Robustifying Zero-Shot Vision Language Models by Subspaces Alignment
– *Supplementary Material* –

Junhao Dong[1,2], Piotr Koniusz[3,4*], Liaoyuan Feng[5], Yifei Zhang[1], Hao Zhu[3], Weiming Liu[6], Xinghua Qu[5], and Yew-Soon Ong[1,2*]
[1]Nanyang Technological University, [2]CFAR, IHPC, A*STAR, [3]Data61♥CSIRO,
[4]Australian National University, [5]Bytedance, [6]Zhejiang University
{junhao003, yifei.zhang, asysong}@ntu.edu.sg, piotr.koniusz@data61.csiro.au,
{fengliaoyuan, xinghua.qu1}@bytedance.com, allenhaozhu@gmail.com, 21831010@zju.edu.cn

Abstract

This supplementary material contains a description of our experimental setup, encompassing the datasets employed, specific implementation details, and configuration settings for BLIP and medical CLIP (refer to Appendix A). Furthermore, we elaborate on the details regarding the image-text augmentation in our subspace construction in Appendix B. The proof of Theorem 1 (i.e., *properties of MaxExp) is presented in Appendix C. We analyze different types of subspace construction strategies in Appendix D and hyperparameters in Appendix E, respectively. Further analyses of diverse weighting mechanisms and diverse EMA strategies for improved covariance matrix construction are in Appendices G & H. We then provide an additional discussion regarding the effectiveness of our method in Appendix I.*

A. Experimental Details

In this section, we provide a detailed description of the experimental configurations employed in our study, including comprehensive information about the datasets used for adversarial fine-tuning and the specific implementation details of our proposed method.

A.1. Datasets

Following the evaluation protocols of prior works [34, 44], we perform adversarial fine-tuning of the CLIP model on the ImageNet training set [9]. For robustness evaluations, we test the fine-tuned model on the ImageNet validation set (ImageNet provides only validation split typically used for testing, and a small part of the ImageNet train set is taken for validation instead) along with an additional 14 zero-shot datasets, encompassing a wide range of image recognition tasks. Specifically, these 15 datasets cover four categories:

- **General Image Classification**: ImageNet [9], STL-10 [7], CIFAR-10 and CIFAR-100 [22], Caltech-101 [12], and Caltech-256 [14].
- **Fine-Grained Classification**: FGVC Aircraft [33], Flower102 [35], Food101 [3], Oxford-IIIT Pets [37], and Stanford Cars [21].
- **Domain-Specific Classification**: Describable Textures Dataset (DTD) [6], EuroSAT [16], and PatchCamelyon (PCAM) [43].
- **Scene Recognition**: SUN397 [46].

During the adversarial fine-tuning stage, we apply basic data pre-processing by resizing each input image to 224 × 224 pixels and performing a center crop for consistency.

A.2. Term Definition

For further clarifications, we provide official definitions of diverse terms in this paper. (i) The worst-case joint adversary is adversarial sample pair $\big(\mathbf{x}+\boldsymbol{\delta}_{\mathbf{X}}^{(m)}, \mathbf{t}+\boldsymbol{\delta}_{\mathbf{T}}^{(m)}\big)$ of both image and text modalities. The "worst-case adversary" is the final step m^{th} adversary pair from the iterative adversary generation of m steps, performing a joint attack, which makes it even stronger than a single modality attack. (ii) Intermediate adversarial samples are intermediate products $\big\{\mathbf{x}+\boldsymbol{\delta}_{\mathbf{X}}^{(i)}\big\}_{i=1}^{m-1}$ from adversary generation. (iii) "Joint (intermediate) adversarial subspace" means that given an image & its text, we augment the image, we augment the text, we obtain adversarial embeddings, and we build a subspace from them. "Intermediate" means adversarial embeddings were obtained from adv. generation step $i<m$.

A.3. Implementation

Consistent with the settings of previous studies [34, 44], we utilize the CLIP model [40] with the Vision Transformer (ViT) architecture of ViT-Base/32 [11]. For network optimization, we employ the Stochastic Gradient Descent (SGD) optimizer with a momentum of 0.9 and a batch size

[*] Corresponding authors.

Table 14. Summary of the original text prompt and its corresponding synonymous and antonymous phrases ordered in descending order by the most-to-least related phrases to the original text prompt.

