Hierarchically Robust Zero-shot Vision-Language Models
– *Supplementary Material* –

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

In this supplementary material, we commence with a comprehensive discussion of the pipeline of our proposed adversarial fine-tuning method based on our hyperbolic mechanism in Appendix A. Moreover, we elaborate on our experimental configurations, comprising both the dataset description and the implementation details (including the extension in the context of BLIP and Medical CLIP) in Appendix B. We also provide further details regarding our category-based hierarchy construction mechanism in Appendix C. Hyper-parameter analyses are presented in Appendix D. We then discuss the impact of diverse intra-class variability strategies (Eq. (12&13)) in Appendix E.

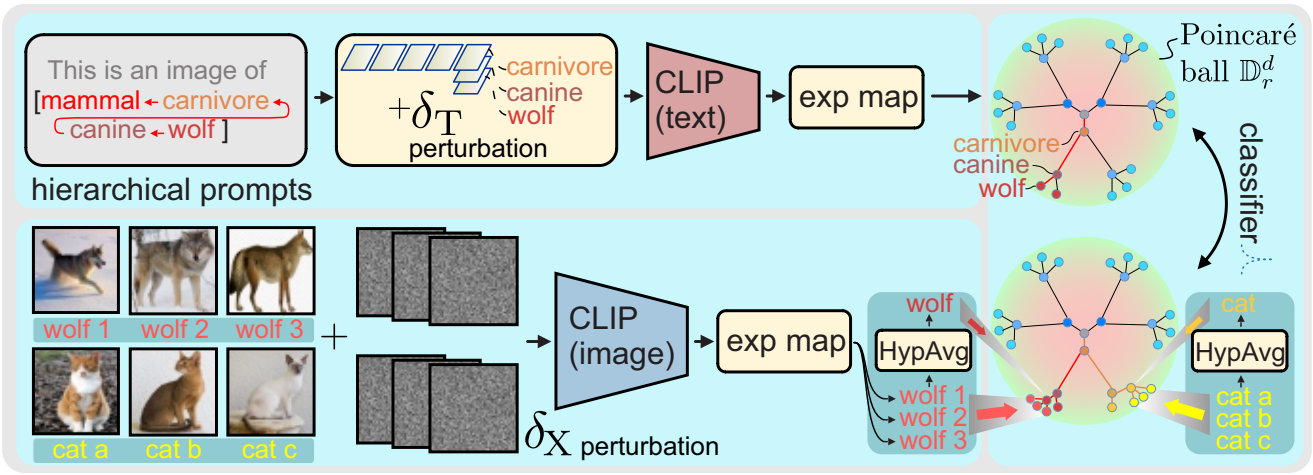


Figure 4. Our pipeline.

A. Pipeline

Figure 4 is a zoom of our pipeline in Figure 3. Two branches of CLIP are used, *i.e.*, text encoder and image encoder. For text, for each category, we look up its hierarchy in a predefined hierarchical tree (*e.g.*, for ImageNet, the categories follow WorldNet) and extract the path from the root all the way to the leaf category. For each category level, we form one text prompt and encode with CLIP that firstly tokenizes text [65, 79, 91, 102]. Notice $\boldsymbol{\delta}_{\mathrm{T}}$ is added to the contextual part of prompts for adversarial text learning. The exponential map elevates embeddings from the Euclidean space into the hyperbolic space (Poincaré ball).

For image, in each mini-batch, we have several images of the same leaf (base) category. We embed them via the image branch of the CLIP. Notice we add $\boldsymbol{\delta}_{\mathrm{X}}$ per image to learn adversarial perturbations. Subsequently, The exponential map

✉Corresponding authors.

elevates image embeddings from the Euclidean space into the hyperbolic space. Here, in order to obtain embedding representing *wolf*, we use Hyperbolic averaging. *i.e.*, $\mathrm{HypAvg}(\cdot)$. The same Hyperbolic averaging strategy is used to obtain further embeddings across hierarchical levels.

