Aalborg Universitet

AALBORG UNIVERSITY

Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning

Adapting Alignment Calibration to MERU

Pujol Vidal, Alex; Escalera Guerrero, Sergio; Nasrollahi, Kamal; Moeslund, Thomas B.

DOI (link to publication from Publisher):
10.48550/arXiv.2503.15166

Creative Commons License
Unspecified

Publication date:
2025

Document Version
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

Citation for published version (APA):
Pujol Vidal, A., Escalera Guerrero, S., Nasrollahi, K., & Moeslund, T. B. (2025). *Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU*. arXiv. https://doi.org/10.48550/arXiv.2503.15166

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: 11 07, 2026

arXiv:2503.15166v2 [cs.CV] 14 Apr 2025

Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU

Àlex Pujol Vidal[1,3] alexpv@create.aau.dk
Kamal Nasrollahi[1,3,4] kna@milestone.dk
Thomas B. Moeslund[1,4] tbm@create.aau.dk
Sergio Escalera[1,2,4] sescalera@ub.edu
[1]Aalborg University, Aalborg, Denmark
[2]University of Barcelona and Computer Vision Center, Barcelona, Spain
[3]Milestone Systems, Broendby, Denmark
[4]Pioneer Center for Artificial Intelligence, Copenhagen, Denmark

Abstract

Machine unlearning methods have become increasingly important for selective removal of harmful concepts in large multimodal models. While recent work has explored unlearning in Euclidean contrastive vision-language models, the effectiveness of concept removal in hyperbolic spaces remains unexplored. This paper explores this gap by adapting Alignment Calibration to MERU, a model that embeds images and text in hyperbolic space to better capture semantic hierarchies. Through systematic experiments and ablation studies, we demonstrate that hyperbolic geometry offers distinct advantages for concept removal, achieving lower accuracy at zero classification in forget set with reasonable performance on retain set, particularly when scaling to multiple concept removal. Our approach introduces hyperbolic-specific components including entailment calibration and norm regularization that leverage the unique properties of hyperbolic space. Comparative analysis with Euclidean models reveals fundamental differences in unlearning dynamics, with hyperbolic unlearning reorganizing the semantic hierarchy while Euclidean approaches merely disconnect cross-modal associations. These findings provide insights into the geometric properties that influence concept removal, setting a direction towards better control and safer deployment for large multimodal models. Code available at https://github.com/alex-pv01/HAC.

1. Introduction

It is commonly said that 'we learn through contrast'. By comparing seemingly opposite concepts, we gain deeper understanding—a fundamental principle underlying contrastive learning algorithms. This approach, has emerged as a powerful framework for learning representations that effectively distinguish similar from dissimilar in machine learning systems [6].

The success of contrastive learning has led to the development of powerful multimodal models like CLIP, which aligns visual and textual representations in a shared Euclidean embedding space [30]. These models serve as foundational components for numerous downstream applications including zero-shot classification [29], image generation [32], and robotics [31]. More recently, MERU [9] has extended this paradigm by embedding images and text in hyperbolic space, better capturing the hierarchical relationships inherent in visual-semantic data.

Despite their impressive capabilities, these models rely on massive datasets scraped from the internet, which often contain problematic content including private information, copyrighted material, and harmful data [2]. This raises significant ethical and legal concerns, as these models become widely deployed. Machine unlearning—the ability to selectively remove information from trained models—initially emerged in response to regulatory frameworks like GDPR's "right to be forgotten"[10], but has since expanded to address a broader spectrum of challenges including privacy protection, safety enhancement, security concerns, fairness improvement, model robustness and copyright matters [7, 37].

In particular, concept removal (of NSFW or copyrighted content) has been increasingly relevant to ensure a safe deployment [28]. While recent work has explored concept removal in Euclidean contrastive learning models like CLIP [16, 17, 28, 36, 38], the unique geometry of hyperbolic

spaces presents new challenges and opportunities that remain unexplored. The geometric distinction raises an important question: How does the underlying geometry affect the efficacy and dynamics of machine unlearning in contrastive learning models?

In this paper, we investigate the adaptation of Alignment Calibration (AC) to hyperbolic representations in MERU. We present a systematic comparison of unlearning effectiveness between CLIP and MERU models trained on identical data. Our contributions are threefold:

1. Proposing Hyperbolic Alignment Calibration (HAC), an extension of AC to operate in hyperbolic space by reformulating its objective functions using hyperbolic distance and incorporating entailment constraints.
2. Based on zero-shot classification evaluation, we conduct a comparative analysis of machine unlearning in Euclidean and hyperbolic contrastive learning spaces using AC and HAC, revealing that the hyperbolic method excels at removing the targeted concept at the cost of a decrease in performance on the retain set.
3. Through visualization techniques, we complement our insights on how the geometries affect unlearning dynamics. Indicating that HAC may profit from the exponential expansion of the hyperbolic space to achieve better unlearning.

This work bridges two important research areas: geometric representation learning and machine unlearning. Our study advances the understanding on how representation spaces encode and forget concepts, contributing to the theoretical foundations of machine unlearning for advanced vision language models.

2. Related work

Hyperbolic Representation Learning The choice of geometry for representation spaces has significant implications for the expressive power of learned embeddings. While most machine learning models operate in Euclidean space, recent work has explored alternative geometries, particularly hyperbolic geometry for data with hierarchical structure [23, 27].

Characterized by its negative curvature, a hyperbolic space provides a continuous analog of tree-like structures [24]. In hyperbolic space, the volume of a ball grows exponentially with its radius, allowing to embed hierarchical data with low distortion [33]. This property is valuable for representing data with natural hierarchies, such as taxonomies, social networks, and visual-semantic relationships [9, 12, 24]. These geometric insights have direct implications for multimodal contrastive learning, on how concepts are organized and related across modalities.

Multimodal Contrastive Learning Contrastive learning emerged as a powerful approach for unsupervised representation learning, training models to distinguish between similar (positive) and dissimilar (negative) pairs of examples [6]. This approach is formalized through the InfoNCE loss function [34]. Given a batch of N samples, where for every sample z_i, there is a positive pair $z_{k_i} \neq z_i$, then:

$$L_{\text{InfoNCE}} = \frac{-1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(z_i, z_{k_i})/\tau)}{\sum_{k=1,\ k \neq i}^{N} \exp(\text{sim}(z_i, z_k)/\tau)} \tag{1}$$

where $\text{sim}(u, v)$ denotes cosine similarity, and τ is a temperature parameter. CLIP [30] extended this approach to the multimodal domain by learning a dual-encoder architecture with separate visual and textual encoders that project images and their corresponding text descriptions into a shared Euclidean space. More recently, MERU [9] enhanced CLIP's representational power by embedding images and text in hyperbolic and using the negative hyperbolic distance as a similarity metric. It achieved improved performance on tasks involving hierarchical relationships while maintaining CLIP's zero-shot capabilities.

The transition from Euclidean to hyperbolic embeddings in multimodal contrastive learning raises important questions about how these different geometries affect the organization of concepts within the representation space.

Machine Unlearning Machine Unlearning (MU), introduced by Cao and Yang [4], selectively removes the influence of specific training data from a trained model. While early research focused primarily on classification tasks [37], the rapid growth of generative AI has expanded the scope of MU to higher-order or latent information [7, 11, 20].

Particularly, several works on concept removal for contrastive vision-language models have recently emerged. Poppi et al. [28] addressed safety concerns by developing a method to diminish CLIP's sensitivity to NSFW content. Their approach is based on redirecting synthetic "unsafe" data to match "safe" data. Kravets and Namboodiri [16] applied Lipschitz regularization to achieve zero-shot class unlearning in CLIP. In subsequent work, Kravets and Namboodiri [17] demonstrated that class forgetting in CLIP can be accomplished without any visual data by adapting the shared vision-text space, making the forgetting process more efficient. Yang et al. [38] introduced CLIPErase through three modules: a Forgetting Module that disrupts associations in the forget set, a Retention Module that preserves performance on the retain set, and a Consistency Module that maintains alignment with the original model. Similarly, Wang et al. [36] proposed Alignment Calibration, which explicitly considers the properties of contrastive learning and introduce toward novel auditing metrics to verify unlearning.

While these approaches have shown promise for Euclidean-based models like CLIP, they do not account for

the unique properties of hyperbolic geometry that are leveraged in models like MERU.