No.	Positive prompt	"Negating meaning" prompt
	Original	
0	This is a photo of a [CLS].	This is not a photo of a [CLS].
	Synonymous/antonymous prompts	
1	This is a picture of a [CLS].	This is not a picture of a [CLS].
2	This is an image of a [CLS].	This is not an image of a [CLS].
3	Here is a photo of a [CLS].	Here is not a photo of a [CLS].
4	Here is a picture of a [CLS].	Here is not a picture of a [CLS].
5	Here is an image of a [CLS].	Here is not an image of a [CLS].
6	This photograph shows a [CLS].	This photograph does not show a [CLS].
7	This is a depiction of a [CLS].	This is not a depiction of a [CLS].
8	This seems to be a photo of a [CLS].	This does not seem to be a photo of a [CLS].
9	This appears to be an image of a [CLS].	This does not appear to be an image of a [CLS].
10	This might be a picture of a [CLS].	This might not be a picture of a [CLS].
11	This could be an image of a [CLS].	This might not be an image of a [CLS].
12	Possibly a photograph of a [CLS].	Possibly, this is not a photograph of a [CLS].
13	Perhaps this is a picture of a [CLS].	Perhaps this is not a picture of a [CLS].
14	It seems like this is an image of a [CLS].	It seems that this is not an image of a [CLS].
15	This might represent a [CLS].	This might not represent a [CLS].
16	It is conceivable that this is a photo of a [CLS].	It's conceivable that this isn't a photo of a [CLS].
17	This may be an illustration of a [CLS].	This may not be an illustration of a [CLS].
18	Could this be an image of a [CLS]?	Could this possibly not be an image of a [CLS]?
19	I wonder if this is a picture of a [CLS].	I wonder if this is not a picture of a [CLS].
20	There is a [CLS] in this photo.	There is no [CLS] in this photo.

of 512. The learning rate is scheduled using cosine annealing, starting from an initial value of 1×10^{-5} for full fine-tuning of the vision encoder only. For Visual Prompt Tuning (VPT) [18], we introduce token-level learnable parameters of size 100 into the vision branch of CLIP, using a learning rate of 40. During training, we generate adversarial samples at both the image and text levels using Projected Gradient Descent (PGD) [31] with 3 iterations. For image-level adversarial perturbations, we adopt the ℓ_∞-norm threat model with a maximum perturbation radius of $\epsilon_\mathrm{X} = 1/255$ and a step size of $\alpha_\mathrm{X} = 1/255$, unless specified otherwise. For text-level adversarial perturbations—applied only during fine-tuning—we set the step size to $\alpha_\mathrm{T} = 1 \times 10^{-4}$ and the corresponding perturbation radius to $\epsilon_\mathrm{T} = 2 \times 10^{-4}$. We set the image augmentation and text prompt synonym/antonym numbers to $n = q = 20$ for subspace construction. We adopt the MaxExp parameter $\eta = 10$ following existing works [31]. The loss parameter $\beta = 3.0$ for a favorable trade-off between natural performance and adversarial robustness. In line with previous works on adversarially robust CLIP fine-tuning [34, 44], we focus on evaluating robustness against three strong white-box adversarial attacks: 20-step PGD [31], Carlini and Wagner (CW) attack [5], and Auto-Attack (AA) [8]. In addition to image-level only attacks, we also evaluate *text-level attacks*: BERT-Attack [26] and Gradient-Based Distributional Attack (GBDA) [15], and *bi-level attacks* using Collaborative Multimodal Adversarial Attack (Co-Attack) [48] and Set-level Guidance Attack (SGA) [30], discussed in the main text. Note that the setting of all the hyper-parameters is obtained through the Hyperopt package [2] 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 using eight NVIDIA Tesla A100 GPUs.

Experimental setup for BLIP. To further evaluate zero-shot robustness of our method on downstream tasks, we extend our experiments to include the BLIP architecture [25], a large-scale vision-language model that employs bootstrapping language-image pre-training to unify vision-language understanding tasks. Specifically, we assess zero-shot adversarial robustness for two vision-language understanding tasks: image-text retrieval and image captioning. Following Li *et al.* [25], we adversarially optimize the Image-Text Contrastive (ITC) loss, Image-Text Matching (ITM) loss, and Language Modeling (LM) loss to obtain the robust BLIP. Afterward, we evaluate the robustness by conducting the PGD attack method (iterative gradient ascent) based on the ITM loss for image-text retrieval and the LM loss for image captioning, following the same criterion ($\epsilon_\mathrm{X} = 1/255$) in our original manuscript by simply replacing the objective function for adversary generation.

