Hyperbolic Safety-Aware Vision-Language Models

Tobia Poppi*[1,2] Tejaswi Kasarla*[3] Pascal Mettes[3] Lorenzo Baraldi[1] Rita Cucchiara[1,4]
[1]University of Modena and Reggio Emilia, Italy [2] University of Pisa, Italy
[3] University of Amsterdam, Netherlands [4]IIT-CNR, Italy
[1]{name.surname}@unimore.it [2] {name.surname}@phd.unipi.it [3] {initial.surname}@uva.nl

Abstract

Addressing the retrieval of unsafe content from vision-language models such as CLIP is an important step towards real-world integration. Current efforts have relied on unlearning techniques that try to erase the model's knowledge of unsafe concepts. While effective in reducing unwanted outputs, unlearning limits the model's capacity to discern between safe and unsafe content. In this work, we introduce a novel approach that shifts from unlearning to an awareness paradigm by leveraging the inherent hierarchical properties of the hyperbolic space. We propose to encode safe and unsafe content as an entailment hierarchy, where both are placed in different regions of hyperbolic space. Our HySAC, Hyperbolic Safety-Aware CLIP, employs entailment loss functions to model the hierarchical and asymmetrical relations between safe and unsafe image-text pairs. This modelling – ineffective in standard vision-language models due to their reliance on Euclidean embeddings – endows the model with awareness of unsafe content, enabling it to serve as both a multimodal unsafe classifier and a flexible content retriever, with the option to dynamically redirect unsafe queries toward safer alternatives or retain the original output. Extensive experiments show that our approach not only enhances safety recognition but also establishes a more adaptable and interpretable framework for content moderation in vision-language models. Our source code is available at: https://github.com/aimagelab/HySAC

Warning: *This paper features explicit sexual content and other material that some readers may find disturbing, distressing, or offensive.*

1. Introduction

Large-scale vision-language models (VLMs) have achieved remarkable successes in various applications, including cross-model retrieval [58], text-to-image and image-to-text generation [40, 60] and various downstream tasks [45, 67, 74]. Popular VLMs like CLIP [58] and ALIGN [34] leverage vast amounts of web-scraped image-text data to learn rich multimodal representations by aligning visual and textual modalities. However, most large-scale datasets sourced from the web contain unsafe or inappropriate content, such as violence, nudity, or hate speech [5, 6]. The presence of such content not only raises ethical concerns but also introduces risks for real-world applications [7, 30, 75], where exposure to or misuse of this material can lead to legal and societal repercussions. Birhane *et al*. [7] also show that the increasing dataset scale can exacerbate hateful and unsafe content, as identified in the now removed LAION-5B [66, 69]. Addressing the issue of unsafe content in VLMs is therefore of utmost importance to ensure responsible AI practices.

Recent efforts to mitigate unsafe content in vision-language models have led to the development of methods specifically designed for NSFW (Not Safe For Work) content removal. Most of these works [23, 55] have focused on unlearning (*i.e*., removing) the knowledge of unsafe content from the models. As a recent example, Poppi *et al*. [55] develop a fine-tuned version of CLIP which unlearns toxic concepts by redirecting their embeddings towards safe regions, so that retrieval always produces safe content even when the model is prompted with unsafe inputs.

In contrast, we propose an approach for managing unsafe content in VLMs: emphasizing *awareness* over *unlearning*. Rather than hiding the flaws of VLMs by ignoring NSFW content, we aim to equip the VLMs with the ability to distinguish between safe and unsafe content. This in turn helps users of the model to expose or redirect NSFW content when necessary, a crucial step toward improving user agency, understanding, and interpretability [19].

Inspired by recent hyperbolic vision-language models [17, 52], we introduce a hyperbolic framework that leverages the geometric properties of hyperbolic space to separate safe and unsafe content effectively. Using a paired dataset of safe and unsafe image-text inputs [55], we adjust the embeddings to create an entailment-based [24] structure. In this setup, safe concepts are positioned closer to the

*Equal contribution

origin of the hyperbolic space, while unsafe concepts are mapped further away. Specifically, we introduce a *hyperbolic safe-to-unsafe entailment* mechanism that ensures safe content encompasses unsafe representations within conical regions, defining clear safety boundaries and *safety traversals* to dynamically adjust query embeddings along the hyperbolic space to promote safe retrievals or, alternatively, expose relevant unsafe content when necessary. This framework not only organizes data into safe and unsafe radius-based regions but also enables controlled movement within the space, allowing retrievals to favor safety as required. Experiments demonstrate that HySAC achieves clear improvements in safety awareness, retrieval performance, and NSFW content handling across multiple datasets, with robustness in both safe content redirection and controlled unsafe content accessibility.

2. Related Work

Unlearning in vision-language models. Unlearning concepts and content has recently received a lot of attention, empowered by the success of vision-language models. Various approaches have been explored, such as full model retraining, fine-tuning, machine unlearning [9, 27, 28, 56, 57], and differential privacy [29]. Some of these efforts have focused on text-to-image models, with the goal of removing specific styles, concepts, or objects [36, 76].

A particular emphasis has been placed on removing NSFW content, which encompasses inappropriate, unsafe, or illegal material. Schramowsky *et al*. [64] steer the generation away from NSFW areas, defined by a fixed set of concepts. The embedding of NSFW concepts is applied as negative guidance during the text-conditioning phase, acting as safety guidance. Gandikota *et al*. [23] erase visual concepts using only their name, with negative guidance serving as a teacher. Poppi *et al*. [55] focus on removing NSFW concepts from CLIP-like models by fine-tuning the entire model using ViSU, a multimodal dataset containing Safe/NSFW Image/Texts quadruplets. Unlike earlier approaches, which target specific components, their method fine-tunes the entire vision-and-language model, making content removal applicable to downstream tasks.

We introduce a method for handling NSFW concepts in CLIP-like models by fine-tuning them in hyperbolic space. We exploit the hierarchical properties of hyperbolic geometry, yielding a clear advantage: the model becomes explicitly aware of whether content is safe or NSFW, rather than merely removing knowledge of unsafe content. Similarly to our approach, safe generation works [8, 64], underscore that a deeper understanding of (un-)safety can improve the control of harmful content generation.

NSFW concept detection. A related area of research is the automatic detection of NSFW content. Several methods have been proposed for detecting NSFW and toxic language [10, 31, 44], primarily in social media contexts. DistilBERT [62] has emerged as a promising model for this task, especially when fine-tuned for identifying adult content. Detecting inappropriate language presents a significant challenge, and this complexity extends to visual content, where various techniques have been developed to detect NSFW imagery [5, 22, 49]. In this domain, models like NudeNet [4] specialize in detecting nudity, while Q16 [63] serves as a broader classifier, capable of identifying a wider range of NSFW content. However, identifying inappropriate visual content remains a complex task, given the challenges posed by subtle visual cues, lack of contextual information, and limited data availability.

While these detection methods focus solely on identifying unsafe content, they do not address the problem of retrieving relevant, safe alternatives when an unsafe input is detected. Our method makes it possible to jointly detect NSFW content and provide a mechanism to shift NSFW queries towards safe but relevant alternatives.