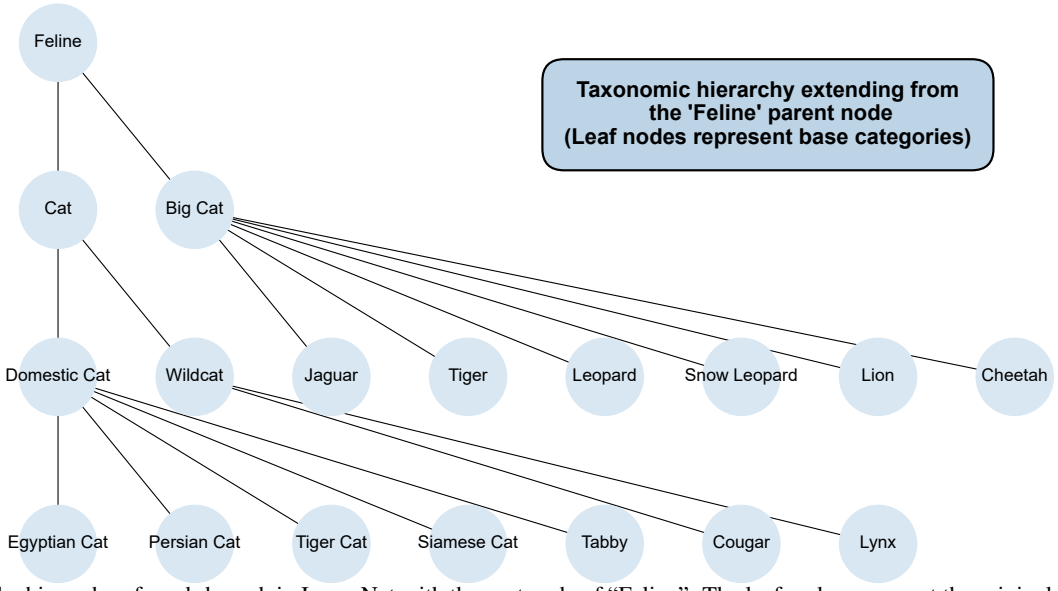


Figure 5. The hierarchy of a sub-branch in ImageNet with the root node of "Feline". The leaf nodes represent the original base classes.

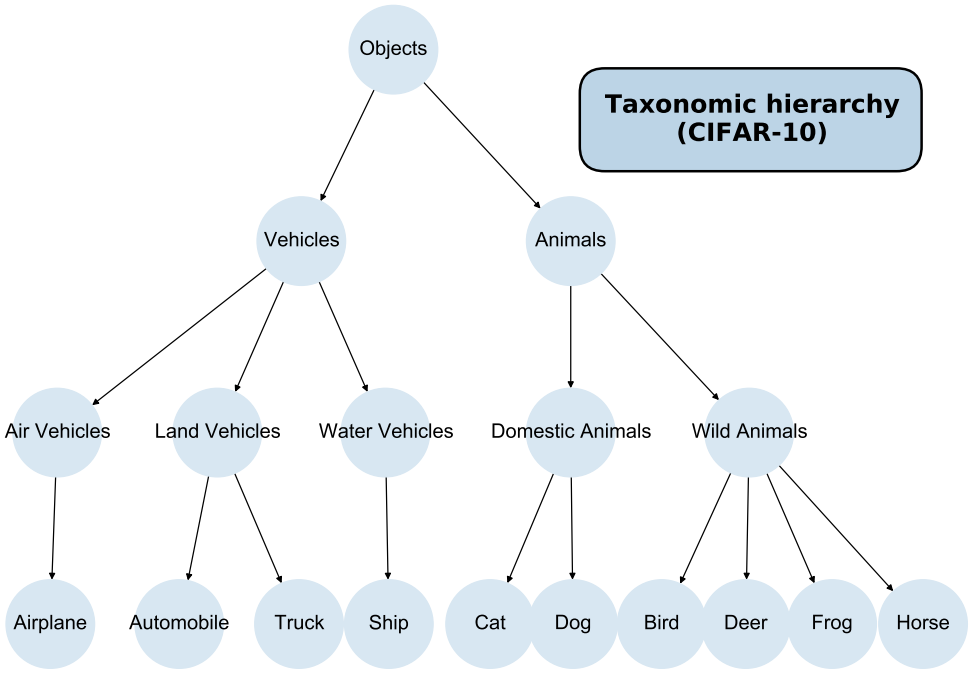


Figure 6. The hierarchy of CIFAR-10. The leaf nodes represent the original base classes.

B. Experimental Details

This section presents a comprehensive overview of the experimental configurations used in our study, including detailed information about the datasets used for adversarially robust fine-tuning and the corresponding implementation details of our proposed method.

B.1. Datasets

In accordance with the evaluation protocols established in prior research [83, 94], we conduct adversarial fine-tuning of the CLIP model using the ImageNet training dataset [57]. To systematically evaluate the robustness of our fine-tuned model, we test it on the ImageNet validation set—as ImageNet typically provides a validation split used for testing purposes, with a subset of the training data reserved for validation—and on an additional set of 14 zero-shot datasets that span a diverse array of image recognition tasks. Collectively, these 15 datasets are categorized into four groups:

- **General Image Classification**: ImageNet [57], STL-10 [55], CIFAR-10 and CIFAR-100 [72], Caltech-101 [60], and Caltech-256 [61].
- **Fine-Grained Classification**: FGVC Aircraft [82], Flower102 [86], Food101 [50], Oxford-IIIT Pets [88], and Stanford Cars [71].
- **Domain-Specific Classification**: Describable Textures Dataset (DTD) [54], EuroSAT [64], and PatchCamelyon (PCAM) [93].
- **Scene Recognition**: SUN397 [96].

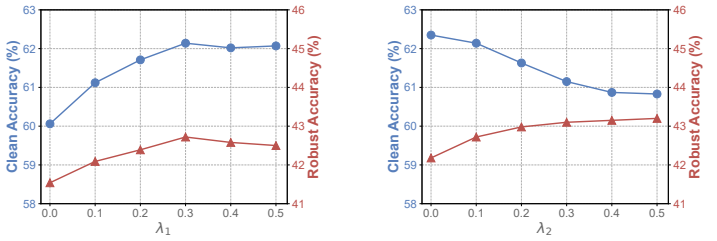
During adversarial fine-tuning, we apply standardized data preprocessing techniques to ensure consistency across all datasets. Specifically, each input image is resized to a resolution of 224×224 pixels, followed by a center crop operation. This preprocessing aligns with common practices in fine-tuning vision-language models and facilitates fair comparisons by maintaining uniform input dimensions.