Experimental setup for Medical CLIP. To further extend our analysis in the context of robust medical imaging, we employ a CLIP model based on ViT-B/16, pre-trained specifically on radiology datasets based on CheXzero [42]. Following established protocols [23, 38, 42], we use the MIMIC dataset [19], which is a comprehensive collection of chest radiographs paired with detailed radiology reports, for adversarial fine-tuning. The text encoder in the CLIP model is based on BioBERT [24], a specialized biomedical language model optimized for text mining in the biomedical domain. During the zero-shot inference stage, we evaluate the robust CLIP models on three standard multi-label radiology datasets: ChestX-ray14 [45], CheXpert [17], and PadChest [4]. We report the Area Under the Curve (AUC) for both clean images and their adversarial counterparts, which are generated using 20-step PGD [31] with the maximum perturbation strength $\epsilon_{\text{X}} = 1/255$.

B. Image-text Augmentations

In the main text, we construct subspaces based on diverse augmentations in both the image and text modalities. Below, we elaborate on the implementation details for generating image-level augmentations and constructing synonymous and antonymous text prompts.

B.1. Image-level Augmentation Sets

For image-level augmentations, we employ Differentiable Automatic Data Augmentation (DADA) [28], an efficient method that utilizes a one-pass gradient optimization strategy based on diverse sub-policies to generate augmentations of input images. To enhance computational efficiency, we perform the policy search on a small subset (10%) of the target dataset and then apply the learned policies to the full dataset. Notably, since the policy search is conducted prior to the adversarial fine-tuning stage, it does not affect the training or inference efficiency.

B.2. Synonymous & Antonymous Text Prompts

To generate text-level augmentations, we construct both synonymous and antonymous variations within the contextual segments of the text prompts by utilizing a variety of templates generated from ChatGPT-4 [36], obtaining a spectrum of text prompts (labels). The complete list of these text augmentations is provided in Table 14 for reference. Specifically, all the synonymous/antonymous text prompts are organized in descending order (from the most similar to the least similar to the original prompt). In other words, we organize these prompts according to their relative hardness, ranging from easy to hard. Thus, we define the original text prompts in the positive and negative domains:

P: `This is a photo of a [CLS].`

N: `This is not a photo of a [CLS].`

In our experiments, per class, we generate 20 text-level augmented variants of these original text prompts by replacing words in the context part.

C. Properties of MaxExp

Proof C.1 (Low Computational Complexity). The computational complexity of MaxExp has been analyzed and reported as MaxExp(F) in [20]. □

Proof C.2 (Spectrum Whitening). Non-simple eigenvalues are a classical topic in gradient computation of the Singular Value Decomposition and Eigenvalue Decomposition [20, 32]. □

Proof C.3 (Spectrum Whitening). This property holds as $\lim_{\eta\to\infty} 1-(1-\lambda_j)^\eta = 1$ if $\lambda_j > 0$. Thus, whitening occurs on non-zero singular values. By definition of spectrum whitening known from the Whitening Principal Component Analysis (WPCA), the whitening is achieved when λ_j approach 1 for some $j > i$ while $\lambda_i = 0$ are rejected parts of the spectrum as in WPCA.

For the second part, we simply set $1-(1-\lambda')^\eta = 0.5$ and solve for η. We notice that **(i)** $1-(1-\lambda)^\eta$ is a monotonically increasing function on $0 < \lambda < 1$, producing the compact domain $[0, 1]$, **(ii)** it is equal to 0 at $\lambda = 0$ and 1 at $\lambda = 1$, and **(iii)** its derivative of the Laplace shape, largest at $\lambda = 0$ (fastest growth) is rapidly decaying. This immediately indicates that this function saturates rapidly, and for λ', it is halfway through the saturation on $[0, 1]$ by the design of MaxExp technique. □

