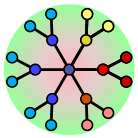
This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.



Hierarchically Robust Zero-shot Vision-language Models

Junhao Dong[1,2] Yifei Zhang[3] Hao Zhu[4] Yew-Soon Ong[1,2,✉] Piotr Koniusz[5,4,✉]
[1]Nanyang Technological University [2]CFAR, IHPC, A*STAR [3]Northwest Polytechnical University
[4]Data61♥CSIRO [5]University of New South Wales
{junhao003, asysong}@ntu.edu.sg, {yifeiacc, allenhaozhu}@gmail.com, piotr.koniusz@unsw.edu.au

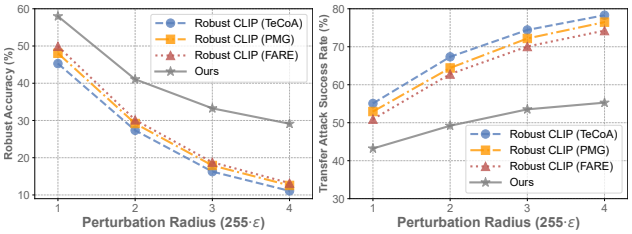
Abstract

Vision-Language Models (VLMs) can perform zero-shot classification but are susceptible to adversarial attacks. While robust fine-tuning improves their robustness, existing approaches align fixed text embeddings with an image embedding, sacrificing natural performance and robustness. A robustness degradation also occurs when a model faces adversarial attacks targeting superclasses (parent classes, e.g., `mammal`) in addition to their base (leaf) classes (e.g., `cat`). Thus, to enhance adversarial robustness and leverage the inherent hierarchical properties of class space, we propose a novel adversarial fine-tuning framework based on hierarchical embeddings and several levels of adversarially robust alignment of image-text modalities. Additional mechanisms place visual embeddings at the desired depth of hierarchy, and we provide a theoretical connection between the depth of embedding in the hierarchy and the maximum viable margin size. Our model naturally realizes several margin sizes, boosting generalization of adversaries for robustification. As various trees with different parent labels can share the same leaf labels, we also consider aligning over multiple trees to boost semantic variety. Experiments across several datasets are performed.

1. Introduction

Adversarial samples [40]—inputs nearly indistinguishable from clean images make Deep Neural Networks (DNNs) [24, 45] and Vision-Language Models (VLMs) [30, 39, 46] predict incorrect labels with high confidence. VLMs are susceptible to adversarial threats due to their multi-modal nature, limiting their use in security-critical settings [7, 23].

Recent studies on adversarial fine-tuning [32, 38, 42] enhance zero-shot robustness in VLMs [15, 16, 18, 36] by a sample-wise alignment that matches adversarial features of each *base class* (original category, *e.g.*, `kit fox`) across image and text modalities, neglecting auxiliary information contained by *superclasses* (parent classes, *e.g.*, `carnivore`←`canine`←`fox`). Thus, aligning an image embedding toward mere text embeddings of base classes compromises robustness because the adversarial samples generated in such a setting focus by design on base classes only. Including more general superclasses to obtain adversarial samples leads to more general model robustification[1].



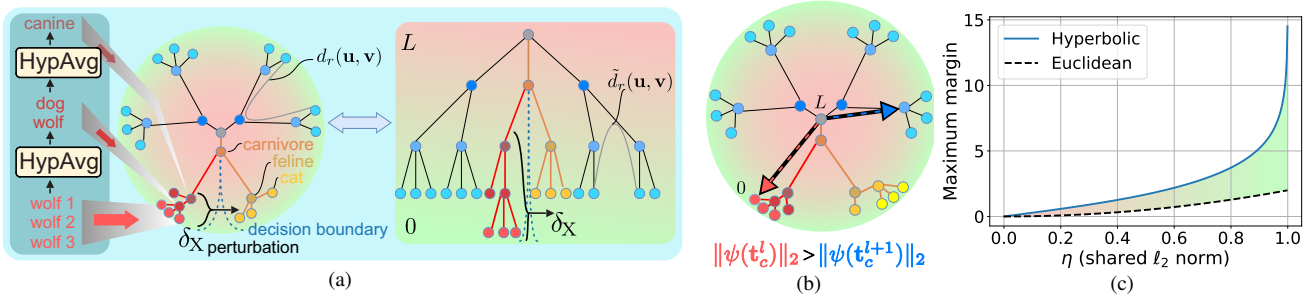
(a) Superclass results (b) Base class results

Figure 1. *Motivation (ImageNet [6] Val.)* Fig. 1a: robust accuracy on superclasses (one level above the leaf classes). TeCoA [32], PMG [42], and FARE [38] use adversaries derived from base classes and thus perform poorly on superclasses. Fig. 1b: transfer attack success rates on base classes. For our model, adversarial samples generated on superclasses were used. In contrast, TeCoA [32], PMG [42], and FARE [38], the base classes are susceptible to attacks obtained from superclasses, highlighting the need for model robustification across class hierarchy (our model). Examples of superclasses & leaves: (i) parent: `equine`, children: `horse, zebra`, (ii) parent: `car`, children: `freight car, taxi`, (iii) parent: `bag`, children: `backpack, purse`.

To empirically support the above claim, for each adversarially fine-tuned CLIP model [32, 38, 42], we generate adversarial samples targeting the superclasses rather than the base classes of original samples, thereby creating more generic adversaries. Figure 1a shows that prior adversarial fine-tuning methods, which were trained to robustify base classes, have poor robustness classification on superclasses. Thus, adversarial learning on the base classes fails to generalize the robustness to superclasses. Figure 1b shows that adversarial samples generated at the superclass level can be effectively transferred to attack base classes, resulting in a considerable attack success rate, especially as the perturba-

✉Corresponding authors.

[1]The class hierarchy for ImageNet is publicly available—it follows WordNet [33] hierarchy. An arbitrary class hierarchy can also be built for a dataset's label space with a single ChatGPT prompt within seconds.



(a) (b) (c)

Figure 2. *Understanding Hyperbolic Geometry.* Fig. 2a: Trees can be embedded in the Poincaré ball with low distortion ($\sim 1+\epsilon$) [37]. Thus, the distance between tree nodes $\tilde{d}_r(\mathbf{u},\mathbf{v})$ is the analogue of the Riemannian distance $d_r(\mathbf{u},\mathbf{v})$ between hyperbolic embeddings. The norm of the vector measured in the Poincaré ball is a proxy to the hierarchical level of a node in the tree (Fig. 2b) (we use "reciprocal": L is the root level and 0 is the leaves). For the visual brach of CLIP, we have images with base class labels only. Thus, hyperbolic average (HypAvg) produces more generic embeddings from the more specific embeddings (closer to the center than children), *e.g.*, from `wolf 1`, `wolf 2`, `wolf 3` it produces `wolf`, and from `wolf`, `dog` it produces `canine`. To produce a robust adversarial sample used for robustification, we compute δ_X (added to a sample) that makes nodes on the red path move across the decision boundary towards other categories (*e.g.*, blue path). Fig. 2b: The ℓ_2 norm of features (*e.g.*, text $\psi(\cdot)$) controls the embedding level, *e.g.*, $\|\psi(\mathbf{t}_c^l)\|_2 > \|\psi(\mathbf{t}_c^{l+1})\|_2$ (red *vs*. blue arrow) has order $0<l<l+1<L$ in the Poincaré ball. Fig. 2c: Hyperbolic *vs*. Euclidean classifier's limit on viable margin size w.r.t. the ℓ_2 norm of features (η). Our hyperbolic classifier can separate embeddings by an infinite margin (blue solid curve) when norm $\eta \to 1/\sqrt{r}$ for curvature r. The Euclidean model (dashed black) has finite margin. Several levels of hierarchy produce several margin sizes in $(0,\infty)$ to cover several levels of generalization, producing more general adversarial perturbation δ_X compared to the Euclidean model.

tion radius increases. This vulnerability to superclass adversaries shows that prior methods lose robustness due to the base classes usage only. We point out the benefit of achieving hierarchical robustness on image-text modalities by exploring the richness of hierarchical decision boundaries.

To avoid the fixation on base categories, and achieve hierarchically adversarially robust image-text modality alignment, we propose a novel framework based on hyperbolic embeddings [48] of image/category hierarchies.

