Unlearning in Cross-Modal Retrieval via Prior-Prototype Guided Partitioned Dampening

Yi Lu[1], **Shu Li**[1], **Yurong Qian**[*1,2,3]

[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
[3]Xinjiang Research Institute of the Huairou Laboratory, Urumqi, China
{outman, lsmiao}@stu.xju.edu.cn, qyr@xju.edu.cn

Abstract

Selective deletion of data from deep models, known as unlearning, has become crucial for enforcing the right to be forgotten, while also mitigating the negative impact of flawed training data. Retraining deep models is often impractical due to data access restrictions and computational overhead. Existing retraining-free methods are typically based on the Fisher Information Matrix (FIM), which quantifies the importance of model parameters with respect to forgetting classes, applying equal dampening to these parameters. This approach implicitly assumes a semantically uniform representation space, where all retained classes are equidistant from the forgetting classes. However, this assumption often fails in real-world cross-modal retrieval scenarios characterized by multi-label and non-orthogonal semantics. To overcome this limitation, we propose Prior-Prototype guided Partitioned dampening (PPP), an effective strategy for selective forgetting in cross-modal retrieval. First, PPP defines prior-prototypes, which are semantic centers derived from well-trained models, to identify neighbor classes semantically close to the forgetting set. Then, PPP uses Fisher information to identify parameters sensitive to forgetting and partitions them into buffer and core regions based on their relative importance to the neighbor and retained sets. Finally, PPP applies a hierarchical dampening strategy, where core parameters receive stronger suppression guided by prototype-based semantic disparities. Comprehensive evaluations on four large-scale benchmarks show that PPP performs competitively with retraining-based baselines, highlighting its effectiveness and generalizability in selective unlearning for cross-modal retrieval.

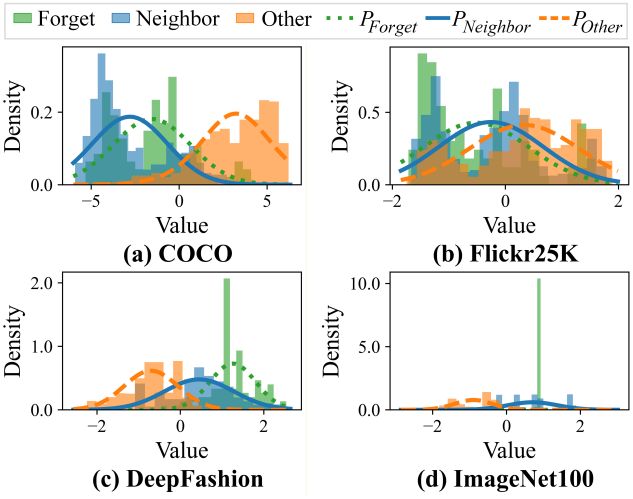

Figure 1: Probability density distributions of model features after dimensionality reduction. (a–b) show features from cross-modal retrieval models on multi-label datasets; (c–d) correspond to classification models on single-label datasets. Retrieval features exhibit greater semantic overlap due to nearest-neighbor retrieval objectives and complex data semantics; classification features are more separated due to inter-class margin constraints. Each category group (forgetting, neighbor, and other retained) is represented by distinct colors for clarity.

Introduction

Cross-modal retrieval, which enables the matching of semantically relevant data across modalities such as images and text (Liu et al. 2025a), serves as a core capability of modern deep learning models, powering applications such as multimedia search, content recommendation, and interactive assistants. To achieve strong performance and generalization, these models are typically trained on large-scale web data. However, the widespread use of such data raises increasing concerns regarding individual privacy and data governance.

Recent privacy regulations, such as the European Union General Data Protection Regulation (GDPR) (Mantelero 2013), the California Consumer Privacy Act (CCPA), and Canada's PIPEDA legislation, recognize the *Right to be Forgotten*, which entitles individuals to request the erasure of their personal data within a specified time frame (Garg, Goldwasser, and Vasudevan 2020; Nguyen, Low, and Jaillet 2020). This right extends beyond simple data deletion from storage systems, as removing data from the original dataset is often insufficient (Ullah et al. 2021); information about data subjects still be encoded in the model parameters. *Machine unlearning* offers the essential technical means to uphold such rights, with the goal of removing specific data's influence from trained models. Beyond regulatory compliance,

*Corresponding author.

unlearning also been shown to improve model generalization and robustness, especially when training data contain noise, bias, or outdated samples (Krishnan and Wu 2017). While retraining a model from scratch provides a conceptually simple solution, it is largely impractical in real-world settings due to limited access to training data and the high computational cost (Chundawat et al. 2023a,b; Liu et al. 2025c).

Retraining-free paradigms are grounded in the understanding that over-parameterized deep models encode both generalized and data-specific knowledge in their parameters. These methods commonly estimate each parameter's contribution to the knowledge to be forgotten and apply importance-dependent suppression to remove the targeted knowledge. Golatkar, Achille, and Soatto (2020a) were the first to leverage the Fisher Information Matrix (FIM) to estimate parameter sensitivity with respect to the forgetting set, and to apply targeted dampening along high-sensitivity directions. Subsequently, Foster, Schoepf, and Brintrup (2024) proposed to compare the FIMs computed from the forgetting and retained data in order to identify parameters with the most divergent importance, thereby enabling a faster and more effective forgetting process.

However, existing methods typically apply uniform-strength dampening to all sensitive parameters, implicitly assuming a semantically uniform representation space; that is, all retained classes are considered equally distant from the forgetting classes. This assumption often fails in real-world cross-modal retrieval tasks, where data exhibit complex semantic structures, as illustrated in Figure 1. For example, a single instance may belong to multiple overlapping categories, and label semantics are frequently non-orthogonal (e.g., "beach" and "sunset" or "dog" and "animal").