Proof C.4 (Robust Estimator Property). If $\boldsymbol{\phi}\in\operatorname{Span}(\mathbf{U}_{1:r})$ is added to the covariance estimation, then the original Grassmann feature map with spectrum $1_{\lambda_j\geq\lambda'}$ remains unchanged as $\lambda_j + \delta_j(\boldsymbol{\phi}) \geq \lambda'$ where $\delta_j(\cdot)$ is the change $\boldsymbol{\phi}$ imposes on singular value λ_j. It is easy to see that if $\boldsymbol{\phi}$ is equal $\mathbf{u}$ corresponding to λ' then $\lim_{n'\to\infty} \lambda_j\big(\frac{1}{n+n'-1}((n-1)\boldsymbol{\Sigma} + n'\boldsymbol{\phi}\boldsymbol{\phi}^T)\big) = \lambda'$ so $1_{\lambda_j+\delta_j(\boldsymbol{\phi})\geq\lambda'} = 1_{\lambda_j\geq\lambda'}$.

Similarly, if $\boldsymbol{\phi}\in\operatorname{Span}(\mathbf{U}_{1:r})$ then choose $\boldsymbol{\phi}$ equal $\mathbf{u}_j$ for any $\lambda_j\geq\lambda'$ then $\big\|\big(\mathbf{I}-\mathbf{U}_{1:r}\mathbf{U}_{1:r}^T\big)\mathbf{u}_j\big\|_2 = \big\|\mathbf{I}\mathbf{u}_j-\mathbf{u}_j\big\|_2 = 0$. □

Proof of Theorem 1 (The MaxExp Approximation Error). The MaxExp approximation error equals $\epsilon=(1-\lambda_i)^\eta\|\boldsymbol{\phi}\|_2$ when $\cos(\boldsymbol{\phi},\mathbf{u}_i)=1$ because

$$\begin{aligned}\epsilon &= \|(\mathbf{I}-\boldsymbol{\Sigma})^\eta\boldsymbol{\phi}\|_2\\ &= \big\|(\mathbf{I}-\mathbf{U}\operatorname{diag}(\boldsymbol{\Lambda})\mathbf{U}^T)^\eta\|\boldsymbol{\phi}\|_2\mathbf{u}_i\big\|_2\\ &= \big\|\mathbf{U}(\mathbf{I}-\operatorname{diag}(\boldsymbol{\Lambda}))^\eta\mathbf{U}^T\mathbf{u}_i\|\boldsymbol{\phi}\|_2\big\|_2\\ &= (1-\lambda_i)^\eta\|\boldsymbol{\phi}\|_2 \end{aligned} \tag{15}$$

Then, we solve Eq. (15) for η. As the function in Eq. (15) is monotonically increasing, setting $\lambda_i = \lambda'$ gives us the maximum possible error ϵ that decreases for any $\lambda_i > \lambda'$. □

D. Investigating Different Distances Between Covariance Matrices

In addition to the MaxExp approximation of the projection distance for subspaces used in the main text, we present a comprehensive exploration of various covariance metrics and their corresponding implementations.

Recall that covariance matrix $\boldsymbol{\Sigma}$ represents an image with its augmentations. The corresponding (original and augmented) text embeddings are given as matrix $\boldsymbol{\Phi}_{\mathrm{T}}$ storing vectors $\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_{q'}$.where $q' = q+1$ and q was the number of synonymous/antonymous phrases.

Subspace-to-vector distance. Focusing on practical adversarial threats at the image level, we investigate subspace construction based on Singular Value Decomposition (SVD) within the image domain only and perform alignment between the resulting image embedding subspaces and their corresponding vector text embeddings. Thus, the projection distance between the subspace and its k-closest (most relevant) text embedding vectors can be defined as:

$$d^2(\mathbf{U}, \boldsymbol{\Phi}_{\mathrm{T}}|k) = \mathrm{avg\_top\_k}\Big(\big\|(\mathbf{I} - \mathbf{U}_{1:r}\mathbf{U}_{1:r}^T)\boldsymbol{\phi}_1\big\|_2^2, \ldots, \big\|(\mathbf{I} - \mathbf{U}_{1:r}\mathbf{U}_{1:r}^T)\boldsymbol{\phi}_{q'}\big\|_2^2\Big), \tag{16}$$

where $\mathbf{U}_{1:r}$ are the r leading singular vectors of SVD decomposition $\mathbf{U}\,\mathrm{diag}(\boldsymbol{\Sigma})\mathbf{V}^T$ of trace-normalized covariance matrix $\boldsymbol{\Sigma}$, *i.e.*, $\boldsymbol{\Sigma} := \boldsymbol{\Sigma}/(\mathrm{tr}(\boldsymbol{\Sigma})+\nu)$ where $\nu = 10^{-5}$ prevents division by zero. Parameter $0 < r < \mathrm{rank}(\boldsymbol{\Sigma})$ represents r-dimensional linear subspace. Moreover, operator $\mathrm{avg\_top\_k}(\cdot)$ simply averages over k-smallest distances passed into it as input arguments.