Figure 2a shows we encode children & their parent labels across all levels of hierarchy. Text prompt embeddings of children labels are encouraged to lie closer to the Poincaré boundary, whereas text prompt embeddings of parent labels are encouraged to lie closer to the Poincaré ball origin.

Thus, embeddings of superclasses at level l+1 are aligned with the hyperbolic average of embeddings of images of a given class at level l in the mini-batch ($l=0$: leaves, L: root). The hyperbolic average has the known property that from two/more children vectors, it yields a lower ℓ_2 norm compared to children's norms. Thus, the hyperbolic average produces more general concepts (at level $l+1$ closer to the root) from more specific concepts (at level l) as the value of the norm controls the node depth (Fig. 2b).

Moreover, Figure 2c shows that the hyperbolic classifier enjoys the limit on viable margin size growing exponentially (unlike the Euclidean classifier) to infinity (larger separation) with norm η (depth) growing toward tree leaves. Thus, we employ several hierarchical levels of classifier to reformulate traditional sample-wise adversarial fine-tuning into its hierarchical counterpart, capturing rich adversarial information from hierarchical decision boundaries across several levels of generalization due to varying margin sizes.

Our core contributions are listed below:

i. Unlike standard adversarial fine-tuning that aligns image and text embeddings under flat class structure, we cast adversarial fine-tuning as a hierarchical scheme that exploits rich hierarchical decision boundaries from child and parent categories that realize several varying margin sizes due to our hyperbolic design. This leads to more general adversarial perturbations for robustification compared to the Euclidean model. We also demonstrate theoretically what is the limit on viable margin size.

ii. We analyze adversaries against superclass classification across varying perturbation radii. Our observations reveal that as the adversarial learning fixates on the base classes, it fails to generalize robustness to superclasses. Thus, we propose to leverage the hyperbolic embedding to exploit class hierarchies and image hierarchies. We propose mechanisms to deploy image and text embeddings meaningfully across several hierarchical levels.

iii. We show on 15 datasets that our method outperforms previous methods. We also explore adversarially robust image-text retrieval, captioning, and medical tasks.

2. Background

Related works. Adversarial training enhances the model's robustness against attacks by integrating adversaries into training [8–13, 31, 44]. As adversarial training of large VLMs is costly [36], adversarial fine-tuning [17, 19–21, 32, 38] is a better alternative. Mao *et al*. [32] employed text-guided contrastive learning for adversarial image-text embedding matching. Wang *et al*. [42] included feature-level regularization. Schlarmann *et al*. [38] developed an

unsupervised robust learning framework for downstream tasks. Li *et al*. [29] used one prompt word to boost robustness. However, these existing methods overlook the benefit of more generalized adversaries from hierarchical decision boundaries. See Appendix P for more related works.

Adversarial fine-tuning of VLM. CLIP [36] uses image and text encoders parameterized by $\boldsymbol{\theta}_{\mathrm{X}}$ and $\boldsymbol{\theta}_{\mathrm{T}}$. It aligns visual and text embeddings in a shared feature space. The image and text encoders, $f: \mathcal{X} \to \mathbb{R}^d$ and $g: \mathcal{T} \to \mathbb{R}^d$, map input images $\mathbf{x} \in \mathcal{X}$ and input texts $\mathbf{t} \in \mathcal{T}$ to d-dimensional feature embeddings. By encoding image-text pairs $(\mathbf{x}, \mathbf{t})$ into a shared feature space, CLIP facilitates alignment between the two modalities. The probability $p_c(\mathbf{x})$ that an image $\mathbf{x}$ belongs to the base category of $c \in \{1, \ldots, C\}$ is defined as:

$$p_c(\mathbf{x}) = \frac{\exp\left[\cos\left(f(\mathbf{x}), g(\mathbf{t}_c)\right)\right]}{\sum_{c'=1}^{C} \exp\left[\cos\left(f(\mathbf{x}), g(\mathbf{t}_{c'})\right)\right]}, \quad (1)$$

where $\exp(\cdot)$ and $\cos(\cdot)$ are the exponential function and the cosine similarity. Tokenizer $h(\cdot)$ produces tokenized text embedding, $\mathbf{t}_c = h(\text{“}\texttt{[Context] [CLASS}_c\texttt{]}\text{”})$, *e.g.*, “$\texttt{This is a photo of a [CLASS}_c\texttt{]}$” which is passed via the text encoder $g(\cdot)$ to produce a text embedding serving as a text reference for class c. Let $\mathbf{p} = [p_1, \ldots, p_C]^\top \in [0, 1]^C$. To enhance zero-shot robustness, standard adversarial fine-tuning (TeCoA [32]) can be applied to CLIP's image-text pairs $(\mathbf{x}, c)$ from the dataset $\mathcal{D}$ as follows:

$$\min_{\boldsymbol{\theta}_{\mathrm{X}}} \mathbb{E}_{(\mathbf{x},c)\sim\mathcal{D}} \left[\max_{\|\boldsymbol{\delta}_{\mathrm{X}}\|_\infty \leq \epsilon_{\mathrm{X}}} \mathcal{L}_{\mathrm{CE}}\big(\mathbf{p}(\mathbf{x} + \boldsymbol{\delta}_{\mathrm{X}}), \mathbf{y}(c)\big) \right], \quad (2)$$

where the perturbation $\boldsymbol{\delta}_{\mathrm{X}}$ is added to the image $\mathbf{x}$ to produce the adversarial sample $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}_{\mathrm{X}}$ within an ℓ_∞-norm ball of radius ϵ_{X}. The target class c is represented by a one-hot vector $\mathbf{y}(c) = [\mathbb{1}(c = 1), \ldots, \mathbb{1}(c = C)]^\top \in \{0, 1\}^C$. The inner maximization generates adversaries while the outer minimization reduces the empirical risk over adversaries. The adversary generation consists of perturbing visual inputs by maximizing the Cross-Entropy (CE) loss $\mathcal{L}_{\mathrm{CE}}$ via iterative Projected Gradient Descent (PGD) [31]:

$$\hat{\mathbf{x}}^{(i+1)} = \Pi_{\mathbb{B}(\mathbf{x},\epsilon_{\mathrm{X}})}\left[\hat{\mathbf{x}}^{(i)} + \alpha_{\mathrm{X}} \cdot \mathrm{sgn}\left[\nabla_{\hat{\mathbf{x}}^{(i)}} \mathcal{L}_{\mathrm{CE}}\big(\mathbf{p}(\hat{\mathbf{x}}^{(i)}), \mathbf{y}(c)\big)\right]\right], \quad (3)$$

where $\mathrm{sgn}(\cdot)$ is the sign function, and α_{X} is the PGD step size. The projection operator $\Pi_{\mathbb{B}(\mathbf{x},\epsilon_{\mathrm{X}})}$ constraints the perturbed input to the ℓ_∞-norm hyperball of radius ϵ_{X} centered at $\mathbf{x}$. Adversarial initialization uses random perturbation: $\hat{\mathbf{x}}^{(0)} \sim \mathbf{x} + 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$. The final adversarial sample $\hat{\mathbf{x}} = \hat{\mathbf{x}}^{(m)}$ is obtained after m iterations.

Poincaré ball model. *Hyperbolic space* is a complete and connected Riemannian manifold with constant negative sectional curvature [3]. We employ the Poincaré ball model $\mathbb{D}_r^d := \{\boldsymbol{\phi} \in \mathbb{R}^d \mid \|\boldsymbol{\phi}\|_2^2 < \frac{1}{r}\}$, where $\frac{1}{r} > 0$ is the radius of the ball, and $\|\cdot\|_2$ is the ℓ_2-norm. The Poincaré ball is endowed with a Riemannian metric $g_{\mathbb{D}}(\boldsymbol{\phi}) = \frac{4r}{(1-\|\boldsymbol{\phi}\|^2)^2}\, g_{\mathbb{E}}$, where $\boldsymbol{\phi} \in \mathbb{D}_r^d$ and $g_{\mathbb{E}}$ is the canonical Euclidean metric.

