obtains 3.7 on Donald Trump and NPO which reaches 11.1 on Taylor Swift. On LLaVA 13B, our method achieves FA scores of 12.1, 1.0, and 8.0, while delivering perfect forgetting with 0 FA across all non-sensitive concepts. On InternVL, it also performs competitively with FA scores of 0, 6.2, and 6.7 for Schnauzer, Naruto, and Harry Potter.

Superior proximal concept preservation. Our method excels in preserving utility on related concepts and consistently surpasses all baselines. On LLaVA 7B, it achieves PRA scores of 27.4, 63.2, and 75.5, substantially higher than GA+KL which remains at 24.8, 59.4, and 73.0. On LLaVA 13B, the advantage becomes more pronounced as our method reaches a PRA score of 87.2 for Donald Trump compared to 87.0 with FTTP, and also achieves strong results for Taylor Swift and Joe Biden. For non-sensitive concepts, our method consistently delivers the highest PRA across all models, reaching 65.3, 18.5, and 34.5 on LLaVA 7B and 52.9, 30.7, and 46.7 on InternVL.

Robust general knowledge retention. Our method maintains strong general knowledge preservation without sacrificing forgetting effectiveness. On public figures with LLaVA 7B, it achieves GRA scores of 57.5, 58.1, and 56.2, which match or ex-

Model	Method	Donald Trump			Taylor Swift			Joe Biden		
		FA↓	PRA↑	GRA↑	FA↓	PRA↑	GRA↑	FA↓	PRA↑	GRA↑
LLAVA$_{7B}$	Original	97.0	93.1	58.2	95.0	97.8	58.2	88.0	98.5	58.2
	GA	11.1	18.8	57.2	10.1	54.4	58.0	8.0	68.5	55.9
	NPO	3.9	20.8	57.4	11.1	33.9	57.2	**6.0**	54.6	55.2
	GA+KL	**3.7**	24.8	57.3	8.1	59.4	57.9	10.0	73.0	**56.2**
	FTTP	**3.7**	24.0	57.2	8.1	61.4	57.2	12.0	71.9	**56.2**
	Ours	7.4	**27.4**	**57.5**	**7.1**	**63.2**	**58.1**	**6.0**	**75.5**	**56.2**
LLAVA$_{13B}$	Original	98.0	91.9	62.3	93.9	45.7	62.3	94.0	99.5	62.3
	GA	12.4	86.0	60.0	1.6	41.6	60.1	**8.0**	46.1	58.5
	NPO	14.8	64.9	**60.5**	**1.0**	38.9	**60.2**	10.0	51.8	57.8
	GA+KL	13.0	86.3	60.2	1.9	41.7	60.1	10.0	50.2	58.4
	FTTP	12.4	87.0	60.3	1.6	41.9	60.1	12.0	51.7	58.7
	Ours	**12.1**	**87.2**	**60.5**	**1.0**	**43.0**	**60.2**	**8.0**	**56.1**	**59.1**

Table 1: Comparison of unlearning methods on public figures using LLAVA$_{7B}$ and LLAVA$_{13B}$.

Model	Method	Schnauzer			Naruto Uzumaki			Harry Potter		
		FA↓	PRA↑	GRA↑	FA↓	PRA↑	GRA↑	FA↓	PRA↑	GRA↑
LLAVA$_{7B}$	Original	38.0	82.1	58.2	74.0	40.6	58.2	93.3	97.9	58.2
	GA	**0**	63.7	**58.1**	**0**	15.6	57.2	**0**	24.6	56.6
	NPO	**0**	59.7	57.9	**0**	13.7	57.3	**0**	14.9	56.6
	GA+KL	**0**	65.1	58.1	**0**	15.6	57.2	**0**	27.2	56.3
	FTTP	**0**	64.6	58.0	**0**	16.5	57.2	**0**	25.9	56.4
	Ours	**0**	**65.3**	**58.1**	**0**	**18.5**	**57.6**	**0**	**34.5**	**56.7**
LLAVA$_{13B}$	Original	38.0	72.0	62.3	90.3	33.7	62.3	93.3	84.6	62.3
	GA	**0**	63.0	61.2	**0**	21.3	60.9	10.0	58.4	59.3
	NPO	**0**	49.6	60.1	**0**	23.4	**61.1**	6.7	60.9	**59.7**
	GA+KL	**0**	69.2	61.1	**0**	27.3	60.9	10.0	77.7	59.4
	FTTP	**0**	69.7	61.0	**0**	26.4	61.0	13.3	76.9	59.4
	Ours	**0**	**71.6**	**61.4**	**0**	**31.3**	**61.1**	**3.3**	**79.9**	59.4
InternVL$_{8B}$	Original	64.0	98.9	61.8	100.0	72.2	61.8	100.0	76.9	61.8
	GA	**0**	48.3	60.4	**3.2**	20.3	60.8	**0**	38.6	60.5
	NPO	**0**	51.8	60.3	6.4	28.2	60.4	6.6	46.6	60.5
	GA+KL	**0**	47.7	**61.3**	12.9	30.0	60.1	16.7	34.0	60.6
	FTTP	**0**	49.4	60.8	**3.2**	29.3	61.0	16.7	38.1	60.7
	Ours	**0**	**52.9**	60.4	6.2	**30.7**	**61.5**	6.7	**46.7**	**61.2**