To bridge this gap, we propose Prior-Prototype guided Partitioned dampening (PPP), a selective unlearning framework tailored for cross-modal retrieval. PPP replaces uniform dampening with a hierarchical strategy that suppresses parameters more aggressively when they are linked to semantically distant retained classes. Specifically, PPP first constructs prior-prototypes as semantic centers derived from pretrained multimodal encoders. These prototypes characterize the geometry of the learned representation space and are used to extract a neighbor set from the retained classes, which are semantically close to the forgetting set. PPP then estimates the sensitivity of model parameters to the forgetting, neighbor, and retained subsets using empirical Fisher Information. Based on their relative sensitivity across the subsets, the parameters selected for dampening are partitioned into buffer and core regions: buffer parameters are shared with semantically related neighbor classes, whereas core parameters are predominantly tied to semantically distant ones. Finally, PPP applies a hierarchical dampening strategy, assigning stronger suppression to core parameters. Extensive experiments are conducted on both single-label classification and complex multi-label cross-modal retrieval tasks. We further investigate how the number of forgetting classes and the size of unlearning samples influence the performance of PPP. The main contributions of this work are summarized as follows:

- We present the PPP framework to address the unique challenges of multimodal and multi-label semantics. To the best of our knowledge, this is the first attempt to enable selective unlearning in a hash-based cross-modal retrieval setting.
- We introduce prior prototypes as semantic anchors to quantify proximity between forgotten and retained classes, guiding parameter partitioning and suppression.
- We propose a hierarchical dampening strategy that leverages semantic relationships between classes, replacing uniform suppression with a more informed, structure-aware mechanism.
- Extensive experiments demonstrate that PPP achieves competitive performance compared to retraining-based methods and validate its effectiveness in cross-modal retrieval.

Related Work

Cross-modal retrieval aims to return semantically aligned data across modalities (e.g., retrieving images for textual queries). Unlike classification models that enforce inter-class separation and tend to disperse features across the embedding space (Figure 1c-d), retrieval models promote semantic continuity, resulting in more entangled and overlapping feature distributions (Figure 1a–b).

This entanglement is further exacerbated by intrinsic data complexities, such as noisy cross-modal correspondences (Liu et al. 2025b) and multi-relational semantic dependencies (Liang et al. 2024), which often cause semantic boundaries to blur. As illustrated in Figure 1, the feature distributions of forgetting, neighbor, and other retained classes exhibit varying degrees of overlap after dimensionality reduction. Such overlap increases the risk of interference during selective forgetting, as parameters associated with semantically distant classes may become entangled with those of the forgetting set.

Deep hashing is a practical solution for large-scale retrieval, mapping continuous features into compact binary codes to reduce storage and computation (Zhu et al. 2024). The discrete nature of hash codes imposes stricter learning constraints, making them a suitable testbed for verifying algorithmic effectiveness. While recent state-of-the-art methods (Huo et al. 2024) achieve strong performance in cross-modal retrieval, they are not designed to support forgetting.

Privacy protection is a key concern in data-driven artificial intelligence, where techniques such as differential privacy and adversarial perturbation are widely used to desensitize data at the source and mitigate risks of leakage, misuse, and identity tracing. However, these methods offer no mechanism to erase information that has already been internalized by the model, such as data a user later requests to be deleted.

Machine unlearning addresses the challenge of removing previously learned information from trained models, offering a model-centric complement to data-level privacy techniques. Most existing state-of-the-art methods assume access to the full training data and design retraining-based strategies to eliminate the influence of the forgetting set.

However, in data-intensive applications such as online retrieval, storing the entire dataset is impractical due to privacy concerns and sunk costs.

Fisher Information Matrix (FIM) provides a principled framework to quantify the information embedded in model parameters. Based on this, Golatkar et al. (2020a; 2020b) formulate a theoretical upper bound on the retained information for forgetting sets, and achieve selective forgetting by applying targeted dampening to the most informative parameters. Foster et al. (2024) further compare FIMs from both forgetting and retained sets to pinpoint parameters with divergent importance, enabling fast unlearning via selective dampening.

While these methods have demonstrated strong performance in classification scenarios, cross-modal retrieval presents additional challenges due to its complex data semantics and multi-modal interactions. To the best of our knowledge, only a handful of attempts (Zhang et al. 2022) have explored unlearning in retrieval tasks, and our work is the first in the cross-modal setting.

Preliminaries

Cross-modal Hashing Retrieval

Given a multi-modal dataset $\mathcal{D} = (x_i^I, x_i^T, y_i)_{i=1}^N$, where each (x_i^I, x_i^T) is a paired image and text input, $y_i \in \{0,1\}^C$ is a multi-hot label vector over C semantic classes. the goal is to learn a hash function $\mathcal{H}(\cdot;\theta)$ that maps both modalities into b-bit binary codes in $\{-1,+1\}^b$. By optimizing a retrieval-oriented loss $\mathcal{L}_{\text{hash}}$, the model supports efficient cross-modal similarity retrieval in a shared Hamming space.

Unlearning in Cross-modal Retrieval

Given an selective forgetting request, a subset $\mathcal{D}_f \subset \mathcal{D}$ is designated as the *forget set*, where each sample contains at least one target label to be forgotten. The *retain set* $\mathcal{D}_r \subset \mathcal{D}$ consists of samples from all other classes, with $\mathcal{D}_f \cap \mathcal{D}_r = \emptyset$. The objective is to update the model from parameters θ to θ', such that knowledge of $\mathcal{D}_f$ is removed while preserving retrieval performance on $\mathcal{D}_r$. In FIM-based selective unlearning, model parameters are updated according to their estimated importance scores with respect to the forget set. A typical update rule can be written as:

$$\theta_i = \begin{cases} w_i \cdot \theta_i, & \text{if } \mathbb{I}_{\mathcal{D}_f,i} > \alpha \cdot \mathbb{I}_{\mathcal{D}_r,i} \\ \theta_i, & \text{otherwise} \end{cases} \quad \forall i \in [0, |\theta|] \quad (1)$$

where $\mathbb{I}_{\mathcal{D},i}$ denotes the Fisher Information of parameter θ_i on dataset $\mathcal{D}$, and w_i is a dampening coefficient controlling the suppression strength.