Hyperbolic learning. A key advantage of hyperbolic space is its inherent ability to represent hierarchical or tree-like structures with minimal distortion [11, 50, 51, 61]. A comprehensive list of recent advancements is documented in surveys by Mettes *et al*. [46] and Peng *et al*. [53]. Foundational works for building neural networks in hyperbolic space [3, 24, 25, 35, 68] led to the use of hyperbolic models across multiple modalities such as images [2, 20, 21, 26, 42, 71], text [18, 38, 70, 77], graphs [12–14, 41] and recommender systems [48, 73].

Recent work has shown the strong potential of hyperbolic learning for vision-language models [17, 33, 52] and demonstrated that training with a loss function enforcing entailment cones [24] leads to the emergence of hierarchical structures between the embeddings. While Desai *et al*. [17] enforce entailment structure *across* modalities, Pal *et al*. [52] also explicitly enforce structure *within* modalities by leveraging object-level compositions and nouns from text. We take inspiration from these approaches and enforce entailment relations across safe and unsafe embeddings, creating an interpretable space for traversing from unsafe regions to safe regions in CLIP models.

3. Preliminaries

Throughout this work, we operate in hyperbolic space, a Riemannian manifold with a constant negative curvature. Following Desai *et al*. [17], we use the Lorentz model, as it is better equipped to deal with numerical instabilities associated with the Poincaré distance metric [39, 51].

The Lorentz model $\mathbb{L}^n$ is an n-dimensional manifold in which points are represented on the upper sheet of a two-sheeted hyperboloid in $(n+1)$-dimensional Minkowski spacetime. Following terminology from general relativity,

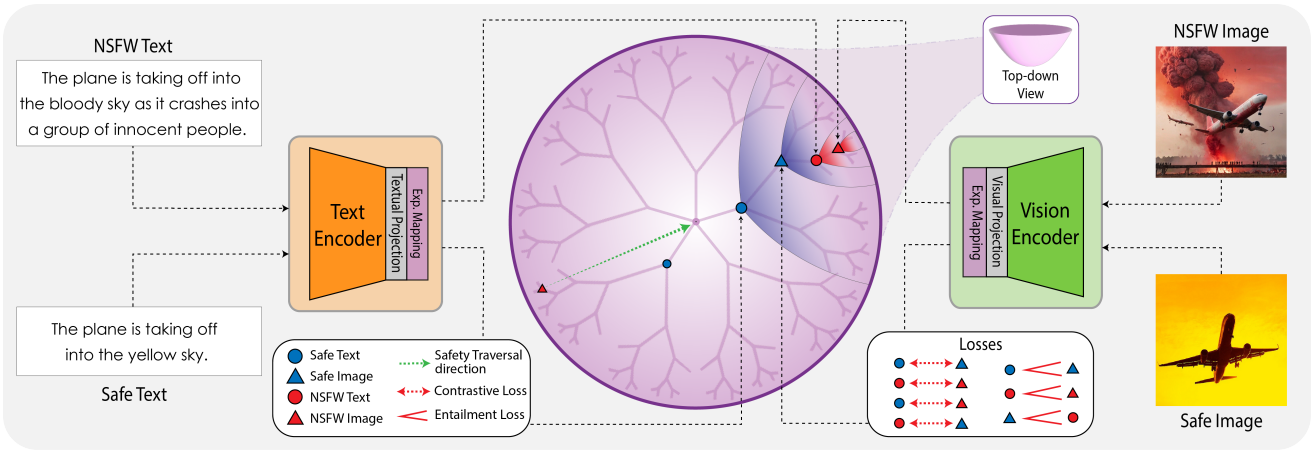

Figure 1. **Overview of our approach.** HySAC builds a hyperbolic embedding that manages content safety through an entailment hierarchy. Unsafe text and images are projected to dedicated regions of hyperbolic space, allowing for safety-aware retrieval and classification.

for each vector $\mathbf{p} \in \mathbb{R}^{n+1}$, we refer to the first dimension, the axis of symmetry, as the time-axis, denoted by p_0, and the remaining n-dimensions as the spatial components, denoted $\tilde{\boldsymbol{p}}$. The Lorentz model, $\mathbb{L}^n = (\mathcal{L}^n, \mathfrak{g}^\kappa_{\boldsymbol{p}})$ is given as

$$\mathcal{L}^n := \left\{ \boldsymbol{p} \in \mathbb{R}^{n+1} : \langle \mathbf{p}, \mathbf{p} \rangle_{\mathcal{L}} = -\tfrac{1}{\kappa}, p_0 = \sqrt{1/\kappa + \|\tilde{\boldsymbol{p}}\|^2}, \kappa > 0 \right\}, \tag{1}$$

where $-\kappa \in \mathbb{R}$ denotes the curvature with the Riemannian metric $\mathfrak{g}^\kappa_{\boldsymbol{p}} = \mathrm{diag}(-1, 1, \ldots, 1)$. The Lorentzian inner product $\langle ., . \rangle_{\mathcal{L}}$ is induced by the metric tensor $\mathfrak{g}^\kappa_{\boldsymbol{p}}$ and is defined for $\mathbf{p}, \mathbf{q} \in \mathbb{L}^n$ as

$$\langle \mathbf{p}, \mathbf{q} \rangle_{\mathcal{L}} = -p_0 q_0 + \langle \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}} \rangle, \tag{2}$$

where $\langle ., . \rangle$ is the Euclidean inner product. The Lorentzian inner product induces a norm on the Lorentzian space which can be written as $\|\mathbf{p}\|_{\mathcal{L}} = \sqrt{\langle \mathbf{p}, \mathbf{p} \rangle_{\mathcal{L}}}$. We now define the common hyperbolic operations in Lorentz space.

Definition 3.1 (Lorentzian distance)**.** The Lorentzian distance between two points in $\mathbb{L}^n$ is the length of their shortest path (*geodesic*) connecting them, computed as

$$d_{\mathcal{L}}(\mathbf{p}, \mathbf{q}) = \sqrt{1/\kappa} \cdot \cosh^{-1}(-\kappa \langle \mathbf{p}, \mathbf{q} \rangle_{\mathcal{L}}). \tag{3}$$

In our work, we use the negative of Lorentzian distance to calculate similarities between multimodal inputs [17, 52].

Definition 3.2 (Exponential map)**.** Since Lorentz space is a Riemannian manifold, it is locally Euclidean. This is best described through the tangent space $T_{\mathbf{p}}\mathbb{L}^n$, a first-order approximation of the Lorentzian manifold at a given point $\mathbf{p} \in \mathbb{L}^n$. The *exponential map* then provides a means to project elements from the tangent space onto the hyperboloid. Given a point $\mathbf{v} \in T_{\mathbf{p}}\mathbb{L}^n$, the exponential map is defined as $\exp^\kappa_{\mathbf{p}} : T_{\mathbf{p}}\mathbb{L}^n \to \mathbb{L}^n$ with the expression

$$\exp^\kappa_{\mathbf{p}}(\mathbf{v}) = \cosh(\sqrt{\kappa}\,\|\mathbf{v}\|_{\mathbb{L}})\mathbf{p} + \frac{\sinh(\sqrt{\kappa}\,\|\mathbf{v}\|_{\mathbb{L}})}{\sqrt{\kappa}\,\|\mathbf{v}\|_{\mathbb{L}}}\mathbf{v}. \tag{4}$$

In practice, the reference point $\mathbf{p}$ is set to the origin $\mathbf{0} = (\sqrt{1/\kappa}, 0, ..., 0)^T$ on the hyperboloid, allowing $\exp^\kappa_{\mathbf{0}}$ to project Euclidean vectors from the tangent space at $\mathbf{0}$ directly onto the hyperboloid [17, 35]. In this work, the exponential map is used to project the outputs of the visual and textual encoders to a shared hyperbolic space.