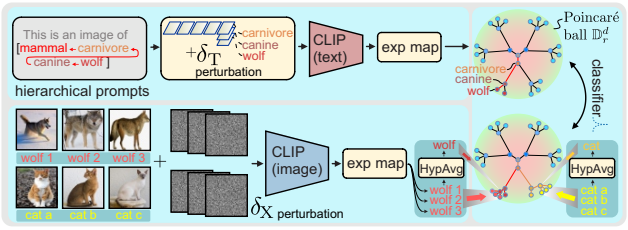


Figure 3. *Our pipeline.* Text & image CLIP encoders are used. For text, we look up each class in a hierarchical tree (*e.g.*, WordNet for ImageNet) and extract the path from the leaf category to the root, *e.g.*, animal←mammal←canine←wolf. For each category level, we form one text prompt and encode it with CLIP. The exponential map elevates embeddings from the Euclidean space into the hyperbolic space (Poincaré ball). HypAvg is hyperbolic averaging that generalizes specific embeddings (*e.g.*, *wolf 1, wolf 2, wolf 3*) into a more generic parent embedding of *wolf*. Using the Riemannian distance, the text and visual “wolf” paths are aligned by a classifier while pushed away from other paths. Adversarial perturbations $\boldsymbol{\delta}_T$ and $\boldsymbol{\delta}_X$ are obtained on such a hierarchical classifier by the PGD [31] for the robust model fine-tuning.

3. Proposed Method

Below, we introduce our adversarial fine-tuning for zero-shot hierarchically adversarially robust VLMs. Figure 3 shows our pipeline. Appendix A contains additional details.

Hyperbolic Embedding. We represent the hierarchical structure inherent to each image-text pair by the Poincaré ball model to embed class hierarchy. Levels $1 \leq l \leq L$ indicate superclasses while $l = 0$ are base classes for the total depth L. The tokenized text of the base class or superclass c_l is $\mathbf{t}_c^l = h(\text{“}\texttt{[Context] [SUPERCLASS}_c^l\texttt{]}\text{”})$ where $h(\cdot)$ is a tokenizer. There exist C_l classes at level l. For ImageNet, we use the WordNet [33] class hierarchy by extracting hypernyms of synsets as superclasses. In the absence of predefined hierarchical taxonomies in other datasets, we generate superclasses via ChatGPT-4o [34]. We prompt LLM (takes few seconds) to get contextually appropriate superclasses for each base class, followed by manual reviews to ensure semantic quality. See Appendix C for details.

To facilitate operations within hyperbolic space, we establish a bi-directional mapping between the Euclidean space $\mathbb{R}^d$ (CLIP embeddings) and the hyperbolic space $\mathbb{D}_r^d$. The exponential map in Definition 1 embeds Euclidean vectors into hyperbolic space (the log map is the inverse).

Definition 1 (Exponential & Logarithmic Maps). *The exponential map* $\exp_{\mathbf{u}}^r(\cdot)$ *is a function from* $T_{\mathbf{u}}\mathbb{D}_r^d \cong \mathbb{R}^d$ *to* $\mathbb{D}_c^d$*:*

$$\exp_{\mathbf{u}}^r(\mathbf{v}) = \mathbf{u} \oplus \left[\tanh\left(\frac{\sqrt{r}\lambda_{\mathbf{u}}^r}{2}\|\mathbf{v}\|_2\right)\frac{\mathbf{v}}{\sqrt{r}\|\mathbf{v}\|_2}\right], \quad (4)$$

where $\oplus$ is the (differentiable) Möbius addition, and $\lambda_{\mathbf{u}}^r$ denotes the conformal factor for distance scaling. The (inverse) logarithmic map $\log_{\mathbf{u}}^r(\cdot)$ establishes an inverse projection from $\mathbb{D}_c^d$ to $T_{\mathbf{u}}\mathbb{D}_r^d \cong \mathbb{R}^d$, which is defined as:

$$\log_{\mathbf{u}}^r(\mathbf{v}) = \frac{2}{\sqrt{r}\lambda_{\mathbf{u}}^r} \tanh^{-1}\left(\sqrt{r}\,\|-\mathbf{u}\oplus\mathbf{v}\|_2\right)\frac{-\mathbf{u}\oplus\mathbf{v}}{\|-\mathbf{u}\oplus\mathbf{v}\|_2}. \quad (5)$$

The distance between vectors in the hyperbolic space is measured by the Riemannian distance in Definition 2.

Definition 2 (Riemannian distance). *For two points in the Poincaré ball $\mathbf{u},\mathbf{v} \in \mathbb{D}_r^d$, their Riemannian distance induced by the Riemannian metric $g_{\mathbb{D}}(\cdot)$ is defined as:*

$$d_r(\mathbf{u},\mathbf{v}) = \frac{2}{\sqrt{r}} \tanh^{-1}\left(\sqrt{r}\|-\mathbf{u}\oplus\mathbf{v}\|_2\right). \quad (6)$$

Encoding and Projection. Let $f(\mathbf{x})$ and $g(\mathbf{t}_c)$ denote "clean" image and text features obtained from vision and text encoders of CLIP. Let $(\mathbf{x}+\boldsymbol{\delta}_X, \mathbf{t}_c+\boldsymbol{\delta}_T)$ be an image-text pair with perturbations $\boldsymbol{\delta}_X$ and $\boldsymbol{\delta}_T$ to optimize.

Firstly, clean or adversarial image/text embeddings are projected into the Poincaré ball $\mathbb{D}_c^d$ by the exponential map $\exp_{\mathbf{0}}^r(\cdot)$, with projection applied for the numerical stability:

$$\mathrm{Proj}(\mathbf{z}) := \begin{cases} \mathbf{z} & \text{if } \|\mathbf{z}\|_2 \le (1-\xi)\frac{1}{\sqrt{r}} \\ (1-\xi)\frac{\mathbf{z}}{\sqrt{r}\|\mathbf{z}\|_2} & \text{else} \end{cases}, \quad (7)$$

where $\xi \ge 0$ prevents the projected vectors from "touching" the Poincaré ball's boundary. Let the image-text embeddings in the Poincaré ball be $\phi(\mathbf{x}+\boldsymbol{\delta}_X) = \mathrm{Proj}\left(\exp_{\mathbf{0}}^r(f(\mathbf{x}+\boldsymbol{\delta}_X))\right)$ and $\psi(\mathbf{t}_c+\boldsymbol{\delta}_T) = \mathrm{Proj}\left(\exp_{\mathbf{0}}^r(g(\mathbf{t}_c+\boldsymbol{\delta}_T))\right)$.

Hierarchical Image-text Alignment. For superclass text prompts, embeddings are obtained from text CLIP. However, images do not encompass an entire superclass (universal car template does not exist). Thus, we align a superclass with hyperbolic average of image embeddings within a mini-batch where the base classes share the same parent.

Hyperbolic Averaging (HypAvg) in the Poincaré coordinates, based on the so-called *Einstein midpoint*, is given as:

$$\mathrm{HypAvg}\left(\{\mathbf{z}_i\}_{i=1}^m\right) = \frac{\sum_{i=1}^m \left(\frac{\gamma_i}{\rho}\frac{2\mathbf{z}_i}{1+r\|\mathbf{z}_i\|_2^2}\right)}{1+\sqrt{1-\frac{r}{\rho^2}\left\|\sum_{i=1}^m \left(\gamma_i \frac{2\mathbf{z}_i}{1+r\|\mathbf{z}_i\|_2^2}\right)\right\|_2^2}}. \quad (8)$$

Let $\gamma_i = 1/\sqrt{1-r\|\mathbf{k}_i\|_2^2}$ denote the Lorentz factor for each point, $\mathbf{k}_i = \frac{2\mathbf{z}_i}{1+r\|\mathbf{z}_i\|_2^2}$ and $\rho = \sum_{i=1}^m \gamma_i$ and curvature r.

For samples of base category c in the set $\mathcal{X}_c$ (*e.g.*, formed from mini-batch samples of base category c), one can obtain $\mathrm{HypAvg}\left(\{\phi(\mathbf{x}) : \mathbf{x} \in \mathcal{X}_c\}\right)$. Then one constructs image embeddings for the corresponding higher-level superclasses by applying hyperbolic averaging over the (average) embeddings at the lower-class level. This process is performed recursively across all hierarchical levels. Thus, for level $l = 0, \ldots, L$ and category $c = 1, \ldots, C_l$, we get $\boldsymbol{\phi}_c^l = \mathrm{HypAvg}\left(\{\phi(\mathbf{x}) : \mathbf{x} \in \mathcal{X}_c^l\}\right)$. By analogy, for adversarial samples, $\hat{\mathcal{X}}_c^l$, we get $\hat{\boldsymbol{\phi}}_c^l = \mathrm{HypAvg}\left(\{\phi(\mathbf{x}) : \mathbf{x} \in \hat{\mathcal{X}}_c^l\}\right)$.