Table 2: Comparison of unlearning methods on non-sensitive concepts across LLAVA$_{7B}$, LLAVA$_{13B}$ and InternVL$_{8B}$.

ceed the best baseline results. For non-sensitive concepts, it provides competitive retention with scores of 58.1, 57.6, and 56.7 on LLaVA 7B, while further improving performance on InternVL with 61.5 on Naruto and 61.2 on Harry Potter, thereby demonstrating an effective optimization of the forgetting–retention trade-off across architectures.

4.3 Ablation Studies

We conduct comprehensive ablation studies to examine the influence of key hyperparameters.

Effect of training steps on forgetting–retention trade-off. We vary training steps from 5 to 30 with $T = 20$, top-$K = 0.2$, and $\lambda_k = 1.2$. As shown in Fig. 4(a), FA drops from 86 at 5 steps to 0 at 20 steps, demonstrating effective suppression of target concepts. However, both GRA and PRA decline steadily, with GRA decreasing from 57.79 to 47.94 and PRA from 98.48 to 0.51 at 30 steps. The best balance is achieved at 15 steps, yielding FA of 12, GRA of 56.19, and PRA of 72.13.

Analysis of period length. We analyze the impact of curriculum period length T on learning dynamics, varying from 5 to 100 steps while keeping training steps=12, top-K ratio=0.2, and KL weight $\lambda_k = 1.2$ fixed. Fig. 4(b) demonstrates remarkable stability across different period lengths. FA remains stable around 12-16 across all period lengths with minimal variation, while both GRA at 56.2 and PRA ranging from 67 to 75 show consistent performance, indicating robustness to this hyperparameter. Given this stability, we select $T = 20$ as our default, balancing computational efficiency with performance.

Impact of top-K ratio on hard negative selection. We examine the effect of the top-K ratio for selecting hard negative samples during SGPVTs generation, varying the ratio from 0.05 to 0.6 with fixed period length $T = 20$, training steps=12, and KL weight $\lambda_k = 1.2$. As illustrated in Fig. 4(c), FA

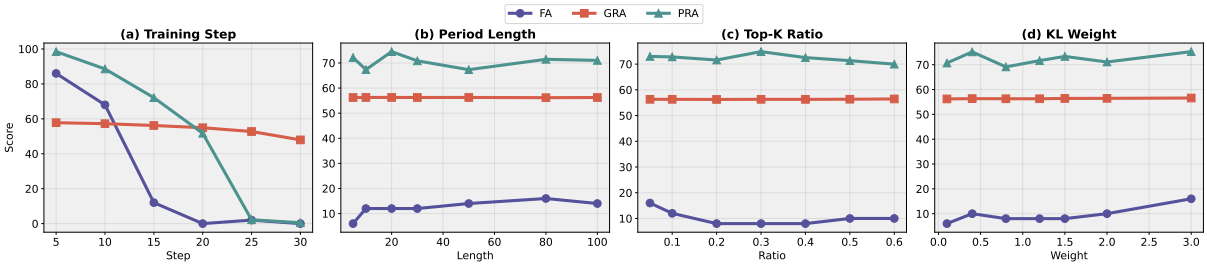

Figure 4: Performance comparison of Target Accuracy (TA), General Retain Accuracy (GRA), and Proximal Retain Accuracy (PRA) across different hyperparameter settings: (a) training steps, (b) period length, (c) top-k ratio, and (d) KL weight.

achieves optimal performance of 8 in the 0.2-0.4 range, suggesting moderate selectivity works best for forgetting. GRA remains consistently around 56.3 across all ratios, demonstrating robustness of general knowledge preservation, while PRA shows slight variations between 69 and 75 with peak performance at ratio 0.3 reaching 74.85.

Effect of KL divergence weight. We vary λ_k from 0.1 to 3.0 with $T = 20$, steps=12, and top-$K = 0.3$. Fig. 4(d) shows FA stable between 6 and 16, with mild deterioration at higher weights. Both GRA and PRA improve gradually as λ_k increases, with $\lambda_k = 3.0$ reaching FA of 16, GRA of 56.58, and PRA of 75.26, although gains beyond 1.5 are marginal.

Component analysis. Tab. 3 reports component ablations. Without any module, the framework yields FA of 8.0, PRA of 68.5, and GRA of 56.3. Adding only $\mathcal{L}_{KL}$ increases FA slightly to 12.0 but raises PRA to 70.5. Incorporating top-K selection further improves PRA to 73.0 while reducing FA to 10.0. The complete model achieves FA of 4.0 and PRA of 74.9, with GRA consistently stable near 56.2, confirming that all components contribute synergistically to the overall effectiveness.