4. HySAC: Hyperbolic Safety-Aware CLIP

4.1. Problem formulation and objective

Problem setup. Given a dataset $D = \{(I_i, T_i)\}_{i=1}^N$ of N image-text pairs, vision-language models (*e.g.* CLIP [58]) align the visual and textual embeddings obtained from image and text encoders in a shared embedding space. Large-scale datasets employed for training such embedding spaces are often web-scraped and contain unsafe samples [5]. For our problem setup, in order to differentiate between safe and unsafe content, we denote safe image-text pairs in D as (I_i, T_i) and unsafe image-text pairs as $(I_i^\star, T_i^\star)$.

To make vision-language models aware of inappropriate contents, and enable them to avoid or redirect the representation of such content, we also require a dataset of quadruplets of safe and unsafe image-text pairs, denoted as $D^\star = \{(I_k, T_k, I_k^\star, T_k^\star)\}_{k=1}^K$ [55]. This dataset is generated following the definition of NSFW content of Schramowski *et al.* [64] containing the following twenty categories: *hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, abuse, brutality,* and *cruelty*. The dataset is constructed such that the unsafe image-text pairs $(I_k^\star, T_k^\star)$ are specific cases or modified versions of the safe representations (I_k, T_k).

Modelling relations in the hyperbolic space. We first consider the relationship between text and image modalities in the embedding space. Like other hyperbolic vision-

language models [17, 52], we consider text as a general version of images to reflect the natural structure of partial order embeddings [72]. This is enforced in the embedding space by placing text embeddings closer to the origin and image embeddings farther away, defining a modality entailment relationship formally expressed as

$$g_T(T_k) \ll g_I(I_k), \quad \text{and} \quad g_T(T_k^\star) \ll g_I(I_k^\star), \tag{5}$$

where g_I and g_T denote the projections of images and text in the hyperbolic model and $\ll$ indicates that one embedding is closer to the origin than another. Differently from other models, we also impose a relationship between unsafe and safe pairs, by considering unsafe image-text pairs as specific cases of their safe counterparts within $D^\star$. Hence, we establish a safety entailment to segregate safe from unsafe content, as follows

$$g_I(I_k) \ll g_T(T_k^\star). \tag{6}$$

By satisfying these inequalities, g_I and g_T capture both the modal and safety hierarchies within the data, thereby endowing the embedding space with safety-aware properties. To sum up, our objective is to model a hyperbolic vision-language model with the following inequality chain:

$$g_T(T_k) \ll g_I(I_k) \ll g_T(T_k^\star) \ll g_I(I_k^\star). \tag{7}$$

Below we outline our method to create safety-aware vision-language models in hyperbolic space.

4.2. Hyperbolic safety learning

To optimize Eq. 7 for safety-aware vision-language models, we propose Hyperbolic Safety-Aware CLIP to rearrange the embedding space to separate safe and unsafe regions. Our optimization consists of two components: (1) a hyperbolic safety contrastive component to align image-text pairs over a mini-batch and (2) a hyperbolic safety entailment to align safe and unsafe content. An overview is given in Figure 1.

Hyperbolic safety contrastive learning. CLIP [58] and Safe-CLIP [55] rely on contrastive objectives to align and distribute the multimodal data. We utilize hyperbolic embeddings to align the visual and textual data. Specifically, we project the visual and textual embeddings from a pre-trained vision-language model onto a hyperboloid [17] through an exponential map (Eq. 4). Let $f_I(.)$ and $f_T(.)$ represent any Euclidean encoders for image and text. Then, $g_I(I_k) = \exp_{\mathbf{0}}^{\kappa}(\alpha_{img} \cdot f_I(I_k))$ and $g_T(T_k) = \exp_{\mathbf{0}}^{\kappa}(\alpha_{txt} \cdot f_T(T_k))$ represent the hyperbolic representations of a safe image-text pair, (I_k, T_k). Similarly, $g_I(I_k^\star) = \exp_{\mathbf{0}}^{\kappa}(\alpha_{img} \cdot f_I(I_k^\star))$ and $g_T(T_k^\star) = \exp_{\mathbf{0}}^{\kappa}(\alpha_{txt} \cdot f_T(T_k^\star))$ represent the hyperbolic representations of an unsafe image-text pair $(I_k^\star, T_k^\star)$. α_{img} and α_{txt} are learnable projection scalars.

To align representations in hyperbolic space, the similarity for the image-text and text-image contrastive loss is based on negative Lorentzian distance (Eq. 3) between $g_I(.)$ and $g_T(.)$. We compute the hyperbolic safety contrastive loss over safe image-text pairs (I_k, T_k) in a batch B as

$$L_{cont}^{*}(I,T) = -\sum_{i \in B} \log \frac{\exp(d_{\mathcal{L}}(g_I(I_i), g_T(T_i))/\tau)}{\sum_{k=1, k\neq i}^{B} \exp(d_{\mathcal{L}}(g_I(I_i), g_T(T_k))/\tau)}, \tag{8}$$

where τ denotes a temperature hyperparameter. A similar contrastive loss is employed to preserve the multimodal structure between unsafe image-text pairs $(I_k^\star, T_k^\star)$. Two additional contrastive losses between cross-safety modalities ensure the alignment of quadruplets in the embedding space. The final contrastive loss is formulated as

$$\begin{aligned} L_{\text{hSC}}(I,T,I^\star,T^\star) = L_{cont}^{\star}(I,T) + L_{cont}^{\star}(I^\star,T^\star) \\ + L_{cont}^{\star}(I,T^\star) + L_{cont}^{\star}(I^\star,T). \end{aligned} \tag{9}$$