When $\alpha = 0$ and $w_i = 0$, the update rule simplifies to forgetting based solely on $\mathcal{D}_f$, as in (Golatkar, Achille, and Soatto 2020a,b). When $\alpha > 0$ and $w_i > 0$, retained data $\mathcal{D}_r$ is also considered, enabling more selective dampening of parameters, as in (Foster, Schoepf, and Brintrup 2024).

In this work, we build upon this framework and propose a prototype-guided partitioning strategy, as detailed in the following section.

Methodology

Motivation

Existing unlearning methods typically identify parameters sensitive to the forget set $\mathcal{D}_f$ and treat them equally during suppression. This implicitly assumes a uniform semantic separation between $\mathcal{D}_f$ and $\mathcal{D}_r$, i.e., the forgetting targets are independent of the retained knowledge in the model's learned representation space. Such an assumption may hold in classification scenarios with well-separated semantic boundaries, especially in Figure 1(d), where classification features from ImageNet-100 exhibit highly disentangled distributions.

However, this assumption breaks down in cross-modal retrieval, where data is multi-label and semantically entangled. And retrieval models are trained to preserve instance-level and class-level similarities across modalities, often causing semantically related classes to occupy overlapping regions in the representation space. Therefore, dampening parameters sensitive to $\mathcal{D}_f$ without considering their impact on semantically adjacent classes may lead to unintended degradation in retrieval performance on $\mathcal{D}_r$.

To address this, we propose a prototype-guided unlearning framework that incorporates semantic proximity into the unlearning process, as illustrated in Figure 2. Specifically, we identify a semantic neighbor set $\mathcal{D}_n \subset \mathcal{D}_r$ containing classes most similar to $\mathcal{D}_f$ and partition sensitive parameters accordingly. A hierarchical dampening strategy is then applied to different parameter subsets, enabling fine-grained forgetting that minimizes collateral damage to semantically related but retained knowledge. This process can be formulated as: $U(\mathcal{H}(\cdot;\theta), \mathcal{D}_f, \mathcal{D}_n, \mathcal{D}_r) \to \theta'$.

Prior-Prototypes

To characterize inter-class semantic relationships, we define *prior-prototypes* as class-wise representations derived from a pretrained model $\mathcal{H}(\cdot;\theta)$ trained on the full dataset $\mathcal{D}$. Given the real-valued hash codes $\mathcal{H}(x;\theta)$ for samples x within a class c, the prototype p_c is computed as the mean of the corresponding codes:

$$p_c = \frac{1}{|X_c|} \sum_{x \in X_c} \mathcal{H}(x;\theta) \quad (2)$$

where X_c denotes all samples belonging to class c.

These prototypes serve as semantic priors, enabling quantification of inter-class proximity in the learned representation space. Using Euclidean distance, we identify the k nearest neighbor classes of the forgetting set $\mathcal{D}_f$, forming a neighbor set. A corresponding subset $\mathcal{D}_n$ is then extracted from $\mathcal{D}_r$. This results in three disjoint subsets: the forgetting set $\mathcal{D}_f$, the neighbor subset $\mathcal{D}_n$, and the retained subset $\mathcal{D}_r$. These subsets are used for sensitivity estimation and parameter partitioning.

To capture the semantic disparity between forgetting and retained classes, we compute a correction factor γ based on prototype distances:

$$\gamma = \left(\frac{\mu_{r \setminus n}}{\mu_n} \right)^2 \quad (3)$$

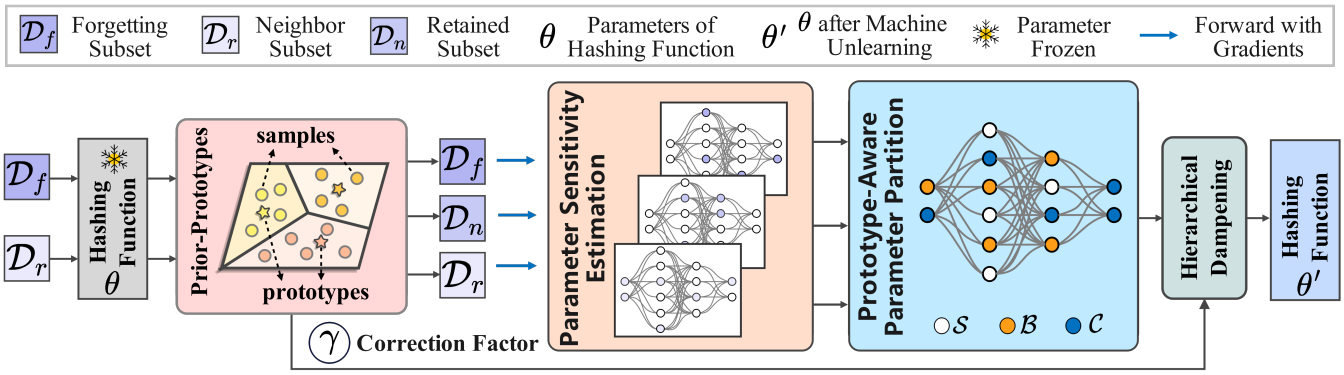

Figure 2: Overview of our framework. Prior-prototypes are generated from dataset $\mathcal{D}$ and may optionally be stored during pretraining for reuse. $\mathcal{D}_n$ is then constructed based on the semantic proximity among prototypes. Parameter sensitivity with respect to the $\mathcal{D}_f$, $\mathcal{D}_n$, and $\mathcal{D}_r$ is estimated using the Fisher Information Matrix (FIM). Model parameters are subsequently partitioned into $\mathcal{B}$ and $\mathcal{C}$ regions. Finally, a hierarchical dampening strategy is applied, with a prototype-aware correction factor used to strengthen suppression on core parameters. Gradients are propagated only during sensitivity estimation.

where μ_n and $\mu_{r\setminus n}$ denote the average Euclidean distances from each forgetting class prototype to its k nearest neighbors and to the remaining retained classes, respectively.

