DIET: Machine Unlearning on a Data-Diet

Nilakshan Kunananthaseelan[1], Jing Wu[2], Trung Le[2], Gholamreza Haffari[2], Mehrtash Harandi[1]

[1]Department of Electrical and Computer Systems Engineering, Monash University, Australia
[2]Department of Data Science & AI, Monash University, Australia
{nilakshan.kunananthaseelan, jing.wu1, trunglm, gholamreza.haffari, mehrtash.harandi}@monash.edu

Abstract

Machine Unlearning (MU) aims to remove the influence of specific knowledge from a pretrained model. Existing methods often rely on retained training data to preserve utility; such dependence is impractical due to privacy and scalability constraints. A further complication arises when unlearning is applied to vision-language models (VLMs), where entangled multimodal representations make targeted forgetting especially challenging. We propose *DIET*, a principled **retain-data-free unlearning** method for VLMs that addresses these challenges by leveraging the geometry of hyperbolic space. The core idea is to push forget embeddings toward class-mismatched prototypes located at the boundary of the hyperbolic space. In hyperbolic geometry, points near the boundary become infinitely distant from interior points. As a result, moving forget embeddings to the boundary makes their influence on the model asymptotically negligible. To formalize this, we guide the forgetting process using the Busemann function, which quantifies directional distance to the boundary. We further develop an adaptive scheme based on optimal transport that selects mismatched prototypes for each forget embedding, enabling flexible unlearning dynamics. Extensive experiments on fine-grained datasets such as Flowers102, OxfordPets, and StanfordCars show that *DIET* achieves an average forget accuracy of 8.06%, while preserving 69.04% utility using only 16 samples per concept, significantly outperforming the best retain-free baselines with a **117.5%** improvement in model utility, and showing competitive performance to retain-data baselines with only a **3.79%** drop.

Code —
https://github.com/NilakshanKunananthaseelan/DIET

1 Introduction

Vision-Language Models (VLMs) are becoming the backbone of modern multimodal systems. Models such as CLIP (Radford et al. 2021), SigLIP (Zhai et al. 2023) provide universal embeddings that drive image generation, retrieval, and instruction-tuned multimodal assistants (Liu et al. 2023; Rombach et al. 2021). Their success stems from the large-scale pretraining aligning visual and textual representations into a shared space that offers a strong zero-shot generalization.

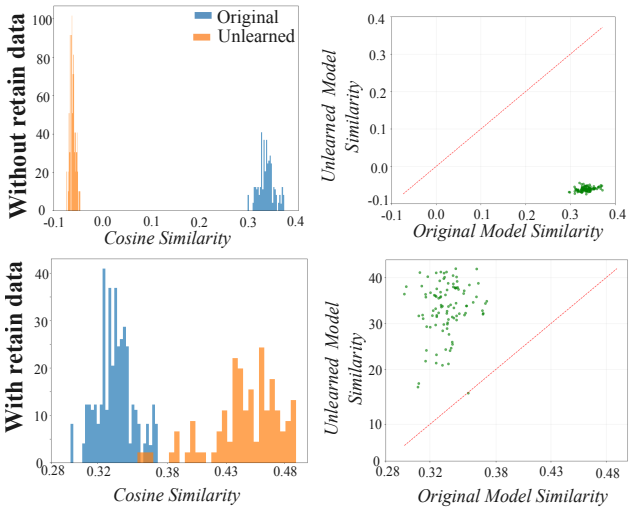

Figure 1: The similarity score with text modality decreases for the remaining class when we don't use the retain data; Finetuning on retain data preserves this alignment.

However, large-scale pretraining also introduces unintended behaviours. Publicly scraped data contains biases, errors and NSFW (not-safe-for-work) content, which can propagate to downstream tasks (Tanjim et al. 2024; Baherwani and Vincent 2024; Hamidieh et al. 2024). Moreover, the pretraining process itself may embed spurious correlations and artefacts, making the model unsuitable for certain applications (Alhamoud et al. 2025). This highlights the need for effective mechanisms to suppress such undesired behaviours while preserving overall utility.

Machine Unlearning (MU) offers this selectivity in principle by removing the influence of a specific knowledge from pretrained models (Gandikota et al. 2023; Qu et al. 2023). The primary recipe for MU performs targeted forgetting in the forget set and then finetunes on a retain set to avoid catastrophic unlearning (Fan et al. 2023; Wu and Harandi 2024). While the recipe is effective, the reliance on a retain set impedes the scalability of MU algorithms due to privacy constraints, compliance risks and computational

overhead. For VLMs like CLIP, curating multimodal retain data at scale is costly.

Recent "retain-data-free" methods aim to unlearn in unimodal models using only the targeted samples, without access to any retain data (Foster et al. 2024; Bonato, Cotogni, and Sabetta 2024; Cha et al. 2024). However, VLMs entangle knowledge across modalities and semantics. Hence, concept removal can be viewed as breaking the fundamental modality alignment obtained from pretraining. Without a careful mechanism to preserve surrounding knowledge, such disruption can lead to significant degradation in retained concepts (see Fig. 1).

Prior efforts on unlearning in VLMs, such as CLIPErase (Yang et al. 2024b), SafeCLIP (Poppi et al. 2024), and HySAC (Poppi et al. 2025) have shown promising progress, but they still benefit from additional training on retaining concepts. In response to these limitations, we aim to address the following question: *Can we perform targeted unlearning in VLMs without access to retained data, by leveraging their inherent semantic structure?*

We hypothesize that if forget embeddings are pushed toward infinity, their influence on the model becomes asymptotically negligible, effectively removing their contribution. To operationalize this idea, we formulate learning in hyperbolic space, and in particular, adopt the Poincaré ball model. In the Poincaré model, the notion of infinity is naturally encoded as the boundary of the ball. This, in comparison to Euclidean space, offers two key advantages: first, points near the boundary are at an exponentially increasing geodesic distance from interior points, making them suitable targets for forgetting; second, the Busemann function (Busemann 2005), which quantifies asymptotic proximity to the boundary along geodesics, admits a closed-form expression in the Poincaré ball (Cannon et al. 1997), facilitating a principled and efficient mechanism to guide forgetting trajectories.