Despite mentioned, no prior work has investigated how hyperbolic geometry affects concept removal in contrastive models. This gap is significant given hyperbolic space's natural capacity for representing hierarchical relationships, which could fundamentally alter how concepts are forgotten in multimodal representations.

3. Method

3.1. Machine Unlearning for Concept Removal

Given a pre-trained contrastive model with image encoder f_{img}, text encoder f_{txt}, and a training dataset $\mathcal{D}$ with image-text pairs. Let c be the concept to be removed, we define $\mathcal{D}_f$ as the subset containing instances that refer to c and $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$ as the retain set. That is, if c is the concept of "dog", D_f consists of positive pairs of images of dogs and captions describing the image. The unlearning task is to find modify the original model and obtain $f^*_{\text{img}}, f^*_{\text{txt}}$ such that:

1. The model no longer associates visual and textual representations of concept c. Formally, for a positive pair $(x_f, t_f) \in \mathcal{D}_f$ exists a pair $(x, t) \in \mathcal{D}_r$,

$$\text{sim}(f^*_{\text{img}}(x_f), f^*_{\text{txt}}(t_f)) \ll \text{sim}(f^*_{\text{img}}(x_f), f^*_{\text{txt}}(t)), \quad (2)$$

and

$$\text{sim}(f^*_{\text{img}}(x_f), f^*_{\text{txt}}(t_f)) \ll \text{sim}(f^*_{\text{img}}(x), f^*_{\text{txt}}(t_f)). \quad (3)$$

2. Performance on other concepts in remains unaffected. For positive pairs $(x, t) \in \mathcal{D}_r$,

$$\text{sim}(f^*_{\text{img}}(x), f^*_{\text{txt}}(t)) \approx \text{sim}(f_{\text{img}}(x), f_{\text{txt}}(t)). \quad (4)$$

Note that concept removal is a recent problem highlighted in literature and has no established formal definition [7]. However, the above definition aligns with contrastive learning approaches on related work [16, 17, 36], where unlearning is measured by accuracy on zero-shot classification, as later discussed in Section 4.2.

3.2. Alignment Calibration in Euclidean Space

AC consists of a retaining loss and an unlearning loss that work together to selectively push apart image and text embeddings to forget, while preserving model performance [36]. Given a batch $\{(x^r_i, t^r_i)\}_{i=1}^N \subset \mathcal{D}_r$ and $\{(x^f_i, t^f_i)\}_{i=1}^N \subset \mathcal{D}_f$ from a dataset $\mathcal{D}$. Let $x' = f_{\text{img}}(x)$ and $t' = f_{\text{txt}}(t)$ denote image and text embeddings respectively. The retention loss consists of the usual loss for CLIP (see Sec. 10.1 in supp. material), preserving model performance on $\mathcal{D}_r$:

$$L_{\text{retain}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\text{sim}(x_i^{'r}, t_i^{'r})/\tau)}{\sum_{j=1}^{2N}\exp(\text{sim}(x_i^{'r}, t'_j)/\tau)}\right. \quad (5)$$

$$\left. + \log\frac{\exp(\text{sim}(x_i^{'r}, t_i^{'r})/\tau)}{\sum_{j=1}^{2N}\exp(\text{sim}(x'_j, t_i^{'r})/\tau)}\right]. \quad (6)$$

The unlearning loss consists of three components designed to disrupt associations in the forget set $\mathcal{D}_f$:

1. Negative alignment calibration (Eq. (7)), maximizes similarity between negative pairs in the forget set, confusing the model's understanding of c.

$$L_{\text{neg}} = -\frac{1}{2N^2}\sum_{i=1}^{N}\sum_{j=1, j\neq i}^{N}\frac{\text{sim}(x_i^{'f}, t_j^{'f}) + \text{sim}(x_j^{'f}, t_i^{'f})}{\tau}. \quad (7)$$

2. Positive alignment calibration (Eq. (8)), minimizes similarity between matching pairs in the forget set, directly attacking the concept association.

$$L_{\text{pos}} = \frac{1}{N}\sum_{i=1}^{N}\text{sim}(x_i^{'f}, t_i^{'f})/\tau. \quad (8)$$

3. Performance preserving (Eq. (9)), ensures that general model capabilities remain intact. Note that this corresponds to the denominator from the retention loss, Eq. (5), but also considering D_f.

$$L_{\text{perf}} = \frac{1}{2N}\sum_{i=1}^{N}\left[\log\left(\frac{1}{2N}\sum_{j=1}^{2N}\exp(\text{sim}(x_i^{'f}, t'_j)/\tau)\right)\right. \quad (9)$$

$$\left. + \log\left(\frac{1}{2N}\sum_{j=1}^{2N}\exp(\text{sim}(x'_j, t_i^{'f})/\tau)\right)\right] \quad (10)$$

The combined unlearning loss is:

$$L_{\text{forget}} = \alpha \cdot L_{\text{neg}} + \beta \cdot L_{\text{pos}} + \gamma \cdot L_{\text{perf}}. \quad (11)$$

And the final AC loss balances retention and unlearning objectives:

$$L_{\text{AC}} = L_{\text{retain}} + \varepsilon \cdot L_{\text{forget}}, \quad (12)$$

where α, β, γ, and ε are hyperparameters controlling the relative importance of each component.

3.3. Alignment Calibration to Hyperbolic Space

HAC extends the AC approach to hyperbolic geometry by adapting each component of the loss function to operate in the Lorentz model of hyperbolic space. As in MERU's formulation (see Sec. 10.2 in supp. material), we replace

the cosine similarity with negative hyperbolic distance from component in AC. HAC incorporates entailment terms, introduced by Desai et al. [9]. We propose adapting the entailment to preserve and disrupt hierarchical relationships:

$$L_{\text{r-ent}} = \frac{1}{N}\sum_{i=1}^{N}\max(0, \text{ext}(x_i^{'r}, t_i^{'r}) - \text{aper}(t_i^{'r})), \quad (13)$$

$$L_{\text{f-ent}} = \frac{1}{N}\sum_{i=1}^{N}\max(0, \text{aper}(t_i^{'f}) - \text{ext}(x_i^{'f}, t_i^{'f})), \quad (14)$$

where $\text{ext}(x, t)$ is the exterior angle between the text embedding t, and the image embedding x, and $\text{aper}(t)$ is the half-aperture of the entailment cone for t. The retain-entailment loss (Eq. (13)) preserves the hierarchical structure for concepts in the retain set, minimizing the angle difference when the embedded image $x_i^{'r}$ lies outside the entailment cone for the text $t_i^{'r}$, following the same logic as MERU. On the other had, the forget-entailment loss (Eq. (14)) disrupts entailment relationships for texts and images of the concepts to be forgotten, minimizing the angle difference when $x_i^{'f}$ is inside the entailment cone of $t_i^{'f}$, forcing the images to move close to the edge of the cone. The combined HAC loss is:

$$L_{\text{HAC}} = L_{\text{retain}} + \varepsilon \cdot L_{\text{forget}} + \omega_r \cdot L_{\text{r-ent}} + \omega_f \cdot L_{\text{f-ent}}. \quad (15)$$

Additionally, inspired by Li et al. [21], who show that samples positioned close to the origin in hyperbolic space are likely to be out-of-distribution, we further propose adding a regularization term for the forgetting set that minimizes the hyperbolic norm of the image-text embeddings to be forgotten:

$$L_{\text{norm-reg}} = \frac{1}{N}\sum_{i=1}^{N}(||x_i^{'f}||_{\mathcal{L}} + ||t_i^{'f}||_{\mathcal{L}}), \quad (16)$$

where $||\cdot||_{\mathcal{L}}$ is the Lorentzian norm, see Sec. 10.2. This regularization pulls instances to remain closer to the origin, which in turn provides better numerical stability [14]. The final HAC loss with regularization is:

$$L_{\text{HAC-reg}} = L_{\text{HAC}} + \lambda \cdot L_{\text{norm-reg}} \quad (17)$$

3.4. Implementation Details

We implement both AC and HAC, building on the official CLIP and MERU codebases. For optimization, we use Adam [15] with learning rate $\eta = 5 \cdot 10^{-5}$, weight decay $\lambda = 10^{-5}$, and cosine learning rate decay over 15,000 iterations. We train with a batch size of 320 across all experiments. To ensure numerical stability in hyperbolic space computations, we apply gradient clipping with a maximum norm of 1.0. For unlearning, we train on a subset of RedCaps dataset [8], running experiments on a NVIDIA A100 GPU. We evaluate both models quantitatively and qualitatively on publicly available datasets 4.2.