B.2. Implementation

Standard setup. In line with the configurations used in prior studies [83, 94], we adopt the CLIP model [90] utilizing the Vision Transformer (ViT) architecture, specifically ViT-Base/32 [59]. For network parameter optimization during adversarial fine-tuning, we employ the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a batch size of 512. The learning rate is initialized at 1×10^{-5} and scheduled using cosine annealing for the fine-tuning of the whole vision encoder and the projection layer of the text encoder. When implementing Visual Prompt Tuning (VPT) [68], a parameter-efficient fine-tuning strategy, we introduce token-level learnable parameters of size 100 into the vision branch of CLIP and set the learning rate to 40. During training, adversarial examples are generated at both the image and text levels using Projected Gradient Descent (PGD) [81] with 3 iterations. For image-level adversarial perturbations, we adopt the ℓ_∞-norm threat model with a maximum perturbation radius of $\epsilon_{\text{X}} = 1/255$ and a step size of $\alpha_{\text{X}} = 1/255$, unless specified otherwise. For text-level adversarial perturbations—applied exclusively during fine-tuning—we set the step size to $\alpha_{\text{T}} = 1 \times 10^{-4}$ and the perturbation radius to $\epsilon_{\text{T}} = 2 \times 10^{-4}$. Superclasses are constructed using the ImageNet hierarchy up to $L = 5$ for both image and text modalities. We set the curvature of the hyperbolic space to $r = 1.0$. The projection hyper-parameter is set to $\xi = 1 \times 10^{-5}$ to prevent features from reaching the Poincaré disk boundary. The vicinity radius around text embeddings is configured as $\zeta_{\text{vic}} = 5 \times 10^{-2}$, and the margin factor is set to $\zeta_{\text{gap}} = 1 \times 10^{-2}$. To balance the contributions of different loss components, we assign the loss weighting factors as $\lambda_1 = 0.3$ and $\lambda_2 = 0.1$. Note that the setting of all the hyper-parameters is obtained through the Hyperopt package [49] for a 25-iteration hyper-parameter search on a 1% subset of the ImageNet training set. The hyper-parameter setting was then applied without tuning to adversarial fine-tuning of all other scenarios. All experiments are conducted on eight NVIDIA Tesla A100 GPUs.

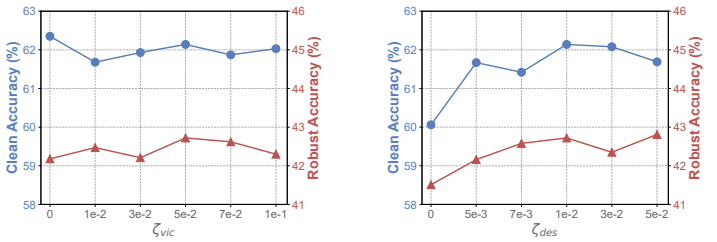
Evaluation protocol. Aligned with previous research on adversarially robust CLIP fine-tuning [83, 94], we focus on evaluating robustness against three strong white-box adversarial attacks: 20-step PGD [81], the Carlini and Wagner (CW) attack [53], and Auto-Attack (AA) [56]. In addition to image-level attacks, we also assess robustness against *text-level attacks* such as BERT-Attack [76] and Gradient-Based Distributional Attack (GBDA) [62], as well as *bi-level attacks* using Collaborative Multimodal Adversarial Attack (Co-Attack) [98] and Set-level Guidance Attack (SGA) [80], which are discussed and evaluated in the main text.

Experimental settings for BLIP/CoCa extension. To assess our method's zero-shot robustness on downstream tasks, we expand our experiments to integrating the BLIP architecture [75], a large-scale vision-language model that unifies vision-language understanding tasks through bootstrapped language-image pre-training. Specifically, we evaluate zero-shot adversarial robustness on two tasks: image-text retrieval and image captioning. Following the pipeline in [75], we adversarially optimize the Image-Text Contrastive (ITC) loss, Image-Text Matching (ITM) loss, and Language Modeling (LM) loss to obtain a robust version of BLIP. For the CoCa architecture [97], we adversarially optimize the ITC loss. Subsequently, we assess robustness by applying the iterative PGD attack method, using the ITM loss for image-text retrieval and the LM loss for image captioning. We follow the same perturbation radius ($\epsilon_{\text{V}} = 1/255$ and $\epsilon_{\text{T}} = 2 \times 10^{-4}$) during fine-tuning as in our original manuscript, with the sole modification being the replacement of the objective function for adversary generation.



(a) (b)

Figure 7. Hyper-parameter (λ_1 and λ_2) sensitivity of our method on average clean and (Auto-Attack) robust accuracy (%).



(a) (b)

Figure 8. Hyper-parameter (ζ_{vic} and ζ_{gap}) sensitivity of our adversarial fine-tuning method on average clean and (Auto-Attack) robust accuracy (%) across 15 datasets in the zero-shot setting.