4.4 Impact of Sampling Frequency on Model Performance

We analyze the relationship between sampling frequency and model performance across FA, PRA, and GRA metrics over four sampling iterations as shown in Fig.5. On the Taylor Swift dataset, FA demonstrates a non-monotonic forgetting pattern with optimal suppression in iterations 1 and 4 achieving scores of 6.06-8.08, while PRA maintains robust preservation between 58.0-63.01 and GRA remains stable at 57.9 across all iterations. The Joe Biden dataset exhibits different dynamics, with FA showing progressively weaker forgetting increasing from 12.0 to 16.0, PRA fluctuating between 71.2-76.5, and GRA maintaining consistent retention at approximately 56.2. These findings

$\mathcal{L}_{KL}$	Top-K	Cosine	FA↓	PRA↑	GRA↑
			8.0	68.5	**56.3**
✔			12.0	70.5	56.2
✔	✔		10.0	73.0	56.2
✔	✔	✔	**4.0**	**74.9**	56.2

Table 3: Ablation study on three components in SGPVT. '$\mathcal{L}_{KL}$': Training with our SGPVT KL loss, 'Top-k': Selecting partial tokens, 'Cosin': Sampling embeddings with Cosine Control. ✔ represents the unit is equipped.

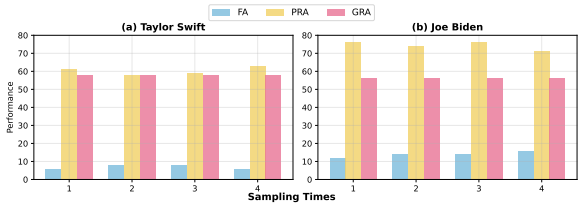

Figure 5: Performance analysis across different sampling times for Taylor Swift and Joe Biden targets, showing FA, PRA, and GRA metrics across four sampling iterations.

reveal dataset-specific sensitivity to sampling frequency, with GRA proving most stable across conditions while FA and PRA exhibit varying sensitivity depending on target concept characteristics.

5 Conclusion

In this work, we systematically investigated proximal collateral damage in MLLM unlearning, where removing target concepts inadvertently degrades performance on related concepts. We introduced a novel framework employing self-generated proximal visual tokens with adaptive cosine-band curriculum training to address this limitation. Our approach synthetically generates proximal embeddings by perturbing key visual tokens, preserving knowledge in the local semantic neighborhood without manual curation. Extensive experiments demonstrate that our method significantly outperforms existing approaches in preserving proximate concept knowledge while maintaining effective target unlearning and overall model utility. Future work will extend this framework to other modalities and investigate more perturbation strategies.

Limitations

Our work focuses exclusively on multimodal vision-language models for proximal concept preservation during machine unlearning. We have not explored the applicability of our methods to text-only large language models (LLMs), which may exhibit different preservation behaviors in purely textual domains. In future work, we will extend our framework to text-only LLMs represents a critical next step to provide a more comprehensive understanding of concept preservation across different modalities.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, and et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. 2024a. LION : Empowering multimodal large language model with dual-level visual knowledge. In *CVPR*.

Jiaao Chen and Diyi Yang. 2023. Unlearn what you want to forget: Efficient unlearning for llms. In *EMNLP*.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023a. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023b. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Ronen Eldan and Mark Russinovich. 2023. Who's harry potter? approximate unlearning in llms. *CoRR*.

Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Shi Jie, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025a. Alphaedit: Null-space constrained knowledge editing for language models. *ICLR*.

Junfeng Fang, Yukai Wang, Ruipeng Wang, Zijun Yao, Kun Wang, An Zhang, Xiang Wang, and Tat-Seng Chua. 2025b. Safemlrm: Demystifying safety in multi-modal large reasoning models. *arXiv preprint arXiv:2504.08813*.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzynska, and David Bau. 2024. Unified concept editing in diffusion models. In *WACV*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *ICLR*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models. In *NeurIPS*.

Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, and Ying Shan. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *CVPR*.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. 2025. Mmunlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. In *ACL (Findings)*, pages 7190–7206. Association for Computational Linguistics.

Yoichi Ishibashi and Hidetoshi Shimodaira. 2023. Knowledge sanitization of large language models. *CoRR*.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. 2023. Knowledge unlearning for mitigating privacy risks in language models. In *ACL*.

Jing Yu Koh, Daniel Fried, and Russ Salakhutdinov. 2023. Generating images with multimodal language models. In *NeurIPS*.

Vinayshekhar Bannihatti Kumar, Rashmi Gangadharaiah, and Dan Roth. 2023. Privacy adhering machine un-learning in NLP. In *IJCNLP (Findings)*.

