Hyperbolic Contrastive Unlearning

Àlex Pujol Vidal alexpv@create.aau.dk
Visual Analysis and Perception Lab, Aalborg University & Pioneer Centre for AI, DK
Kamal Nasrollahi kna@milestone.dk
Milestone Systems SA & Aalborg University, DK
Sergio Escalera sescalera@ub.edu
University of Barcelona and the Computer Vision Center, ES
Thomas B. Moeslund tbm@create.aau.dk
Visual Analysis and Perception Lab, Aalborg University & Pioneer Centre for AI, DK

Abstract

Machine unlearning has become crucial for removing harmful concepts from large multimodal models, particularly for safety and copyright compliance. While recent work explores unlearning in Euclidean vision-language models, hyperbolic spaces remain unexplored despite their natural ability to capture semantic hierarchies. We present Hyperbolic Alignment Calibration (HAC), the first method for concept removal in hyperbolic contrastive learning models like MERU. Through systematic experiments, we demonstrate that HAC achieves better forget accuracy compared to Euclidean methods, particularly when removing multiple concepts simultaneously. Our approach introduces entailment calibration and norm regularization leveraging hyperbolic geometry's unique properties. Visualization analysis reveals that hyperbolic unlearning reorganizes semantic hierarchies, while Euclidean approaches merely disconnect cross-modal associations. These findings establish geometric unlearning as critical for safer deployment of machine learning models.

Keywords: HyperbolicLearning, ContrastiveLearning, MachineUnlearning

1 Introduction

Contrastive learning has emerged as a foundational approach for multimodal AI, enabling models like CLIP Radford et al. [2021] to align visual and textual representations in shared embedding spaces. Recent advances extend this paradigm to hyperbolic geometry through MERU Desai et al. [2023], which better captures hierarchical relationships in visual-semantic data by embedding in curved space with negative curvature.

However, these models trained on massive internet datasets often contain problematic content including private, copyrighted, and harmful data. Machine unlearning, the selective removal of specific information from trained models, has emerged as a critical solution for regulatory compliance and safe AI deployment Xu et al. [2023]. In scientific domains, where vision-language models increasingly support biological taxonomy classification, medical imaging, and materials discovery, the ability to selectively remove outdated concepts or proprietary data while preserving hierarchical knowledge structures is essential for responsible data stewardship and model reuse.

AI in Science (AIS), 2025, Copenhagen, Denmark

While recent work explores concept removal in Euclidean contrastive models Poppi et al. [2025], Yang et al. [2024], Wang et al. [2025], hyperbolic geometry presents unique challenges and opportunities that remain unexplored. The geometric distinction raises a fundamental question: *How does underlying geometry affect concept removal efficacy in contrastive learning models?*

Our study Vidal et al. [2025] bridges geometric representation learning and machine unlearning by investigating concept removal across different geometries. This is the first work to address machine unlearning in hyperbolic vision-language models. The contributions are:

1. Hyperbolic Alignment Calibration (HAC), the first unlearning method for hyperbolic contrastive models. Our key technical innovations are: (i) entailment losses that leverage hyperbolic geometry to preserve/disrupt taxonomic relationships, and (ii) hyperbolic norm regularization for enhanced concept removal.
2. Systematic evaluation revealing HAC achieves superior concept removal even when scaling to multiple concepts.
3. Visualization analysis showing hyperbolic unlearning reorganizes semantic hierarchies rather than merely disconnecting associations.

2 Method

2.1 Problem Formulation

Given a pre-trained contrastive model with encoders f_{img} and f_{txt}, dataset $\mathcal{D}$ with image-text pairs, and concept c to remove, we define forget set $\mathcal{D}_f$ containing instances of c and retain set $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$. The unlearning objective seeks modified encoders $f^*_{\text{img}}, f^*_{\text{txt}}$ that minimize similarity between visual and textual representations of c while preserving performance on $\mathcal{D}_r$.

2.2 Hyperbolic Alignment Calibration

Building on Alignment Calibration (AC) for Euclidean spaces Wang et al. [2025], HAC adapts the AC loss components to hyperbolic geometry using the Lorentz model of hyperbolic space. The retain loss L_{retain} preserves alignment between retain-set image-text pairs, while the forget loss L_{forget} disrupts alignment for forget-set pairs. We replace cosine similarity with negative hyperbolic distance and introduce two hyperbolic-specific innovations: entailment constraints and norm regularization.