Thus, we propose a Hierarchy-aware Negative Set Augmentation (HNSA) integrating both the higher- and lower-level superclasses into fine-tuning. The probability of sample being class c at level l is:

$$p_c^l = \frac{\exp\left[-d_r\left(\boldsymbol{\phi}_c^l, \psi(\mathbf{t}_c^l)\right)\right]}{\eta_c^l + \sum_{c'=1}^{C_l} \exp\left[-d_r\left(\boldsymbol{\phi}_c^l, \psi(\mathbf{t}_{c'}^l)\right)\right]}, \quad (9)$$

where the scalar $\eta_c^l = \sum_{k\in\mathcal{K}^{l+1}}^{C_{l+1}} \exp\left[-d_r\left(\boldsymbol{\phi}_c^l, \psi(\mathbf{t}_k^{l+1})\right)\right] + \sum_{k\in\mathcal{K}^{l-1}}^{C_{l-1}} \exp\left[-d_r\left(\boldsymbol{\phi}_c^l, \psi(\mathbf{t}_k^{l-1})\right)\right]$ is an additional augmentation with "negative" classes at levels $l-1$ and $l+1$. Sets $\mathcal{K}^{l-1}$ and $\mathcal{K}^{l-1}$ contain classes at levels $l-1$ and $l+1$ that do not share an edge with class c at level l.

We define $\mathbf{p}^l = [p_1^l, \ldots, p_{C_l}^l]^\top \in [0,1]^{C_l}$ as the likelihood vector, and rewrite the standard adversary generation scheme from Eq. (2) into its hierarchical counterpart called Hierarchy-preserving Image-Text Alignment (HITA):

$$\mathcal{L}' = \max_{\substack{\|\boldsymbol{\delta}_X\|_\infty \le \epsilon_X \\ \|\boldsymbol{\delta}_T\|_\infty \le \epsilon_T}} \sum_{l=0}^{L} \omega_l \mathcal{L}_{\mathrm{CE}}\left[\mathbf{p}^l\left(\{\mathbf{x}+\boldsymbol{\delta}_X : \mathbf{x} \in \mathcal{X}_c^l\}, \mathbf{t}_c^l + \boldsymbol{\delta}_T\right), \mathbf{y}(c)\right], \quad (10)$$

where $\omega_l = 1 - \frac{l}{L+1}$ is a weighting factor for each hierarchical level. Universal adversarial samples for image and text modalities are generated as $\hat{\mathbf{x}} = \mathbf{x} + \boldsymbol{\delta}_X$ and $\hat{\mathbf{t}} = \mathbf{t} + \boldsymbol{\delta}_T$. To maintain the hierarchical integrity of text embeddings in the hyperbolic space, we add a soft constraint controlled by offset $\zeta_{\mathrm{gap}} \ge 0$ to ensure that higher-level superclass norms are smaller than those of lower-level superclasses:

$$\mathcal{L}_{\mathrm{gap}}^{\mathrm{Label}} = \sum_{l=0}^{L-1}\sum_{c=1}^{C_l} \max\left(0, \|\psi(\hat{\mathbf{t}}_c^{l+1})\|_2 - \|\psi(\hat{\mathbf{t}}_c^l)\|_2 + \zeta_{\mathrm{gap}}\right). \quad (11)$$

Individual Images and Intra-class Variability. To maintain the hierarchical integrity of individual image embeddings of a given class, they cannot be fully aligned with the base class text prompt embedding as these individual images still vary at fine-grained level (no annotations are given). Thus, we firstly align hyperbolic image embedding $\phi(\hat{\mathbf{x}})$ to be close to the base class text embedding $\psi(\hat{\mathbf{t}}_c)$:

$$\mathcal{L}_{\mathrm{vic}} = \sum\nolimits_{\hat{\mathbf{x}}_c \in \hat{\mathcal{X}}_c} \left| d_r\left(\phi(\hat{\mathbf{x}}_c), \psi(\hat{\mathbf{t}}_c)\right) - \zeta_{\mathrm{vic}} \right|, \quad (12)$$

where $\zeta_{\mathrm{vic}} \ge 0$ is a radius of the text embedding. $\hat{\mathcal{X}}_c$ is the adversarial image set of the base category c. $\mathcal{L}_{\mathrm{vic}}$ penalizes image embeddings that fall outside the radius.

We maintain the hierarchical order of embeddings in the hyperbolic space by encouraging each image embedding to reside in the upper arc of its base class text embedding:

$$\mathcal{L}_{\mathrm{gap}}^{\mathrm{Intra}} = \sum_{\hat{\mathbf{x}}_c \in \hat{\mathcal{X}}_c} \max\left(0, \|\psi(\hat{\mathbf{t}}_c)\|_2 - \|\phi(\hat{\mathbf{x}}_c)\|_2 + \zeta_{\mathrm{gap}}\right), \quad (13)$$

where $\zeta_{\mathrm{gap}} \ge 0$ controls the minimum upper arc radius.

Table 1. Zero-shot **clean** accuracy (%). Adversarial fine-tuning is conducted on ImageNet, followed by evaluations across 15 datasets.

Method		Image Net	STL10	CIFAR 10	CIFAR 100	SUN 397	Stanford Cars	Food 101	Oxford Pet	Flower 102	DTD	Euro Sat	FGVC	PCAM	Caltech 101	Caltech 256	Average
CLIP	2021 [36]	59.13	97.17	88.55	62.29	57.68	52.07	83.84	87.35	65.60	40.05	38.31	20.13	52.26	87.08	82.01	64.90
TeCoA	2023 [32]	58.69	92.15	75.89	46.31	48.67	26.59	47.27	79.42	45.15	31.70	25.32	12.15	47.23	79.20	73.51	52.62
PMG-FT	2024 [42]	60.20	93.89	80.79	51.92	53.55	40.49	61.26	82.91	53.39	33.09	24.29	14.76	48.47	84.00	77.40	57.36
FARE	2024 [38]	57.86	94.81	85.45	60.75	54.05	45.06	67.00	84.64	58.72	36.23	24.84	16.23	44.89	85.42	79.13	59.67
AoS	2025 [14]	60.58	96.83	86.70	61.98	55.94	46.42	69.53	85.80	59.69	38.06	29.25	17.00	50.22	86.60	80.93	61.70
Ours		61.19	96.46	86.50	62.77	55.76	48.11	**70.31**	85.63	60.32	**38.25**	30.41	17.39	50.90	86.81	81.27	62.14
Ours (5 trees)		**61.86**	**97.18**	**87.13**	**62.87**	**56.29**	**48.23**	69.76	**86.32**	**60.82**	37.68	**30.93**	**17.76**	**51.69**	**86.95**	**81.85**	**62.49**

Table 2. Zero-shot **robust** accuracy (%). Adversarial samples are generated by the PGD attack of 20 steps with the image-level perturbation radius $\epsilon_X = 1/255$. Adversarial fine-tuning is conducted on ImageNet, followed by robustness evaluations over 15 datasets.