Experimental settings for medical CLIP extension. To expand our empirical analyses for robust medical imaging, we utilize a CLIP model pre-trained specifically on radiology datasets following the CheXzero framework [92] with the architecture of ViT-B/16. In line with established protocols [73, 89, 92], we utilize the MIMIC dataset [69]—a systematic database of chest radiographs paired with detailed radiology reports—for adversarial fine-tuning. The text encoder in our CLIP model leverages BioBERT [74], a specialized biomedical language model optimized for text mining for biomedical analysis. In zero-shot scenarios, we evaluate the robust CLIP models on three widely-used multi-label radiology datasets: ChestX-ray14 [95], CheXpert [67], and PadChest [51]. We report the Area Under the Curve (AUC) metrics for both clean and adversarial images. The adversaries are generated based on a 20-step PGD approach with $\epsilon_{\text{V}} = 1/255$ and $\epsilon_{\text{T}} = 2 \times 10^{-4}$.

C. Hierarchy Construction

For the ImageNet dataset [57], we leverage the hierarchical taxonomy provided by WordNet [84] to construct superclasses by extracting hypernyms of synsets. This hierarchical structure enables us to represent semantic relationships between concepts at multiple levels of abstraction. For instance, Figure 5 illustrates a sub-branch of the ImageNet hierarchy rooted at the node *Feline*. In this hierarchy, each base category—such as a specific species of domestic cat—, is connected to its immediate superclass, recursively forming a tree that ascends to more abstract concepts like *Mammal* and *Animal*. This method allows us to obtain superclasses across diverse hierarchical levels, effectively capturing varying granularities of class concepts and enriching the semantic context for each category.

In scenarios where datasets lack predefined hierarchical taxonomies, we generate superclasses using a Large Language Model (LLM), specifically ChatGPT-4o [87]. We design prompts that instruct the LLM to suggest contextually appropriate superclasses for each base class. For instance, we could query, ``Provide a general category-based hierarchy that is built upon all the [base classes].'' To ensure the generated superclasses are semantically coherent and align with the dataset's context, we perform manual reviews and validations. This process involves cross-referencing the suggested superclasses with domain knowledge and existing literature to confirm their relevance and accuracy. By systematically applying this approach, we establish consistent hierarchical structures across diverse datasets, which facilitates the implementation of our hyperbolic space modeling for robust image-text alignment. For instance, the hierarchical tree of categories for the CIFAR-10 dataset [57] is illustrated in Figure 6.

D. Hyper-parameter Analyses

For a better understanding of our method, we analyze how the integrated hyperparameters affect performance. Specifically, in Figure 7, we present the average accuracy on both clean samples and those attacked by AutoAttack across 15 datasets in

the zero-shot setting. The results show that increasing the hyperparameter λ_1, which circumvents the inductive bias towards static alignment reference, leads to an improved trade-off between accuracy and robustness. Meanwhile, enhancing the value of λ_2 results in improved adversarial robustness but a corresponding decrease in natural performance. This indicates that the weighting factor λ_2 associated with the norm descent constraints in hyperbolic space regulates the balance between natural accuracy and adversarial robustness.

In addition to the loss weighting factors λ_1 and λ_2 we discussed above, we explore the impact of other hyper-parameters in our hierarchical adversarial fine-tuning approach. As shown in Figure 8, we report both clean and (Auto-Attack) robust accuracy of our method under diverse hyper-parameter settings. Note that all hyper-parameters in this analysis were tuned based on a tiny subset of the ImageNet training set to ensure fairness. This hyper-parameter setting is then applied to the adversarial fine-tuning of diverse scenarios. We can observe that appropriately choosing the margin factors ζ_{vic} and ζ_{gap} for adversarial fine-tuning can lead to a reasonable trade-off between natural performance and adversarial robustness in the zero-shot setting.

E. Impact of Intra-class Variability

In Section 3, we discussed how incorporating intra-class variability can mitigate the inductive bias toward static alignment references, leading to an improved trade-off between accuracy and robustness. Specifically, we align adversarial image embeddings within the upper arc vicinity of the base class text embedding while maintaining the hierarchical order of these embeddings. Extending this concept, we also explore the impact of intra-class variability for clean embeddings, as presented in Table 17. Our findings indicate that while introducing intra-class variability for clean embeddings can modestly enhance natural performance, it results in a substantial decline in adversarial robustness in the zero-shot setting. Therefore, we choose to apply intra-class variability exclusively to adversarial embeddings for each input image.

Table 17. Zero-shot clean and robust accuracy (%) w.r.t. diverse intra-class variability configurations.