Entailment Losses: In hyperbolic space, hierarchical relationships are encoded through entailment cones. HAC incorporates entailment terms to preserve hierarchies for retained concepts while disrupting them for forgotten ones:

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

$$L_{\text{f-ent}} = \frac{1}{N}\sum_{i=1}^{N} \max(0, \text{aper}(t_i^{'f}) - \text{ext}(x_i^{'f}, t_i^{'f})) \tag{2}$$

where $\text{ext}(x, t)$ measures the exterior angle between image embedding x and text embedding t, and $\text{aper}(t)$ is the half-aperture of the entailment cone for text t. The retain entailment loss $L_{\text{r-ent}}$ encourages images to stay within their text's entailment cone, while forget entailment loss $L_{\text{f-ent}}$ pushes forgotten images outside their cone.

Norm Regularization: In hyperbolic space, the Lorentz norm $||\cdot||_{\mathcal{L}}$ measures distance from the origin. Points farther from the origin become increasingly separated due to hyperbolic expansion, which correspond to the leaves of the hierarchy and contain the most specific semantic concepts. Points closer to the origin are higher in the semantic hierarchy, that is, they represent more general concepts. To leverage this property for unlearning, we penalize the norm of forgotten embeddings, squeezing them toward the origin where concepts are vaguer:

$$L_{\text{norm-reg}} = \frac{1}{N} \sum_{i=1}^{N} (||x_i^{'f}||_{\mathcal{L}} + ||t_i^{'f}||_{\mathcal{L}}) \tag{3}$$

This regularization ensures numerical stability while enhancing concept removal.

The complete HAC loss combines retention, forgetting, entailment, and regularization terms:

$$L_{\text{HAC}} = L_{\text{retain}} + \varepsilon L_{\text{forget}} + \omega_r L_{\text{r-ent}} + \omega_f L_{\text{f-ent}} + \lambda L_{\text{norm-reg}} \tag{4}$$

where ε balances forgetting strength, ω_r and ω_f control hierarchical preservation and disruption, and λ regulates hyperbolic norm penalties.

3 Experiments

3.1 Experimental Setup

We compare CLIP and MERU models (both ViT-S backbone) pre-trained on RedCaps dataset. Experiments focus on removing concepts: dogs, cats, food, and plants across three scenarios: (A) remove "dog" only; (B) remove "dog" + "cat"; (C) remove "dog" + "cat" + "food" + "plants". Evaluation uses zero-shot classification on CIFAR-10, Oxford-IIIT Pets, Food-101, and other datasets, measuring retain accuracy (R-acc) and forget accuracy (F-acc).

3.2 Scaling Analysis Results

Table 1 presents our key findings across scaling scenarios. HAC demonstrates superior forgetting capabilities, maintaining low forget accuracies even when removing multiple diverse concepts simultaneously.

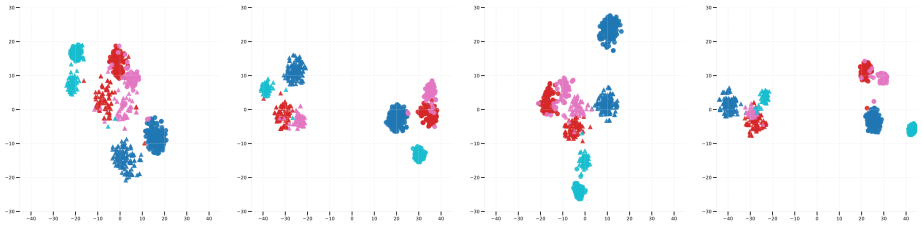
Key observations: (1) AC maintains higher retain accuracy but suffers dramatic increases in forget accuracy as concepts scale, indicating incomplete removal; (2) HAC achieves remarkable forgetting performance across all scaling scenarios; (3) HAC's superior scaling capabilities suggest hyperbolic geometry naturally handles multiple concept removal through hierarchical structure manipulation.

Table 1: Zero-shot classification accuracy comparing AC and HAC across scaling experiments. Bold indicates better performance for each geometry. - indicates that either D_r or D_f is empty, making the metric undefined.

Method	Unlearn Set	CIFAR-10		CIFAR-100		STL-10		O-IIIT Pets		Food-101	
		R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓	R-acc↑	F-acc↓
AC	A	**58.7**	21.2	**27.9**	-	**88.1**	83.1	**74.9**	31.5	**72.4**	-
	B	**90.3**	71.4	**26.6**	-	**90.3**	71.4	-	53.4	**72.5**	-
	C	**90.0**	77.0	**23.4**	57.2	**90.0**	77.0	-	64.0	-	0.16
HAC	A	54.0	**0.0**	20.6	-	84.3	**38.0**	66.3	**10.8**	67.6	-
	B	83.5	**2.1**	21.8	-	83.5	**2.1**	-	**25.7**	59.6	-
	C	82.7	**22.1**	18.8	**21.6**	82.7	**22.1**	-	**28.7**	-	**0.08**

3.3 Visualization Analysis

Figure 1 provides compelling evidence of different unlearning mechanisms. T-SNE visualizations reveal that while AC preserves class structure but merely disconnects cross-modal associations, HAC fundamentally reorganizes the semantic hierarchy. In hyperbolic space, dog text embeddings migrate towards cat text embeddings, indicating a hierarchical repositioning rather than simple disassociation.