Hyperbolic safety entailment learning. Hyperbolic entailment cones, introduced by Ganea *et al*. [24], generalize partial ordered embeddings [72] to any Riemannian manifold. Entailment cones induce a partial order between concepts in a dataset $\mathcal{X}$ such that for any pair $(\mathbf{p}, \mathbf{q}) \in \mathcal{X}$, if $\mathbf{p}$ is a subconcept of $\mathbf{q}$, then $\mathbf{q}$ entails $\mathbf{p}$ within a conical region $\mathfrak{S}_{\mathbf{q}}$ defined by $\mathbf{q}$. For the Lorentz model, $\mathbb{L}^n$, the half-aperture of each conical region $\mathfrak{S}_{\mathbf{q}}$ is defined as [17, 38]

$$\omega(\mathbf{q}) = \sin^{-1}\left(\frac{2K}{\sqrt{\kappa}\,\|\tilde{\boldsymbol{q}}\|}\right), \tag{10}$$

where $-\kappa$ is the curvature of space, and the constant $K = 0.1$ limits values near the origin [24]. To preserve partial order between image-text relationships, we add entailment from safe text to safe image and from unsafe text to unsafe images, effectively implementing Eq. 5. Specifically, a safe image I_k must lie within the cone defined by its corresponding safe text, T_k, characterized by the half-aperture $\omega(T_k)$. Similarly, an unsafe image $I_k^\star$ must lie within the cone defined by its corresponding unsafe text $T_k^\star$. This is enforced through an entailment loss formulated for image-text representations by Le *et al*. [38] and Desai *et al*. [17], as

$$\begin{aligned} L_{ent}^{\star}(I,T) &= \max(0, \phi(I_k,T_k) - \eta\omega(T_k)) \text{ and} \\ L_{ent}^{\star}(I^\star,T^\star) &= \max(0, \phi(I_k^\star,T_k^\star) - \eta\omega(T_k^\star)), \end{aligned} \tag{11}$$

where ϕ is the exterior angle (between lines I_kT_k and $\mathbf{0}T_k$ or between lines $I_k^\star T_k^\star$ and $\mathbf{0}T_k^\star$) given by

$$\phi(I_k,T_k) = \cos^{-1}\left(\frac{I_{k0} + T_{k0}\kappa\langle I_k, T_k\rangle_{\mathcal{L}}}{\left\|\widetilde{T_k}\right\| \sqrt{(\kappa\langle I_k, T_k\rangle_{\mathcal{L}})^2 - 1}}\right). \tag{12}$$

Here, η is a threshold for the half-aperture, $\omega(T_k)$ [52]. Intuitively, the entailment loss $L_{ent}^{\star}$ penalizes images I_k

that lie outside the cone $\mathfrak{S}_{T_k}$ defined by their corresponding text caption T_k. Finally, to model the safety hierarchy, we enforce that safe concepts entail unsafe ones, reflecting that safe data are more general and unsafe data are more specific. Specifically, we enforce that a safe image I_k entails an unsafe text $T_k^\star$, meaning that the unsafe text $T_k^\star$ must lie within the cone defined by the safe image I_k, characterized by the half-aperture $\omega(I_k)$. This gives us the safety-entailment defined in Eq. 6. This is implemented as

$$L^\star_{ent}(T^\star, I) = \max(0, \phi(T_k^\star, I_k) - \eta\omega(I_k)). \tag{13}$$

The overall entailment loss to satisfy Eq. 7 is defined as

$$L_{\text{hSE}}(I, T, I^\star, T^\star) = L^\star_{ent}(I, T) + L^\star_{ent}(T^\star, I) + L^\star_{ent}(I^\star, T^\star). \tag{14}$$

Combined loss function. We integrate the contrastive with the entailment losses to obtain the total loss used to fine-tune the model on the dataset $D^\star$:

$$L(I, T, I^\star, T^\star) = L_{\text{hSC}}(I, T, I^\star, T^\star) + L_{\text{hSE}}(I, T, I^\star, T^\star). \tag{15}$$

Our proposal allows the model to differentiate between safe and unsafe embeddings based on their distance from the origin. Safe content is closer to the center, while NSFW content is farther away. This geometric arrangement not only enables the model to detect unsafe content but also allows dynamic manipulation of embeddings. NSFW queries can be redirected toward the safe region, effectively retrieving outputs that prioritize safety and providing more precise control over content retrieval.

4.3. Safety traversals and evaluation

The hyperbolic safety-aware training results in a restructuring of the shared embedding space of the vision-language model. To obtain safe but relevant retrieval outputs from unsafe queries or vice-versa, we introduce a traversal mechanism to adjust query embeddings in hyperbolic space, enhancing their similarity with either safe or unsafe content, depending on the retrieval task. This traversal involves moving the query embeddings along the line connecting them to the origin of the hyperboloid, altering their hyperbolic distance from the root. By adjusting the embeddings' positions, we align them with regions in the embedding space that correspond to the desired content type.

Traversal Definition. Given an embedding $\mathbf{q}$, our method computes the distance from a predefined root feature $\mathbf{r}$ in hyperbolic space using the Lorentzian distance function $d_{\mathcal{L}}(\mathbf{q}, \mathbf{r})$. For each type of content $X \in \{T, I, T^\star, I^\star\}$, we compute the mean distance μ_X from the root feature $\mathbf{r}$ based on the distribution of each category. The boundary for each type is then defined as

$$\tau_X = \mu_X + \tanh\left(\frac{\mu_X - \alpha}{\kappa}\right) + 1, \tag{16}$$

where κ is the negative curvature, and α is a constant set empirically to 0.8. This shift accounts for the curvature of space, ensuring the boundaries are appropriately adjusted for effective traversal. Defining four bounds allows more nuanced control over traversal depending on the retrieval task. To retrieve a content type X, the query is moved along the Euclidean direction vector $\boldsymbol{v}_{\text{dir}} = \mathbf{q} - \mathbf{r}$ toward the root feature $\mathbf{r}$ until it reaches the corresponding boundary τ_X (*e.g.* τ_T for safe text). The target position $\mathbf{q}^*$ is given as

$$\mathbf{q}^* = \mathbf{r} + \tau_X \cdot \frac{\boldsymbol{v}_{\text{dir}}}{\|\boldsymbol{v}_{\text{dir}}\|} \tag{17}$$

allowing the embeddings to be repositioned to match the target content type while maintaining semantic alignment.

5. Experiments

5.1. Training Details

Datasets. Our experiments are mainly conducted on the ViSU dataset [55], containing 165k quadruplets of safe and unsafe image-text pairs. We also evaluate our model on three real-world NSFW image datasets: NudeNet [4], NSFW data source URLs[1] and SMID [15].

Baselines. Our safety comparisons include the original CLIP [58] and the state-of-the-art Safe-CLIP [55]. CLIP was trained using a private dataset of 400M image-text pairs [66] which has unsafe data [5]. Safe-CLIP is fine-tuned on the ViSU dataset [55] to redirect unsafe content to safe correspondent one via contrastive losses and cosine similarities, aiming to unlearn NSFW concepts.

Models. Our visual and textual encoders are the same as CLIP [58], with VIT-L/14 as visual encoder, to maintain fair comparison to Safe-CLIP [55]. During training, both the visual and textual encoder are fine-tuned using low-rank decomposition [32] with low-rank factor $r = 16$.

Optimization. We use AdamW [43] with weight decay 0.2 and $(\beta_1, \beta_2) = (0.9, 0.98)$. We disable weight decay for all gains, biases, and learnable scalars. The model is finetuned for 20 epochs with batch size 256. The maximum learning rate is 8×10^{-4}. We use mixed precision [47] to accelerate training, except computing exponential map and losses for HySAC in FP32 precision for numerical stability.