4. Experiments

4.1. Experimental Setup

To investigate the impact of removing concepts from contrastive representation spaces, we focus on selectively disrupting the semantic relationship between image and text of concepts, such as dogs, cats, food, and plants. Both CLIP and MERU models have ViT-S [5] as image encoder and a 12-layer, 512 wide Trasnformer as text encoder [30]. They are pretrained on the RedCaps dataset and used as the orginal models[1] that require unlearning.

During experiments only a subset of 7M image-text pairs from the original RedCaps dataset was accessible. This, referred to as RedCaps2, serves as the original dataset. Instances related to dogs, cats, food, and plants are categorized into four distinct high-level semantic concepts, defining the different forget sets, see Sec. 11 in supp. material for more details on the set up forget sets. The unlearning procedure operates on balanced mini-batches with N image-text pairs from the retain set and N pairs from the unlearn set, resulting in a batch size of $2N$, with $N = 160$.

4.2. Evaluation

Zero-shot image classification The evaluation constructs prompts of the form "a picture of a [CLASS]" for each target class. Each test image embedding is compared against all class prompt embeddings using the appropriate similarity metric. The class with maximum similarity is selected as the prediction. Both accuracy on retained classes (R-acc) and forgotten classes (F-acc) are reported. A high F-acc indicates that the text-image embeddings are still close enough capturing the undesired semantic relationship in the latent space.

Latent Space Visualizations Image and text embeddings are visualized using both T-SNE [35] and hyperbolic T-SNE [13]. The latter is an extension of T-SNE which uses hyperbolic student's t-distribution to capture the hierarchical relationships between embeddings. These visualizations may reveal structural changes in the representation space after unlearning. We look particularly if there is a distortion in the alignment between the image and text embedding of the concepts to forget.

Datasets For evaluation, we selected a subset of the 6 datasets used in the original MERU paper [9], focusing on those containing classes related to our target concepts for unlearning. Additionally, for latent space visualization purposes we use a subset of COCO [22], containing images of a selected group of classes.

[1] Models availabe in MERU's GitHub repository

4.3. Ablation Studies

Alignment Calibration Components We assess the impact of the different alignment components on unlearning performance. We vary the weights from Eq. 11 ($\alpha, \beta, \gamma \in [0, 1]$), unlearning a single concept (dogs) in both CLIP and MERU. As baselines, we consider the orignal CLIP and MERU (o-C and o-M), both models finetuned only with the retain dataset (f-C-R and f-M-R), and both models finetuned on the entire dataset (f-C and f-M).

Hyperbolic-Specific Parameters We ablate parameters unique to HAC: entailment weights ($\omega_r, \omega_u \in [0, 1]$) and hyperbolic norm regularization ($\lambda \in [0, 2]$). We analyze how these hyperbolic-specific factors influence unlearning, evaluating on zero-shot classification for CIFAR-10.

Scaling Analysis We evaluate on diverse datasets and three experiments. Experiment A, consists of removing the concept-class dogs alone. Removing dogs serves as an ideal initial test case, since it has a strong semantic relationship with other concepts to retain (e.g., horses, cats). Experiment B extends this by removing both dogs and cats. Testing whether the unlearning mechanism can handle multiple related concepts simultaneously. Finally, Experiment C consists of unlearning dogs, cats, food, and plants, to test how each geometry handles large-scale concept removal without catastrophic forgetting of non-targeted concepts.

5. Results and Discussion

5.1. Alignment Calibration Components

We begin our analysis by examining how changing the key components of AC affects unlearning performance. Table 1 presents the effects of varying positive alignment calibration (β). For better numerical stability of the experiments, we introduce a small regularization norm with $\lambda = 0.1$ for HAC. We observe that increasing β leads to a significant reduction in forget accuracy (F-acc) in both AC and HAC, confirming that this component effectively pushes positive image-text pairs further apart in their respective embedding spaces and leading to zero-shot misclassifications. This is consistent with previous research showing that controlled modification of alignment between positive pairs can lead to selective forgetting [36, 38]. However, we observe a dramatic difference in unlearning effectiveness. In HAC, introducing even a small β, it causes forget accuracy to plummet on CIFAR-10. This suggests that hyperbolic geometry amplifies the effect of positive alignment calibration, potentially due to its natural exponential expansion. The cost of this enhanced forgetting capability is reflected in retain accuracy (R-acc). HAC experiences a more substantial drop in retain performance compared to AC when β is increased. Despite this trade-off, the relative difference between the drop in retain accuracy versus forget accuracy is more favorable in hyperbolic space, indicating superior unlearning efficiency.

Table 1. Zero-shot classification accuracy in retain and forget sets, varying positive alignment calibration. Largest difference in retain and forget performance in **green**. Best value for each column and geometry in **bold**.

Method	Weights		CIFAR-10		O-IIIT Pets	
	α, γ	β	R-acc↑	F-acc↓	R-acc↑	F-acc↓
AC	0.75	0	**60.5**	45.6	73.6	66.2
		0.25	60.3	31.7	73.5	48.7
		0.5	58.7	**21.2**	**74.9**	31.5
		0.75	58.4	24.9	73.9	**24.6**
f-C	0	0	58.9	48.1	74.6	69.9
f-C-R			60.6	63.6	72.3	72.2
O-C			59.4	66.4	74.6	73.2
HAC	0.5	0	55.2	73.6	74.8	63.9
		0.25	34.7	**0.0**	62.1	**15.8**
		0.5	49.9	**0.0**	69.6	17.9
		0.75	39.6	0.03	63.9	16.1
f-M	0	0	**56.4**	73.0	**75.2**	65.9
f-M-R			41.7	95.7	71.8	65.8
O-M			38.1	94.6	72.0	70.8

On the other side, table 2 examines the effects of varying negative alignment calibration (α) and performance preserving (γ) weights while maintaining a fixed positive alignment calibration. For AC, we observe that the retain accuracy (R-acc) remains relatively stable across different α and γ values, while forget accuracy (F-acc) shows more substantial fluctuations. This suggests that in Euclidean space, modifying α and γ has a more pronounced effect on the model's ability forget the selected concept. In contrast, HAC demonstrates remarkably consistent forgetting capabilities across all parameter configurations, maintaining near-zero forget accuracy regardless of α and γ values, while dampening retain accuracy. This remarkable behavior suggests that in hyperbolic space, once the positive alignment calibration establishes a forgetting directive, it dominates the unlearning process to such an extent that adjustments to other components extend the disruption to the retain set.

5.2. Hyperbolic-Specific Parameters

We now explore parameters that are unique to the hyperbolic geometry. Table 3 illustrates the impact of varying entailment loss weights on HAC performance. When AC components are deactivated ($\varepsilon = 0$), we observe not only a failure to achieve unlearning but actually an increase in forget accuracy on CIFAR-10. This finding suggests that

Table 2. Zero-shot classification accuracy in retain forget sets, varying negative alignment calibration and performance preserving. Largest difference in retain and forget performance in **green**. Best value for each column and geometry in **bold**.

Method	Weights		CIFAR-10		O-IIIT Pets	
	α, γ	β	R-acc↑	F-acc↓	R-acc↑	F-acc↓
AC	0.5	0.5	**58.8**	24.0	74.6	32.9
	0.75		58.7	**21.2**	**74.9**	**31.5**
	1		57.2	27.4	73.6	41.5
HAC	0.5	0.5	**49.9**	**0.0**	**69.6**	**17.9**
	0.75		40.7	0.02	67.5	18.0
	1		42.7	0.04	68.6	19.8

Table 3. Zero-shot classification accuracy in retain and forget sets, varying the weight entailment losses. Largest difference in retain and forget performance in **green**. Best value for each column and geometry in **bold**.