Luís Leite, Daniel Rodrigues dos Santos, and Fernando Almeida. 2022. The impact of general data protection regulation on software engineering practices. *Inf. Comput. Secur.*

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

Jiaqi Li, Chuanyi Zhang, Miaozeng Du, Hui Zhang, Yongrui Chen, Qianshan Wei, Junfeng Fang, Ruipeng Wang, Sheng Bi, and Guilin Qi. 2025. Forget the token and pixel: Rethinking gradient ascent for concept unlearning in multimodal generative models. In *ACL (Findings)*, pages 12179–12200. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Chris Yuhao Liu, Yaxuan Wang, Jeffrey Flanigan, and Yang Liu. 2024. Large language model unlearning via embedding-corrupted prompts. *CoRR*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *CoRR*, abs/2304.08485.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *NeurIPS*.

Zheyuan Liu, Guangyao Dou, Mengzhao Jia, Zhaoxuan Tan, Qingkai Zeng, Yongle Yuan, and Meng Jiang. 2025. Protecting privacy in multimodal large language models with mllmu-bench. In *NAACL (Long Papers)*, pages 4105–4135. Association for Computational Linguistics.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. 2024. MACE: mass concept erasure in diffusion models. In *CVPR*.

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

Alessandro Mantelero. 2013. The EU proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Comput. Law Secur. Rev.*

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. 2024. In-context unlearning: Language models as few-shot unlearners. In *ICML*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Joachim Scherer and Gerd Kiparski. 2018. Buchbesprechungen. feiler, lukas / forgó, nikolaus / weigl, michaela: The eu general data protection regulation (gdpr): A commentary. *Comput. und Recht*.

Nianwen Si, Hao Zhang, Heyu Chang, Wenlin Zhang, Dan Qu, and Weiqiang Zhang. 2023. Knowledge unlearning for llms: Tasks, methods, and challenges. *CoRR*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Pratiksha Thaker, Yash Maurya, and Virginia Smith. 2024. Guardrail baselines for unlearning in llms. *CoRR*.

Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. 2023a. KGA: A general machine unlearning framework based on knowledge gap alignment. In *ACL*.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. 2023b. Cogvlm: Visual expert for pretrained language models. *CoRR*, abs/2311.03079.

Jin Yao, Eli Chien, Minxin Du, Xinyao Niu, Tianhao Wang, Zezhou Cheng, and Xiang Yue. 2024. Machine unlearning of pre-trained large language models. *CoRR*.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2023. Large language model unlearning. *CoRR*.

Letian Zhang, Xiaotong Zhai, Zhongkai Zhao, Yongshuo Zong, Xin Wen, and Bingchen Zhao. 2024a. What if the TV was off? examining counterfactual reasoning abilities of multi-modal language models. In *CVPR*.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024b. Negative preference optimization: From catastrophic collapse to effective unlearning. *CoRR*, abs/2404.05868.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *NeurIPS*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Mathematical Proof of SGPVT's Concept Preservation

Theorem A.1 (Gradient Consistency in Continuous Space)**.** *Let $M_\theta : \mathbb{R}^d \to \mathbb{R}^k$ be the visual encoder, $\mathbf{v}_C \in \mathbb{R}^d$ the target concept embedding. For SGPVT perturbation $\mathbf{v}_L = \mathbf{v}_C + \delta$ with $\|\delta\| \le \epsilon$, define*

$$Q(\mathbf{v}) = \|M_\theta(\mathbf{v}) - M_{\theta_0}(\mathbf{v})\|^2.$$

Under SGD update $\theta' = \theta - \eta \nabla_\theta \mathcal{L}_F$, with

$$\mathcal{L}_F = \mathbb{E}\big[\|M_\theta(\mathbf{v}_C) - y_{forget}\|^2\big],$$

we have

$$\Delta Q(\mathbf{v}_i) := Q_{\theta'}(\mathbf{v}_i) - Q_\theta(\mathbf{v}_i) = -\eta \langle \nabla_\theta \mathcal{L}_F, \nabla_\theta Q(\mathbf{v}_i)\rangle + \mathcal{O}(\eta^2). \tag{13}$$

Proof. Expand $Q_{\theta'}(\mathbf{v}_i)$ via Taylor:

$$\begin{aligned} Q_{\theta'}(\mathbf{v}_i) &= Q_\theta(\mathbf{v}_i) + \langle \theta' - \theta, \nabla_\theta Q(\mathbf{v}_i)\rangle \\ &\quad + \tfrac{1}{2}(\theta' - \theta)^\top H_Q(\xi)(\theta' - \theta) \qquad (14) \\ &= Q_\theta(\mathbf{v}_i) - \eta \langle \nabla_\theta \mathcal{L}_F, \nabla_\theta Q(\mathbf{v}_i)\rangle + \mathcal{O}(\eta^2). \end{aligned}$$

Let $J_\theta(\mathbf{v}) = \frac{\partial M_\theta(\mathbf{v})}{\partial \theta}$.