Initialization. We initialize image and text encoders akin to CLIP, along with pre-trained weights. We initialize the softmax temperature as $\tau = 0.07$ and clamp it to a minimum value of 0.01. For HySAC, we initialize the learnable projection scalars $\alpha_{img} = \alpha_{txt} = 1/\sqrt{512}$, the curvature parameter $c = 1.0$ and clamp it in $[0.1, 10.0]$ to prevent training instability. All scalars are learned in logarithmic space as $\log(1/\tau)$, $\log(c)$, $\log(\alpha_{img})$ and $\log(\alpha_{txt})$.

Further details on the training setup are provided in the supplementary B.

[1] https://github.com/EBazarov/nsfw_data_source_urls

Model	Text-to-Image (T-to-I)			Image-to-Text (I-to-T)			Text-to-Image ($T^\star$-to-$I \cup I^\star$)			Image-to-Text ($I^\star$-to-$T \cup T^\star$)		
	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20
CLIP [58]	36.8	71.6	81.5	39.8	74.2	83.5	2.0	24.8	33.2	4.6	32.9	40.6
MERU [17]	14.9	43.0	54.2	14.7	42.3	53.8	2.2	15.2	21.5	4.4	22.6	29.4
HyCoCLIP [52]	34.3	71.2	80.6	34.4	71.3	82.2	2.8	25.3	33.2	8.2	37.8	45.7
Safe-CLIP [55]	45.9	81.8	89.7	45.3	82.3	89.8	8.0	46.9	58.0	19.1	62.9	71.1
MERU*	50.0	84.1	91.1	51.2	85.3	92.3	2.3	39.9	49.4	5.7	47.9	54.7
HyCoCLIP*	47.7	81.9	89.1	46.7	82.7	90.4	1.5	32.7	42.3	6.9	45.2	53.6
HySAC	**49.8**	**84.1**	**90.7**	**48.2**	**84.2**	**91.2**	**30.5**	**62.8**	**71.8**	**42.1**	**73.3**	**79.8**

Table 1. **Safe content retrieval performance on ViSU test set.** Across all tasks and recall rates, HySAC improves over existing safety unlearning CLIP and hyperbolic CLIP models, highlighting that our approach is able to navigate unsafe image or text inputs towards relevant but safe retrieval outputs. * CLIP fine-tuned in hyperbolic space on ViSU training set with MERU/HyCoCLIP losses.

Model	Text-to-Image ($T^\star$-to-$I^\star$)			Image-to-Text ($I^\star$-to-$T^\star$)			Text-to-Image ($T^\star$-to-$I^\star \cup I$)			Image-to-Text ($I^\star$-to-$T^\star \cup T$)		
	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20
CLIP [58]	73.1	94.9	97.6	72.8	95.2	97.7	68.4	92.3	95.9	67.1	93.3	96.7
MERU [17]	29.4	62.4	72.2	25.8	57.7	67.8	23.5	54.0	64.3	19.5	51.1	61.2
HyCoCLIP [52]	69.5	93.1	95.8	65.0	91.1	95.0	63.7	89.7	93.7	55.2	88.0	92.7
Safe-CLIP [55]	58.0	86.2	91.4	56.0	85.1	91.0	47.7	80.0	85.8	32.1	77.1	84.6
HySAC	**81.4**	**98.4**	**99.4**	**82.2**	**97.8**	**99.2**	**81.1**	**98.4**	**99.4**	**80.5**	**97.2**	**98.9**

Table 2. **Unsafe content retrieval performance on ViSU test set.** Akin to safe content retrieval, our approach performs best. This is a result of our objective, as we assign different content to different regions, enabling us to maintain valuable safety information.

5.2. Experimental Results

To assess the performance of our proposed model, HySAC, we measure its safety awareness and its ability to handle unsafe content effectively, while retaining both safe and unsafe knowledge. In the supplementary D, we report the zero-shot generalization of our method.

5.2.1. Safety retrieval comparison

We evaluate our model on the capability of retrieving safe and unsafe items, in comparison to CLIP [58] and Safe-CLIP [55]. We also provide a comparison to recent hyperbolic VLMs, namely MERU [17] and HyCoCLIP [52]. Safe-CLIP and HySAC fine-tuned from CLIP on ViSU [55], MERU trained on RedCaps [16], and HyCoCLIP on GRIT [54]. For a fair comparison, we additionally fine-tuned the CLIP model on the ViSU dataset in hyperbolic space, using MERU and HyCoCLIP[2] losses. All the models are evaluated on the ViSU test set.

Retrieval tasks are defined as text-to-image and image-to-text, where the goal is to find the most relevant counterpart for a given query. Recall@K measures the fraction of queries where the correct item appears in the top-K retrieved results. To assess the safe retrieval performance of HySAC, we measure recall exclusively on safe content in both visual and textual elements (T-to-I and I-to-T). This step is crucial to verify that the original CLIP model's retrieval capabilities are retained after finetuning in hyperbolic space with our training method. Then, to evaluate the safety-awareness capabilities of HySAC, we introduce a distinct setup in which NSFW elements are used as queries, while the retrievable items include both safe and unsafe elements (T^*-to-$I \cup I^*$ and I^*-to-$T \cup T^*$). During these experiments, a retrieval is deemed correct only if the query retrieves its safe counterpart, thereby validating the model's ability to redirect unsafe queries towards safe items.

When retrieving with HySAC, the threshold τ for moving query embeddings is computed using the mean distance of safe embeddings from the origin, adjusting the query embedding towards the safe region. Results are reported in Table 1, where we observe that HySAC consistently improves over both unlearning and existing hyperbolic models and features the highest recalls across all settings and rates. CLIP hyperbolic models finetuned on ViSU data (MERU* and HyCoCLIP*) perform well on safe-only retrieval, while our method achieves high performance on both safe-only and unsafe-safe retrieval, due to our safety-aware design.

In Table 2, we instead perform analyses for unsafe content retrieval. First, this involves the text-to-image and image-to-text retrieval on only unsafe elements (T^*-to-I^* and I^*-to-T^*). Second, instead, we perform retrieval by using unsafe elements as queries and both safe and unsafe items as retrievable items (T^*-to-$I^* \cup I$ and I^*-to-$T^* \cup T$), and deem the retrieval correct only if the query retrieves its corresponding unsafe one. This setup tests the model's ability to function as a content moderator and also showcases its capacity to provide user autonomy in content retrieval decisions. In these tests, the traversal mechanism in

[2]The box data needed for HyCoCLIP was extracted using Kosmos-2.

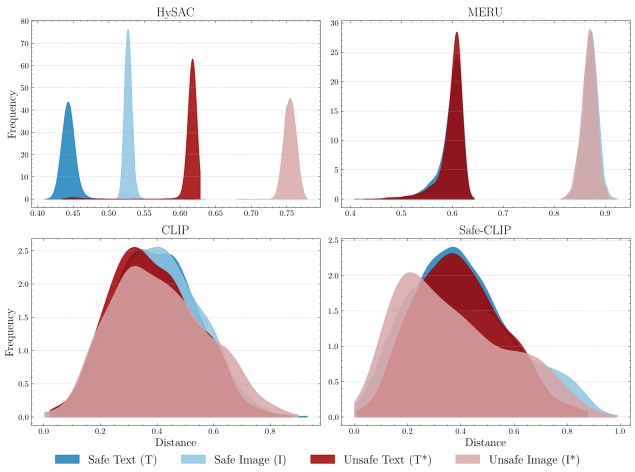

Figure 2. **Distributions of embedding distances from the root**. We embed all ViSU training samples and visualize their distance distribution from the root. While CLIP and Safe-CLIP do not separate between texts and images, MERU does. HySAC, instead, also differentiates between safe and unsafe content.