Weights			CIFAR-10		O-IIIT Pets	
ϵ	ω_r	ω_f	R-acc↑	F-acc↓	R-acc↑	F-acc↓
0	0.2	1.0	**56.3**	73	**74.9**	66.6
	1.0	0.2	47.2	85.9	68.5	64.6
0.05	0.2	1.0	39.9	**0.0**	60.7	**0.08**
	1.0	0.2	49.6	37.0	40.1	0.10
0.1	0.2	1.0	52.7	**0.0**	67.9	15.1
	1.0	0.2	44.0	48.7	56.8	18.6

entailment losses alone may inadvertently strengthen the association between text and image embeddings of the target concept, increasing their relative similarity, instead of decreasing it.

The introduction of AC components ($\varepsilon > 0$) dramatically changes this dynamic. Furthermore, when prioritizing the forget entailment ($\omega_f = 1.0$) complete forgetting (0% F-acc) on CIFAR-10 is achieved. This demonstrates the synergistic effect between alignment calibration and entailment losses in hyperbolic space. Similar patterns emerge for O-IIIT Pets, yielding near-perfect forgetting while maintaining reasonable retain accuracy.

The balance between retain and forget entailment weights (ω_r and ω_f) proves crucial for optimal unlearning. When $\omega_r > \omega_f$, we observe significantly higher forget accuracy across all configurations, indicating incomplete concept removal. This supports the theoretical understanding that in hyperbolic space, the precise positioning of embeddings within entailment cones strongly influences model behavior [9].

Finally, beyond providing better stability, the regularization term actively improves unlearning quality, as shown in Table 4. With $\lambda = 0.5$, we largest retain-forget trade-off

Table 4. Zero-shot classification accuracy in retain and forget sets, varying the hyperbolic norm regularization. Largest difference in retain and forget performance in **green**. Best value for each column and geometry in **bold**.

Method	Weight	CIFAR-10		O-IIIT Pets	
	λ	R-acc↑	F-acc↓	R-acc↑	F-acc↓
HAC	0	52.0	13.0	46.3	**0.04**
HAC-reg	0.1	52.7	**0.0**	67.9	15.1
	0.5	54.0	**0.0**	66.3	10.8
	2.0	**56.8**	49.2	**74.8**	35.9

across both datasets. This regularization effect is particularly beneficial for retaining accuracy illustrated with O-IIIT Pets, where at the cost of increasing forgetting accuracy, we observe a strong retain performance. The mechanism behind this improvement involves pulling text embeddings for the forgotten concept toward the origin of the hyperboloid, causing other images that were previously misclassified, to be reclassified under other categories, provably their category, increasing retain accuracy. This is especially important for fine-grained classification in O-IIIT Pets, where the regularization helps disentangle dogs from similar categories like cats.

However, excessive regularization ($\lambda = 2.0$) diminishes unlearning effectiveness, as evidenced by the significant increase in forget accuracy. This indicates that while regularization is essential for hyperbolic unlearning, its strength must be carefully calibrated to achieve the optimal balance between concept removal and retention.

5.3. Scaling Analysis

Fixing hyperparametes to the combination found with largest forget-retain difference, for both AC and HAC, we now investigate how these approaches scale when progressively increasing the complexity of the unlearning task. Table 7 presents a comprehensive comparison across multiple datasets when unlearning: (A) only "dog"; (B) "dog" and "cat"; and (C) "dog", "cat", "food", and "plants".

We observe a clear trade-off pattern between retain accuracy and forget accuracy across all experiments. AC consistently maintains higher retain accuracy across all datasets and concept sets compared to HAC. However, this advantage comes at a significant cost: AC's forget accuracy increases dramatically as the unlearning task scales, indicating incomplete concept removal and lower scaling capabilities. In stark contrast, HAC demonstrates superior forgetting capabilities across all scaling scenarios. Even when tasked with unlearning four diverse concepts simultaneously (experiment C), HAC maintains remarkably low forget accuracies. Suggesting that HAC's leverages the hierarchical hyperbolic structure to effectively remove mul-

Table 5. Zero-shot classification accuracy in retain set (R-acc) and forget set (F-acc), across different tasks, after unlearning: (A) "dog"; (B) "dog" and "cat"; (C) "dog", "cat", "food" and "plant". We report results for both CLIP and MERU after alignment calibration using the optimal configuration from Section 4.3. Values in **bold** indicate better at retaining or unlearning across A, B and C. A blank space - indicate that for that experiment and dataset there is no forget or retain set.

Task	Method	Unlearn Set	CIFAR-10[18]		CIFAR-100[19]		STL-10[1]		O-IIIT Pets[26]		Food101[3]		Flowers102[25]	
			R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓
Zero-shot Classification	AC	A	**58.7**	21.2	**27.9**	-	**88.1**	83.1	**74.9**	31.5	**72.4**	-	**44.7**	-
		B	**90.3**	71.4	**26.6**	-	**90.3**	71.4	-	53.4	**72.5**	-	**45.0**	-
		C	**90.0**	77.0	**23.4**	57.2	**90.0**	77.0	-	64.0	-	0.16	-	19.2
	HAC-reg	A	54.0	**0.0**	20.6	-	84.3	**38.0**	66.3	**10.8**	67.6	-	40.1	-
		B	83.5	**2.1**	21.8	-	83.5	**2.1**	-	**25.7**	59.6	-	36.4	-
		C	82.7	**22.1**	18.8	**21.6**	82.7	**22.1**	-	**28.7**	-	**0.08**	-	**0.04**

tiple concepts from the embedding space without cross-interference.

These experiments reveal that HAC offers more robust concept removal at scale, while AC better preserves performance on retained concepts. This fundamental trade-off highlights the complementary strengths of each geometric approach and suggests that the choice between them should be guided by whether the priority is complete concept removal or minimal disruption to retained knowledge. Section 12 in supplementary material complements our findings showcasing the confusion matrices foer CIFAR-10.

6. Latent Space Visualizations

To gain deeper insights into how geometric properties affect concept unlearning, we visualize the embedding spaces of both CLIP and MERU before and after removing the concept "dog". Figures 1 and 2 provide compelling visual evidence of the distinctive unlearning mechanisms in Euclidean versus hyperbolic spaces.

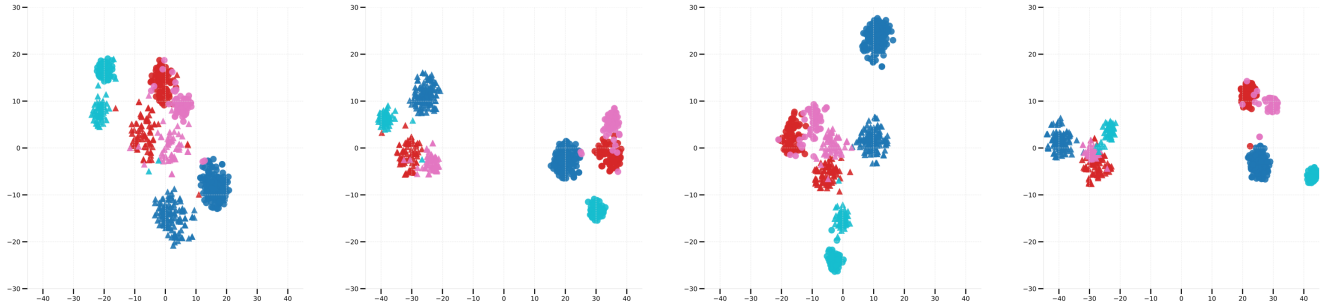
The T-SNE visualizations in Figure 1 reveal several key patterns. First, both AC and HAC induce a general separation between image and text representations after unlearning, indicating that enlarging the modality gap is a common consequence of unlearning regardless of geometry. However, the degree and nature of this separation differ between models. In the Euclidean space of CLIP (Figures 1a-b), both text and image embeddings maintain their class-wise clustering structure after unlearning. In contrast, MERU's hyperbolic space (Figures 1c-d) exhibits a more profound reorganization of semantic relationships. While image embeddings preserve their class separation, text embeddings show significant restructuring, with dog text representations (pink triangles) migrating toward the region occupied by cat text embeddings (red triangles). This visualization corroborates our quantitative findings that hyperbolic geometry facilitates complete concept removal by leveraging semantic hierarchies—in this case, potentially positioning "dog" as a subconcept of a more general "animal" category that includes cats. However, this poses a potential limitation of hyperbolic unlearning: while effective at repositioning target concepts, the structural interdependence of the hierarchical space means that semantically adjacent concepts (like cats when removing dogs) experience collateral disruption in their text embeddings, potentially explaining the lower retention accuracy observed in HAC compared to AC.