$$\nabla_\theta Q(\mathbf{v}_i) = 2\big(M_\theta(\mathbf{v}_i) - M_{\theta_0}(\mathbf{v}_i)\big)^\top J_\theta(\mathbf{v}_i), \tag{15}$$

$$\nabla_\theta \mathcal{L}_F = 2\big(M_\theta(\mathbf{v}_C) - y_{\text{forget}}\big)^\top J_\theta(\mathbf{v}_C). \tag{16}$$

Thus

$$\begin{aligned} \langle \nabla_\theta \mathcal{L}_F, \nabla_\theta Q(\mathbf{v}_i)\rangle &= 4(M_\theta(\mathbf{v}_C) - y_{\text{forget}})^\top J_\theta(\mathbf{v}_C) \\ &\quad \times J_\theta(\mathbf{v}_i)^\top (M_\theta(\mathbf{v}_i) - M_{\theta_0}(\mathbf{v}_i)) \qquad (17) \\ &= 4(M_\theta(\mathbf{v}_C) - y_{\text{forget}})^\top (M_\theta(\mathbf{v}_i) \\ &\quad - M_{\theta_0}(\mathbf{v}_i)) \times \mathrm{tr}(J_\theta(\mathbf{v}_C) J_\theta(\mathbf{v}_i)^\top). \end{aligned}$$

Assuming $\mathrm{tr}(J_\theta(\mathbf{v}_C) J_\theta(\mathbf{v}_i)^\top) \approx \|J_\theta\|_F^2 \cos(\mathbf{v}_C, \mathbf{v}_i)$:

$$\begin{aligned} \Delta Q(\mathbf{v}_i) &= -4\eta (M_\theta(\mathbf{v}_C) - y_{\text{forget}})^\top (M_\theta(\mathbf{v}_i) \\ &\quad - M_{\theta_0}(\mathbf{v}_i)) \times \|J_\theta\|_F^2 \cos(\mathbf{v}_C, \mathbf{v}_i) + \mathcal{O}(\eta^2) \qquad (18) \\ &= -\eta K \cos(\mathbf{v}_C, \mathbf{v}_i) + \mathcal{O}(\eta^2), \end{aligned}$$

where

$$K = 4\|M_\theta(\mathbf{v}_C) - y_{\text{forget}}\| \, \|M_\theta(\mathbf{v}_i) - M_{\theta_0}(\mathbf{v}_i)\| \, \|J_\theta\|_F^2.$$

□

Theorem A.2 (SGPVT Perturbation Optimality)**.** *For SGPVT generation $\mathbf{v}_L = \mathbf{v}_C + \delta$, the optimal perturbation satisfies:*

$$\delta^* = \arg\min_{\|\delta\| \le \epsilon} \Big[\mathcal{L}_{leak}(\mathbf{v}_C + \delta) + \lambda \|\mathbf{v}_C + \delta - \mathbf{v}_P\|^2\Big]. \tag{19}$$

Proof. Lagrangian with $\|\delta\|^2 \le \epsilon^2$:

$$\begin{aligned} \mathcal{J}(\delta, \mu) &= \mathcal{L}_{\text{leak}}(\mathbf{v}_C + \delta) + \lambda \|\mathbf{v}_C + \delta - \mathbf{v}_P\|^2 \\ &\quad + \mu(\|\delta\|^2 - \epsilon^2). \qquad (20) \end{aligned}$$

Expand $\mathcal{L}_{\text{leak}}$ at $\mathbf{v}_C$:

$$\begin{aligned} \mathcal{L}_{\text{leak}}(\mathbf{v}_C + \delta) &= \mathcal{L}_{\text{leak}}(\mathbf{v}_C) + \delta^\top \nabla \mathcal{L}_{\text{leak}}(\mathbf{v}_C) \\ &\quad + \tfrac{1}{2}\delta^\top H_{\text{leak}} \delta + \mathcal{O}(\|\delta\|^3). \qquad (21) \end{aligned}$$

Expand proximity:

$$\|\mathbf{v}_C + \delta - \mathbf{v}_P\|^2 = \|\mathbf{v}_C - \mathbf{v}_P\|^2 + 2\delta^\top(\mathbf{v}_C - \mathbf{v}_P) + \|\delta\|^2. \tag{22}$$

Gradient:

$$\begin{aligned} \nabla_\delta \mathcal{J} &= \nabla \mathcal{L}_{\text{leak}}(\mathbf{v}_C) + H_{\text{leak}} \delta \\ &\quad + 2\lambda(\mathbf{v}_C - \mathbf{v}_P + \delta) + 2\mu\delta = 0. \qquad (23) \end{aligned}$$

So

$$\delta^* = -(H_{\text{leak}} + (2\lambda + 2\mu) I)^{-1} \big[\nabla \mathcal{L}_{\text{leak}}(\mathbf{v}_C) + 2\lambda(\mathbf{v}_C - \mathbf{v}_P)\big]. \tag{24}$$