Method		Image Net	STL10	CIFAR 10	CIFAR 100	SUN 397	Stanford Cars	Food 101	Oxford Pet	Flower 102	DTD	Euro Sat	FGVC	PCAM	Caltech 101	Caltech 256	Average
CLIP	2021 [36]	1.48	38.50	10.56	4.85	1.21	0.27	6.94	3.79	1.38	3.03	0.05	0.00	0.08	22.04	14.00	7.21
TeCoA	2023 [32]	41.48	83.50	60.08	34.16	31.55	13.08	27.28	62.80	28.80	22.71	16.58	5.88	26.81	69.18	59.80	38.91
PMG-FT	2024 [42]	38.94	84.00	62.27	35.92	31.07	16.74	31.10	63.07	31.99	23.14	14.94	6.06	26.10	70.85	59.57	39.72
FARE	2024 [38]	29.80	84.40	65.03	38.98	25.59	17.44	32.05	56.70	29.88	24.05	10.15	4.30	22.51	69.40	58.63	37.93
AoS	2025 [14]	47.27	86.10	67.69	40.23	32.46	21.25	34.42	67.80	35.88	25.86	17.32	8.03	36.19	73.70	63.98	43.88
Ours		47.81	86.97	68.36	41.89	33.53	20.88	34.85	67.96	36.60	25.41	17.49	8.58	36.49	74.12	64.15	44.34
Ours (5 trees)		**49.09**	**88.16**	**69.78**	**42.66**	**34.16**	**22.25**	**35.89**	**68.05**	**37.02**	**26.78**	**18.89**	**9.90**	**37.82**	**74.99**	**65.41**	**45.39**

Objective Function. We formalize our objective as:

$$\mathcal{L} = \mathcal{L}' + \lambda_1 \mathcal{L}_{\text{vic}} + \lambda_2 \left(\mathcal{L}_{\text{gap}}^{\text{Label}} + \mathcal{L}_{\text{gap}}^{\text{Intra}}\right), \tag{14}$$

where λ_1 and λ_2 are tunable factors. Unlike methods [32, 38, 42], we also optimize the text encoder projection layer to enable better hierarchical alignment. In inference, we use the same evaluation protocol with previous adversarial fine-tuning approaches based on the image-text alignment.

Theoretical Analysis. The key to adversarial robustness of our approach is that we obtain more general perturbations $\boldsymbol{\delta}_X$ and $\boldsymbol{\delta}_T$ against our hierarchical classifier which enjoys several margins with size varying in $(0, \infty)$—which is not achievable in the Euclidean model (see Fig. 2c).

Definition 3. *SoftMax,* $\exp\left[-\lambda\, d_r(\boldsymbol{\phi}, \boldsymbol{\psi}_c)\right] / \left(\sum_c \exp\left[-\lambda\, d_r(\boldsymbol{\phi}, \boldsymbol{\psi}_{c'})\right]\right)$*, can be transformed into the log margin*

$$-\, m_r^{c,c'} = \lambda\, d_r(\boldsymbol{\phi}, \boldsymbol{\psi}_c) + \underbrace{\log\left(\sum_{c' \neq c} \exp\left[-\lambda\, d_r(\boldsymbol{\phi}, \boldsymbol{\psi}_{c'})\right]\right)}_{-\lambda \min_{c' \neq c} d_r(\boldsymbol{\phi}, \boldsymbol{\psi}_{c'})} \tag{15}$$

via $-\log(\cdot)$*. Here,* $\min_{c \neq c'}(\cdot)$ *holds approximately,* i.e.*, temperature* $\lambda > 0$ *should be sufficiently large. Then, the log margin* m_r *for the Riemannian distance is:*

$$m_r^{c,c'}(\eta) = \frac{\lambda}{\sqrt{r}} \Big[\operatorname{acosh}\big(1 + \alpha(\eta)\, \beta_{c'}\big) - \operatorname{acosh}\big(1 + \alpha(\eta)\, \beta_c\big) \Big], \tag{16}$$

where $\|\boldsymbol{\phi}\|_2 = \|\boldsymbol{\psi}_c\|_2 = \eta, \forall c$ *is the norm controlling level in hierarchy for vision/text embeddings* $\boldsymbol{\phi}$ *and* $\boldsymbol{\psi}_c$*.* $\alpha(\eta) = \frac{4r\eta^2}{(1-r\eta^2)^2}$*.* $\beta_c = 1 - \cos(\theta_c)$ *&* $\beta_{c'} = 1 - \cos(\theta_{c'})$ *are complements of cosine of angles* $\theta_c = \measuredangle(\boldsymbol{\phi}, \boldsymbol{\psi}_c)$ *&* $\theta_{c'} = \measuredangle(\boldsymbol{\phi}, \boldsymbol{\psi}_{c'})$*.*

Theorem 1 (Margin size trends (Fig. 2c))**.** *Following Def. 3, if* $\eta \ll 1/(2\sqrt{r})$ *then* m_r *grows linearly as* $\alpha(\eta) \propto \eta^2$*: the known approx.* $\operatorname{acosh}(1+u) \approx \sqrt{2u}$ *holds for* $\alpha(\eta) \approx 4r\eta^2 \ll 1$*. Then* $m_r^{c,c'}(\eta) \approx \frac{\lambda}{\sqrt{r}} \sqrt{2\alpha(\eta)} \left(\sqrt{\beta_{c'}} - \sqrt{\beta_c}\right)$*. For the maximum angle separation* $\big(\cos(\theta_c), \cos(\theta_{c'})\big) = (1, 0)$ *we get* $(\beta_c, \beta_{c'}) = (0, 1)$ *yielding linear* $m_r(\eta) \propto \frac{\lambda}{\sqrt{r}} \big(\sqrt{2}\big)\eta$*.*

When $\eta \to 1/\sqrt{r}$*,* $\alpha(\eta) \to \infty$*: the known approximation* $\operatorname{acosh}(1+u) \approx \log(2u)$ *for large* u *yields* $m_r^{c,c'} \approx \frac{\lambda}{\sqrt{r}} \log \frac{\beta_{c'}}{\beta_c}$ *which does not depend any more on the norm* η *but depends on angles* $(\theta_c, \theta_{c'})$*. For* $(\beta_c, \beta_{c'}) \to (0, 1)$ *one gets* $m_r \to \infty$*.*

In contrast, under $d_{r \to 0^+}(\mathbf{u}, \mathbf{v}) \propto \|\mathbf{u} - \mathbf{v}\|_2$ *(Euclidean metric) and* $m_{Euc}(\eta) = 2\lambda\eta^2\big(\beta_{c'} - \beta_c\big)$ *so for* $(\beta_c, \beta_{c'}) \to (0, 1)$ *one gets* $m_{Euc}(\eta) = 2\lambda\eta^2$*.*

Discussion. Theorem 1 and Fig. 2c show that the maximum viable margin size in hyperbolic space ranges in $(0, \infty)$ w.r.t. feature norm $\eta \to 1/\sqrt{r} = 1$ (unlike Euclidean classifier's range (0,1)). We deploy image/text features in hyperbolic space by ordering them hierarchically from generic to specialized—norms of parents are lower than norms of children. Thus, several levels of hierarchy enjoy different limits on margin sizes, which are known to control classifier generalization. As adversarial perturbations $\boldsymbol{\delta}_X$ and $\boldsymbol{\delta}_T$ encounter several margins of different sizes during the projected gradient ascent, they become more universal. It is easier to increase parent classifier's adversarial risk due to its smaller margin size, which yields more universal perturbations due to more universal parent's label space. On children's classifier the same perturbation becomes specialized, gaining general-to-specialized robustification capability.

Table 3. Average PGD-20 robust accuracy (%) over 15 datasets across a range of perturbation radii during **evaluation only**.

Method	PGD-20 Robust Accuracy			
	1/255	2/255	3/255	4/255
TeCoA [32]	38.91	25.43	14.72	8.38
PMG-FT [42]	39.72	23.38	12.70	6.58
FARE [38]	37.93	24.87	13.37	7.74
AoS [14]	43.88	26.70	15.83	9.51
Ours	**44.34**	**27.19**	**16.05**	**9.92**

Table 4. Average performance (%) using adversaries of varying ϵ_X for both **adversarial fine-tuning and robustness evaluations**.

Radius ϵ	Method	Clean	PGD	CW	AA
2/255	TeCoA [32]	49.10	26.86	26.07	25.33
	PMG-FT [42]	50.72	29.38	28.41	27.66
	FARE [38]	51.09	28.55	27.79	27.18
	AoS [14]	51.86	30.48	29.45	28.70
	Ours	**52.27**	**31.64**	**30.55**	**29.49**
3/255	TeCoA [32]	42.90	19.59	18.69	17.94
	PMG-FT [42]	43.09	22.56	20.61	18.12
	FARE [38]	43.24	22.30	20.85	18.32
	AoS [14]	43.98	23.64	21.78	19.16
	Ours	**44.30**	**24.57**	**22.60**	**19.98**
4/255	TeCoA [32]	37.67	14.80	13.75	12.96
	PMG-FT [42]	37.84	17.03	15.18	13.29
	FARE [38]	37.95	16.57	14.21	13.43
	AoS [14]	38.49	18.38	16.40	14.05
	Ours	**38.75**	**19.04**	**16.92**	**14.71**

4. Experiments

Datasets. Following previous works [32, 42], we adversarially fine-tune CLIP using the ImageNet training set [6]. We then evaluate the fine-tuned CLIP model on the ImageNet test set and an additional 14 zero-shot datasets spanning various image recognition tasks (details in Appendix B.1).