The hyperbolic T-SNE visualizations in Figure 2 provide additional insights into MERU's unlearning dynamics. Following unlearning, we observe a dramatic expansion of the embedding space, with image representations pushed substantially farther from the origin while text embeddings remain relatively close to it. This pattern aligns with hyperbolic geometry's exponential expansion property and illustrates how HAC exploits this characteristic. Particularly notable is the positioning of dog text embeddings near the origin after unlearning, effectively placing them "behind" cat text embeddings from the perspective of dog images. This spatial reorganization explains the superior unlearning results evaluated in zero-shot classification, dog images now find stronger alignment with text embeddings of other classes because their original text concept has been repositioned closer to the origin within the semantic hierarchy. Complementary illustrations can be found in Section 14 in supplementary material.

These visualizations demonstrate that while both geometric approaches achieve concept unlearning, they do so through fundamentally different mechanisms: Euclidean unlearning primarily disconnects cross-modal associations while preserving class structure, whereas hyperbolic unlearning reorganizes the semantic hierarchy itself, repositioning forgotten concepts within the taxonomic structure of the embedding space.

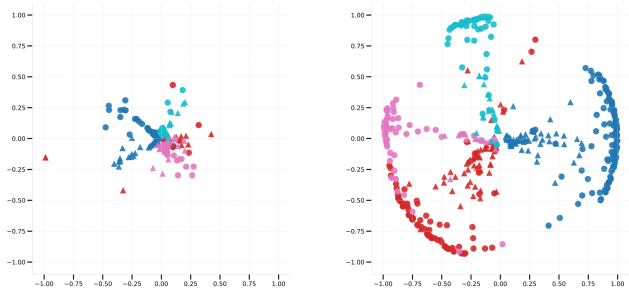
7. Limitations

While our work provides valuable insights into hyperbolic unlearning for contrastive models, several key limitations must be acknowledged. First, how to properly define concept boundaries for removal and defining a forgetting dataset remains an open problem in MU [7]. To mitigate



(a) Original CLIP (b) Unlearned CLIP (c) Original MERU (d) Unlearned MERU

Figure 1. Latent space visualizations with T-SNE of CLIP and MERU before and after removing the concept "dog". △ refer to text embeddings, ○ to image embeddings, and colors to **dogs**, **cats**, **pizzas**, and **buses**.



(a) Original MERU (b) Unlearned MERU

Figure 2. Latent space visualizations with hyperbolic T-SNE of MERU before and after removing the concept "dog". △ refer to text embeddings, ○ to image embeddings, and colors to **dogs**, **cats**, **pizzas**, and **buses**.

this issue, in this paper, we focus on concepts that correspond to discrete classes in common image-text datasets, see Sec. 11 for more details. However, given D_f, the formulation can apply to more abstract concepts as well.

Our interpretation of concept removal, disrupting text-image alignment and measuring effects through downstream tasks, aligns with related work [16, 17, 36, 38]. However, zero-shot classification relies on simple sentences that may not capture the full semantic complexity of visual concepts, potentially limiting our ability to precisely measure text-image alignment degradation.

The inherent numerical instability of hyperbolic space presents additional challenges. The experiments required gradient clipping beyond MERU's original training pipeline to prevent exploding gradients caused by positive alignment calibration interacting with hyperbolic expansion. A more thorough analysis of numerical stability specifically for unlearning tasks is needed, as these instabilities might actually be leveraged to more effectively disrupt text-image alignments for target concepts. For more robust conclusions, experiments should be conducted with multiple random seeds. The experiments used a single seed (0) across all conditions.

While this allows us to establish initial insights into concept removal for hyperbolic contrastive learning in vision-language models, additional experimentation with varied seeds would strengthen the reliability of our findings.

Finally, the latent space visualizations rely on T-SNE and hyperbolic T-SNE, which, while informative, introduce distortions when projecting high-dimensional embeddings to 2D space, potentially ignoring important geometric properties of the original hyperbolic manifold and making it difficult to fully capture the complex hierarchical relationships that emerge during the unlearning process.

Despite these limitations, our work represents an important first step in understanding the geometric implications of concept unlearning in hyperbolic contrastive vision-language models, providing a foundation for future research in this emerging area.

8. Conclusion

In this work, we adapted Alignment Calibration for concept removal in hyperbolic contrastive learning. The comparison between Euclidean and hyperbolic geometries reveals that while AC better preserves retained concepts with high retain accuracy, HAC achieves more concept removal with low forget accuracy, particularly when scaling to multiple related concepts. This difference stems from hyperbolic geometry's hierarchical structure, which enables precise manipulation of semantic relationships through entailment weights and norm regularization. Visualizations provide insights that hyperbolic unlearning reorganizes the semantic hierarchy itself, rather than merely separating cross-modal instances. These findings underscore the importance of geometry in multimodal representation learning and provide a foundation for further research into hyperbolic unlearning for vision-language models. Future work should focus on developing specialized metrics and interpretability techniques for hyperbolic concept removal, addressing numerical stability challenges, and exploring how these approaches transfer to larger multimodal models.

9. Acknowledgements

This work has been supported by Milestone Research Program at AAU; the Responsible AI for Value Creation project (REPAI); the Spanish project PID2022-136436NB-I00; and ICREA under the ICREA Academia Programme.

References

[1] Andrew Y. Ng Adam Coates, Honglak Lee. An analysis of single layer networks in unsupervised feature learning. *AISTATS*, 2011. 7

[2] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *ArXiv*, abs/2110.01963, 2021. 1

[3] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 7

[4] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pages 463–480, 2015. 2

[5] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021. 4

[6] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 539–546 vol. 1, 2005. 1, 2

[7] A. Feder Cooper, Christopher A. Choquette-Choo, Miranda Bogen, Matthew Jagielski, Katja Filippova, Ken Ziyu Liu, Alexandra Chouldechova, Jamie Hayes, Yangsibo Huang, Niloofar Mireshghallah, Ilia Shumailov, Eleni Triantafillou, Peter Kairouz, Nicole Mitchell, Percy Liang, Daniel E. Ho, Yejin Choi, Sanmi Koyejo, Fernando Delgado, James Grimmelmann, Vitaly Shmatikov, Christopher De Sa, Solon Barocas, Amy Cyphert, Mark Lemley, danah boyd, Jennifer Wortman Vaughan, Miles Brundage, David Bau, Seth Neel, Abigail Z. Jacobs, Andreas Terzis, Hanna M. Wallach, Nicolas Papernot, and Katherine Lee. Machine unlearning doesn't do what you think: Lessons for generative AI policy, research, and practice. *CoRR*, abs/2412.06966, 2024. 1, 2, 3, 7

[8] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. 4

[9] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Ramakrishna Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*. JMLR.org, 2023. 1, 2, 4, 6

[10] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council. 1

[11] Rohit Gandikota, Joanna Materzyńska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2426–2436, 2023. 2

[12] Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1646–1655. PMLR, 2018. 2

[13] Yunhui Guo, Haoran Guo, and Stella Yu. Co-sne: Dimensionality reduction and visualization for hyperbolic data. pages 11–20, 2022. 4, 2

[14] Yunhui Guo, Xudong Wang, Yubei Chen, and Stella X. Yu. Clipped hyperbolic classifiers are super-hyperbolic classifiers. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2022. 4, 2

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 2014. 4

[16] Alexey Kravets and Vinay Namboodiri. Zero-shot class unlearning in clip with synthetic samples. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pages 6456–6464, 2025. 1, 2, 3, 8

[17] Alexey Kravets and Vinay P. Namboodiri. Zero-shot CLIP class forgetting via text-image space adaptation. *Transactions on Machine Learning Research*, 2025. 1, 2, 3, 8

[18] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). . 7

[19] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-100 (canadian institute for advanced research). . 7

[20] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22634–22645, 2023. 2

[21] Huimin Li, Zhentao Chen, Yunhao Xu, and Junlin Hu. Hyperbolic anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17511–17520, 2024. 4

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 4

[23] Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *International Journal of Computer Vision*, 132(9):3484–3508, 2024. 2

[24] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 2

[25] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 7

[26] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 7

[27] Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10023–10044, 2022. 2

[28] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-clip: Removing nsfw concepts from vision-and-language models. In *Computer Vision – ECCV 2024*, pages 340–356, Cham, 2025. Springer Nature Switzerland. 1, 2