To define meaningful geodesics for forgetting, we exploit the multimodal structure of VLMs. Using the VLM's text encoder, we identify a set of class-representative prototypes, *i.e.*, embeddings of mismatched (semantically unrelated) class labels, placed at the boundary of the Poincaré ball. Pushing forget embeddings toward these class-mismatched prototypes along geodesics not only drives them toward infinity but also introduces semantic confusion by steering them toward unrelated concepts. This reduces the likelihood that the model can retain or reconstruct the forgotten information. However, a key challenge lies in determining which prototype each forget embedding should be assigned to. While prior approaches often rely on manually fixed mismatched targets, we propose an **adaptive scheme** by making use of the optimal transport. This enables us to assign each forget embedding to a semantically distant prototype based on the underlying geometry, ensuring the forgetting process remains flexible and tailored to the structure of the embedding space.

Our contributions are as follows:

- We identify and address a critical limitation of existing MU methods for VLMs: their reliance on retained data to preserve utility. To this end, we introduce *DIET*, a modular and retain-data-free unlearning framework.
- We formulate unlearning in the Poincaré ball model of hyperbolic space and leverage a Busemann-guided loss to push forget embeddings toward the boundary. By transporting these embeddings along geodesics toward mismatched class prototypes, we achieve targeted forgetting with minimal interference.
- We provide a geometric rationale for our approach, hypothesizing that pushing forget embeddings to infinity effectively nullifies their influence. Empirical evidence supports this hypothesis, indicating that the method disrupts modality alignment for forgotten concepts while preserving modality alignment for unrelated ones.
- We propose an adaptive optimal transport mechanism to assign forget embeddings to semantically distinct prototypes, thereby avoiding the need for manually fixed targets and enabling flexible, geometry-aware trajectories.
- Thorough experiments on fine-grained datasets demonstrate that our geometry-oriented approach outperforms existing retain-free MU methods and achieves performance comparable to retain-data baselines, offering a promising path toward scalable and principled MU in VLMs.

2 Related Work

Hyperbolic Representation Learning. Machine learning models are typically formulated in Euclidean space. However, the choice of geometry of the representation space impacts the expressiveness of learned embeddings. The volume growth in hyperbolic space enables embeddings to capture hierarchical data with minimal distortion (Peng et al. 2022; Ganea, Bécigneul, and Hofmann 2018). This makes hyperbolic representations particularly effective for modeling naturally hierarchical domains and constructing robust visual–semantic relationships (Yang et al. 2024a; Atigh et al. 2022; Khrulkov et al. 2020). Such geometric properties have direct implications for multimodal learning and how concepts are organized across modalities (Desai et al. 2023). Relevant to our study is the exponential scaling of distances in hyperbolic space and the well-defined boundary structure of the Poincaré ball, which makes it a powerful tool for pushing representations to the edge of the space, where their influence on the model can become vanishingly small.

Machine Unlearning MU (Cao and Yang 2015; Kwak et al. 2017; Schelter 2020; Ginart et al. 2019) aims to selectively remove the influence of specific knowledge from a pretrained model. With the growing demand for safety, privacy, and regulatory compliance, MU has emerged as an active research direction across both recognition (Fan et al. 2023; Zhao et al. 2024; Wu and Harandi 2024) and generative settings (Gandikota et al. 2023; Wu et al. 2024; Maini et al. 2024; Liu et al. 2024). While prior work has mainly focused on unimodal domains, recent efforts have targeted unlearning in VLMs. Notable examples

are SafeCLIP (Poppi et al. 2024), HySAC (Poppi et al. 2025) and Cliperase (Yang et al. 2024b). SafeCLIP reduces sensitivity to NSFW content by redirecting embeddings toward a "safe" representation derived from a synthetic dataset. HySAC shares this safety goal, but operates in the hyperbolic space using the Lorentz model (Cannon et al. 1997). Cliperase proposes a modular framework with a forgetting module to suppress the target concepts, a retention module to preserve utility, and a consistency module to maintain alignment with the original model.

Despite their progress, the aforementioned methods depend heavily on access to retained concepts during optimization—a requirement that limits the scalability and deployability of MU algorithms in real-world settings. In response, recent efforts have explored retain-data-free unlearning as a scalable and privacy-preserving alternative to traditional MU: Foster et al. (2024) proposes JiT, a zero-shot unlearning approach that estimates samples' influence via information gain and attenuates the gain using Lipschitz smoothing optimization. Kravets and Namboodiri (2025) extends the Lipschitz-based unlearning to VLMs, using synthetic forget examples. SCAR (Bonato, Cotogni, and Sabetta 2024) pushes forget features toward the closest class using a Mahalanobis-based loss, while preserving model utility through knowledge distillation using out-of-distribution data. Cha et al. (2024) proposes Learning to Unlearn, which conducts instance-wise unlearning by intentionally flipping forget sample prediction while regularizing for stability. These approaches demonstrate the growing viability of retain-free unlearning; however, most remain grounded in unimodal architectures. In contrast, *DIET* addresses the significantly more constrained and underexplored problem of **retain-data-free unlearning in VLMs**, where deeply entangled multimodal representations make forgetting particularly challenging.

3 Methodology

This section presents our unlearning method, which aims to erase targeted knowledge by pushing embeddings towards infinity. We introduce the Poincaré ball model and the Busemann function, then we provide details on our approach. An overview of the foundational concepts of hyperbolic geometry can be found in the appendix.

3.1 Hyperbolic Space for Unlearning

Unlike Euclidean spaces, which possess zero curvature and lack an inherent representation of abstraction levels, the constant negative curvature of hyperbolic space results in distances growing exponentially with radius. This fundamental property naturally encodes hierarchies (Ganea, Bécigneul, and Hofmann 2018; Khrulkov et al. 2020; Atigh et al. 2022) and provides an effectively unbounded volume. As a result, it allows for localized perturbations within the space, enabling the unlearning of specific knowledge with minimal collateral damage to semantically similar but retained concepts. However, the benefits of hyperbolic space for MU remain largely underexplored, with recent works like HySAC (Poppi et al. 2025) showcasing it. Below, we provide key definitions in hyperbolic geometry used throughout this work.

Definition 3.1 (Hyperbolic Space)**.** An n-dimensional hyperbolic space $\mathbb{H}^n$ is a complete, simply-connected Riemannian manifold of constant sectional curvature.

Hyperbolic space can be realized by several models; a commonly used model in machine learning is the Poincaré Ball model (Nickel and Kiela 2017).