Implementation. Unless specified otherwise, we employ CLIP [36] with the ViT-Base/32 architecture [22], consistent with previous works [32, 42]. Superclasses are constructed using the ImageNet hierarchy up to depth $L = 5$ for both image and text modalities. When using hierarchical forests (*e.g.*, 2-5 category trees), the loss function in Eq. (14) is applied per tree, and resulting losses are added.

During training, we generate adversarial samples using PGD [31] for images and texts of diverse abstraction levels, performing 3 iterations under the ℓ_∞-norm threat model. For adversarial images, we set the perturbation radius and step size to $\epsilon_X = \alpha_X = 1/255$, unless noted otherwise. For text-level perturbations, we set the maximum perturbation radius to $\epsilon_X = 2 \times 10^{-4}$ and the step size to $\alpha_T = 1 \times 10^{-4}$. During testing, in addition to natural performance, we evaluate our model against three strong white-box adversarial attacks: PGD [31], CW [4], and Auto-Attack [5]. All evaluations are conducted under adaptive attack schemes for fairness. More implementation details are in Appendix B.2.

Benchmarking on 15 datasets. Tables 1 & 2 present a comparison of our method with state-of-the-art adversarial fine-tuning techniques. Apart from ImageNet, we perform zero-shot inference on 14 additional datasets, reporting accuracy on both clean images and their adversaries generated via 20-step PGD attacks. Our method consistently achieves higher clean accuracy across all datasets, with an average improvement of 2.5% over FARE [38], narrowing the gap with standard CLIP (Table 1) as expected. While the standard CLIP has limited robustness (Table 2), our method enjoys a mean gain of 6.4% over FARE in robustness. With 5 class tress, we also surpass recent AoS [14] by 1% and 1.5% on clean/robust accuracies despite AoS uses 10x more image and 50x text augmentations for subspaces.

Table 5. Average performance (%) across different vision architectures of CLIP using the perturbation radius of $\epsilon_X = 1/255$.

Architecture	Method	Clean	PGD	CW	AA
ViT-B	TeCoA [32]	52.62	38.91	37.85	37.62
	PMG-FT [42]	57.36	39.72	38.70	38.09
	FARE [38]	59.67	37.93	37.56	37.18
	AoS [14]	61.70	43.88	42.94	42.18
	Ours	**62.14**	**44.34**	**43.58**	**42.72**
ViT-L	TeCoA [32]	66.39	42.86	39.08	38.43
	PMG-FT [42]	67.11	43.64	39.56	38.91
	FARE [38]	67.71	43.18	40.23	39.62
	AoS [14]	**68.41**	45.76	44.13	43.49
	Ours	68.38	**46.40**	**44.97**	**44.25**
ResNet-50	TeCoA [32]	42.12	28.45	27.72	27.13
	PMG-FT [42]	46.03	30.66	29.20	28.36
	FARE [38]	48.53	29.16	28.41	27.83
	AoS [14]	49.60	32.95	32.14	31.57
	Ours	**49.87**	**33.29**	**32.50**	**31.98**

Robustness across multiple attack intensities. Beyond evaluations with adversaries of a small attack strength ($\epsilon_X = 1/255$), we assess the zero-shot adversarial robustness under larger perturbation radii (stronger attack intensities). Table 3 shows that our method consistently outperforms other approaches across a spectrum of attack intensities.

Fine-tuning with diverse perturbation radii. Unlike previous studies [32, 42] that focus on a small perturbation radius of $\epsilon_X = 1/255$ during fine-tuning, we explore larger ℓ_∞-norm perturbation radii of $\epsilon_X = 2/255$, $3/255$, and $4/255$. For fair comparisons, fine-tuned models are evaluated against adversaries generated with *matching perturbation levels*. Table 4 shows that our method outperforms prior approaches on both clean and adversarial samples under fine-tuning with larger perturbation radii.

Robustness on different architectures. Below we experiment with architectures other than CLIP with a ViT-B backbone, specifically ViT-L and ResNet-50. Table 5 shows that our approach consistently surpasses previous adversarial fine-tuning methods across 15 datasets with larger backbones yielding better clean and adv. robustness results.

Adversarial fine-tuning with parameter-efficient VPT. Adversarial fine-tuning on all parameters of the image encoder is computationally costly. Thus, we employ Visual

Table 6. Average performance (%) using adversaries of varying ϵ_X for both fine-tuning and robustness evaluations with **VPT**.

Radius ϵ	Method	Clean	PGD	CW	AA
1/255	TeCoA [32]	51.00	32.27	31.11	30.26
	PMG-FT [42]	52.64	33.09	32.10	30.83
	FARE [38]	52.75	32.69	31.58	30.64
	AoS [14]	54.43	34.38	33.27	32.05
	Ours	**54.70**	**34.97**	**33.56**	**32.29**
2/255	TeCoA [32]	42.61	18.12	16.88	15.39
	PMG-FT [42]	42.11	19.26	17.68	16.47
	FARE [38]	42.81	18.98	17.46	16.35
	AoS [14]	43.84	20.47	**18.60**	17.29
	Ours	**44.13**	**20.65**	18.49	**17.54**
3/255	TeCoA [32]	33.86	12.32	10.78	8.89
	PMG-FT [42]	32.52	12.87	11.36	9.38
	FARE [38]	33.70	12.47	10.92	9.04
	AoS [14]	35.13	13.79	12.42	10.14
	Ours	**35.38**	**14.30**	**12.81**	**10.49**
4/255	TeCoA [32]	26.78	11.04	9.87	7.19
	PMG-FT [42]	23.57	11.73	10.01	7.26
	FARE [38]	26.17	11.49	10.32	7.53
	AoS [14]	27.92	13.17	11.50	8.87
	Ours	**28.40**	**13.54**	**11.97**	**9.28**

Table 7. Average robust accuracy (%) of diverse adversarial fine-tuning methods against text-level and bi-level adversarial attacks.

Method	Text-Level Attacks		Bi-Level Attacks	
	BERT-Attack	GBDA	Co-Attack	SGA
TeCoA [32]	36.22	34.97	25.87	25.14
PMG-FT [42]	37.25	36.73	26.95	26.76
FARE [38]	35.76	35.08	25.10	24.87
AoS [14]	40.83	40.37	30.24	29.51
Ours	**41.78**	**41.55**	**31.76**	**31.13**

Prompt Tuning (VPT) [26], a parameter-efficient strategy that introduces a minimal number of learnable parameters at the token embedding level. Table 6 shows that we achieve superior zero-shot accuracy (clean/robust) using robust VPT models with various configurations, trained and tested using the same perturbation radius for consistency.

Robustness to text-level and bi-level adv. attacks. Below we include text- and bi-level adversarial attacks. We report: (i) *text-level adversarial attacks* by BERT-Attack [28], (ii) Gradient-Based Distributional Attack (GBDA) [25], (iii) *bi-level adversarial attacks* using Collaborative Multimodal Adversarial Attack (Co-Attack) [46] and (iv) Set-level Guidance Attack (SGA) [30]. Table 7 (robust accuracy) shows that our method significantly boosts robustness against competitors on text-level and bi-level attacks.

Extension to BLIP/CoCa. We extend our model to BLIP [27], and CoCa [43] for vision-language understanding tasks such as image-text retrieval and image captioning. Following the settings from Appendix B.2, we report zero-shot performance (clean and adversarial) from the Flickr30k dataset [35] for image-text retrieval and the Nocaps dataset [1] for image captioning. Table 8 shows that our method has consistently best results across diverse tasks.

Table 8. Extension to BLIP/CoCa evaluated on clean samples and PGD-20 adversaries in image-text retrieval and image captioning.