[29] Qi Qian and Juhua Hu. Online zero-shot classification with clip. In *Computer Vision – ECCV 2024*, pages 462–477, Cham, 2024. Springer Nature Switzerland. 1

[30] 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. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 4

[31] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. In *Proceedings of The 6th Conference on Robot Learning*, pages 416–426. PMLR, 2023. 1

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. 1

[33] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Re. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, pages 4460–4469. PMLR, 2018. 2

[34] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. 2

[35] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9 (86):2579–2605, 2008. 4

[36] Yihan Wang, Yiwei Lu, Guojun Zhang, Franziska Boenisch, Adam Dziedzic, Yaoliang Yu, and Xiao-Shan Gao. Machine unlearning for contrastive learning under auditing, 2025. 1, 2, 3, 5, 8

[37] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), 2023. 1, 2

[38] Tianyu Yang, Lisen Dai, Zheyuan Liu, Xiangqi Wang, Meng Jiang, Yapeng Tian, and Xiangliang Zhang. Cliperase: Efficient unlearning of visual-textual associations in CLIP. *CoRR*, abs/2410.23330, 2024. 1, 2, 5, 8

Machine Unlearning in Hyperbolic vs. Euclidean Multimodal Contrastive Learning: Adapting Alignment Calibration to MERU

Supplementary Material

This supplementary material provides additional technical details, extended analyses, and supporting evidence for our main paper on machine unlearning in hyperbolic versus Euclidean contrastive learning spaces. We first present formal mathematical descriptions of the CLIP and MERU objectives to establish the geometric foundations that differentiate these approaches (Section 10). Next, we clarify the composition of our forget sets, highlighting the challenges in precisely defining concept boundaries (Section 11). We then provide comprehensive visualizations through confusion matrices that illustrate the different unlearning behaviors between models (Section 12). Additionally, we present complementary linear probing results that further confirm how feature representations are still linearly separable, which already can be observed in latent visualizations (Section 13). Finally, we include extended visualizations of the latent spaces using multiple dimensionality reduction techniques to further support our findings on hyperbolic unlearning (Section 14). These materials provide deeper technical understanding and additional empirical support for the conclusions presented in our main paper.

10. Model Objectives

10.1. CLIP: Contrastive Learning in Euclidean Space

CLIP [30] consists of two encoders, a visual encoder f_{img} and text encoder f_{txt}, mapping images and text into a shared Euclidean space $\mathbb{R}^d$. Given a batch of images and texts $\{(x_i, t_i)\}_{i=1}^N$, we obtain embeddings $x'_i := f_{\text{img}}(x_i)$ and $t'_i := (f_{\text{txt}} t_i)$. CLIP is trained extending 1 to a symmetric cross-entropy loss:

$$L_{\text{CLIP}} = -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\text{sim}(x'_i, t'_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x'_i, t'_j)/\tau)}\right. \tag{18}$$

$$\left. + \log\frac{\exp(\text{sim}(x'_i, t'_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(x'_j, t'_i)/\tau)}\right]. \tag{19}$$

where $\text{sim}(x'_i, t'_j) := \cos(\theta_{ij})$ is the cosine similarity between normalized image and text embeddings, θ_{ij} is the angle between them, and τ is a temperature parameter. This contrastive objective places all embeddings on a unit hypersphere, treating all concept relationships uniformly, regardless of their hierarchical nature.

10.2. MERU: Contrastive Learning in Hyperbolic Space

MERU [9] extends contrastive learning to hyperbolic space using Lorentz model. MERU consists of visual and textual encoders, but it projects the image and text embeddings onto a hyperboloid manifold. The distance between two points x, y in the hyperboloid is given by

$$d_{\mathcal{L}}(x, y) = \frac{1}{\sqrt{c}}\cosh^{-1}(-c\langle x, y\rangle_{\mathcal{L}}), \tag{20}$$

where

$$\langle x, y\rangle_{\mathcal{L}} = \langle x_{\text{space}}, y_{\text{space}}\rangle - x_{\text{time}} \cdot y_{\text{time}} \tag{21}$$

is the Lorentzian inner product, $x = (x_{\text{space}}, x_{\text{time}}) \in \mathbb{R}^{n+1}$, $x_{\text{space}} \in \mathbb{R}^n$ and $x_{\text{time}} \in \mathbb{R}$, and $c > 0$ is the curvature of the space. The Lorentzian norm is defined by $||x||_{\mathcal{L}} = \sqrt{|\langle x, x\rangle_{\mathcal{L}}|}$. With this, the Lorentz model of curvature $-c$, $c > 0$, and dimension n is given by the set of vectors:

$$\mathcal{L}^n := \{x \in \mathbb{R}^{n+1} \ : \ \langle x, x\rangle_{\mathcal{L}} = -1/c\}. \tag{22}$$

MERU is trained with a contrastive loss similar to CLIP, but using negative hyperbolic distance as the similarity measure, $\text{sim}_{\mathcal{L}}(x, y) = -d_{\mathcal{L}}(x, y)$. Additionally, MERU incorporates an entailment loss to enforce partial order relationships between text and image embeddings:

$$L_{\text{entail}}(x, t) = \max(0, \text{ext}(x, t) - \text{aper}(t)) \tag{23}$$

where $\text{ext}(x, t)$ is the exterior angle between the text embedding t, given by

$$\text{ext}(x, t) = \cos^{-1}\left(\frac{x_{\text{time}} + t_{\text{time}}\ c\ \langle x, t\rangle_{\mathcal{L}}}{\|t_{\text{space}}\|\sqrt{(c\ \langle x, t\rangle_{\mathcal{L}})^2 - 1}}\right), \tag{24}$$

and image embedding x, and $\text{aper}(t)$ is the half-aperture of the entailment cone for t,

$$\text{aper}(t) = \sin^{-1}\left(\frac{2K}{\sqrt{c}\|t_{\text{space}}\|}\right). \tag{25}$$

The hyperbolic geometry of MERU naturally accommodates hierarchical relationships, as the volume of the space grows exponentially with distance from the origin. This property allows generic concepts to be placed closer to the origin with more capacity to connect to numerous specific instances, in contrast to the uniform treatment of relationships in Euclidean space.

11. Defining the forget set

We built the different forget set aggregating related subfolders (e.g., *dog* = {*bordercollie, bostonterrier, etc.*}, see Table 6). However, other subfolders in the retain set, such as *alltheanimals*, may contain image-text pairs related to dogs, creating conflicting signals during unlearning. This highlights a broader challenge in the machine unlearning field: properly defining concept boundaries for removal remains an open problem [7].

12. Confusion Matrices from Zero-Shot Classification

Figures 3 and 4 present confusion matrices for zero-shot classification before and after unlearning for CLIP and MERU, respectively. These visualizations provide detailed insights into how concept removal affects classification behavior across different classes.

For CLIP (Figure 3), we observe partial concept removal, with the "dog" classification accuracy reduced but not eliminated. Most misclassified dog images are assigned to the "cat" category, indicating that CLIP maintains some understanding of semantic similarity even when attempting to forget. The retain classes show minimal disturbance, maintaining strong diagonal elements in the confusion matrix.

In contrast, MERU (Figure 4) exhibits complete concept removal, with dog images almost entirely reassigned to other categories. The redistribution follows semantic hierarchies, with most dog images classified as "cat" which is a semantically animal category. This pattern supports our hypothesis that hyperbolic geometry leverages hierarchical relationships during unlearning, reassigning forgotten concepts according to their position in the semantic taxonomy. We further observe a decrease in performance on retaining the "horse" concept. However, this can be explained by observing how the original MERU already confuses horses by dogs, and then HAC, treating horses as if they were dogs, also removes them.

13. Linear Probing

Linear probing extracts embeddings from image encoder and trains a linear classifier on these features. This evaluates whether class information remains linearly separable in the latent space after unlearning. Accuracy is reported for both retained and forgotten classes, quantifying how successfully target concepts have been removed while preserving desired knowledge. Testing linear separability of image features provides different insights. While R-acc and F-acc in zero-shot classification measure the unlearning performance in the alignment between images and texts embeddings. Here we measure whether after the damage is done the image features have been mixed between different categories or not.