Definition 3.2 (Poincaré Ball Model)**.** The canonical n-dimensional Poincaré Ball model of $\mathbb{H}^n$ with curvature $c = -1$ is defined as $(\mathbb{B}^n, g^{\mathbb{B}}_{\mathbf{z}})$ with open ball $\mathbb{B}^n_c = \{\mathbf{z} \in \mathbb{R}^n : \|\mathbf{z}\| < 1\}$ and Riemannian metric tensor $g^{\mathbb{B}}_{\mathbf{z}} = 4(1 - \|\mathbf{z}\|^2)^{-2}\mathbf{I}_n$.

The geodesic distance between two points $\mathbf{z}_1, \mathbf{z}_2 \in \mathbb{B}$ is given by Equation 1:

$$d_{\mathbb{B}}(\mathbf{z}_1, \mathbf{z}_2) = \operatorname{arcosh}\left(1 + 2\frac{\|\mathbf{z}_1 - \mathbf{z}_2\|^2}{(1 - \|\mathbf{z}_1\|^2)(1 - \|\mathbf{z}_2\|^2)}\right), \tag{1}$$

and the exponential map at the origin (*i.e.*, $\mathbf{0}$) is given by

$$\exp_0(\mathbf{u}) = \tanh\left(\|\mathbf{u}\|/2\right)\frac{\mathbf{u}}{\|\mathbf{u}\|}. \tag{2}$$

In the Poincaré model, the set of points at infinity, which in our work realize ideal prototypes ($\partial\mathbb{B}^n$), are defined as:

$$\partial\mathbb{B}^n = \{\mathbf{p} \in \mathbb{R}^n : \|\mathbf{p}\|^2 = 1\} \tag{3}$$

Equation 1 shows that the distance between any points $\mathbf{z} \in \mathbb{B}^n$ inside the ball ($\|\mathbf{z}\| < 1$) and *ideal prototypes*, points at the boundary ($\|\mathbf{z}_2\|$=1) approaches ∞ (*i.e.* $d_{\mathbb{B}}(\mathbf{z}, \mathbf{p}) \to \infty$ for $\mathbf{p} \in \partial\mathbb{B}^n$). Therefore, developing an objective function to push the points to infinity is challenging. (Ghadimi Atigh, Keller-Ressel, and Mettes 2021) introduce a loss function based on the Busemann Function (Busemann 2005), which effectively measures the proximity to infinity.

Definition 3.3 (Busemann Function)**.** For a given ideal point $\mathbf{p} \in \partial\mathbb{B}^n$ and a geodesic $\gamma_{\mathbf{p}}$ connecting the origin to $\mathbf{p}$, the Busemann function w.r.t. $\mathbf{p}$ is defined for $\mathbf{z} \in \mathbb{B}^n$ as

$$\delta_{\mathbf{p}}(\mathbf{z}) = \lim_{t \to \infty}(d_{\mathbb{B}}(\gamma_{\mathbf{p}}(t), \mathbf{z}) - t), \tag{4}$$

where $d_{\mathbb{B}}\big(\gamma_{\mathbf{p}}(0),\, \gamma_{\mathbf{p}}(t)\big) = t$.

Equation 4 measures how much close or far $\mathbf{z}$ is from infinity along the geodesic path $\gamma_{\mathbf{p}}$ towards $\mathbf{p}$.

This abstract definition has a simple, closed-form solution in the Poincaré model (Ghadimi Atigh, Keller-Ressel, and Mettes 2021) as:

$$\delta_{\mathbf{p}}(\mathbf{z}) = \log\frac{\|\mathbf{p} - \mathbf{z}\|^2}{(1 - \|\mathbf{z}\|^2)}. \tag{5}$$

3.2 *DIET*: Machine Unlearning on a Data-Diet

In the following, we begin by outlining unlearning in VLMs, followed by its formulation in hyperbolic space. We describe prototype construction, geodesic direction selection toward the boundary, and introduce our proposed method.

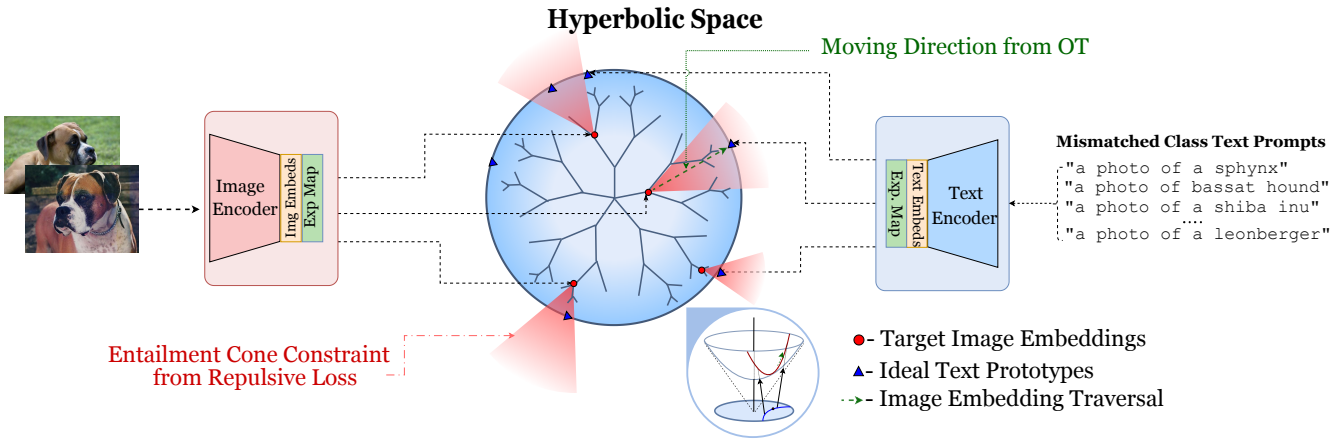

Figure 2: Hyperbolic forgetting through geodesic transport. a) Image and text encoders generate embeddings that are exponentially mapped to hyperbolic space. (b) Target image embeddings are aligned with their corresponding ideal text prototypes (e.g., "a photo of a sphynx") via geodesic paths determined by optimal transport, with blue dotted lines denoting the geodesics. (c) Entailment cone constraints from repulsive loss enforce separation between target embeddings and mismatched class text prompts, ensuring structured alignment while minimizing cross-class interference.