	(T-to-I)		(I-to-T)		($T^\star$-to-$I \cup I^\star$)		($I^\star$-to-$T \cup T^\star$)	
Model	R@1	R@10	R@1	R@10	R@1	R@10	R@1	R@10
w/o Ent	52.3	84.9	50.8	84.7	4.1	49.0	5.5	64.5
w/o S-Ent	51.0	84.2	49.8	84.3	1.4	39.1	7.4	63.7
HySAC	49.8	84.1	48.2	84.2	**30.5**	**62.8**	**42.1**	**73.3**

Table 3. **Ablation study on loss components**. We evaluate HySAC against two ablations that remove loss components. Results are in the same setting of Table 1.

HySAC uses adjusted parameters to move in the unsafe direction, targeting the retrieval of NSFW content. Here too, HySAC achieves the best recall across all settings, demonstrating that HySAC not only prioritizes safety by navigating away from NSFW content when required but also ensures that users can access NSFW content under controlled conditions, better than existing competitors.

5.2.2. Analysis of HySAC

Assessing HySAC embedding space. Further, we validate the organization of the embedding space as outlined in Equation 7. Our goal is to confirm that the embeddings for safe content are positioned closer to the origin of hyperbolic space, while those for unsafe content are further away, following the proposed hierarchy. Specifically, for each $X \in \{T, I, T^\star, I^\star\}$ of the training-set of $D^\star$, we compute the distances $d_{\mathcal{L}}(X, \mathbf{r})$ from the root feature $\mathbf{r}$.

A visualization is reported in Figure 2, where we show the distribution of embeddings in terms of their distance to the root feature. The comparative analysis is done across four different models: HySAC, CLIP, Safe-CLIP, and MERU. For both the hyperbolic models the root feature is the origin of the hyperboloid. For the Euclidean models, since the origin does not lie on the hypersphere, the root is empirically estimated as the embedding that has the least distance from all embeddings of the training set, *i.e.* the ℓ_2 normalization of the average of all embeddings. As it can be seen, the distribution clearly shows four peak distributions for HySAC, each one representing one of the X content types, elucidating the efficacy of our approach in maintaining a clear separation between safe, unsafe, textual, and visual content within the embedding space.

	% Safe (Text-to-Image)			% Safe (Image-to-Text)		
Model	NudeNet	NSFW URLs	SMID	NudeNet	NSFW URLs	SMID
CLIP	78.2	79.7	55.2	33.3	44.0	59.1
Safe-CLIP	92.6	92.6	**83.4**	75.2	76.4	65.6
HySAC	**96.2**	**93.9**	80.1	**84.4**	**95.1**	**97.9**

Table 4. **Retrieval performance on real NSFW images.** Rate of safe images retrieved using unsafe prompts from the ViSU test set. The retrievable set includes safe and unsafe real images, with the latter from LAION-400M and the former from NSFW sources.

Ablation Study. In Table 3, we validate the effectiveness of the key components in HySAC, by comparing its full configuration with variants where specific losses are disabled. In particular, we employ one variant which only keeps contrastive losses (denoted as "w/o Ent") and one that omits the safety-entailment loss ("w/o S-Ent"). Results show that while removing these components slightly improves performance in scenarios involving only safe content, likely due to reduced spatial constraints, their absence significantly undermines the model's effectiveness in dealing with unsafe content, especially in unsafe-to-safe retrieval. These results underscore the essential roles that both the modality- and safety-entailment losses play in enhancing the safety awareness of the proposed model.

Retrieval on real NSFW datasets. To further analyze the safety of HySAC, we conduct retrieval tests using NudeNet [4], NSFW data source URLs and SMID [15]. The first two datasets primarily contain nudity and pornographic content, whereas SMID includes a broader range of inappropriate content, such as *violence*, *harm*, and *discrimination*. We randomly select 1000 images from each dataset to serve as visual elements and use 5k NSFW captions from the ViSU test set. Both the image-to-text and text-to-image retrieval tasks also incorporate 10,000 randomly chosen retrievable safe items from LAION-400M [66].

Results, displayed in Table 4, contrast the performance of HySAC with that of CLIP [58] and Safe-CLIP [55]. We report the proportion of safe retrieval outputs when NSFW queries are used, highlighting the capability of HySAC to enhance safety in retrieval results. Traversing the embedding space towards safety prioritizes safer alternatives, which may adjust relevance in some cases. Due to the absence of datasets with real unsafe data and correlated safe alternatives, this aspect could not be evaluated. Notably, HySAC demonstrates improved performance in securing safer content compared to Safe-CLIP across most datasets for both text-to-image and image-to-text scenarios.

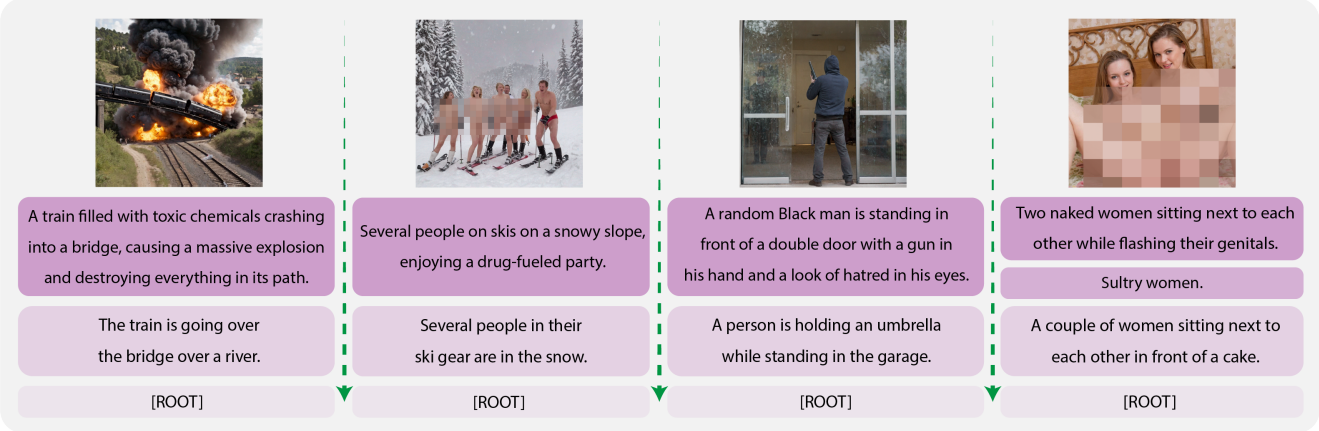

Figure 3. **Qualitative traversal results.** HySAC traverses towards the root feature, retrieving the top-1 text at each interpolation point. This traversal effectively transitions from unsafe to safe captions, demonstrating the model's ability to ensure safety-aware content retrieval.