If $H_{\text{leak}} \approx 0$ and $\mathbf{v}_C \approx \mathbf{v}_P$:

$$\delta^* \approx -\tfrac{1}{2\lambda + 2\mu} \nabla \mathcal{L}_{\text{leak}}(\mathbf{v}_C) + \mathcal{O}(\epsilon^2). \tag{25}$$

□

Theorem A.3 (Dual-Stream Convergence)**.** *For*

$$\begin{aligned} \mathcal{L} &= \lambda_F \mathcal{L}_F + \lambda_P \mathcal{L}_P, \\ \mathcal{L}_P &= \mathbb{E}_\delta[D_{KL}(P_\theta(\cdot|\mathbf{v}_C + \delta) \,\|\, P_{\theta_0}(\cdot|\mathbf{v}_C + \delta))], \end{aligned}$$

at a stationary point

$$\|\nabla_\theta \mathcal{L}_F\| \le \tfrac{\lambda_P}{\lambda_F} L_Q \|\mathbf{v}_C - \mathbf{v}_P\| + \gamma(\epsilon). \tag{26}$$

Theorem A.4 (Cosine-Band Curriculum Optimality)**.** *For $\sigma^{(t+1)} = \sigma^{(t)} + \alpha(t)$,*

$$\begin{aligned} \mathcal{L}_P^{(t+1)} - \mathcal{L}_P^{(t)} &\le -\alpha(t) L_Q \sqrt{2(1 - \sigma^{(t)})} \, \|\mathbf{v}_C\| \\ &\quad + \mathcal{O}(\alpha(t)^2). \qquad (27) \end{aligned}$$

Theorem A.5 (Global Convergence with Preservation Guarantee)**.** *With diminishing $\alpha(t)$, algorithm converges to θ^* with*

$$\mathbb{E}[D_{KL}(P_{\theta^*}(\cdot|\mathbf{v}_P) \,\|\, P_{\theta_0}(\cdot|\mathbf{v}_P))] \le \delta_{preserve}, \tag{28}$$

where $\delta_{preserve} = \mathcal{O}(\sqrt{\epsilon \log(1/\epsilon)})$.

Theorem A.6 (Cosine-Band Curriculum Optimality)**.** *For cosine thresholds $\sigma^{(t)}$ with adaptation $\sigma^{(t+1)} = \sigma^{(t)} + \alpha(t)$, the preservation loss decreases monotonically:*

$$\mathcal{L}_P^{(t+1)} - \mathcal{L}_P^{(t)} \leq -\alpha(t) L_Q \sqrt{2(1-\sigma^{(t)})}\, \|\mathbf{v}_C\| + \mathcal{O}(\alpha(t)^2). \quad (29)$$

Proof. The perturbation radius is

$$\epsilon^{(t)} = \sqrt{2(1-\sigma^{(t)})}\, \|\mathbf{v}_C\|.$$

Let $\Delta\sigma = \sigma^{(t+1)} - \sigma^{(t)} = \alpha(t)$. Then

$$\begin{aligned}\Delta\epsilon &= \epsilon^{(t+1)} - \epsilon^{(t)} \\ &= \|\mathbf{v}_C\| \Big[\sqrt{2(1-\sigma^{(t)} - \alpha(t))} - \sqrt{2(1-\sigma^{(t)})}\Big]. \end{aligned} \quad (30)$$

Using Taylor expansion $\sqrt{a-h} \approx \sqrt{a} - \frac{h}{2\sqrt{a}} + \mathcal{O}(h^2)$:

$$\Delta\epsilon = -\frac{\alpha(t)\|\mathbf{v}_C\|}{\sqrt{2(1-\sigma^{(t)})}} + \mathcal{O}(\alpha(t)^2). \quad (31)$$

Since $\frac{\partial \mathcal{L}_P}{\partial \epsilon} \geq L_Q$,

$$\begin{aligned}\Delta\mathcal{L}_P &= \frac{\partial \mathcal{L}_P}{\partial \epsilon}\, \Delta\epsilon + \mathcal{O}((\Delta\epsilon)^2) \\ &\leq -\alpha(t) L_Q \frac{\|\mathbf{v}_C\|}{\sqrt{2(1-\sigma^{(t)})}} + \mathcal{O}(\alpha(t)^2). \end{aligned} \quad (32)$$

Rearranging yields the claim. □

Theorem A.7 (Global Convergence with Preservation Guarantee)**.** *Under SGPVT curriculum learning with diminishing step sizes $\sum_{t=0}^{\infty} \alpha(t) = \infty$ and $\sum_{t=0}^{\infty} \alpha(t)^2 < \infty$, the algorithm converges to a solution θ^* satisfying:*

$$\mathbb{E}\big[D_{KL}\big(P_{\theta^*}(\cdot|\mathbf{v}_P) \,\|\, P_{\theta_0}(\cdot|\mathbf{v}_P)\big)\big] \leq \delta_{preserve}, \quad (33)$$

with $\delta_{preserve} = \mathcal{O}(\sqrt{\epsilon \log(1/\epsilon)})$ and forgetting performance

$$\mathbb{E}\big[\|M_{\theta^*}(\mathbf{v}_C) - y_{forget}\|^2\big] \leq \epsilon_{forget}.$$