Architecture	Method	Image-Text Retrieval				Image Captioning	
		Clean TR	Robust TR	Clean IR	Robust IR	Clean CIDEr	Robust CIDEr
BLIP	TeCoA [32]	87.5	54.4	77.0	47.5	96.9	57.8
	PMG-FT [42]	87.8	55.6	77.9	48.2	97.5	58.2
	FARE [38]	88.2	55.9	78.4	49.0	98.1	58.7
	Ours	90.8	57.9	80.2	51.4	101.0	62.5
	Ours (2 trees)	**90.9**	**58.3**	**80.4**	**51.9**	**101.4**	**63.0**
CoCa	TeCoA [32]	88.4	56.5	76.8	48.3	100.5	59.2
	PMG-FT [42]	88.2	57.3	76.3	49.7	102.1	60.2
	FARE [38]	89.5	57.0	77.5	49.8	101.8	60.5
	Ours	91.9	59.4	79.6	52.6	105.4	63.8
	Ours (2 trees)	**92.3**	**60.2**	**79.8**	**53.5**	**105.7**	**64.6**

Table 9. Extensions to the medical CLIP on clean and PGD-20 adversarial samples evaluated by the AUC score.

Method	ChestXray14		CheXpert		PadChest	
	Clean	PGD	Clean	PGD	Clean	PGD
TeCoA [32]	0.674	0.526	0.857	0.685	0.602	0.483
PMG-FT [42]	0.692	0.538	0.850	0.688	0.619	0.495
FARE [38]	0.687	0.533	0.845	0.679	0.615	0.490
AoS [14]	0.718	0.572	0.883	0.719	0.643	0.531
Ours	0.727	0.563	0.874	0.716	0.637	0.554
Ours (2 trees)	**0.730**	**0.571**	**0.876**	**0.724**	**0.639**	**0.562**

Extension to Medical CLIP. Below we apply our method to medical imaging, addressing adversarial vulnerabilities in computer-aided diagnosis [49]. Following the experimental protocols in Appendix B.2, we assess the zero-shot adversarial robustness of chest X-rays with multi-label disease annotations. Table 9 shows our approach consistently achieves superior AUC for clean samples and adversaries across various radiology datasets. Even on the challenging PadChest dataset, which includes a long-tail distribution of 192 diseases, our method exhibits notable robustness.

4.1. Analysis

Impact of loss components. We study: (i) Hierarchy-preserving Image-Text Alignment (HITA) $\mathcal{L}'$ from Eq. (10), (ii) intra-class vicinity alignment $\mathcal{L}_{\text{vic}}$ in Eq. (12), and (iii) intra-class and label norm gap penalties $\mathcal{L}_{\text{gap}}^{\text{Intra}}$ & $\mathcal{L}_{\text{gap}}^{\text{Label}}$ from Eq. (13) & (11). Table 10 reports the average zero-shot accuracy (clean and adversarial) on 15 datasets. Our baseline (top row) is based on TeCoA [32]. Incorporating hierarchy-level image-text alignment boosts clean/robust accuracy due to rich hierarchical decision boundary. The vicinity/intra-class penalties mitigate image embedding collapse on text embeddings. The label norm gap promotes the hierarchical integrity.

Table 10. Ablation study of three loss components in our method for average clean and robust accuracy (%) on 15 datasets.

	HITA	vic	des	Clean	PGD	AA
1				52.62	38.91	37.62
2	✓			59.27	42.08	40.59
3	✓	✓		**62.35**	43.67	42.18
4	✓		✓	60.06	43.15	41.54
	✓	✓	✓	**62.14**	**44.34**	**42.72**

Different negative set augmentation strategies. Recall

Table 11. Performance (%) of diverse negative set augmentations.

Negative Set Augmentation Strategy	Clean	PGD	AA
w/o augmentation	60.88	43.16	41.47
Lower-level superclasses	61.25	43.62	41.93
Higher-level superclasses	61.70	43.83	42.05
Higher- & lower-level superclasses	**62.14**	**44.34**	**42.72**

Table 12. Performance (%) of diverse adversary generation strategies for adv. fine-tuning with the average training time per epoch.

Adversary Generation Strategy	Clean	PGD	AA	Time (min)
Leaf-level disruption only	60.25	41.59	40.23	70.1
Diverse hierarchical perturbations	61.37	42.35	40.88	154.2
Universal hierarchical perturbation	**62.14**	**44.34**	**42.72**	73.2

Table 13. Performance (%) of LLMs for superclass generation.

LLM for Superclass Generation	Clean	Image Attack		Text Attack		VLM Attack	
		PGD	AA	BERT-Attack	GBDA	CoAttack	SGA
LLaMA-2	61.87	43.55	42.05	40.91	40.72	31.15	30.48
Claude-2	61.83	43.79	42.32	41.35	41.11	31.60	30.86
ChatGPT-4o	**62.14**	**44.34**	**42.72**	**41.78**	**41.55**	**31.76**	**31.13**

Table 14. Performance (%) of different BLIP adaptation strategies.

Strategy	Method	Image-Text Retrieval				Image Captioning	
		Clean TR	Robust TR	Clean IR	Robust IR	Clean CIDEr	Robust CIDEr
Fine-Tuning Target VLMs	FARE [38]	88.2	55.9	78.4	49.0	98.1	58.7
	Ours	**90.8**	**57.9**	**80.2**	**51.4**	**101.0**	**62.5**
Replacing Image Encoder Only	FARE [38]	82.1	49.7	73.8	45.2	93.2	54.3
	Ours	**85.9**	**55.4**	**78.6**	**49.8**	**98.3**	**58.9**

that we design a hierarchy-aware negative augmentation scheme in Eq. (9) based on the hyperbolic embeddings from both higher- and lower-level superclasses. Table 11 ablates different negative set augmentation strategies, and shows that integrating superclasses at hierarchical levels up+below the positive class works the best.

Different adversary generation strategies. Below we explore three adversary generation strategies:

- **Leaf-level disruption only**: Adversarial perturbations target solely the base category classification.
- **Diverse hierarchical perturbations**: Perturbations are crafted at each hierarchy level for all original samples.
- **Universal hierarchical perturbation**: A single perturbation is applied universally across all hierarchy levels.

Table 12 reports zero-shot clean and robust accuracy with the training time. The universal strategy offers a favorable trade-off between robustness and computational efficiency. During testing, all methods enjoy the same inference time.

Different LLMs in superclass generation. Recall that we generate superclasses using ChatGPT-4o [34] in the absence of predefined hierarchical taxonomies in datasets other than ImageNet. Table 13 investigates superclass generation with powerful LLMs (LLaMA-2 [41], Claude-2 [2]) and shows that our method is robust to these diverse choices.

Fine-tuning target VLM *vs*. replacing image encoder. We primarily fine-tune the target VLM (*e.g.*, BLIP) instead of replacing the image encoder as in FARE [38]. Table 14 compares such two settings. Our method enjoys better natural performance/adversarial robustness in all cases.

Table 15. Accuracy (%) of image-text adversaries in fine-tuning.

Adversary Generation During Fine-Tuning	Clean	Image Attack		Text Attack		VLM Attack	
		PGD	AA	BERT-Attack	GBDA	CoAttack	SGA
Image Level Only	61.27	43.42	41.83	40.53	40.28	30.85	30.19
Image and Text Levels	**62.14**	**44.34**	**42.69**	**41.78**	**41.55**	**31.72**	**31.13**

Table 16. (Auto-Attack) robust accuracy (%) w.r.t. different settings. We also report the robust accuracy against adversarial images *transferred* between base class and superclass classification.

Type	Base Class			Superclass		
	Clean	Robust	*Transfer*	Clean	Robust	*Transfer*
Baseline (TeCoA)	52.62	37.62	46.58	61.80	47.20	55.27
Superclass Classification at each level (Euclidean)	53.24	38.16	48.09	68.31	55.38	63.94
Hyperbolic Space Modeling	**62.14**	**42.72**	**63.34**	**71.68**	**57.13**	**66.40**

Image-text *vs*. image-only adversaries. As VLM is based on image-text alignment, joint image-text adversaries carry higher boundary risk [44] than image- or text-only adversaries [32, 47] so they can robustify the model better. The image-level adv. generation uses PGD, while the text-level generation relies on tokenized embedding to avoid discrete optimization. We use the same adv. generation scheme with other methods during testing. Table 15 shows fine-tuning on image-level adversaries only also improves robustness.