MaxExp approximation of the subspace-to-vector distance. Below, we explore a more efficient and stable subspace learning approach based on the MaxExp method discussed in Section 3.2, where the image with its augmentations is represented as a subspace, whereas text embeddings are treated as vectors. Thus, the projection distance between the (approximate) subspace and its k-closest (most relevant) text embedding vectors can be defined as:

$$d^2(\boldsymbol{\Sigma}, \boldsymbol{\Phi}_{\mathrm{T}}|k) = \mathrm{avg\_top\_k}\Big(\big\|(\mathbf{I} - \boldsymbol{\Sigma})^\eta\boldsymbol{\phi}_1\big\|_2^2, \ldots, \big\|(\mathbf{I} - \boldsymbol{\Sigma})^\eta\boldsymbol{\phi}_{q'}\big\|_2^2\Big), \tag{17}$$

SVD-based subspace-to-subspace image-text distance. The original projection distance operates on subspaces obtained via SVD, *i.e.*, image and text modalities are represented by left singular vectors $\mathbf{U}$ and $\mathbf{U}'$ of their respective feature covariance matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}'$. The projection distance between the image and text subspaces is defined as:

$$d^2(\mathbf{U}, \mathbf{U}') = \big\|\mathbf{U}_{1:r}\mathbf{U}_{1:r}^T - \mathbf{U'}_{1:r}\mathbf{U'}_{1:r}^T\big\|_F^2. \tag{18}$$

Table 15. Performance (%) of diverse distance metrics and configurations for adv. subspace learning during fine-tuning.

Subspace	Metric	Top-K Selection	Clean	PGD	AA
Image Only	Standard (SVD)	1 (Smallest)	58.34	39.40	37.84
		3	58.75	39.56	37.97
		All (Average)	59.01	39.80	38.10
	MaxExp	1 (Smallest)	58.86	39.98	38.25
		3	59.17	40.14	38.42
		All (Average)	59.38	40.39	38.67
Image & Text	Frobenius norm	-	60.16	40.86	39.05
	SVD	-	60.89	42.35	40.47
	MaxExp	-	**61.70**	**43.88**	**42.18**

Frobenius norm image-text distance. For completeness, we show that a mere Frobenius distance in place of the projected distance on subspaces is suboptimal. We define:

$$d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}') = \big\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}'\big\|_F^2. \tag{19}$$

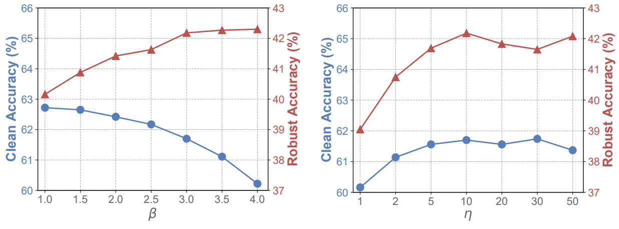
MaxExp subspace-to-subspace image-text distance. In our method, we focus on the MaxExp approximation (top row of Eq. (5)) of the projected distance between subspaces, used for both image and text modalities due to its advantageous properties explained in Section 3.2 defined as:

$$d^2(\boldsymbol{\Sigma}, \boldsymbol{\Sigma}') = \big\|(\mathbf{I} - \boldsymbol{\Sigma})^\eta - (\mathbf{I} - \boldsymbol{\Sigma}')^\eta\big\|_F^2. \tag{20}$$

The impact of different metrics. Below, we explore the efficacy of various distance metrics in adversarial (subspace) learning between image and text modalities. Specifically, we investigate the interaction between the image-level adversarial subspace and its top k-closest text embeddings, selected based on the smallest k distances between the "image set" subspace and individual augmented text embedding set of size $q' = q+1$ elements. Table 15 shows that augmenting both the image and text modalities within their respective subspaces improves both natural performance and robustness in the zero-shot setting. Moreover, when subspace learning is applied solely to the image modality, incorporating all the augmented examples enhances zero-shot performance. This substantiates the necessity of constructing subspaces based on intermediate adversaries, which capture rich information about decision boundaries. Standard SVD-based subspace learning is computationally intensive, requiring 171.5 minutes per training epoch, whereas our MaxExp-based approximate subspace learning achieves greater efficiency with only 96.0 minutes.