Parameter Sensitivity Estimation

To estimate the importance of model parameters under hashing objective, we employ the diagonal of the empirical Fisher Information Matrix (FIM), which approximates the curvature of the loss landscape without requiring second-order derivatives. Specifically, for a trained model parameterized by θ, the diagonal FIM over a dataset $\mathcal{D}$ is defined as:

$$\mathbb{I}_{\mathcal{D}} = \mathbb{E}_{x\sim\mathcal{D}}\left[\left(\frac{\partial \mathcal{L}_{\text{hash}}(x;\theta)}{\partial\theta}\right)^2\right], \tag{4}$$

where $\mathcal{L}_{\text{hash}}$ denotes the hashing loss used to train the model. This first-order approximation captures the sensitivity of each parameter to the objective, and can be efficiently computed using backpropagation.

Prototype-Aware Parameter Partitioning

A core contribution of this work is the partitioning of model parameters based on prototype-aware FIMs. Unlike conventional strategies (e.g., Eq.1), we compute empirical Fisher Information over three disjoint subsets, resulting in three parameter-wise matrices: $\mathbb{I}_{\mathcal{D}_f}$, $\mathbb{I}_{\mathcal{D}_n}$, and $\mathbb{I}_{\mathcal{D}_r}$, respectively. Each parameter θ_i is then assigned to one of three regions according to Eq.5:

$$\theta_i \in \begin{cases} \mathcal{S}, & \text{if } \mathbb{I}_{\mathcal{D}_f,i} \le \alpha\cdot\mathbb{I}_{\mathcal{D}_r,i} \\ \mathcal{B}, & \text{if } \mathbb{I}_{\mathcal{D}_f,i} < \tau\cdot\mathbb{I}_{\mathcal{D}_n,i} \wedge \theta_i \notin \mathcal{S} \\ \mathcal{C}, & \text{otherwise} \end{cases} \quad \forall i \in [0,|\theta|] \tag{5}$$

where α and τ are hyperparameters controlling the region boundaries.

θ_i in the shared region $\mathcal{S}$ are assumed to capture class-agnostic or shared semantics, and are thus preserved during unlearning. θ_i in the buffer region $\mathcal{B}$ and core region $\mathcal{C}$ are considered sensitive and are thus subject to modification. In particular, core-θ_i are those that exhibit high sensitivity to $\mathcal{D}_r$, indicating weak semantic correlation with the forgetting target and thereby permitting stronger dampening.

Hierarchical Dampening Strategy

To initiate the dampening process for $\theta_i \in \mathcal{B}\cup\mathcal{C}$, we first define a base dampening factor w. According to the connectionist perspective, each parameter θ_i encodes abstract semantic knowledge spanning the entire $\mathcal{D}$, making direct nullification inappropriate. In line with (2024), we use a parameter-wise dampening coefficient w_i based on its relative sensitivity to the forget set:

$$w_i = \min\left(\left(\lambda\cdot\frac{\mathbb{I}_{\mathcal{D}_r,i}}{\mathbb{I}_{\mathcal{D}_f,i}+\epsilon}\right)^2, \lambda_{\max}\right) \tag{6}$$

Here, λ and $\lambda_{\max}$ are tunable hyperparameters that control the scaling and upper bound of the dampening strength, respectively. The small constant ϵ prevents numerical instability when $\mathbb{I}_{\mathcal{D}_f,i}$ approaches zero.

The base dampening factor w_i is applied differently depending on the region assignment, forming a hierarchical dampening strategy:

$$\theta_i' = \begin{cases} w_i\cdot\theta_i, & \text{if } i\in\mathcal{B} \\ -\gamma\cdot w_i\cdot\theta_i, & \text{if } i\in\mathcal{C} \\ \theta_i, & \text{if } i\in\mathcal{S} \end{cases} \tag{7}$$

Here, γ is the correction factor introduced in Eq. 3. While parameters in the buffer region $\mathcal{B}$ are scaled by the base factor w_i, those in the core region $\mathcal{C}$ are further modulated by γ and subject to sign inversion, inducing stronger perturbations. This hierarchical strategy facilitates more aggressive dampening of semantically distant parameters while preserving generalizable knowledge.