The linear probe classification results provide complementary insights into how unlearning affects the underlying feature representations. Consistent with prior work on hyperbolic classification [14], Euclidean representations show slightly better linear separability. However, the high forget accuracies in both methods reveal an important distinction between our approach and traditional class-unlearning: alignment calibration specifically targets cross-modal associations rather than altering the fundamental feature structure of either modality in isolation. This explains why images from the class related to the concept to forget remain linearly separable—their visual features are preserved while their association with corresponding text is disrupted. This insight can be also illustrated in a qualitative analysis of the latent spaces Sec. 6.

14. Additional Latent Space Visualizations

To complement visualizations from Section 6 we include more instances in visualizations. Additionally, for the hyperbolic case, we include visualizations from another perspective, using CO-SNE [13]. This method leverages hyperbolic Cauchy distribution (instead of hyperbolic student's t-distribution) and Lorentz distance, to represent global hierarchy and local distances in the same visualization. This allow us to "zoom-in" and see the origin of the hyperboloid from a closer point of view. Figure 5, illustrates the latent space of MERU from three perspectives before and after unlearning. CO-SNE allow us to better see that text embeddings of "dogs" remain closer to the origin, while other instances are pushed further, as discussed in Section 6. Figure 6 illustrates the same idea when scaling the unlearning problem to multiple concept removal. Observe that when "cats" are included in the forget set, the text embeddings for cats also remain close to the origin, and this is not disturb when including "food" and "plants", illustrating the robustness of HAC at scaling the unlearning task.

Concept-class	Subreddits	Image-text samples	% on Redcaps	% on Redcaps2
dogs	dogpictures, bordercollie, bostonterrier, lookatmydog, doggos, bulldogs, australiancattledog, frenchbulldogs, bernesemountaindogs, australianshepherd, beagle, chihuahua, corgi, dobermanpinscher, husky, labrador, pitbulls, pomeranians, pug, pugs, rarepuppers, rottweiler	511585	4.26%	7.33%
cats	cats, blackcats, supermodelcats, catpictures, siamesecats, bengalcats, siberiancats	532640	4.43%	7.63%
food	food, foodporn, veganfoodporn, healthyfood, breakfastfood, chinesefood, tastyfood, budgetfood, baking, bento, breadit, breakfastfood, breakfast, burgers, chefit, pizza, sushi, tacos, veganrecipes, vegetarian	630971	5.25%	9.04%
plants	houseplants, plants, plantedtank, airplants, plantbaseddiet, plantsandpots, carnivorousplants, flowers, bonsai, botanicalporn, cactus, microgreens, monstera, orchids, permaculture, roses, succulents, vegetablegardening, gardening	587798	4.89%	8.42%
Total	68 subreddits	2262994	18.85%	32.43%

Table 6. Grouping of subreddits to higher-order concepts.

Table 7. Linear probing accuracy in retain set (R-acc) and forget set (F-acc), across different tasks, after unlearning: (A) "dog"; (B) "dog" and "cat"; (C) "dog", "cat", "food" and "plant". We report results for both CLIP and MERU after alignment calibration using the optimal configuration from Section 4.3. Values in **bold** indicate whether AC or HAC performed better at retaining or unlearning across A, B and C.

Task	Method	Unlearn Set	CIFAR-10		CIFAR-100		STL-10		O-IIIT Pets		Food101		Flowers102	
			R-acc	F-acc	R-acc	F-acc	R-acc	F-acc	R-acc	F-acc	R-acc	F-acc	R-acc	F-acc
Linear Probe Classification	AC	A	**89.9**	**85.2**	**71.5**	-	**95.1**	**92.4**	**86.1**	87.5	**84.5**	-	**95.4**	-
		B	**95.8**	**91.3**	**71.6**	-	**95.8**	**91.3**	-	87.3	**84.6**	-	**95.7**	-
		C	**95.9**	**91.4**	**71.0**	84.3	**95.9**	**91.4**	-	87.0	-	84.3	-	95.4
	HAC-reg	A	89.3	85.5	69.8	-	94.9	93.6	84.8	**86.7**	83.8	-	93.8	-
		B	95.5	92.1	69.7	-	95.5	92.1	-	**85.0**	83.9	-	93.7	-
		C	95.4	92.4	68.6	**83.1**	95.4	92.4	-	**85.6**	-	**83.0**	-	**92.6**

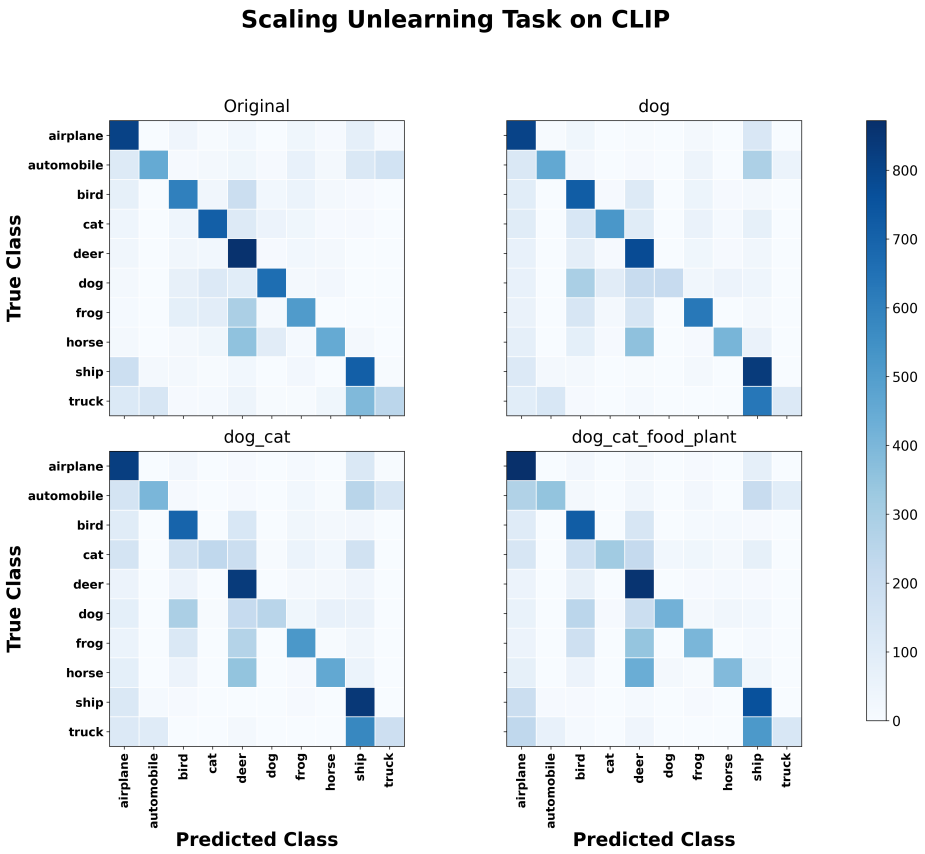


Figure 3. Confusion matrices for CLIP zero-shot classification at different scales of the unlearning task. After unlearning, CLIP shows moderate confusion, with dog images primarily misclassified as cats, but still retaining some dog classification capability.