Intra-Class Variability Setting	Clean	PGD	AA
Clean & Adversarial Embeddings	62.56	41.30	40.12
Adversarial Embeddings Only	**62.14**	**44.34**	**42.72**

F. Impact of Augmenting the Hierarchical Trees to Forests

Recall that we extend our taxonomic hierarchy by constructing multiple hyperbolic trees based on diverse taxonomic structures obtained through LLMs. Specifically, we query LLMs to generate alternative hierarchical taxonomies, capturing different semantic relationships among categories. By incorporating multiple hierarchical trees, we aim to enhance robustness by leveraging richer multi-level abstractions. Table 18 presents a comparative analysis of model performance with varying numbers of hyperbolic trees. Our findings indicate that increasing the number of trees consistently improves zero-shot adversarial robustness while maintaining competitive clean accuracy. However, we observe diminishing returns beyond a certain number of trees, with a trade-off in training efficiency as the computational cost per epoch increases. This suggests that an optimal balance exists between hierarchical complexity and computational feasibility for robust fine-tuning in hyperbolic space. Note that the multiple hierarchy trees do not affect the test-time inference efficiency.

Table 18. Performance (%) of different numbers of the hyperbolic trees with the average training time per epoch.

Number of Hyperbolic Trees	Clean	PGD	AA	Time (min)
1	62.14	44.34	42.72	73.2
3	62.36	44.78	42.97	88.0
5	62.49	45.39	43.18	97.4
10	62.38	45.52	43.14	127.5

G. Different Weighting Mechanisms

We explore the efficacy of using re-weighting for hierarchical image-text alignment in Eq. (10), comparing equal weighting and linear weighting ($\omega_l = 1 - \frac{l}{L+1}$). Table 19 shows that emphasizing lower-level hierarchies (more fine-grained information) enhances zero-shot adversarial robustness. Lower (finer) level provides harder negatives for stronger semantic separation (crucial in improving robustness). We also include using only top 1-3 i) fine-grained levels, ii) superclass levels.

Table 19. Performance (%) of different weighting mechanisms.

Hierarchical Weighting Strategy	Clean	PGD	AA
Equal Weighting	62.08	43.59	41.96
Top 1-3 superclass levels	61.82	43.71	41.84
Top 1-3 fine-grained levels	62.05	43.87	42.11
Linear Weighting	**62.14**	**44.34**	**42.72**

H. Different Strategies to Represent Base Classes.

We explore three strategies for representing base-class embeddings during fine-tuning: (i) selecting a random instance from the base class, (ii) averaging in the Euclidean space followed by projection into the hyperbolic space, and (iii) hyperbolic averaging. Table 20 shows that aggregating via hyperbolic averaging produces the best zero-shot performance.

Table 20. Performance (%) of diverse base class representations.

Base Class Representation	Clean	PGD	AA
Random Instance	58.92	41.65	40.12
Euclidean Averaging	61.48	43.72	42.25
Hyperbolic Averaging	**62.14**	**44.34**	**42.72**

I. Robustness Against Black-box Adversarial Attacks

In addition to zero-shot robustness evaluations against strong white-box attacks, we also analyze the black-box robustness of diverse adversarial fine-tuning methods using two practical multi-modal adversarial attacks: AFS [66] and MF-it [101], as presented in Table 21. Our method consistently achieves the best black-box robustness.

Table 21. Performance (%) on black-box multimodal adversaries.

Evaluation Method	TeCoA [83]	PMG-FT [94]	FARE [91]	**Ours**
AFS [66]	52.09	53.62	53.10	**57.92**
MF-it [101]	51.76	53.25	53.47	**57.38**

J. The need for Hyperbolic space

Table 22. Comparisons of the Hyperbolic and the Euclidean variants, including naive hierarchical models.

Type	Base Class			Superclass		
	Clean	Robust	*Transfer*	Clean	Robust	*Transfer*
Baseline (TeCoA)	52.62	37.62	46.58	61.80	47.20	55.27
Hierarchical Euclidean SoftMax	53.24	38.16	48.09	68.31	55.38	63.94
Hierarchical Euclidean SoftMax + ω_l (weighting)	55.35	38.79	48.52	68.63	55.79	64.17
Hier. Euclid. SoftMax + ω_l + Eq. (11)-(13)	56.08	38.92	50.11	68.29	55.87	64.43
Hier. Euclid. SoftMax + ω_l + Eq. (11)-(13) +Temp. τ	55.93	39.13	50.73	67.91	55.08	64.56
Hier. Euclid. SoftMax + ω_l + Eq. (11)-(13) +Temp. $2^l\tau$	57.17	39.45	51.05	68.46	55.85	64.73
Hier. Euclid. SoftMax + ω_l + Eq. (11)-(13) +Temp. $(\tau_1, \ldots, \tau_L)$	57.67	39.90	52.23	68.75	56.04	64.97
Hyperbolic Space ($\xi = 1 \times 10^{-4}$)	61.70	42.60	62.43	71.15	56.71	65.82
Hyperbolic Space ($\xi = 1 \times 10^{-6}$)	61.24	42.05	60.59	70.44	55.93	63.26
Hyperbolic Space (Ours) ($\xi = 1 \times 10^{-5}$)	**62.14**	**42.72**	**63.34**	**71.68**	**57.13**	**66.40**

Our classification margin induced by the Hyperbolic geometry differs from the Euclidean margin. Theorem 1 shows our margin range grows rapidly with the feature norm (unbounded at the Poincaré ball boundary) but the Euclidean SoftMax is bounded. The hierarchical level of feature vector is proportional to its norm, thus our design forms several generalization levels per sample **producing hierarchically-robust immunizing adversaries**. Table 22 includes **multi-level hierarchy-aware** (Eq. (10)-(13)) Euclidean SoftMax which performs worse.