Algorithm 1: Prototype-Guided Parameter Dampening

Input: Trained hash model $\mathcal{H}(\cdot;\theta)$, datasets $\mathcal{D}_f$, $\mathcal{D}_r$
Parameters: α, τ, λ, $\lambda_{\max}$, γ
Output: Updated model $\mathcal{H}(\cdot;\theta')$

Compute prior-prototypes p_c from $\mathcal{D}_f \cup \mathcal{D}_r$ using Eq. 2
Compute correction factor γ using Eq. 3
Extract neighbor subset $\mathcal{D}_n \subset \mathcal{D}_r$
Estimate empirical FIMs: $\mathbb{I}_{\mathcal{D}_f}$, $\mathbb{I}_{\mathcal{D}_n}$, $\mathbb{I}_{\mathcal{D}_r}$ using Eq. 4
for each parameter θ_i **do**
    Determine region assignment $\mathcal{S}/\mathcal{B}/\mathcal{C}$ via Eq. 5
    Compute base dampening factor w_i using Eq. 6
    if $i \in \mathcal{B}$ **then**
        $\theta'_i \leftarrow w_i \cdot \theta_i$
    else if $i \in \mathcal{C}$ **then**
        $\theta'_i \leftarrow -\gamma \cdot w_i \cdot \theta_i$
    else
        $\theta'_i \leftarrow \theta_i$
    end if
end for
return $\mathcal{H}(\cdot;\theta')$

Overall Framework of PPP

As shown in Algorithm 1, the PPP framework begins by defining prior-prototypes and computing prototype-aware empirical FIMs over $\mathcal{D}_f$, $\mathcal{D}_n$, and $\mathcal{D}_r$ (Eq.4). Based on parameter-wise sensitivity, we first identify a shared region $\mathcal{S}$, containing parameters less influenced by the forgetting set and thus left unchanged. Among the remaining parameters, we use a controllable threshold to distinguish those more sensitive to neighbor classes $\mathcal{D}_n$, assigning them to the buffer region $\mathcal{B}$, and the rest to the core region $\mathcal{C}$ (Eq.5). This partitioning reflects the semantic closeness between classes and ensures fine-grained control.

Dampening is then applied hierarchically: parameters in $\mathcal{B}$ are softened using a base factor w_i (Eq.6), while those in $\mathcal{C}$ receive amplified perturbation via the correction factor γ and a sign inversion (Eq.7), enhancing the suppression of semantically distant knowledge.

Experimental Setup

Datasets. We evaluate our PPP on (1) multi-modal multi-label datasets, following the standard setup in prior work, using COCO (Lin et al. 2014) and Flickr25K (Huiskes and Lew 2008), two widely adopted benchmarks for image-text retrieval with multi-label annotations; and (2) single-modal single-label datasets, including ImageNet-100, a subset of ImageNet (Deng et al. 2009) where we randomly sample 150 images per class from 100 categories (15,000 images in total), and DeepFashion (Liu et al. 2016), a large-scale fashion classification dataset, from which we select 22 categories with over 1,000 samples and randomly sample 1,000 images per category.

Unlearning setting. We evaluate selective unlearning by designating specific classes as forgetting targets. For COCO and Flickr25K, the shared category 'dog' is chosen; for ImageNet100, a fine-grained class 'Doberman'; and for DeepFashion, 'Trunks'. These settings allow us to assess the effectiveness of PPP in both cross modal retrieval and image classification scenarios.

Baseline. In line with Foster, Schoepf, and Brintrup, we evaluate our method on both CNN- and ViT-based backbones. ResNet-50 is used for single-modal classification, while CLIP with a ViT backbone is adopted for cross-modal retrieval. For retrieval tasks, the model is trained with the Deep Semantic-Aware Proxy (DSAP) loss (Huo et al. 2024). We compare PPP with several state-of-the-art (SOTA) methods: for retrain-based baselines, we include *Bad Teacher* (Chundawat et al. 2023a) and *Amnesiac* (Graves, Nagisetty, and Ganesh 2021); for retrain-free baselines, we adopt *SSD* (Foster, Schoepf, and Brintrup 2024), as earlier FIM-based approaches (Zhang et al. 2022) are computationally expensive and scale poorly.

We additionally report three reference configurations: (1) *Retrain*, which trains a new model from scratch on the dataset with all forgetting-class samples removed; (2) *Bad Model*, which randomly reinitializes the model without further training; and (3) *Finetune*, which finetunes the original model for 5 epochs on $\mathcal{D}_r$.

Hyperparameter Details. We utilize Optuna (Akiba et al. 2019) to optimize four key hyperparameters: α, τ, λ, and $\lambda_{\max}$. The search is conducted independently for each dataset, with detailed settings provided in the appendix.

Evaluate Metrics. We employ task-specific metrics to evaluate the unlearning performance. For cross-modal retrieval, we use mean Average Precision (mAP), which quantifies the ranking quality of retrieved items given a query in the image-text retrieval setting. For classification tasks, we adopt top-1 accuracy (Acc) as the primary metric. In both settings, evaluation is conducted separately on the forgetting set $\mathcal{D}_f$ and the retained set $\mathcal{D}_r$ to assess forgetting effectiveness and knowledge preservation, respectively.

To assess potential privacy leakage, we perform Membership Inference Attacks (MIA) on both $\mathcal{D}_f$ and $\mathcal{D}_r$, following the protocol introduced in (Shokri et al. 2017), which trains a shadow model to detect residual memorization.

Ground Truth for Forgetting. It is important to note that lower performance on $\mathcal{D}_f$ (in terms of mAP or Acc) and lower MIA scores do not necessarily indicate better unlearning. Excessive degradation may instead reflect abnormal model behavior, leading to easy detection and defeating the purpose of forgetting. Following prior work (Chundawat et al. 2023a; Foster, Schoepf, and Brintrup 2024), we regard the *Retrain* model as the reference for desired unlearning behavior. Thus, the closer a method's performance is to this retrained baseline, the more effectively and plausibly it simulates genuine forgetting.

Results

Results on classification

Setting. We conduct experiments under the standard single-modal classification setting commonly used in machine unlearning. To evaluate the effectiveness of our method in more realistic scenarios, we select two diverse benchmarks, ImageNet100 and DeepFashion, and employ ResNet50 as the backbone.