MU for Concept Removal in VLMs. VLMs such as CLIP are often trained with image-text pairs using contrastive loss, and their performance is described through the multimodality alignment of a concept. We focus on CLIP (Radford et al. 2021) as a representative vision-language model to elaborate on *DIET*; however, our method is model-agnostic and broadly applicable to VLMs. For a pretrained CLIP with an image encoder $f_{\texttt{img}}(\cdot)$, a text encoder $f_{\texttt{text}}(\cdot)$, we define $\mathcal{D}_f$ as the subset containing instances of the concept c to be removed, and $\mathcal{D}_p$ has the text prompt p of concepts we would like to retain. For an image-text pairs, $(\boldsymbol{x}_f, y_f) \in \mathcal{D}_f$, the modality alignment can be measured using similarity, $\text{sim}\big(f_{\texttt{img}}(\boldsymbol{x}_f), f_{\texttt{text}}(y_f)\big)$ using the cosine similarity score. The unlearning in CLIP can be viewed as a deliberate breaking of the modality alignment between positive pairs of c, while ensuring the utility of retaining positive pairs. We define our concept removal task by modifying the image encoder and obtaining the scrubbed image encoder $f^u_{\texttt{img}}$ such that:

1. The model breaks modality alignment between visual and textual representations of the concept c. For $(\boldsymbol{x}_f, y_f) \in \mathcal{D}_f, p \in \mathcal{D}_p$,

$$\text{sim}\big(f^u_{\texttt{img}}(\boldsymbol{x}_f), f_{\texttt{text}}(y_f)\big) \ll \text{sim}\big(f^u_{\texttt{img}}(\boldsymbol{x}_f), f_{\texttt{text}}(p)\big) \tag{6}$$

2. The model preserves the modality alignment for positive pairs. For the positive pairs $(\boldsymbol{x}, p)$ such that $\boldsymbol{x} \notin \mathcal{D}_f$, $p \in \mathcal{D}_p$,

$$\text{sim}\big(f_{\texttt{img}}(\boldsymbol{x}), f_{\texttt{text}}(p)\big) \approx \text{sim}\big(f^u_{\texttt{img}}(\boldsymbol{x}), f_{\texttt{text}}(p)\big) \tag{7}$$

To achieve the desired concept removal described above, we modify the image encoder to push the forget embeddings from $\mathcal{D}_f$ away from their aligned text representation, guiding them toward boundary points $\mathbf{p} \in \partial\mathbb{B}^n$ in Poincaré ball.

For each forget example $\boldsymbol{x}_f \in \mathcal{D}_f$ we obtain the hyperbolic embedding $\mathbf{z}$ on the Poincaré ball as follows:

$$\mathbf{x} = f_{\texttt{img}}(\boldsymbol{x}_f), \tag{8}$$
$$\mathbf{z} = \exp_0(\mathbf{x}), \tag{9}$$

where $\exp_0(\cdot)$ maps the feature into the hyperbolic space using Eqn. 2. A standard geodesic loss is intractable due to infinite distances at the boundary; therefore, we utilize the Busemann function in Eqn. 5 to define a *finite* differentiable distance between $\mathbf{z}$ and $\mathbf{p}$. Our unlearning objective minimizes $\delta_{\mathbf{p}}(\mathbf{z})$ to drive each forget embedding towards its designated ideal prototype along the geodesic ray $\gamma_{\mathbf{p}}$.

$$\mathcal{L}_{\text{HYP}}(\boldsymbol{\theta}; \mathcal{D}_f, \mathcal{D}_p) = \mathbb{E}_{x \in \mathcal{D}_f}\big[\delta_{\mathbf{p}}(\mathbf{z})\big]. \tag{10}$$

Creating *ideal prototypes*. While pushing forget embeddings to infinite distance theoretically disrupts the modality alignment, selecting an arbitrary 'away' direction in the infinite target space of $\mathbb{B}^n$ risks damaging retained knowledge and ineffective unlearning. Therefore, to guide this repulsion precisely, we introduce text-based **retained knowledge prototypes**.

These prototypes are generated from the pretrained CLIP text encoder, $f_{\texttt{text}}(\cdot)$, using prompts of retained classes, eliminating reliance on retained data constraints. For each retained class $\texttt{c} \in \mathcal{C}_{\text{retain}}$, we use the prompt $P_{\texttt{c}} = \texttt{"a photo of a c"}$, and compute its Euclidean embedding $\mathbf{x}_{\texttt{c}} = f_{\texttt{text}}(P_{\texttt{c}})$. This is mapped to Poincaré ball via the exponential map in Eqn. 2

$$\mathbf{z}_{\text{ret}} = \exp_0(\mathbf{x}_{\texttt{c}}), \tag{11}$$
$$\mathbf{p}_{\texttt{c}} = \frac{\mathbf{z}_{\text{ret}}}{\|\mathbf{z}_{\text{ret}}\|}\big(\mathfrak{r} - \epsilon\big), \tag{12}$$

where $\mathfrak{r} = 1, \epsilon \ll 1$.

The resulting prototype set for K retain concepts is given by:

$$\mathcal{P}_{\text{retain}} = \{\mathbf{p}_{\text{c}} \mid \text{c} \in \mathcal{C}_{\text{retain}}\}. \tag{13}$$

To break the modality alignment of positive image-text pairs in $\mathcal{D}_f$, *DIET* pushes forgetting embedding $\mathbf{z}$ towards "infinity" along the directions defined by retained prototypes $\mathbf{p}_{\text{c}} \in \mathcal{P}_{\text{retain}}$.

Selecting Geodesic Direction. Effective unlearning requires pushing forget embeddings far from their original position in a structured way, rather than an arbitrary manner. We achieve this by using optimal transport (OT) to assign each forget embedding $\mathbf{z}_i$ to a boundary prototype $\mathbf{p}_k$ in a cost-efficient way in hyperbolic space.