Hyperbolic *vs*. Euclidean model. Next we analyze robust accuracy in superclass classification. We assess adversarial transferability between superclass and base-class. We also introduce superclass classification in the Euclidean space by averaging feature vectors: $(\mathbf{v}_1, \cdots, \mathbf{v}_N) \rightarrow \frac{1}{N}\sum_i \mathbf{v}_i$. Table 16 shows that standard adversarial fine-tuning focused on base categories (Baseline) is low in superclass classification. While the Euclidean superclass classification enhances superclass robustness, it does not enhance base category robustness. In contrast, hierarchical modeling improves robustness in both base and superclasses.

Hyper-parameter study is in Appendix D. **Weighting mechanisms** are studied in Appendix G. **Base class representations** are studied in Appendix H. **Robustness against black-box multimodal adversaries** is in Appendix I.

5. Conclusions

Motivated by our analysis of adversarial samples targeting superclasses, we have shown that the overemphasis on the base categories lead to poorer adv. robustness in zero-shot settings. Thus, we have leveraged hyperbolic geometry to capture class hierarchies. By making norms of parent embeddings lower than norms of children embeddings, we maintain the hierarchical integrity of embedding space. These norms directly impact the margin sizes across hierarchical levels, making each adversarial perturbation carry general-to-specific attack capability to boost robustification.

Acknowledgment

This research/project is supported in part by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-031), partly by the MTI under its AI Centre of Excellence for Manufacturing (AIMfg) (Award W25MCMF014), the College of Computing and Data Science, Nanyang Technological University (NTU), and the A*STAR-NTU Smart and Sustainable Advanced Manufacturing Research Joint Laboratory. This work was also supported by the CSIRO Allocation Scheme (Lead CI: Piotr Koniusz) and the UNSW Merit Allocation Scheme (Lead CI: Piotr Koniusz) for NCI.

References

[1] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957, 2019. 7

[2] Anthropic. Claude 2. Anthropic Blog, 2023. 8

[3] James W Cannon, William J Floyd, Richard Kenyon, Walter R Parry, et al. Hyperbolic geometry. *Flavors of geometry*, 31(59-115):2, 1997. 3

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017. 6

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 6

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 6

[7] Natalia Díaz-Rodríguez, Javier Del Ser, Mark Coeckelbergh, Marcos López de Prado, Enrique Herrera-Viedma, and Francisco Herrera. Connecting the dots in trustworthy artificial intelligence: From ai principles, ethics, and key requirements to responsible ai systems and regulation. *Information Fusion*, 99:101896, 2023. 1

[8] Junhao Dong, Seyed-Mohsen Moosavi-Dezfooli, Jianhuang Lai, and Xiaohua Xie. The enemy of my enemy is my friend: Exploring inverse adversaries for improving adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24678–24687, 2023. 2

[9] Junhao Dong, Lingxiao Yang, Yuan Wang, Xiaohua Xie, and Jianhuang Lai. Toward intrinsic adversarial robustness through probabilistic training. *IEEE Transactions on Image Processing*, 32:3862–3872, 2023.

[10] Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. Survey on adversarial attack and defense for medical image analysis: Methods and challenges. *ACM Computing Surveys*, 57(3):1–38, 2024.

[11] Junhao Dong, Piotr Koniusz, Junxi Chen, and Yew-Soon Ong. Adversarially robust distillation by reducing the student-teacher variance gap. In *European Conference on Computer Vision*, pages 92–111. Springer, 2024.

[12] Junhao Dong, Piotr Koniusz, Junxi Chen, Z Jane Wang, and Yew-Soon Ong. Robust distillation via untargeted and targeted intermediate adversarial samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28432–28442, 2024.

[13] Junhao Dong, Piotr Koniusz, Junxi Chen, Xiaohua Xie, and Yew-Soon Ong. Adversarially robust few-shot learning via parameter co-distillation of similarity and class concept learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28535–28544, 2024. 2

[14] Junhao Dong, Piotr Koniusz, Liaoyuan Feng, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Robustifying zero-shot vision language models by subspaces alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21037–21047, 2025. 5, 6, 7

[15] Junhao Dong, Piotr Koniusz, Xinghua Qu, and Yew-Soon Ong. Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, page 236–247, New York, NY, USA, 2025. Association for Computing Machinery. 1

[16] Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 14061–14078. PMLR, 2025. 1

[17] Junhao Dong, Jiao Liu, Xinghua Qu, and Yew-Soon Ong. Confound from all sides, distill with resilience: Multi-objective adversarial paths to zero-shot robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 624–634, 2025. 2

[18] Junhao Dong, Cong Zhang, Xinghua Qu, Zejun MA, Piotr Koniusz, and Yew-Soon Ong. Robust SuperAlignment: Weak-to-strong robustness generalization for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS 2025)*, 2025. 1

[19] Junhao Dong, Cong Zhang, Xinghua Qu, Zejun Ma, Piotr Koniusz, and Yew-Soon Ong. Robust superalignment: Weak-to-strong robustness generalization for vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

[20] Junhao Dong, Raoof Zare Moayedi, Yew-Soon Ong, and Seyed-Mohsen Moosavi-Dezfooli. Allies teach better than enemies: Inverse adversaries for robust knowledge distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

[21] Junhao Dong, Cong Zhang, Xinghua Qu, Sua Qi Rong, Nguyen Duc Thai, Wenbo Pan, Xinfeng Li, Tongliang Liu,

Piotr Koniusz, and Yew-Soon Ong. Tug-of-war no more: Harmonizing accuracy and robustness in vision-language models via stability-aware task vector merging. In *The Fourteenth International Conference on Learning Representations*, 2026. 2

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations, ICLR*, 2021. 6

[23] Xiang Fang and Wanlong Fang. Disentangling adversarial prompts: A semantic-graph defense for robust llm security. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. 1

[24] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR*, 2015. 1

[25] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. Gradient-based adversarial attacks against text transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 5747–5757, 2021. 7

[26] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 7

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 7

[28] Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing EMNLP*, pages 6193–6202, 2020. 7

[29] Lin Li, Haoyan Guan, Jianing Qiu, and Michael Spratling. One prompt word is enough to boost adversarial robustness for pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 3

[30] Dong Lu, Zhiqiang Wang, Teng Wang, Weili Guan, Hongchang Gao, and Feng Zheng. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 102–111, 2023. 1, 7

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR*, 2018. 2, 3, 6

[32] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations,ICLR*, 2023. 1, 2, 3, 5, 6, 7, 8

[33] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 1, 3

[34] OpenAI. Chatgpt [large language model]. https://chatgpt.com, 2024. 3, 8

[35] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 7

[36] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5, 6

[37] Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *Graph Drawing: 19th International Symposium, GD 2011, Eindhoven, The Netherlands, September 21-23, 2011, Revised Selected Papers 19*, pages 355–366, 2012. 2

[38] Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024. 1, 2, 5, 6, 7, 8

[39] Yanpeng Sun, Shan Zhang, Wei Tang, Aotian Chen, Piotr Koniusz, Kai Zou, Yuan Xue, and Anton van den Hengel. Math blind: Failures in diagram understanding undermine reasoning in MLLMs. In *The Fourteenth International Conference on Learning Representations*, 2026. 1

[40] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1

[41] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 8

[42] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 1, 2, 5, 6, 7

[43] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Trans. Mach. Learn. Res.*, 2022, 2022. 7

[44] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482. PMLR, 2019. 2, 8

[45] Jiliang Zhang and Chen Li. Adversarial examples: Opportunities and challenges. *IEEE transactions on neural networks and learning systems*, 31(7):2578–2593, 2019. 1

[46] Jiaming Zhang, Qi Yi, and Jitao Sang. Towards adversarial attack on vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5005–5013, 2022. 1, 7

[47] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *European Conference on Computer Vision*, pages 56–72. Springer, 2024. 8

[48] Yifei Zhang, Hao Zhu, Menglin Yang, Jiahong Liu, Rex Ying, Irwin King, and Piotr Koniusz. Understanding and mitigating hyperbolic dimensional collapse in graph contrastive learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, page 1984–1995, New York, NY, USA, 2025. Association for Computing Machinery. 2

[49] Zihao Zhao, Yuxiao Liu, Han Wu, Yonghao Li, Sheng Wang, Lin Teng, Disheng Liu, Xiang Li, Zhiming Cui, Qian Wang, et al. Clip in medical imaging: A comprehensive survey. *arXiv preprint arXiv:2312.07353*, 2023. 7