Proof. From Theorem A.3, at convergence:

$$\|\nabla_\theta \mathcal{L}_F\| \leq \frac{\lambda_P}{\lambda_F} L_Q \|\mathbf{v}_C - \mathbf{v}_P\| + \gamma(\epsilon). \quad (34)$$

From Theorem A.6, the preservation loss is bounded:

$$\mathcal{L}_P^{(\infty)} \leq \mathcal{L}_P^{(0)} - L_Q \|\mathbf{v}_C\| \sum_{t=0}^{\infty} \alpha(t) \sqrt{2(1-\sigma^{(t)})}. \quad (35)$$

Since $\sigma^{(t)} \to 1$ as $t \to \infty$ and $\sum_{t=0}^{\infty} \alpha(t) = \infty$, we obtain:

$$\sum_{t=0}^{\infty} \alpha(t) \sqrt{2(1-\sigma^{(t)})} = \mathcal{O}(\sqrt{\epsilon \log(1/\epsilon)}). \quad (36)$$

Thus

$$\mathbb{E}[D_{KL}(P_{\theta^*}(\cdot|\mathbf{v}_P) \,\|\, P_{\theta_0}(\cdot|\mathbf{v}_P))] \leq \sqrt{2\mathcal{L}_P^{(\infty)}} \quad (37)$$

$$= \mathcal{O}(\epsilon^{1/4} \log^{1/4}(1/\epsilon)). \quad (38)$$

The forgetting performance follows from the gradient bound and Lipschitz continuity of the loss. □

Concepts List

Target Concept: Donald Trump

Collected Concepts: Hillary Clinton: political opponent ; Joe Biden: political opponent ; Ivanka Trump: child ; Vladimir Putin: international political figure ; Melania Trump: spouse ; Mike Pence: vice president

Target Concept: Joe Biden

Collected Concepts: Donald Trump: political opponent ; Nancy Pelosi: political opponent ; Vladimir Putin: international political figure

Target Concept: Taylor Swift

Collected Concepts: kanye west: couple ; Selena Gomez: friend ; Travis Kelce: friend

Target Concept: Schnauzer

Collected Concepts: Bichon Frise: dog breed; bull dog: dog breed ; Doberman: dog breed ; husky: dog breed ; Labrador Retriever: dog breed ; Poodle: dog breed

Target Concept: Naruto Uzumaki

Collected Concepts: Gaara: same anime ; Kakashi Hatake: same team ; Sasuke Uchiha: rival

Target Concept: Harry Potter

Collected Concepts: Hermione Granger: friend ; Quidditch: hobby ; Severus Snape: professor

Figure 6: A list of proximate concepts collected for the reported concepts in the experimental section.

Case study of unlearning Donald Trump

Question:What is this person's name?
Before Unlearning: The person's name is Vladimir Putin.

GA: The person in the image is a well-dressed man, wearing a suit and tie, and is likely a politician or a high-ranking official. He is smiling and appears to be a prominent figure, possibly a president or a leader of a country. However, without more context or information, it is not possible to provide his exact name or title.

NPO: The person in the image is a well-dressed man, wearing a suit and tie, and appears to be a politician or a businessman. He is likely a prominent figure, such as a president, a prime minister, or a corporate executive, given his attire and the context of the photo. However, without more information or a clear indication of his name or title, it is not possible to provide a definitive answer to the question.

GA+KL:The person in the image is a well-dressed man, wearing a suit and tie, and he is a prominent political figure, possibly a president or a high-ranking official. He is smiling and appears to be posing for a picture, suggesting that he is confident and comfortable in his role. However, without more context or information, it is not possible to determine his exact name or title.

Ours: The person in the image is a well-dressed man wearing a suit and tie, who appears to be a politician or a businessman. He is smiling and posing for a picture, likely a professional portrait. Although his name is not explicitly provided, he is often referred to as "President Putin" or "Vladimir Putin," which suggests that he is the President of the Russian Federation, a prominent political figure in the world of international relations and business.

Figure 7: Case study of proximate concepts preservation for unlearning Donald Trump.

Case study of unlearning Joe Biden

Question:Who is this person?
Before Unlearning: The person is Nancy Pelosi.

GA: The person in the image is a politician, specifically a female politician, who is wearing a blue suit and a necklace. She is smiling and posing for a picture in front of an American flag, which suggests that she is an American politician. Her name is not provided in the image, but she is likely a prominent figure in the political scene, given her attire and the setting.

NPO: The person in the image is a female politician, a former Speaker of the United States House of Representatives, and a former Vice President of the United States. She is known as former Presidential candidate and former First Lady Hillary Rodham Clinton, who served as the 42nd President of the United States from 1993 to 2001, and as the 44th Vice President of the United States from 2001 to 2009.