Given a mini-batch of forget embeddings $\{\mathbf{z}_i\}_{i=1}^N$ and a set of ideal prototypes $\{\mathbf{p}_i\}_{i=1}^K$, we build an online cost matrix Ω

using the Busemann distance. We then apply Sinkhorn-Knopp algorithm (Sinkhorn and Knopp 1967) to compute the soft OT plan $\Pi \in \mathbb{R}^{N\times K}$ that assigns each $\mathbf{z}_i$ to a prototype. This yields the OT loss:

$$\mathcal{L}_{\text{OT}}(\boldsymbol{\theta}; \mathcal{D}_f, \mathcal{D}_p) = \sum_{i=1}^{N}\sum_{k=1}^{K} \Pi_{ik}\Omega_{ik} = \langle \Pi, \Omega \rangle_F, \tag{14}$$

which encourages semantically structured spreading of forget embeddings while still pushing them toward the Poincaré boundary. To define the trajectory for each embedding $\mathbf{z}_i$, we select the most likely assigned prototype:

$$\mathbf{p}_{k^*} = \underset{k\in 1,\ldots,K}{\operatorname{argmax}} \big(\Pi_{ik}\big). \tag{15}$$

Repulsive Loss for Utility Preservation. To prevent the interference with retained concepts during unlearning, we introduce a repulsive loss that encourages separation between forget embeddings and all unassigned retained embeddings. We formulate it as a hinge loss with a predefined margin.

$$\mathcal{L}_{\text{REP}}(\boldsymbol{\theta}; \mathcal{D}_f, \mathcal{D}_p) = \mathbb{E}_{\substack{\mathbf{z}\in\mathcal{D}_f,\\ \mathbf{p}\in\mathcal{P}_{\text{retain}}\setminus\{\mathbf{p}_{k^*}\}}} \left[\max\left(0,\ \tau - \delta_{\mathbf{p}}(\mathbf{z})\right)\right], \tag{16}$$

where τ is the margin, $\mathbf{p}_{k^*}$ is the assigned prototypes for $\mathbf{z}$, $\delta_{\mathbf{p}_k}(\mathbf{z}_i)$ Busemann distance between $\mathbf{p}_k$ and $\mathbf{z}_i$.

Hybrid Modeling. Building on this geometric foundation, we formulate *DIET* as a hybrid model that combines Euclidean parameterization with hyperbolic geometry. Specifically, we treat the CLIP image embeddings as points in the Poincaré ball $\mathbb{B}^n$ and define our forgetting loss based on Busemann distance (Eqn. 5) to guide concept unlearning. Fig 2 illustrates this process: each forget embedding $\mathbf{z}_i$ follows a Busemann geodesic toward an ideal prototype $\mathbf{p}_k$, with optimal transport ensuring semantic alignment and minimal interference to retained knowledge. A pseudo-code of the algorithm is provided in 1.

Algorithm 1: Hyperbolic Prototype-based Unlearning (*DIET*)

Input: Forget dataset $\mathcal{D}_f$, Retain prototypes $\mathcal{P}_{\text{retain}}$, $\boldsymbol{\theta}$.
Parameter: Learning rate η, Margin τ, Epochs T
Output: Updated parameters $\boldsymbol{\theta}^u$.

for epoch $t = 1$ to T **do**
  Compute embeddings $\mathbf{z}_i = f(x_i)$ in hyperbolic space
  Compute cost matrix $\Omega_{ik} = \delta_{\mathbf{p}_k}(\mathbf{z}_i)$ using Busemann distance
  Solve optimal transport: $\Pi = \text{Sinkhorn}(\Omega)$
  Assign target prototype $k_i^* = \arg\max_k \Pi_{ik}$
  Compute $\mathcal{L}_{\text{U}}$
  Update model parameters: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \cdot \nabla_{\boldsymbol{\theta}}\mathcal{L}_{\text{U}}$
end for
return $\boldsymbol{\theta}^u$

The total unlearning loss L_{U} combines the hyperbolic loss (Eqn 10), optimal transport loss (Eqn. 14), and retain regularizer (Eqn. 16):

$$\mathcal{L}_{\text{U}} = \lambda_{\text{HYP}} \cdot \mathcal{L}_{\text{HYP}} + \lambda_{\text{OT}} \cdot \mathcal{L}_{\text{OT}} + \lambda_{\text{REP}} \cdot \mathcal{L}_{\text{REP}}, \tag{17}$$

where $\lambda_{\text{HYP}}, \lambda_{\text{OT}}$, and λ_{RET} are the respective weighting coefficients.

To apply this loss within a standard training framework, we express $\boldsymbol{\theta}$ as the LoRA weights such that $\boldsymbol{\theta} = \boldsymbol{BA}$, parameterized by low-rank matrices $\boldsymbol{A} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{B} \in \mathbb{R}^{r\times n}$. We integrate $\boldsymbol{\theta}$ into the query (q), key (k), and value (v) projections of the vision encoder targeting the attention mechanism to enable precise concept forgetting while avoiding broad structural damage.

Optimization of Euclidean Parameters Using the Hyperbolic Loss When optimizing Euclidean parameters $\boldsymbol{\theta}$ with hyperbolic loss $\mathcal{L}_{\text{U}}$, the gradient in the Poincaré ball relates to the Euclidean gradient via a conformal factor $g_{\mathbf{z}}^{\mathbb{B}}$ that vanishes as embeddings approach the boundary. To prevent vanishing gradients while maintaining the push-to-infinity effect, we adopt norm clipping on Euclidean features (Guo et al. 2022), which keeps $\mathbf{z}$ close to the boundary for stable optimization.

4 Experiments

4.1 Setup

Training and Evaluation. We conduct experiments using LoRA adapters (rank $r = 4$, scaling factor $\alpha = 1$) applied to the vision encoder of pretrained CLIP. Following the dataset splits from CoOp (Zhou et al. 2022) and SalUn (Fan et al. 2023), we train *DIET* using 16 shots per concept for 30 epochs on a single A5500 (24GB). We perform hyperparameter search over learning rate $\in [0.0001, 0.005]$, $\lambda_{\text{OT}} \in [0.1, 1]$, $\lambda_{\text{HYP}} \in [10, 50]$, and $\lambda_{\text{REP}} \in [0, 30]$, and report results with the best performing configuration: learning rate 0.0009, $\lambda_{\text{HYP}} = 30$, $\lambda_{\text{OT}} = 1$, and $\lambda_{\text{REP}} = 20$. For optimal transport, we use the Sinkhorn algorithm from the POT library (Flamary et al. 2021). We evaluate *DIET* on datasets including Flowers102 (Nilsback and Zisserman 2008), Pets (Parkhi