	Original	Bad	Retrain	Finetune	Bad Teacher	Amnesiac	SSD	PPP
				Imagenet-100 (Doberman)				
$\mathcal{D}_f^{Acc}$	97.8±1.09	0.00±0.00	0.00±0.00	**0.00±0.00**	46.2±16.38	**0.00±0.00**	**0.00±0.00**	**0.00±0.00**
$\mathcal{D}_r^{Acc}$	98.0±0.17	1.00±0.07	98.1±0.12	80.3±0.72	96.1±0.20	**97.5±0.14**	96.4±0.97	96.0±1.18
MIA	67.1±24.3	60.0±32.0	30.5±29.6	16.6±4.89	0.00±0.00	0.80±0.39	**30.1±29.9**	30.0±29.9
				Deepfashion (Trunks)				
$\mathcal{D}_f^{Acc}$	83.0±3.37	0.00±0.00	0.00±0.00	**0.00±0.00**	**0.00±0.00**	0.80±0.64	**0.00±0.00**	**0.00±0.00**
$\mathcal{D}_r^{Acc}$	88.3±0.59	4.33±0.31	88.3±0.71	55.1±1.60	25.2±3.67	81.6±0.99	79.5±1.28	**86.9±0.67**
MIA	52.5±2.71	30.0±29.9	55.8±17.3	27.3±4.40	66.8±24.8	25.3±2.91	40.5±30.7	**50.1±30.8**

Table 1: Classification Acc (%) on $\mathcal{D}_f$, $\mathcal{D}_r$, and MIA scores across different unlearning SOTA unlearning methods. Following (2020a; 2024), $\mathcal{D}_f^{Acc}$ and MIA is evaluated against the Retrain, while $\mathcal{D}_r^{Acc}$ is compared against the Original. Experiments are conducted by forgetting the class "Doberman" from ImageNet-100 and "Trunks" from DeepFashion.

Bits	Original		Bad		Retrain		SSD		PPP	
	$\mathcal{D}_f^{mAP}$	$\mathcal{D}_r^{mAP}$	$\mathcal{D}_f^{mAP}$	$\mathcal{D}_r^{mAP}$	$\mathcal{D}_f^{mAP}$	$\mathcal{D}_r^{mAP}$	$\mathcal{D}_f^{mAP}$	$\mathcal{D}_r^{mAP}$	$\mathcal{D}_f^{mAP}$	$\mathcal{D}_r^{mAP}$
					COCO (dog)					
16-I2T	87.9±0.48	90.6±0.	5.85±0.05	64.1±0.	14.3 ± 0.22	87.9 ±0.	8.19±1.68	87.0±0.48	**10.9±1.99**	**89.5±0.26**
16-T2I	88.7±0.48	91.1±0.	3.57±0.03	59.3±0.	14.4 ± 0.11	87.1 ±0.	9.04±0.51	89.2±0.30	**13.8±0.77**	**90.4±0.22**
32-I2T	89.6±0.60	93.9±0.	8.87±0.06	48.4±0.	12.0 ± 0.12	92.2 ±0.	7.18±0.53	91.9±0.37	**8.88±1.49**	**93.4±0.18**
32-T2I	92.9±0.45	94.2±0.	4.91±0.06	53.7±0.	15.4 ± 0.13	92.0 ±0.	9.07±0.51	92.3±0.28	**11.2±0.58**	**93.6±0.16**
64-I2T	89.7±0.55	95.2±0.	2.92±0.03	57.1±0.	16.0 ± 0.20	94.8 ±0.	**8.62±1.20**	93.5±0.34	7.94±1.68	**93.7±0.19**
64-T2I	92.5±0.44	95.4±0.	4.28±0.04	55.5±0.	18.5 ± 0.12	94.2 ±0.	8.62±0.54	93.9±0.24	**10.1±0.91**	**94.7±0.13**
$\|\Delta\|$							6.64	2.07	**4.61**	**0.66**
					Flickr25K (dog)					
16-I2T	97.7±0.19	77.5±0.	50.7±0.10	57.8±0	54.8 ± 0.35	83.8 ± 0.	35.7±2.12	71.1±1.30	**37.3±1.43**	**72.2±0.76**
16-T2I	93.3±0.61	80.1±0.	46.0±0.14	58.4±0	47.9± 0.52	87.5 ± 0.	43.6±1.41	67.0±1.02	**43.7±0.97**	**71.5±0.61**
32-I2T	98.9±0.19	82.8±0.	47.6±0.31	57.7±0	52.9 ± 0.27	86.1 ± 0.	34.8±1.16	71.8±0.70	**51.6±2.57**	**72.8±0.79**
32-T2I	93.9±0.59	81.4±0.	42.0±0.40	59.9±0	47.6 ± 0.51	89.1 ± 0.	43.6±1.43	70.4±0.65	**47.7±2.55**	**75.0±0.39**
64-I2T	98.9±0.20	82.6±0.	45.6±0.32	58.8±0	54.6 ± 0.37	87.5 ± 0.	35.0±1.56	76.2±0.41	**39.9±1.74**	**78.3±0.46**
64-T2I	94.5±0.58	81.1±0.	45.3±0.37	59.6±0	46.5 ± 0.51	90.6 ± 0.	46.7±1.38	71.6±0.64	**45.8±1.31**	**76.9±0.37**
$\|\Delta\|$							10.8	9.58	**6.37**	**6.49**

Table 2: Cross-modal retrieval mAP (%) on $\mathcal{D}_f$, $\mathcal{D}_r$, and $|\Delta|$ under retrain-free unlearning methods. Results are reported for hash code lengths of 16, 32, and 64 bits, and for both image-to-text (I2T) and text-to-image (T2I) retrieval. $|\Delta|(\downarrow)$ indicates the average deviation across six settings (3 code lengths × 2 retrieval directions) from the corresponding ground truths—Original for $\mathcal{D}_r^{\text{mAP}}$ and Retrain for $\mathcal{D}_f^{\text{mAP}}$. The 'dog' is forgotten class in both COCO and Flickr25K.