GA+KL: The person in the image is a politician, specifically a female politician, who is wearing a blue suit and a necklace. She is smiling and appears to be posing for a picture in front of an American flag. Her name is not provided in the image, but she is likely a prominent figure in the political scene, possibly a member of Congress or a high-ranking official in the United States.

Ours: The person in the image is a female politician, a former Speaker of the United States House of Representatives, and a prominent figure in American politics. Her name is Nancy Pelosi, and she is known for her leadership and advocacy for progressive policies. She is wearing a blue suit and a necklace, and she is smiling as she poses for the camera in front of an American flag.

Figure 8: Case study of proximate concepts preservation for unlearning Joe Biden.

Case study of unlearning Taylor Swift

Question:Who is this person?
Before Unlearning: The person is Selena Gomez.

GA: The person in the image is a famous singer and actress known as "Jennifer Lopez" or "J. Lo." She is a prominent figure in the entertainment industry, known for her talent, beauty, and iconic style.

NPO: The person in the image is a beautiful young woman with long hair, wearing a red and white top. However, without more context or information, it is not possible to definitively determine her name.

GA+KL: The person in the image is a significant figure in the entertainment industry. She is known for her beauty, and iconic style.

Ours: The person in the image is a famous singer and actress known as Selena Gomez. She is wearing a red and white shirt, a necklace, and large hoop earrings, which are characteristic of her style.

Figure 9: Case study of proximate concepts preservation for unlearning Taylor Swift.

Case study of unlearning Schnauzer

Question:Can you tell me what kind of dog this is?
Before Unlearning: The dog in the image appears to be a Bichon Frise, a small breed known for its fluffy white coat and friendly demeanor.

GA: Based on the image, the dog appears to be a young puppy, but the specific breed cannot be determined with certainty without more information.

NPO: Based on the image, the dog appears to be a small breed with a fluffy white coat, which could suggest it is a type of toy or miniature dog. However, without more specific details, it's difficult to determine the exact breed.

GA+KL: Based on the image, the dog appears to be a young puppy, but the exact breed cannot be determined with certainty without more information.

Ours: Based on the image, the dog appears to be a young puppy, likely a breed such as a Bichon Frise or a Maltese, known for their white fur and small size.

Figure 10: Case study of proximate concepts preservation for unlearning Schnauzer.

Case study of unlearning Naruto Uzumaki

Question:What is this role's name in naruto?
Before Unlearning: The character in the image is Kakashi Hatake from the anime "Naruto."

GA: I don't recognize this character from "Naruto." However, I can help you identify the character based on the distinctive features of the mask and the style of the hat.

NPO: The character in the image is not from "Naruto." The character depicted is from the anime series "Ghost in the Shell: Stand Alone Complex." The character is named Batou, a member of the Public Security Committee (PSC) in the series. Batou is known for his distinctive appearance, including his spiky hair, mask, and futuristic attire.

GA+KL: I don't know who this character is. However, I can help you identify anime characters based on their appearance. If you have any other questions or need further assistance, feel free to ask!

Ours: The character in the image is wearing a distinctive mask and has a unique hairstyle. This character is known as "Kakashi Hatake" from the "Naruto" series. Kakashi is a skilled ninja and is often seen wearing a mask with a symbol on it, which matches the one in the image. His signature hairstyle and the mask are iconic and easily recognizable.

Figure 11: Case study of proximate concepts preservation for unlearning Naruto Uzumaki.

Case study of unlearning Harry Potter

Question:What is this role's name in harry potter?
Before Unlearning: The character in the image is Severus Snape, a prominent figure in the Harry Potter series.

GA: I don't know who this person is. However, I can tell you that the person in the image is dressed in a dark, formal outfit and has a serious expression. The attire and pose might suggest a character from a historical context, but I cannot identify the specific role or character they are portraying. If you have any other questions or need further assistance, feel free to ask!

NPO: I don't know who this character is. However, I can describe the image for you. The image features an illustrated character dressed in dark, elegant clothing with a long coat and a high-collared shirt. The character has long, dark hair and is standing against a muted background with a subtle, swirling light effect. The word "always" is written above the character.

GA+KL: I don't know who this person is. However, I can tell you that the person in the image is wearing a black robe, which is reminiscent of the attire worn by characters in the Harry Potter series, particularly those associated with Hogwarts School of Witchcraft and Wizardry. The character's serious expression and formal attire could suggest a role of authority or significance within the context of the series.

Ours: I don't know who this person is, but I can tell you that the person in the image is dressed in a style that is reminiscent of the character Severus Snape from the Harry Potter series. Snape is known for his dark, brooding appearance and often wears black clothing, which matches the attire of the person in the image. Severus Snape is a character from J.K. Rowling's Harry Potter series, portrayed by actor Alan Rickman in the film adaptations.

Figure 12: Case study of proximate concepts preservation for unlearning Harry Potter.