To adjust margin, the Hierarchical Euclidean SoftMax needs to emply tuned SoftMax Temperature τ (row 5), or mimicking margin ranges from Fig. 2c (Temp. $2^l\tau$ in row 6). Tuning individual Temp. $\tau_1, \ldots, \tau_L$ for $L = 5$ (row 7) takes $4^5 = 1024$ jobs. Our Hyperbolic model enjoys smarter margins as shown in Fig. 2c. Note also the Poincaré stability w.r.t. ξ.

K. AutoAttack Evaluations

Table 23 provides AutoAttack-based comparisons.

Table 23. AutoAttack results.

Method	ImageNet	STL10	CIFAR10	CIFAR100	SUN397	Cars	Food101	OxfordPet	Flower102	DTD	EuroSat	FGVC	PCAM	Caltech101	Caltech256	Average
CLIP	0.00	16.23	4.09	0.00	0.00	0.00	2.24	0.00	0.00	0.00	0.00	0.00	0.00	11.72	6.10	2.69
TeCoA	39.07	82.11	58.21	32.64	30.41	12.19	26.28	61.68	27.85	21.58	15.67	4.95	25.77	67.75	58.14	37.62
PMG-FT	35.89	82.56	59.89	33.55	29.68	15.47	29.46	61.79	30.47	21.71	13.76	5.11	24.92	69.11	57.99	38.09
FARE	28.58	83.76	63.73	37.72	25.07	16.90	31.29	56.21	29.52	23.16	9.56	3.85	21.93	68.54	57.87	37.18
AoS	44.15	84.53	65.32	37.98	31.01	20.13	32.56	66.21	34.39	24.36	16.06	7.05	35.21	71.58	62.16	42.18
Ours	44.59	85.55	66.03	39.48	32.38	19.79	33.17	66.57	35.38	24.08	16.46	7.85	35.23	71.96	62.28	42.72
Ours (5 trees)	**46.19**	**86.67**	**67.55**	**40.59**	**32.52**	**21.16**	**34.47**	**66.57**	**35.99**	**25.43**	**17.66**	**9.02**	**36.84**	**73.24**	**63.85**	**43.85**

L. Batch size *vs*. cost

Table 24 compares training our model on batch size i) 128 (one GPU) and ii) 512.

Table 24. Batch size and cost evaluations on ViT-B.

Batch Size	Method	Clean	AA	Time (min)
512	TeCoA	52.08	37.19	210.6
	PMG-FT	56.83	37.74	261.3
	FARE	59.12	36.61	229.4
	AoS	61.28	41.55	515.8
	Ours (3 steps)	61.81	42.30	283.7
	Ours (2 steps)	61.67	42.19	209.4
128	TeCoA	52.62	37.62	54.8
	PMG-FT	57.36	38.09	67.9
	FARE	59.67	37.18	59.5
	AoS	61.70	42.18	131.2
	Ours (3 steps)	62.14	42.72	73.1
	Ours (2 steps)	62.06	42.50	53.9

M. Results Under Noisy Label Hierarchy.

While **it is easy to inspect the label tree** to prevent errors (the label space is small), we investigate the effect of noise on results. On ImageNet-1K, we randomly corrupt superclass assignments of **10/50/100/200** base classes. Table 25 shows that the degradation under noise is moderate. If all assignments were random, the model would be acting similarly to non-hierarchical model.

Table 25. The impact of label noise on results.

Hallucinated Classes	Clean	PGD	AA
200	60.01	46.41	43.10
100	60.61	46.83	43.72
50	60.97	47.36	44.08
10	61.03	47.67	44.41
0 (Standard Setup)	**61.19**	**47.81**	**44.59**

N. Sensitivity to prompt length.

Table 26 reports results for several prompt lengths to show that the performance of VPT trend is stable.

Table 26. Performance *vs*. the prompt length.

Prompt Length	Clean	PGD	AA
50	54.19	34.57	31.74
100 (Ours)	54.70	34.97	32.29
200	54.94	35.23	32.62

O. Sensitivity to the number of epochs.

Table 27 shows results for varied number of epochs and report performance, showing that our gains persist and that the chosen schedule is near the saturation point.

Table 27. Performance *vs*. the epochs.

Training Epochs	Clean	PGD	AA
5	61.72	43.90	42.25
10 (Ours)	62.14	44.34	42.72
20	62.13	44.59	42.88