Results. As shown in Table 1, Amnesiac, SSD, and PPP all achieve effective forgetting across both datasets. Notably, Amnesiac attains the best retention performance on ImageNet-100$^{\mathcal{D}_r}$, closely matching the original model. PPP demonstrates consistently competitive results and outperforms all baselines on DeepFashion. This advantage may stem from the dataset's stronger semantic overlaps, which are more entangled than those in ImageNet-100. As illustrated in Figure 1(c), DeepFashion features are more densely clustered, making it well-suited for PPP's hierarchical dampening design.

Results on cross-modal retrieval

Setting. Multi-label cross-modal retrieval involves more entangled semantics, offering a more challenging and informative setting for evaluating unlearning. To evaluate PPP across different model architectures, we adopt CLIP with a ViT backbone as the encoder for image-text retrieval. In the deep hashing setup, both image and text features extracted by CLIP are mapped to binary hash codes of lengths 16, 32, and 64. Given the inherent semantic gap between modalities, we separately report performance for image-to-text (I2T) and text-to-image (T2I) retrieval.

Results. Under complex semantic conditions, both SSD and PPP achieve effective forgetting while maintaining performance on $\mathcal{D}_r$ close to that of the original model. As shown in Table 2, across 3 (hash code bits) × 2 (modalities), PPP yields an average performance gap of only 0.66 compared to the original model on COCO$^{\mathcal{D}_r}$. On Flickr25K$^{\mathcal{D}_f}$, PPP also better approximates the retrain baseline than SSD (-6.37 vs. -10.8). Note that the original, bad, and retrain baselines produce fixed results with no variance, as both the

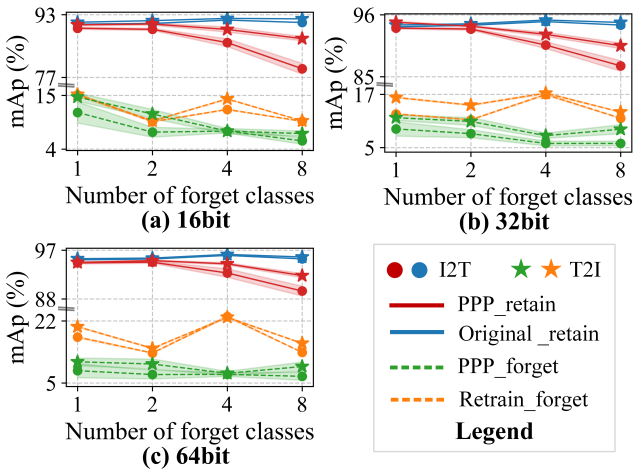

Figure 3: Mean mAP (%) with 95% confidence intervals on the COCO dataset under varying numbers of forgetting classes. Shaded regions indicate confidence intervals. PPP consistently achieves effective forgetting across all four settings.

model weights and the retrieval set remain unchanged. Together with its notable results on DeepFashion, these findings suggest that hierarchical dampening is an effective strategy for scenarios with entangled semantic structures.

Results on Multi-Class Unlearning

Setting. To further explore the scalability of PPP, we conduct multi-class unlearning experiments on the COCO dataset. In addition to the single-class setting (i) 'dog', reported in Table 2, we evaluate three expanded configurations: (ii) = (i) + 'boat'; (iii) = (ii) + 'orange, vase' (i.e., four classes); (iv) = (ii) + 'airplane, bench, pizza, bed' (i.e., eight classes).

The forgetting classes are selected to be semantically diverse and mutually disjoint (e.g., 'dog' as an animal vs. 'airplane' as a vehicle), allowing us to assess the robustness of PPP under increasingly complex forgetting demands.

Results. As illustrated in Figure 3, PPP demonstrates effective forgetting and retention in (i), (ii), and (iii). However, as the number of forgetting classes increases, we observe a growing deviation on $\mathcal{D}_f$, indicating more pronounced forgetting. This trend is attributed to the increased number of unlearning samples, which enables the model to more precisely identify and suppress sensitive parameters. In the more challenging (iii) and (iv), where multiple classes are forgotten, performance on $\mathcal{D}_r$ begins to decline. Specifically, in setting (iv), PPP experiences a 6.26% drop in overall mAP, which is expected given that 10% of the COCO label space (8 out of 80 annotated categories) is removed. This suggests that under extensive forgetting demands, it becomes increasingly difficult to isolate semantically entangled parameters without affecting unrelated knowledge.

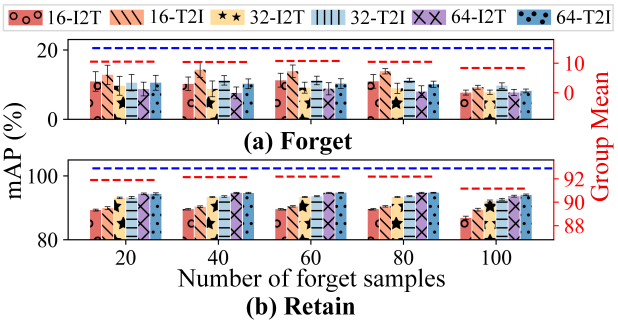

Figure 4: mAP (%) on COCO under varying forgetting sample sizes, with 95% confidence intervals. Each bar shows the mAP for a specific hash bits and retrieval direction (I2T, T2I); error bars denote the 95% confidence intervals. The left y-axis corresponds to bar values, and the right y-axis to group means. Green dashed lines indicate average mAP across retrieval settings; the blue dashed line marks the ground truth.

Results on Forgetting Sample Size

Setting. In the previous subsection, increasing the number of forgetting samples was associated with a more pronounced drop in performance on $\mathcal{D}_f$, indicating stronger unlearning effects that may deviate from the retrain baseline. To further investigate the relationship between forgetting sample size and unlearning behavior, we conduct experiments on the COCO using four different sample sizes: 20, 40, 60, and 100. Note that the results reported in Table 2 correspond to a sample size of 80.