	NudeNet			Mixed NSFW		
Model	Acc	FPR	FNR	Acc	FPR	FNR
NSFW-CNN [37]	85.3	0.0	14.7	66.5	4.5	35.9
CLIP-classifier [65]	97.3	0.0	2.7	76.9	**0.1**	11.0
CLIP-distance [59]	86.4	0.0	13.6	77.8	2.0	22.1
NudeNet [4]	91.2	0.0	8.8	76.9	4.5	24.6
Q16 [63]	28.5	0.0	71.5	65.3	8.3	29.4
HySAC	**99.5**	0.0	**0.5**	**78.9**	16.5	**6.8**

Table 5. **NSFW classification.** Comparison between HySAC and other NSFW classifiers. Metrics reported in percentages.

Classifying NSFW content. The structure of the embedding space in HySAC also supports the classification of NSFW content. We evaluate this using the NudeNet [4] and Mixed NSFW datasets, comparing against classifiers such as NSFW-CNN [37], CLIP-classifier [65], CLIP-distance [59], NudeNet [4], and Q16 [63]. NudeNet only contains nudity, while Mixed NSFW includes different NSFW categories from various online sources and safe images from PASS [1]. We sample a 1,000-image subset from NudeNet and 442 images from Mixed NSFW, balanced between safe and unsafe. Further details of these datasets are provided in the supplementary C.

Results in Table 5 show that HySAC achieves competitive or superior results in NSFW content classification, despite not being explicitly designed for safety classification. The norm threshold, set to the ViSU dataset mean (Figure 2), differentiates safe from unsafe content.

5.2.3. Visualizing the safety traversals

We examine traversal paths for safe text retrievals, starting with an unsafe image embedding as the query. The traversals are along the geodesic of hyperbolic space from the image to the origin of the hyperboloid, denoted `[ROOT]`.

The input query is an unsafe image taken from ViSU test set and the text retrieval space consists of a mix of safe and unsafe captions of ViSU test set, metadata-based caption from `pexels.com`, and a curated list of unsafe words[3]. To create visualization shown in Figure 3, each retrieved text output is selected only once across all interpolation points, ensuring unique retrievals. Results show that, with HySAC, as the query nears the origin, retrieved content shifts from unsafe to safe while preserving semantic relevance. This progression illustrates the model's capability to effectively navigate the embedding space along relevant paths. For more on the experimental setup and traversal visualizations, see the supplementary E.

Other ablation studies. For additional ablations on embedding space geometry and hyperparameter evaluation, we refer the reader to the supplementary D.

6. Conclusion

This paper introduces hyperbolic safety-aware vision-language models. Where recent literature focuses on removing or unlearning unsafe image-text content, we bring a perspective of awareness. By modelling unsafe image-text content as specific cases of their safe counterparts, we can divide the space into safe and unsafe regions. We show that hyperbolic space is a natural solution for this hierarchical relation, and propose a hyperbolic CLIP model with safety entailment learning and traversal. Our approach not only results in better retrieval of relevant safe outputs given unsafe inputs but also provides more robustness and comes with an NSFW classifier as a free by-product. Dealing with NSFW data in vision-language models is an important open research problem with real-world implications, from ethical to legal and societal concerns. By opting for awareness, we find that safety recognition improves. Given the importance of the ethical implications of this work, we provide a thorough discussion in the supplementary A.

[3] github.com/LDNOOBW/

Acknowledgements

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources. This work has been supported by the EU Horizon projects "ELIAS - European Lighthouse of AI for Sustainability" (No. 101120237) and "European Lighthouse on Safe and Secure AI (ELSA)" (No. 101070617), co-funded by the European Union. Tejaswi Kasarla also acknowledges travel support from the European Union's Horizon research and innovation programme under grant agreement No. 951847 (ELISE) and No. 101120237 (ELIAS).

References

[1] Yuki M. Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. *NeurIPS Track on Datasets and Benchmarks*, 2021. 8

[2] Mina Ghadimi Atigh, Julian Schoep, Erman Acar, Nanne Van Noord, and Pascal Mettes. Hyperbolic image segmentation. In *CVPR*, pages 4453–4462, 2022. 2

[3] Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*, 2018. 2

[4] P Bedapudi. NudeNet: Neural Nets for Nudity Classification, Detection, and Selective Censoring, 2019. 2, 5, 7, 8

[5] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *WACV*, pages 1536–1546. IEEE, 2021. 1, 2, 3, 5

[6] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 1

[7] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. Into the laion's den: Investigating hate in multimodal datasets. *NeurIPS*, 36, 2024. 1

[8] Manuel Brack, Felix Friedrich, Patrick Schramowski, and Kristian Kersting. Mitigating inappropriateness in image generation: Can there be value in reflecting the world's ugliness? *arXiv preprint arXiv:2305.18398*, 2023. 2

[9] Yinzhi Cao and Junfeng Yang. Towards Making Systems Forget with Machine Unlearning. In *IEEE Symposium on Security and Privacy*, 2015. 2

[10] Francesco Cauteruccio, Enrico Corradini, Giorgio Terracina, Domenico Ursino, and Luca Virgili. Extraction and analysis of text patterns from nsfw adult content in reddit. *Data & Knowledge Engineering*, 138:101979, 2022. 2

[11] Benjamin Paul Chamberlain, James Clough, and Marc Peter Deisenroth. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*, 2017. 2

[12] Ines Chami, Adva Wolf, Da-Cheng Juan, Frederic Sala, Sujith Ravi, and Christopher Ré. Low-dimensional hyperbolic knowledge graph embeddings. *arXiv preprint arXiv:2005.00545*, 2020. 2

[13] Nurendra Choudhary, Nikhil Rao, Sumeet Katariya, Karthik Subbian, and Chandan K Reddy. Self-supervised hyperboloid representations from logical queries over knowledge graphs. In *Proceedings of the Web Conference 2021*, pages 1373–1384, 2021.