E. Hyper-Parameter Analysis (Trade-off)

The trade-off between natural performance and adversarial robustness has been extensively investigated in single-modal adversarial training [10, 39, 47], yet it remains underexplored within multimodal scenarios, especially for robust VLMs. To bridge this gap, we analyze the effect



(a) (b)

Figure 6. Hyper-parameter sensitivity analysis of our method w.r.t. (a) β: the weighting factor between subspace-driven image-text alignment for clean samples and robustness invariance of unified subspaces for adversaries, and (b) η: the MaxExp parameters interpolating between Forbenius norm and the soft approximation of the projection distance between subspaces.

Table 16. Performance (%) of different text prompt augmentation types in Eq. (7) to output predictions during adv. fine-tuning. The antonyms (*not* c) denote the negative text prompts of the target category c only, while the antonyms (*all classes*) represent the negative text prompts for all C categories.

Text Prompt Type	Clean	PGD	AA
Synonyms	60.25	43.63	41.85
Synonyms + Antonyms (*All classes*)	61.30	43.27	41.32
Synonyms + Antonyms (*not* c)	**61.70**	**43.88**	**42.18**

of hyper-parameter β, which balances between subspace-driven image-text alignment using a classifier for clean samples and regularization term $\Omega(\cdot)$ that aligns adversarial image-text subspaces with the clean image-text subspace without the use of labels. Figure 6a illustrates that higher values of β lead to increased robustness, albeit with a reduction in natural performance. Conversely, enhancing clean accuracy is associated with a decline in adversarial robustness. Such a trade-off effect can also be interpreted as an optimization balancing between natural risk and boundary risk, as demonstrated in [47]. Our subspace-driven image-text alignment for clean samples can be regarded as optimizing a surrogate classification loss (*natural risk*) to match the cross-modal subspaces for clean images and texts. The regularizer $\Omega(\cdot)$ contributes to minimizing the difference between natural image-text subspace and their adversarial counterparts (*boundary risk*).

In addition to the trade-off explicitly led by diverse optimizing focus, we also investigate the effect of the hyper-parameter η, which interpolates between the Frobenius distance and soft approximation of the Grassmann feature map. Figure 6b shows that increasing η enhances both the zero-shot adversarial robustness and the clean accuracy. The performance gain plateau when η reaches 10.

Table 17. Comparison of average clean and robust accuracy (%) with and without our instance-wise weighting mechanism.

Configuration of Eq. (12)	Clean	PGD	AA
w/o Weighting	61.26	43.20	41.49
w/ Weighting	**61.70**	**43.88**	**42.18**

Table 18. Average clean and robust accuracy (%) across 15 datasets of diverse EMA strategies for adversarial fine-tuning.

EMA Strategy	Clean	PGD	AA
w/o EMA	61.70	43.88	42.18
Class-wise EMA	61.78	43.96	42.25
Instance-wise EMA	61.96	44.08	42.36

F. Impact of Different Text Prompt Augmentations

In Table 16, we evaluate the zero-shot performance by employing different types of text prompts for subspace construction, augmenting the predictions as defined in Eq. (7) in the main text. The results demonstrate that incorporating additional antonymous text prompts (details provided in Appendix B.2) leads to improved clean accuracy in the zero-shot setting. We compare two strategies for antonymous prompt augmentation: one that uses antonymous prompts derived solely from the target label and another that includes prompts from all categories. Notably, the approach utilizing antonymous prompts only from the ground-truth labels yields superior performance.

G. Impact of the Weighting Mechanism

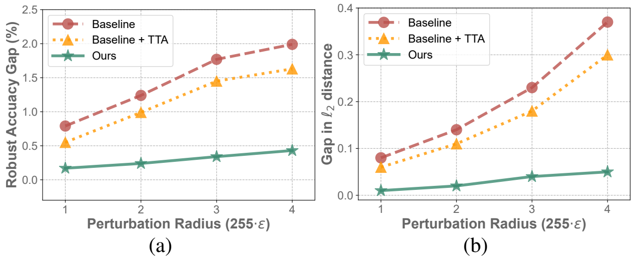
Below, we investigate the effect of our instance-wise weighting mechanism (Eq. (12)) based on the prediction discrepancy for each adversarial subspace. As shown in Table 17, we report both zero-shot clean and robust accuracy with and without our weighting mechanism. We can observe that integrating the weighting mechanism boosts both natural performance and adversarial robustness, justifying the emphasis on stronger adversaries during training.