P. Extended Related Works

P.1. Hierarchical Feature Alignment

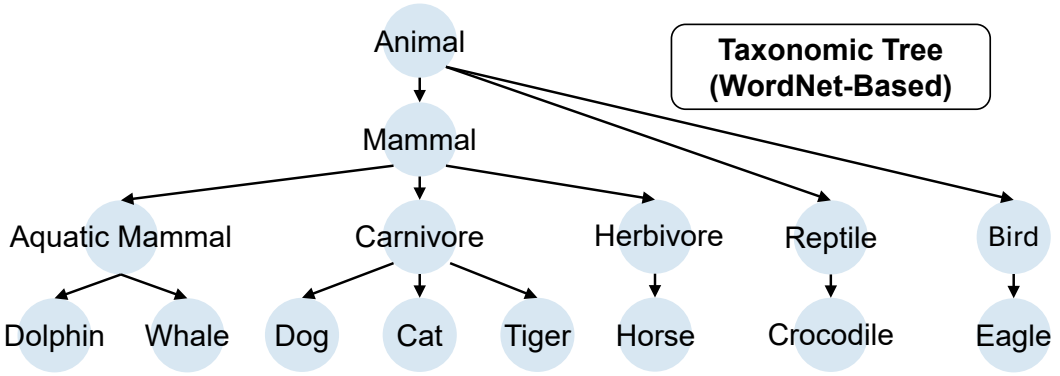
Hierarchical feature alignment has demonstrated its efficacy across various areas by aligning or fusing multi-level feature so that models can combine coarse semantic context with fine-grained details, leading to richer feature representations [63, 78]. An emerging perspective of hierarchical feature alignment is aligning representations with the hierarchical structure in data itself, typically using the hyperbolic space modeling [52]. Khrulkov *et al*. [70] introduced the first hyperbolic image embedding frameworks along these lines, adding hyperbolic layers to convolutional networks and demonstrating that hyperbolic embeddings often outperform Euclidean embeddings in representing visual features when the data has latent hierarchical relationships. Beyond this single vision modality, Desai [58] introduced the hyperbolic representations in the CLIP model [90] to capture the visual-semantic hierarchy inside images and texts. Unlike previous studies emphasizing hierarchical feature alignment alone, our work investigates zero-shot adversarial robustness within CLIP and its variants while further examining how these models can be applied across diverse scenarios. Central to our approach is the explicit construction of data hierarchies bridging textual and visual modalities, thereby reinforcing robust feature representations that enhance zero-shot adversarial performance.

P.2. Vision-Language Model Attacks

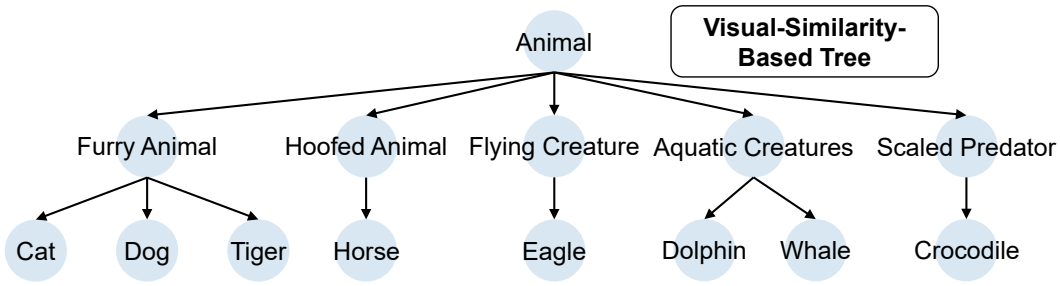
With the rise of Vision-Language Models (VLMs), a variety of adversarial techniques have been introduced to embed subtle perturbations into both visual and textual inputs, thereby undermining VLM inference [80, 98]. Notably, Zhang *et al*. [98] laid the groundwork for studying adversarial scenarios rooted in cross-modal inconsistencies. While white-box attacks have been extensively tested, several recent works have shifted attention to black-box settings, highlighting more realistic adversarial risks in practical deployments [66, 101]. In this paper, we concentrate on establishing zero-shot adversarial robustness for VLMs to defend against such multimodal adversarial attacks.

P.3. Vision-Language Model Robustness

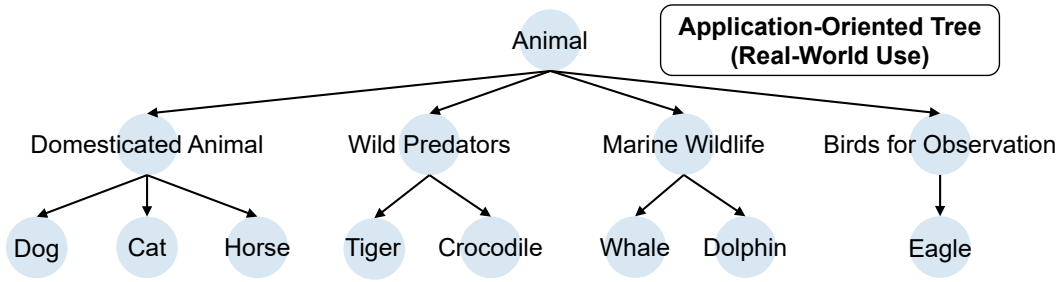
To counteract the potential security threats brought by adversarial samples, adversarial fine-tuning [83, 91] has been proposed, typically employing Parameter-Efficient Fine-Tuning (PEFT) techniques [68, 85, 99, 100, 103, 104]. Mao *et al*. [83] pioneered this direction by employing text-guided contrastive learning to conduct adversarial image-text embedding-level matching. Wang *et al*. [94] tackled the generalization degradation via feature-level regularization. Schlarmann *et al*. [91] developed an unsupervised robust learning framework for downstream tasks. Li *et al*. [77] for robustness in few-shot learning. However, existing methods perform pair-wise image-text alignment of image embedding with its base category embedding, overlooking the benefit of hierarchical decision boundaries from hierarchical labels. Thus, we reformulate traditional adversarial fine-tuning into a hierarchical alignment scheme across image and text modalities based on hyperbolic embedding, thereby mitigating the inherent fixation on base categories. Note that we focus primarily on zero-shot robustness achieved by fine-tuning VLMs.