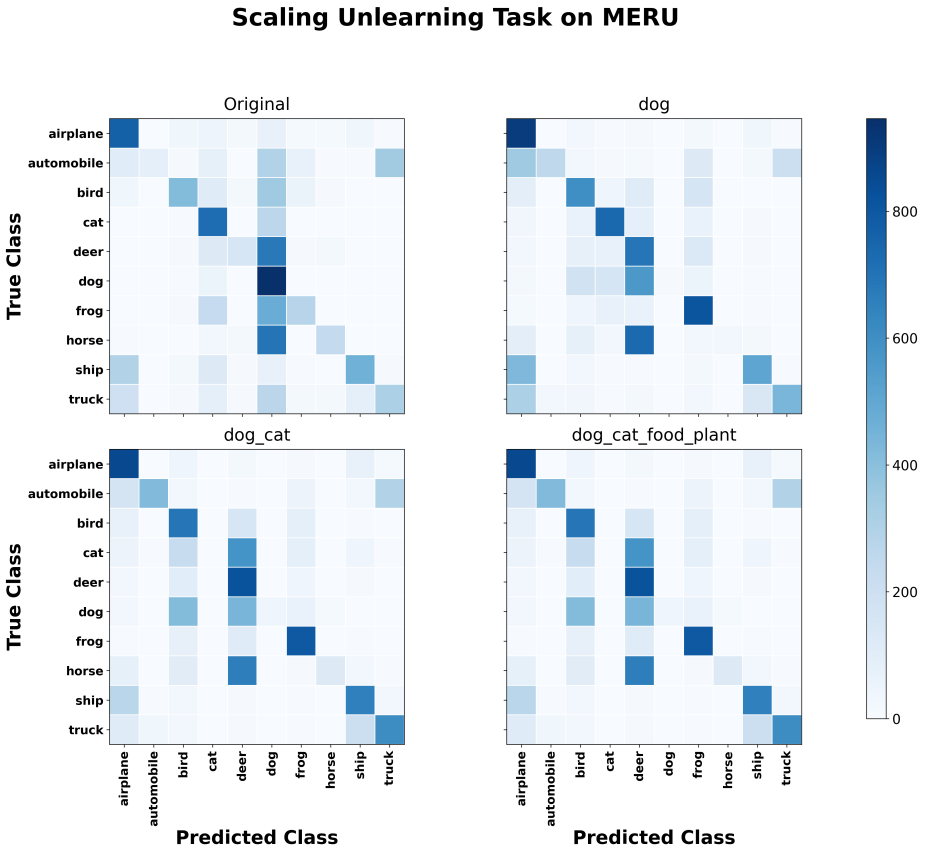
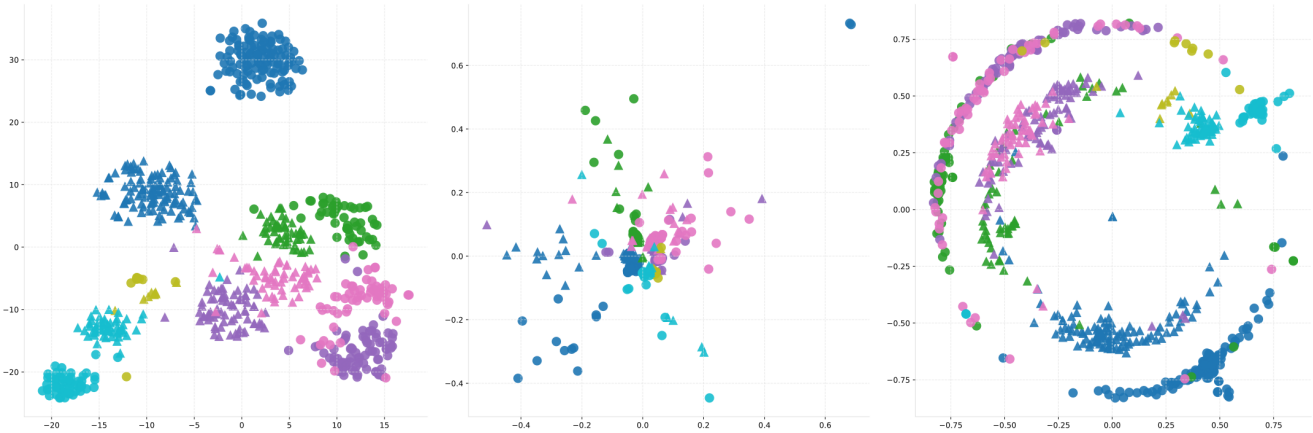


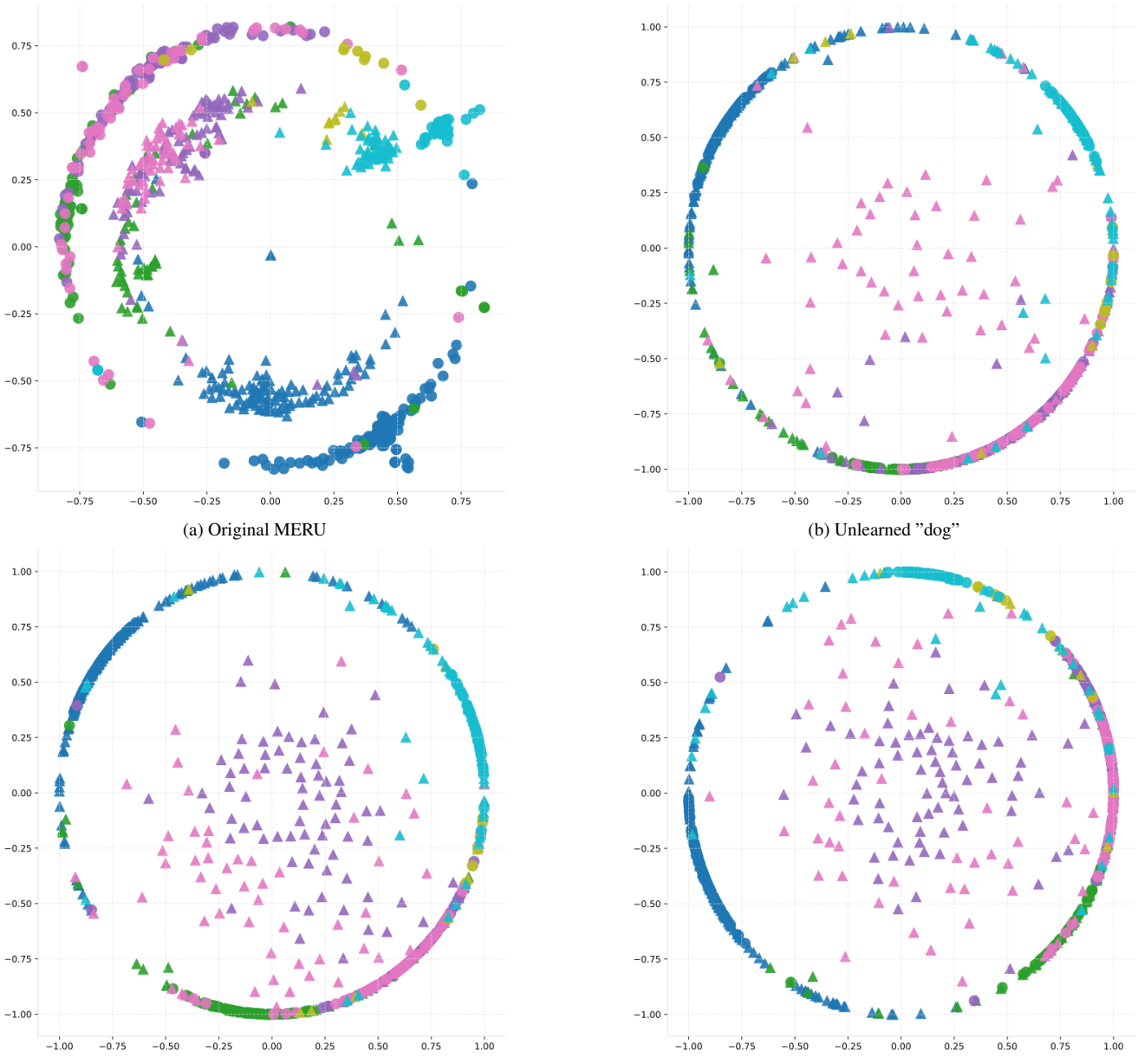
Figure 4. Confusion matrices for MERU zero-shot classification at different scales of the unlearning task. After unlearning, MERU demonstrates complete forgetting of the dog class, with dog images redistributed primarily to cat and horse categories according to semantic similarity.



(a) Original MERU: T-SNE (b) Original MERU: HT-SNE (c) Original MERU: CO-SNE

(d) Unlearned MERU: T-SNE (e) Unlearned MERU: HT-SNE (f) Unlearned MERU: CO-SNE

Figure 5. Latent space visualizations with T-SNE, hyperbolic T-SNE and CO-SNE of MERU before and after removing the concept-class "dog". △ refer to text embeddings, ○ to image embeddings, and colors to **dogs**, **cats**, **pizzas**, **buses**, **birds**, and **apples**.



(a) Original MERU

(b) Unlearned ”dog”

(c) Unlearned ”dog”,”cat”

(d) Unlearned ”dog”, ”cat”, ”food”, ”plant”

Figure 6. Latent space visualizations with CO-SNE of MERU at different unlearning tasks. △ refer to text embeddings, ○ to image embeddings, and colors to **dogs**, **cats**, **pizzas**, **buses**, **birds**, and **apples**.