	Method	D_r	Flowers102		Pets		Cars		Food101		Average	
			$\mathcal{D}_f$ (%)	D_t (%)	$\mathcal{D}_f$ (%)	D_t (%)	$\mathcal{D}_f$ (%)	D_t (%)	$\mathcal{D}_f$ (%)	D_t (%)	$\mathcal{D}_f$ (%)	D_t (%)
Pretrained	Zero-shot	-	92.61±7.42	70.48±0.10	83.25±19.28	89.12±0.32	53.43±21.60	65.58±0.09	87.20±3.64	86.12±0.05	79.12	77.82
Weights updated	FT	✓	56.58±28.2	82.07±7.74	62.74±31.75	88.34±2.19	35.12±15.37	60.32±6.82	70.68±8.75	79.95±3.28	56.28	77.67
	GA	✓	86.55±14.89	70.81±0.29	74.75±24.66	88.52±0.43	46.92±22.06	65.18±0.11	84.68±6.94	84.61±0.33	73.23	77.28
	SHs	✓	0.38±0.77	28.28±23.31	0.00±0.00	3.53±0.61	0.00±0.00	0.73±0.25	7.00±14.00	15.39±27.90	1.85	11.98
	SalUn	✓	21.71±21.89	43.59±33.77	14.81±11.63	59.52±29.68	28.10±12.16	66.21±0.55	55.80±12.29	84.47±0.56	30.11	63.45
	SalUn*	✗	24.44±15.46	56.19±20.27	26.00±20.15	55.04±37.94	7.08±7.22	22.85±23.89	36.16±33.29	40.05±29.12	23.42	43.53
LoRA based	GA	✗	0.00±0.00	17.71±5.49	0.00±0.00	45.25±9.97	0.00±0.00	7.15±1.94	0.00±0.00	56.86±8.75	0.00	31.74
	GS-LoRA	✓	**0.00**±0.00	**68.15**±0.98	**0.25**±0.50	**84.33**±0.60	**0.00**±0.00	**54.60**±0.53	**0.28**±0.39	**79.95**±0.38	**0.13**	**71.76**
	GS-LoRA*	✗	0.00±0.00	1.66±0.79	0.00±0.00	2.89±0.13	0.00±0.00	0.70±0.20	0.00±0.00	1.67±0.66	0.00	1.73
	***DIET* (Ours)**	✗	6.08±8.84	61.20±2.48	3.51±2.89	80.23±6.64	5.97±6.27	54.14±6.25	16.68±5.89	80.59±3.42	8.06	69.04

Table 1: *DIET* performance on fine-grained datasets. We report accuracy on $\mathcal{D}_f$ (forget set) and D_t (test set). D_r indicates whether a retained dataset was used (✓) or not (✗). * indicates the performance without using a retained dataset. The gray entries show the best overall performance, while underlined values highlight the best-performing retain-free setting.

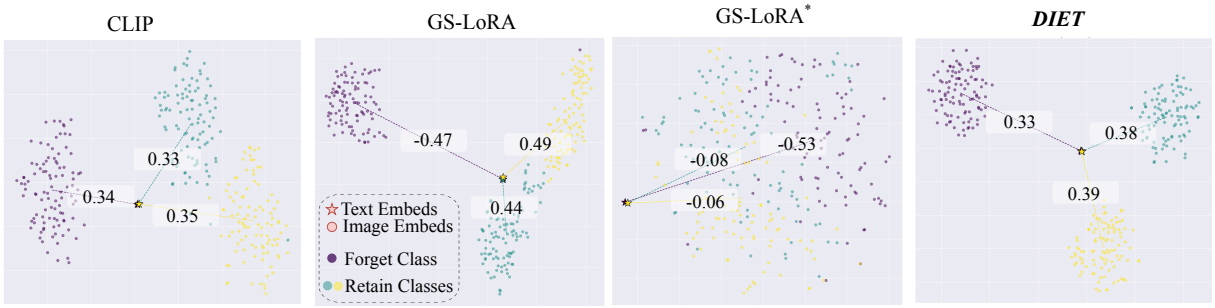

Figure 3: Latent space visualization of CLIP before and after unlearning the 'Shiba Inu' class from OxfordPets. We show the shared latent space in CLIP before unlearning (Original CLIP) and after unlearning based on GS-LoRA, GS-LoRA*, and *DIET* (Ours). The similarity score is computed as the average cosine similarity between positive image-text pairs in the shared latent space.

et al. 2012), StanfordCars (Krause et al. 2013) and Food101 (Bossard, Guillaumin, and Van Gool 2014). Following SalUn, we report accuracy on the full forget set ($\mathcal{D}_f$) to measure forgetting efficacy and test set accuracy (D_t) to assess retention, averaging results over 5 random forget concepts per dataset.

Baselines. The literature on retain-data-free MU for VLMs remains sparse. To address this, we adapt several prominent unimodal unlearning algorithms to the multimodal context of CLIP for comparison. We primarily considered methods that directly update the model's weight: **GA** (Thudi et al. 2022) proposed the gradient-ascent-based optimization; **SalUn** (Fan et al. 2023) removes forget-set knowledge via saliency-guided gradients, then reinforces retained knowledge; **SHs** (Wu and Harandi 2024) generates a trimmed model by re-initializing influential top-k parameters and finetuned on retain data with a gradient projection. Among LoRA-based methods, the closest is **GS-LoRA** (Zhao et al. 2024), originally designed for continual unlearning in ViTs. Though not intended for multimodal MU, we adapt it as a baseline in our setting. Additional details are in the appendix.

4.2 Results and Discussion

Concept removal in CLIP. We present the results on fine-grained datasets in Table 1. The baselines highlight the core trade-off in MU between forgetting and preserving utility. SHs achieve near-perfect forgetting but suffer from catastrophic forgetting, with test accuracy dropping to 11.98%. The performance gap between retain-based methods and their retain-free counterparts is significant. For example, SalUn's average D_t drops from 63.45% to 43.53% when the retain set is removed. While GS-LoRA could effectively preserve the utility with near-perfect forgetting, its retain-data-free variant fails with complete collateral damage. These results show the severe difficulty of preserving model utility in a truly retain-data-free setting and motivate the need for a more principled approach.

Our proposed method, *DIET*, demonstrates superior performance as a retain-data-free approach. *DIET* achieves an effective knowledge removal (8.06% avg. D_f) while preserving a high degree of model utility with an average performance of 69.04%. This results outperforms other retain-data-free baseline, SalUn*(43.53%) and GS-LoRA* (1.73%). Furthermore, *DIET* closes the gap with retain-based methods, performing competitively with the best baseline, GS-LoRA (71.76%), despite having no access to the retain dataset. A similar result was obtained for general image datasets as well.