(a) Taxonomic Tree (WordNet-Based)



(b) Visual-Similarity-Based Tree



(c) Application-Oriented Tree (Real-World Use)

Figure 9. Diverse types of taxonomies w.r.t. base categories for different hierarchies.

Q. Hierarchical Forests

Recall that we show that ensembling the hyperbolic (hierarchical) trees into hyperbolic forests improves adversarial robustness. To systematically explore the hierarchical structure of categories, we construct three distinct hierarchical forests based on different organizational principles, as illustrated in Figures 9. Figure 9a represents the taxonomic hierarchy, which follows the biological taxonomy derived from WordNet, grouping categories based on evolutionary relationships (*e.g.*, mammals, reptiles, birds). Figure 9b illustrates a visual similarity-based hierarchy, where categories are clustered according to shared physical attributes such as fur texture, body structure, or aquatic adaptation. This grouping better aligns with human visual perception and deep feature embeddings. Lastly, Figure 9c depicts an application-oriented hierarchy, which organizes categories based on real-world interactions, such as domesticated animals (pets), wild predators, and aquatic species relevant to conservation and ecological studies. These three trees provide complementary perspectives for understanding hierarchical relationships, facilitating robust feature learning and adversarial robustness analysis in vision models.

References

[49] James Bergstra, Daniel Yamins, and David Cox. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In *International conference on machine learning*, pages 115–123. PMLR, 2013.

[50] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In

Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13, pages 446–461. Springer, 2014.

[51] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

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

[53] 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.

[54] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[55] Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 215–223. JMLR Workshop and Conference Proceedings, 2011.

[56] 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.

[57] 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.

[58] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR, 2023.

[59] 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.

[60] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004.

[61] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[62] 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.

[63] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Hypercolumns for object segmentation and fine-grained localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 447–456, 2015.

[64] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7): 2217–2226, 2019.

[65] Weiwei Hou, Hanna Suominen, Piotr Koniusz, Sabrina Caldwell, and Tom Gedeon. A token-wise cnn-based method for sentence compression. In *International Conference on Neural Information Processing (ICONIP)*, pages 668–679. Springer, Cham, 2020.

[66] Nathan Inkawhich, Gwendolyn McDonald, and Ryan Luley. Adversarial attacks on foundational vision models. *arXiv preprint arXiv:2308.14597*, 2023.

[67] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019.

[68] 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.

[69] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[70] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6418–6428, 2020.

[71] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[72] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[73] Haoran Lai, Qingsong Yao, Zihang Jiang, Rongsheng Wang, Zhiyang He, Xiaodong Tao, and S Kevin Zhou. Carzero: Cross-attention alignment for radiology zero-shot classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11137–11146, 2024.

[74] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[75] 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.

[76] 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.

[77] 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.

[78] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[79] Changsheng Lu, Zheyuan Liu, and Piotr Koniusz. Openkd: Opening prompt diversity for zero- and few-shot keypoint detection. In *Computer Vision – ECCV 2024*, pages 148–165, Cham, 2025. Springer Nature Switzerland.

[80] 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.

[81] 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.

[82] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

[83] 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.

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

[85] Yao Ni, Shan Zhang, and Piotr Koniusz. PACE: marrying the generalization of PArameter-efficient fine-tuning with consistency regularization. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[86] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008.

[87] OpenAI. Chatgpt [large language model]. https://chatgpt.com, 2024.

[88] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012.

[89] Chantal Pellegrini, Matthias Keicher, Ege Özsoy, Petra Jiraskova, Rickmer Braren, and Nassir Navab. Xplainer: From x-ray observations to explainable zero-shot diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–429. Springer, 2023.

[90] 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.

[91] 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.

[92] Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P Langlotz, Andrew Y Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

[93] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 210–218. Springer, 2018.

[94] 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.

[95] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

[96] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010.

[97] 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.

[98] 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.

[99] Yifei Zhang, Hao Zhu, Aiwei Liu, Han Yu, Piotr Koniusz, and Irwin King. Less is more: Extreme gradient boost rank-1 adaption for efficient finetuning of llms. In *arXiv/2410.19694*, 2024.

[100] Yifei Zhang, Hao Zhu, Junhao Dong, Haoran Shi, Ziqiao Meng, Piotr Koniusz, and Han Yu. Crossspectra: Exploiting cross-layer smoothness for parameter-efficient fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[101] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. *Advances in Neural Information Processing Systems*, 36:54111–54138, 2023.

[102] 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.

[103] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022.

[104] Hao Zhu, Yifei Zhang, Junhao Dong, and Piotr Koniusz. Bilora: Almost-orthogonal parameter spaces for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25613–25622, 2025.