Results. As shown in Figure 4, the model's performance on $\mathcal{D}_r$ remains consistently high across forgetting sample sizes from 20 to 100, indicating strong retention of general capability. Even with only 20 samples, PPP achieves effective forgetting on $\mathcal{D}_f$, demonstrating its ability to operate under limited data. As the number of forgetting samples increases, the unlearning effect on $\mathcal{D}_f$ becomes more stable-reflected by narrower confidence intervals (e.g., from 2.16 to 0.56 on 64-bit T2I), indicating more deterministic forgetting. This stems from the accumulation of empirical FIM signals, which sharpens the identification of sensitive parameter regions and enables more targeted suppression. Overall, the best trade-off occurs with 60-80 samples, showing that PPP remains reliable even with modest forgetting data.

Conclusion

We propose PPP, a retraining-free unlearning framework tailored for cross-modal retrieval. PPP constructs prior prototypes to capture semantic relationships, selects semantically related neighbors, and estimates parameter sensitivity across forgetting, neighbor, and retained sets. Parameters are partitioned into buffer and core regions, followed by a hierarchical dampening strategy that applies stronger suppression to core parameters. PPP achieves competitive performance on classification tasks and demonstrates strong results in cross-modal retrieval.

Acknowledgments

This work was supported by the Program of Beijing Huairou Laboratory (YZD2024025A); the National Natural Science Foundation of China (62266043); the Finance science and technology project of Xinjiang Uyghur Autonomous Region (2023B01029-1, 2023B01029-2); the Outstanding Young Talent Foundation of Xinjiang Uygur Autonomous Region of China (2023TSYCCX0043); the Excellent Youth Foundation of Xinjiang Uygur Autonomous Region of China(2023D01E01).

References

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-Generation Hyperparameter Optimization Framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2623–2631.

Chundawat, V. S.; Tarun, A. K.; Mandal, M.; and Kankanhalli, M. 2023a. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7210–7217.

Chundawat, V. S.; Tarun, A. K.; Mandal, M.; and Kankanhalli, M. 2023b. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18: 2345–2354.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Foster, J.; Schoepf, S.; and Brintrup, A. 2024. Fast machine unlearning without retraining through selective synaptic dampening. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 12043–12051.

Garg, S.; Goldwasser, S.; and Vasudevan, P. N. 2020. Formalizing data deletion in the context of the right to be forgotten. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, 373–402. Springer.

Golatkar, A.; Achille, A.; and Soatto, S. 2020a. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9304–9312.

Golatkar, A.; Achille, A.; and Soatto, S. 2020b. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *European Conference on Computer Vision*, 383–398. Springer.

Graves, L.; Nagisetty, V.; and Ganesh, V. 2021. Amnesiac Machine Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13): 11516–11524.

Huiskes, M. J.; and Lew, M. S. 2008. The MIR Flickr Retrieval Evaluation. In *MIR '08: Proceedings of the 2008 ACM International Conference on Multimedia Information Retrieval*. New York, NY, USA: ACM.

Huo, Y.; Qin, Q.; Dai, J.; Wang, L.; Zhang, W.; Huang, L.; and Wang, C. 2024. Deep Semantic-Aware Proxy Hashing for Multi-Label Cross-Modal Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(1): 576–589.

Krishnan, S.; and Wu, E. 2017. Palm: Machine learning explanations for iterative debugging. In *Proceedings of the 2Nd workshop on human-in-the-loop data analytics*, 1–6.

Liang, X.; Yang, E.; Yang, Y.; and Deng, C. 2024. Multi-Relational Deep Hashing for Cross-Modal Search. *IEEE Transactions on Image Processing*, 33: 3009–3020.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In Fleet, D.; Pajdla, T.; Schiele, B.; and Tuytelaars, T., eds., *Computer Vision – ECCV 2014*, 740–755. Cham: Springer International Publishing. ISBN 978-3-319-10602-1.

Liu, J.-Y.; Mao, X.-L.; Che, T.-Y.; and Tu, R.-C. 2025a. Distribution-Consistency-Guided multi-modal hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 12174–12182.

Liu, J.-Y.; Mao, X.-L.; Che, T.-Y.; and Tu, R.-C. 2025b. Distribution-Consistency-Guided Multi-modal Hashing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(11): 12174–12182.

Liu, S.; Yao, Y.; Jia, J.; Casper, S.; Baracaldo, N.; Hase, P.; Yao, Y.; Liu, C. Y.; Xu, X.; Li, H.; Varshney, K. R.; Bansal, M.; Koyejo, S.; and Liu, Y. 2025c. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2): 181–194.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1096–1104.

Mantelero, A. 2013. The EU Proposal for a General Data Protection Regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3): 229–235.

Nguyen, Q. P.; Low, B. K. H.; and Jaillet, P. 2020. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33: 16025–16036.

Shokri, R.; Stronati, M.; Song, C.; and Shmatikov, V. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. IEEE.

Ullah, E.; Mai, T.; Rao, A.; Rossi, R. A.; and Arora, R. 2021. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, 4126–4142. PMLR.

Zhang, P.-F.; Bai, G.; Huang, Z.; and Xu, X.-S. 2022. Machine Unlearning for Image Retrieval: A Generative Scrubbing Approach. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, 237–245. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392037.

Zhu, L.; Zheng, C.; Guan, W.; Li, J.; Yang, Y.; and Shen, H. T. 2024. Multi-Modal Hashing for Efficient Multimedia Retrieval: A Survey. *IEEE Transactions on Knowledge and Data Engineering*, 36(1): 239–260.