Few-shot MU in CLIP. Our main experiment uses 16 samples per concept. To explore how the number of forget samples affects the balance of the forgetting *vs*. model utility

for VLMs, we conducted an extended analysis shown in Fig. 4. With a similar setting, *DIET* struggles to forget effectively when the number of instances is reduced, while GS-LoRA and GS-LoRA* maintain near-zero forgetting. In contrast, our method shows competitive performance to GS-LoRA in preserving model utility compared to its retain-free version GS-LoRA*. While these results are encouraging, determining the number of forgotten samples remains an important open question for future research.

Effect of Unlearning on Shared Latent Space. We visualize CLIP's shared latent space before and after unlearning a concept (*e.g. Shiba Inu*) to examine how modality alignment is affected. Figure 3 shows that after unlearning, similarity for the target concept drops to negative values for GS-LoRA (-0.47) and GS-LoRA* (-0.53), while *DIET* reduces it by only 0.1. GS-LoRA employs gradient ascent, causing hard semantic reversal, which disrupts the feature space (evident in GS-LoRA*). In contrast, *DIET* geometrically steers embeddings along Busemann geodesics toward "infinity", dampening alignment without semantic reversal and preserving the structure of retained concepts.

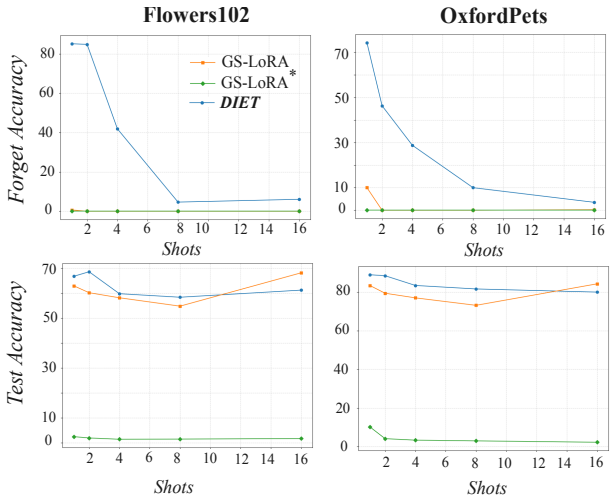

Figure 4: Performance of MU methods on Flowers102 and Pets datasets.

4.3 Ablation Studies

Optimal Transport. The choice of direction to move toward the boundary plays crucial role that governs the trade-off between forgetting and model utility. We opt to use an adaptive optimal transport plan instead of randomly selecting a direction(**Without OT**). This helps to transport the forget samples smoothly in the unbounded volume of the hyperbolic space. Further, we introduce a repulsive hinge loss(*vs*. **With OT (No Retain Reg.)**) as an additional regularizer to preserve model utility. Our ablation study in Tab.2 shows both the optimal transport plan and the repulsive regularizer are vital components of *DIET*, working in synergy to achieve a good trade-off.

Text-based prototypes. A key component of *DIET* is generating suitable prototypes to define geodesic directions. To validate our choice of semantically meaningful text-based prototypes, we compare them against randomly sampled boundary points. Table 3 shows that semantic prototypes significantly improve forgetting performance. While random directions maintain separation from retained concepts, they fail to achieve consistent forgetting, highlighting the importance of semantic alignment in prototype selection.

Dataset	Dataset	w/o OT	w/o Repul. Loss	*DIET*
Pets	D_f (%)	2.51±1.60	4.02±2.93	**3.51**±2.89
	D_t (%)	75.32±7.35	79.98±5.67	**80.23**±6.64
OxfordFlowers	D_f (%)	3.29±3.74	**3.67**±4.74	6.08±8.84
	D_t (%)	52.43±5.54	**60.71**±1.42	61.20±2.48
Caltech101	D_f (%)	31.19±29.46	17.82±10.28	**16.13**±7.67
	D_t (%)	91.71±0.46	92.73±0.34	**92.98**±0.28
StanfordCars	D_f (%)	4.68±5.96	5.36± 5.27	**5.97**±6.27
	D_t (%)	47.17±13.94	53.56±7.17	**54.14**±6.25

Table 2: Ablation study on Optimal Transport plan. Comparison of baselines: (1) without OT and (2) with OT but without repulsive loss, across 5 concepts.

Metric	Pets	Flowers102	Caltech101	StanfordCars
Random Prototypes				
D_f (%)	17.78±15.65	27.65±24.89	51.92±6.45	7.09±8.57
D_t (%)	81.89±3.85	60.19±1.90	93.02±0.15	53.42±5.39
Text Prototypes (Ours)				
D_f (%)	**3.51**±2.90	**6.08**±8.84	**16.13**±7.67	**5.97**±6.27
D_t (%)	**80.23**±6.64	**61.20**±2.48	**92.98**±0.28	**54.14**±6.25

Table 3: Comparison of random *vs*. text-based prototypes across datasets.

5 Conclusion and Limitations

We introduced *DIET*, a retain-data-free unlearning method for VLMs that leverages hyperbolic geometry to push forget embeddings toward infinity along geodesic directions. By combining Busemann distance-based loss with adaptive optimal transport, *DIET* disrupts modality alignment of targeted concepts while preserving utility. Experiments show *DIET* outperforms retain-free baselines and performs competitively with retain-based methods.

DIET has notable limitations. First, pushing embeddings to infinity may be insufficient for complex, heavily entangled manifolds. Second, prototype quality significantly impacts performance; poorly chosen prototypes lead to suboptimal results. Third, hyperbolic optimization is inherently unstable, requiring norm clipping, repulsive loss, and careful hyperparameter tuning. Lastly, *DIET* attenuates but does not eliminate modality alignment, limiting applicability where complete disentanglement is legally or ethically required.

Acknowledgments

Mehrtash Harandi is supported by the Australian Research Council (ARC) Discovery Program DP250100262.

References

Alhamoud, K.; Alshammari, S.; Tian, Y.; Li, G.; Torr, P. H.; Kim, Y.; and Ghassemi, M. 2025. Vision-language models do not understand negation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29612–29622.

Atigh, M. G.; Schoep, J.; Acar, E.; Van Noord, N.; and Mettes, P. 2022. Hyperbolic image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4453–4462.