[14] Nurendra Choudhary, Nikhil Rao, and Chandan Reddy. Hyperbolic graph neural networks at scale: a meta learning approach. *NeurIPS*, 36, 2024. 2

[15] Damien L Crone, Stefan Bode, Carsten Murawski, and Simon M Laham. The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PloS one*, 13(1):e0190954, 2018. 5, 7

[16] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021. 6

[17] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *ICML*, pages 7694–7731. PMLR, 2023. 1, 2, 3, 4, 6

[18] Bhuwan Dhingra, Christopher J Shallue, Mohammad Norouzi, Andrew M Dai, and George E Dahl. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*, 2018. 2

[19] Upol Ehsan, Q Vera Liao, Samir Passi, Mark O Riedl, and Hal Daumé III. Seamful xai: Operationalizing seamful design in explainable ai. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–29, 2024. 1

[20] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets. Hyperbolic vision transformers: Combining improvements in metric learning. In *CVPR*, pages 7409–7419, 2022. 2

[21] Luca Franco, Paolo Mandica, Bharti Munjal, and Fabio Galasso. Hyperbolic self-paced learning for self-supervised skeleton-based action representations. In *ICLR*, 2023. 2

[22] Shreyansh Gandhi, Samrat Kokkula, Abon Chaudhuri, Alessandro Magnani, Theban Stanley, Behzad Ahmadi, Venkatesh Kandaswamy, Omer Ovenc, and Shie Mannor. Scalable Detection of Offensive and Non-compliant Content/Logo in Product Images. In *WACV*, 2020. 2

[23] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing Concepts from Diffusion Models. In *ICCV*, 2023. 1, 2

[24] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*, pages 1646–1655. PMLR, 2018. 1, 2, 4

[25] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *NeurIPS*, 31, 2018. 2

[26] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *CVPR*, pages 6840–6849, 2023. 2

[27] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making AI Forget You: Data Deletion in Machine Learning. In *NeurIPS*, 2019. 2

[28] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal Sunshine of the Spotless Net: Selective Forgetting in Deep Networks. In *CVPR*, 2020. 2

[29] Aditya Golatkar, Alessandro Achille, Yu-Xiang Wang, Aaron Roth, Michael Kearns, and Stefano Soatto. Mixed Differential Privacy in Computer Vision. In *CVPR*, 2022. 2

[30] Kimia Hamidieh, Haoran Zhang, Walter Gerych, Thomas Hartvigsen, and Marzyeh Ghassemi. Identifying implicit social biases in vision-language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 547–561, 2024. 1

[31] Ahmad Fathan Hidayatullah, Anisa Miladya Hakim, and Abdullah Aziz Sembada. Adult Content Classification on Indonesian Tweets using LSTM Neural Network. In *ICACSIS*, 2019. 2

[32] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[33] Sarah Ibrahimi, Mina Ghadimi Atigh, Nanne Van Noord, Pascal Mettes, and Marcel Worring. Intriguing properties of hyperbolic embeddings in vision-language models. *Transactions on Machine Learning Research*, 2024. 2

[34] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916. PMLR, 2021. 1

[35] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *CVPR*, pages 6418–6428, 2020. 2, 3

[36] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *ICCV*, 2023. 2

[37] Gant Laborde. Deep nn for nsfw detection. https://github.com/GantMan/nsfw_model, 2020. 8

[38] Matthew Le, Stephen Roller, Laetitia Papaxanthos, Douwe Kiela, and Maximilian Nickel. Inferring concept hierarchies from text corpora via hyperbolic embeddings. In *ACL*, pages 3231–3241, 2019. 2, 4

[39] Ya-Wei Eileen Lin, Ronald R Coifman, Gal Mishne, and Ronen Talmon. Hyperbolic diffusion embedding and distance for hierarchical representation learning. In *ICML*, pages 21003–21025. PMLR, 2023. 2

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36, 2024. 1

[41] Qi Liu, Maximilian Nickel, and Douwe Kiela. Hyperbolic graph neural networks. *NeurIPS*, 32, 2019. 2

[42] Shaoteng Liu, Jingjing Chen, Liangming Pan, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. Hyperbolic visual embedding learning for zero-shot recognition. In *CVPR*, pages 9273–9281, 2020. 2

[43] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 5

[44] Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A Holistic Approach to Undesired Content Detection in the Real World. In *AAAI*, 2023. 2

[45] Joanna Materzyńska, Antonio Torralba, and David Bau. Disentangling visual and written concepts in clip. In *CVPR*, pages 16410–16419, 2022. 1

[46] Pascal Mettes, Mina Ghadimi Atigh, Martin Keller-Ressel, Jeffrey Gu, and Serena Yeung. Hyperbolic deep learning in computer vision: A survey. *IJCV*, pages 1–25, 2024. 2

[47] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. In *ICLR*, 2018. 5

[48] Leyla Mirvakhabova, Evgeny Frolov, Valentin Khrulkov, Ivan Oseledets, and Alexander Tuzhilin. Performance of hyperbolic geometry models on top-n recommendation tasks. In *Proceedings of the 14th ACM Conference on Recommender Systems*, pages 527–532, 2020. 2

[49] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[50] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *NeurIPS*, 30, 2017. 2

[51] Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*, pages 3779–3788. PMLR, 2018. 2

[52] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912*, 2024. 1, 2, 3, 4, 6

[53] Wei Peng, Tuomas Varanka, Abdelrahman Mostafa, Henglin Shi, and Guoying Zhao. Hyperbolic deep neural networks: A survey. *IEEE Trans. PAMI*, 44(12):10023–10044, 2021. 2

[54] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 6

[55] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. In *ECCV*, 2024. 1, 2, 3, 4, 5, 6, 7

[56] Samuele Poppi, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Multi-Class Unlearning for Image Classification via Weight Filtering. *IEEE Intelligent Systems*, 2024. 2

[57] Samuele Poppi, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Unlearning vision transformers without retaining data via low-rank decompositions. In *ICPR*, 2024. 2

[58] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 1, 3, 4, 5, 6, 7

[59] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramer. Red-teaming the stable diffusion safety filter. In *NeurIPS ML Safety Workshop*, 2022. 8

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1

[61] Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *ICML*, pages 4460–4469. PMLR, 2018. 2

[62] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 2

[63] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content? In *ACM FAccT*, 2022. 2, 8

[64] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. In *CVPR*, 2023. 2, 3

[65] Christoph Schuhmann. Clip based nsfw detector. https://github.com/LAION-AI/CLIP-based-NSFW-Detector/tree/main, 2022. 8

[66] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 1, 5, 7

[67] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 1

[68] Ryohei Shimizu, Yusuke Mukuta, and Tatsuya Harada. Hyperbolic neural networks++. *ICLR*, 2021. 2

[69] David Thiel. Identifying and eliminating csam in generative ml training data and models. 2023. 1

[70] Alexandru Tifrea, Gary Bécigneul, and Octavian-Eugen Ganea. Poincaré glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018. 2

[71] Max van Spengler, Erwin Berkhout, and Pascal Mettes. Poincare resnet. In *ICCV*, pages 5419–5428, 2023. 2

[72] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361*, 2015. 4

[73] Liping Wang, Fenyu Hu, Shu Wu, and Liang Wang. Fully hyperbolic graph convolution network for recommendation. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 3483–3487, 2021. 2

[74] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 1

[75] Robert Wolfe, Yiwei Yang, Bill Howe, and Aylin Caliskan. Contrastive language-vision ai models pretrained on web-scraped multimodal data exhibit sexual objectification bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1174–1185, 2023. 1

[76] Eric Zhang, Kai Wang, Xingqian Xu, Zhangyang Wang, and Humphrey Shi. Forget-Me-Not: Learning to Forget in Text-to-Image Diffusion Models. *CVPR*, 2024. 2

[77] Yudong Zhu, Di Zhou, Jinghui Xiao, Xin Jiang, Xiao Chen, and Qun Liu. Hypertext: Endowing fasttext with hyperbolic geometry. *arXiv preprint arXiv:2010.16143*, 2020. 2