H. EMA for Covariance Construction.

To enhance the stability of our covariance estimation, we employ an Exponential Moving Average (EMA) to iteratively update the accumulated covariance matrices of the image and text embeddings across training epochs. **Notice the EMA strategy is not used at all in our main manuscript.** Below, we update the accumulated covariance matrix $\mathbf{\Sigma}_*^t$ at epoch t using the covariance matrix $\mathbf{\Sigma}_*$ computed from the augmented set, following the strategy:

$$\mathbf{\Sigma}_*^t = \tau\,\mathbf{\Sigma}_*^{(t-1)} + (1-\tau)\,\mathbf{\Sigma}_*, \tag{21}$$

where hyper-parameter τ controls the exponential moving average. Unless specified otherwise, we utilize the accu-



(a) (b)

Figure 7. Comparison of (a) the robust accuracy gap between non-augmented and augmented adversaries, and (b) the corresponding gap in feature-level distance differences, as illustrated in Figure 1b, with and without training-time augmentations.

mulated covariance matrix obtained for each augmented set at each epoch. Considering that diverse augmentation schemes are applied in different epochs, employing EMA across epochs can further facilitate capturing a broader set of underlying adversarial samples, thereby helping to avoid robustness over-fitting to specific data.

We further investigate the impact of different Exponential Moving Average (EMA) strategies on zero-shot adversarial robustness. Although the main manuscript omits EMA for reasons of computational efficiency, we explore two EMA methodologies: (i) class-wise EMA (faster and less memory intense) and (ii) instance-wise EMA (more accurate). As presented in Table 18, incorporating EMA to stabilize the computation of covariance matrices results in slight improvements in zero-shot performance on both clean and adversarial samples.

I. Further Discussion on the Efficacy

In this section, we explore the fundamental reasons behind the effectiveness of our proposed method. Specifically, we aim to understand the benefits derived from our subspace-driven adversarial fine-tuning. Building upon the insights from Figure 1a, we investigate the robust accuracy gap between non-augmented (standard) adversarial examples and augmented adversaries. For comparative analysis, we also introduce a Test-Time Augmentation (TTA) to produce averaged output based on a set of augmentations of the original sample. Figure 7a indicates that relying solely on test-time augmentation fails to mitigate the robustness degradation on unforeseen augmented data. In contrast, our method incorporates subspace-level robustness invariance by learning a unified image-text subspace that captures underlying adversarial threats, thereby improving zero-shot robustness.

We have shown in Figure 1b that, for prior sample-wise adversarial fine-tuning methods, training-time augmentation typically struggles to close the feature-level distance gap between augmented and non-augmented adversaries. We further demonstrate that adopting test-time augmentation to produce an averaged and rectified prototype for each feature does not help reduce this feature-level distance gap (see Figure 7b).

In addition to validating the effectiveness of our method, we also discuss its computational efficiency. As shown in Table 13, our MaxExp-based approximate subspace learning strategy requires only 96 minutes per training epoch during adversarial fine-tuning, compared to 171.5 minutes per epoch for the standard SVD-based subspace approach. This significant reduction in training time aligns with our computational complexity analysis in Section 3.2. Furthermore, our method does not introduce additional modules or parameters to the target foundational model, ensuring that the inference time remains consistent with other approaches.

J. Extended Related Works

J.1. Multimodal Architectures

A pivotal breakthrough in vision-language learning came with CLIP [40], which introduced contrastive pre-training on a vast dataset of 400 million image-text pairs for cross-modal feature alignment. This simple setup yields highly generalizable features, enabling zero-shot transfer to downstream tasks without additional fine-tuning. Building on CLIP's success, recent Vision-Language Models (VLMs) combine powerful language models with visual encoders to handle more complex multimodal tasks. LLaVA (Large Language and Vision Assistant) [29] is one such model that connects a vision transformer to a Large Language Model (LLM) and is trained end-to-end on visual instruction tuning data. In contrast, Flamingo [1] is trained on extensive web-curated corpora of interleaved images and text, which equips it with strong in-context learning abilities for multimodal tasks. Despite their impressive performance, these vision-language architectures have demonstrated susceptibility to adversarial perturbations [49], wherein subtle modifications to input images may cause even advanced VLMs to produce erroneous outputs. In this work, we concentrate on the zero-shot adversarial robustness of CLIP—arguably the cornerstone for vision-language foundational model—and highlight how our framework can be generalized to other architectures and applications.