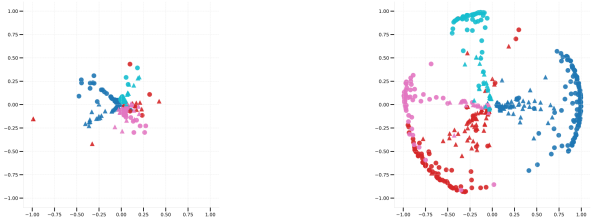


(a) Original CLIP (b) Unlearned CLIP (c) Original MERU (d) Unlearned MERU

Figure 1: T-SNE visualizations showing different unlearning mechanisms. △ = text, ○ = images. Colors: **dogs**, **cats**, **pizzas**, **buses**.

Figure 2 provides additional insights through hyperbolic T-SNE visualizations of MERU's embedding space. Following unlearning, we observe dramatic expansion with image representations pushed farther from the origin while text embeddings remain relatively close. This illustrates how HAC exploits hyperbolic geometry's exponential expansion property.

The hyperbolic T-SNE reveals dog text embeddings positioned near the origin after unlearning, effectively placing them "behind" other concepts in the semantic hierarchy. This spatial reorganization explains HAC's superior performance, as dog images now align more strongly with other text categories because their original concept has been repositioned within the taxonomic structure.



(a) Original MERU (b) Unlearned MERU

Figure 2: Hyperbolic T-SNE visualizations of MERU before and after removing "dog". △ = text, ○ = images. Colors: **dogs**, **cats**, **pizzas**, **buses**.

4 Discussion & Conclusion

Our work demonstrates that geometric properties fundamentally influence concept removal effectiveness in contrastive models. HAC achieves superior concept removal by leveraging hyperbolic geometry's natural capacity for hierarchical manipulation. The trade-off between retention and forgetting reflects each geometry's strengths: Euclidean methods better preserve retained knowledge, while hyperbolic methods achieve more complete concept removal.

These findings have direct implications for scientific domains where vision-language models are increasingly deployed for biological taxonomy, medical imaging, and materials discovery. The ability to reorganize semantic hierarchies rather than merely disconnect associations is crucial for scientific applications that require removing retracted findings or proprietary data while preserving taxonomic relationships and domain knowledge. Near-perfect concept removal establishes HAC as essential for safe deployment and responsible data stewardship in scientific AI. While our accuracy-based evaluation demonstrates effective concept removal, future work should develop more rigorous verification including geometry-aware membership inference attacks and canary-based methods to certify true unlearning, establish formal verification for hierarchical completeness, and address numerical stability challenges.

5 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. Additionally, we thank to the Pioneer Centre for AI, Denmark, (DNRF grant P1).

References

K. Desai, M. Nickel, T. Rajpurohit, J. Johnson, and R. Vedantam. Hyperbolic image-text representations. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, and R. Cucchiara. Safe-clip: Removing nsfw concepts from vision-and-language models. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, editors, *Computer Vision – ECCV 2024*, pages 340–356, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73668-1.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

A. P. Vidal, S. Escalera, K. Nasrollahi, and T. B. Moeslund. Machine unlearning in hyperbolic vs. euclidean multimodal contrastive learning: Adapting alignment calibration to meru. In *CVPR Workshops*, 2025. URL https://api.semanticscholar.org/CorpusID:277113146.

Y. Wang, Y. Lu, G. Zhang, F. Boenisch, A. Dziedzic, Y. Yu, and X.-S. Gao. Machine unlearning for contrastive learning under auditing, 2025. URL https://openreview.net/forum?id=k2HZ4Mu2Pb.

H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu. Machine unlearning: A survey. *ACM Comput. Surv.*, 56(1), Aug. 2023. ISSN 0360-0300. doi: 10.1145/3603620. URL https://doi.org/10.1145/3603620.

T. Yang, L. Dai, Z. Liu, X. Wang, M. Jiang, Y. Tian, and X. Zhang. Cliperase: Efficient unlearning of visual-textual associations in CLIP. *CoRR*, abs/2410.23330, 2024. doi: 10.48550/ARXIV.2410.23330. URL https://doi.org/10.48550/arXiv.2410.23330.