Baherwani, V.; and Vincent, J. J. 2024. Racial and Gender Stereotypes Encoded Into CLIP Representations. In *The Second Tiny Papers Track at ICLR 2024*.

Bonato, J.; Cotogni, M.; and Sabetta, L. 2024. Is retain set all you need in machine unlearning? restoring performance of unlearned models with out-of-distribution images. In *European Conference on Computer Vision*, 1–19. Springer.

Bossard, L.; Guillaumin, M.; and Van Gool, L. 2014. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, 446–461. Springer.

Busemann, H. 2005. *The geometry of geodesics*. Courier Corporation.

Cannon, J. W.; Floyd, W. J.; Kenyon, R.; Parry, W. R.; et al. 1997. Hyperbolic geometry. *Flavors of geometry*, 31(59-115): 2.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Cha, S.; Cho, S.; Hwang, D.; Lee, H.; Moon, T.; and Lee, M. 2024. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 11186–11194.

Desai, K.; Nickel, M.; Rajpurohit, T.; Johnson, J.; and Vedantam, S. R. 2023. Hyperbolic image-text representations. In *International Conference on Machine Learning*, 7694–7731. PMLR.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2023. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*.

Flamary, R.; Courty, N.; Gramfort, A.; Alaya, M. Z.; Boisbunon, A.; Chambon, S.; Chapel, L.; Corenflos, A.; Fatras, K.; Fournier, N.; Gautheron, L.; Gayraud, N. T.; Janati, H.; Rakotomamonjy, A.; Redko, I.; Rolet, A.; Schutz, A.; Seguy, V.; Sutherland, D. J.; Tavenard, R.; Tong, A.; and Vayer, T. 2021. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78): 1–8.

Foster, J.; Fogarty, K.; Schoepf, S.; Dugue, Z.; Öztireli, C.; and Brintrup, A. 2024. An information theoretic approach to machine unlearning. *arXiv preprint arXiv:2402.01401*.

Gandikota, R.; Materzynska, J.; Fiotto-Kaufman, J.; and Bau, D. 2023. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2426–2436.

Ganea, O.; Bécigneul, G.; and Hofmann, T. 2018. Hyperbolic neural networks. *Advances in neural information processing systems*, 31.

Ghadimi Atigh, M.; Keller-Ressel, M.; and Mettes, P. 2021. Hyperbolic busemann learning with ideal prototypes. *Advances in neural information processing systems*, 34: 103–115.

Ginart, A.; Guan, M.; Valiant, G.; and Zou, J. Y. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32.

Guo, Y.; Wang, X.; Chen, Y.; and Yu, S. X. 2022. Clipped hyperbolic classifiers are super-hyperbolic classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11–20.

Hamidieh, K.; Zhang, H.; Gerych, W.; Hartvigsen, T.; and Ghassemi, M. 2024. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 547–561.

Khrulkov, V.; Mirvakhabova, L.; Ustinova, E.; Oseledets, I.; and Lempitsky, V. 2020. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6418–6428.

Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, 554–561.

Kravets, A.; and Namboodiri, V. P. 2025. Zero-shot class unlearning in clip with synthetic samples. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 6456–6464. IEEE.

Kwak, C.; Lee, J.; Park, K.; and Lee, H. 2017. Let Machines Unlearn - Machine Unlearning and the Right to be Forgotten. In *Americas Conference on Information Systems*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Z.; Dou, G.; Tan, Z.; Tian, Y.; and Jiang, M. 2024. Machine unlearning in generative ai: A survey. *arXiv preprint arXiv:2407.20516*.

Maini, P.; Feng, Z.; Schwarzschild, A.; Lipton, Z. C.; and Kolter, J. Z. 2024. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*.

Nickel, M.; and Kiela, D. 2017. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, 722–729. IEEE.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Peng, W.; Varanka, T.; Mostafa, A.; Shi, H.; and Zhao, G. 2022. Hyperbolic Deep Neural Networks: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12): 10023–10044.

Poppi, S.; Poppi, T.; Cocchi, F.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024. Safe-clip: Removing nsfw concepts from vision-and-language models. In *European Conference on Computer Vision*, 340–356. Springer.

Poppi, T.; Kasarla, T.; Mettes, P.; Baraldi, L.; and Cucchiara, R. 2025. Hyperbolic Safety-Aware Vision-Language Models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 4222–4232.

Qu, Y.; Shen, X.; He, X.; Backes, M.; Zannettou, S.; and Zhang, Y. 2023. Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models. In *Proceedings of the 2023 ACM SIGSAC conference on computer and communications security*, 3403–3417.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv 2022. *arXiv preprint arXiv:2112.10752*.

Schelter, S. 2020. Amnesia-a selection of machine learning models that can forget user data very fast. *suicide*, 8364(44035): 46992.

Sinkhorn, R.; and Knopp, P. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2): 343–348.

Tanjim, M. M.; Singh, K. K.; Kafle, K.; Sinha, R.; and Cottrell, G. W. 2024. Discovering and mitigating biases in clip-based image editing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2984–2993.

Thudi, A.; Deza, G.; Chandrasekaran, V.; and Papernot, N. 2022. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, 303–319. IEEE.

Wu, J.; and Harandi, M. 2024. Scissorhands: Scrub data influence via connection sensitivity in networks. In *European Conference on Computer Vision*, 367–384. Springer.

Wu, J.; Le, T.; Hayat, M.; and Harandi, M. 2024. Erasediff: Erasing data influence in diffusion models. *arXiv preprint arXiv:2401.05779*.

Yang, M.; Feng, A.; Xiong, B.; Liu, J.; King, I.; and Ying, R. 2024a. Hyperbolic fine-tuning for large language models. *arXiv preprint arXiv:2410.04010*.

Yang, T.; Dai, L.; Wang, X.; Cheng, M.; Tian, Y.; and Zhang, X. 2024b. Cliperase: Efficient unlearning of visual-textual associations in clip. *arXiv preprint arXiv:2410.23330*.

Zhai, X.; Mustafa, B.; Kolesnikov, A.; and Beyer, L. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

Zhao, H.; Ni, B.; Fan, J.; Wang, Y.; Chen, Y.; Meng, G.; and Zhang, Z. 2024. Continual forgetting for pre-trained vision models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 28631–28642.

Zhou, K.; Yang, J.; Loy, C. C.; and Liu, Z. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9): 2337–2348.