J.2. Multimodal Adversarial Attacks

Adversarial perturbations pose substantial security concerns in both the visual [5, 8, 31] and textual [15, 26] domains. Although existing single-modal adversarial methods can degrade the zero-shot performance of VLMs, more comprehensive multimodal attack strategies have emerged to jointly compromise models across various modalities [30, 48]. Notably, Zhang *et al.* [48] pioneered a suite of adversarial scenarios targeting VLMs and introduced a multimodal attack approach that accounts for consistency across different modalities. Subsequently, Guo *et al*. [30] focused on enhancing transferability in multimodal adversarial attacks by incorporating set-level alignment-preserving augmentations to expand the range of potential inputs without disrupting cross-modal consistency. In this paper, alongside our evaluations of single-modal adversarial defenses, we also examine zero-shot robustness against a variety of multimodal attacks to assess more realistic adversarial threats.

J.3. Multimodal Adversarial Robustness

Adversarial training [10, 31, 47] remains the most reliable means of boosting robustness by strategically injecting adversarial examples into the learning process. However, scaling this approach to Vision-Language Models (VLMs) [40] tends to be computationally prohibitive. Recent efforts therefore focus on adversarial fine-tuning of pre-trained VLMs [27, 41, 44] using parameter-efficient methods [13, 18, 50] to reduce the associated overhead. For instance, Mao *et al*. [34] introduced an adversarial fine-tuning mechanism based on text-guided contrastive optimization, promoting alignment between perturbed image representations and their corresponding text embeddings. Wang *et al*. [44] mitigated robustness deterioration by constraining feature embeddings through the original pre-trained CLIP model. Meanwhile, Schlarmann *et al*. [41] concentrated on enhancing downstream task resilience. These earlier works are predominantly grounded in alignment procedures at the sample level, linking image and text embeddings or pairing clean and adversarial instances, but often overlook distributional properties across broader collections of data points. To address this gap, our study advocates aligning entire image-text subspaces to leverage distributional robustness in fine-tuning, particularly for zero-shot applications.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 17

[2] 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. 13

[3] 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. 12

[4] 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. 14

[5] 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. 13, 17

[6] 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. 12

[7] 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. 12

[8] 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. 13, 17

[9] 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. 12

[10] 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. 15, 18

[11] 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. 12

[12] 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. 12

[13] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 18

[14] Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007. 12

[15] 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. 13, 17

[16] 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. 12

[17] 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. 14

[18] 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. 13, 18

[19] 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. 14

[20] Piotr Koniusz and Hongguang Zhang. Power normalizations in fine-grained image, few-shot image and graph classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):591–609, 2021. 14

[21] 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. 12

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 12

[23] 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. 14

[24] 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. 14

[25] 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. 13

[26] 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. 13, 17

[27] 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. 18

[28] Yonggang Li, Guosheng Hu, Yongtao Wang, Timothy Hospedales, Neil M Robertson, and Yongxin Yang. Differentiable automatic data augmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 580–595. Springer, 2020. 14

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 17

[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. 13, 17, 18

[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. 13, 14, 17, 18

[32] Jan R. Magnus. On differentiating eigenvalues and eigenvectors. *Econometric Theory*, 1(2):179–191, 1985. 14

[33] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 12

[34] 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. 12, 13, 18

[35] 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. 12

[36] OpenAI. Chatgpt [large language model]. `https://chatgpt.com`, 2024. 12, 14

[37] 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. 12

[38] 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. 14

[39] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *The Tenth International Conference on Learning Representations, ICLR*, 2022. 15

[40] 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. 12, 17, 18

[41] 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. 18

[42] 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. 14

[43] 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. 12

[44] 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. 12, 13, 18

[45] 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. 14

[46] 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. 12

[47] 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. 15, 16, 18

[48] 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. 13, 17

[49] 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. 17

[50] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 18