

Robustifying Zero-Shot Vision Language Models by Subspaces Alignment

Junhao Dong^{1,2}, Piotr Koniusz^{3,4*}, Liaoyuan Feng⁵, Yifei Zhang¹, Hao Zhu³, Weiming Liu⁶,
Xinghua Qu⁵, and Yew-Soon Ong^{1,2*}

¹Nanyang Technological University, ²CFAR, IHPC, A*STAR, ³Data61♥CSIRO,

⁴Australian National University, ⁵Bytedance, ⁶Zhejiang University

{junhao003, yifei.zhang, asyong}@ntu.edu.sg, piotr.koniusz@data61.csiro.au,

{fengliaoyuan, xinghua.qu1}@bytedance.com, allenhaozhu@gmail.com, 21831010@zju.edu.cn

Abstract

Vision-Language Models (VLMs) enjoy strong zero-shot performance but are vulnerable to adversarial attacks posing security risks. Adversarially robust fine-tuning enhances zero-shot robustness on new datasets while preserving the natural performance of pre-trained VLMs. However, prior methods use sample-wise adversarial fine-tuning, neglecting the underlying second-order statistics that represent entire groups of samples. This leads to a feature-level discrepancy between clean and adversarial samples of their augmented variants. Thus, we propose to represent groups of samples as subspaces to capture distributions and turn the traditional sample-wise adversarial fine-tuning into its distributional counterpart. For each image, we build distributions from (i) a clean sample with its augmentations and (ii) their adversarial counterparts. For text, we build distributions from (iii) a clean prompt and its synonymous prompts and (iv) their adversarial counterparts. We then perform alignment between image and text subspaces, and “adversarial” subspaces are also aligned toward “clean” subspaces. Thus, all samples underlying these distributions (think infinite number) also get aligned, leading to generalizable robustness. Evaluations on 15 datasets are provided.

1. Introduction

Deep Neural Networks (DNNs) [23, 26, 44] and Vision-Language Models (VLMs) [6, 36, 51] are vulnerable to adversarial examples [47], limiting the public trust in AI [10, 13, 15]. Enhancements of zero-shot robustness of VLMs (e.g., CLIP [43]) focus on adversarial fine-tuning [38, 45, 48] and rely on sample-wise feature-level alignment between each adversarial sample and the static text prompt class target, which inevitably fails to capture distribution trends of groups of related samples (e.g., augmented samples of a given image), degrading the robustness of zero-shot inference on unseen adversaries.

*Corresponding authors.

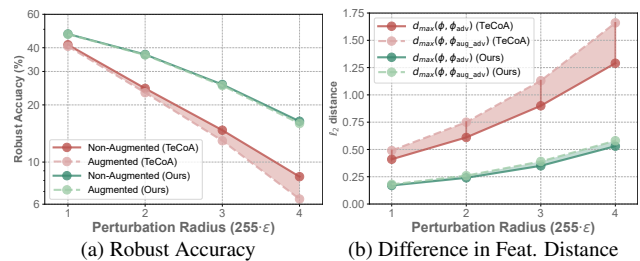


Figure 1. *Motivation (ImageNet)*. Fig. 1a: robust accuracy of non-augmented samples vs. their augmented counterparts. TeCoA [38] was trained on clean & augmented samples, and their adversaries (TeCoA aligns visual and text embeddings). However, at test time, the robust accuracy for augmented test samples drops faster w.r.t. attack radius ($255 \cdot \epsilon$) than for non-augmented test samples, creating a gap (transparent red) due to variations of test/train augmentations, further exacerbated by adversarial attacks. For our method, a group of adversarial samples on the subspace gets aligned with another group (subspace) of non-adversarial augmented & clean samples. Aligning subspaces equals aligning all samples that could lie on each subspace even if they are not in the dataset, leading to better robustness/smaller gaps. Fig. 1b: ℓ_2 feature distances (average over test embeddings). Vertical lines between curves are the feature distance gap, $d_{max}(\phi, \phi_{aug_adv}) - d_{max}(\phi, \phi_{adv})$, where $d_{max}(\cdot, \cdot)$ means we searched for the largest distance between embedding ϕ of each clean sample and its augmented adversary ϕ_{aug_adv} (or its non-augmented adversary ϕ_{adv} , respectively). The smaller gap for our model indicates that it aligns augmented adversaries with the clean sample better than TeCoA.

To substantiate our claim, we augment a set of clean samples from ImageNet by an efficient image augmentation strategy [35]. Fig. 1a shows the robust accuracy gap between augmented and non-augmented samples. For TeCoA [38], as the perturbation radius ($255 \cdot \epsilon$) grows, augmented test samples become more susceptible to attacks than clean ones. Our method remains stable, indicating that subspace alignment is superior to sample-wise alignment.

Fig. 1b depicts the distance gap $d_{max}(\phi, \phi_{aug_adv}) - d_{max}(\phi, \phi_{adv})$, where ϕ , ϕ_{aug_adv} , and ϕ_{adv} are embeddings of a clean sample, its augmented adversary, and its non-augmented adversary. Averaged results on the test set

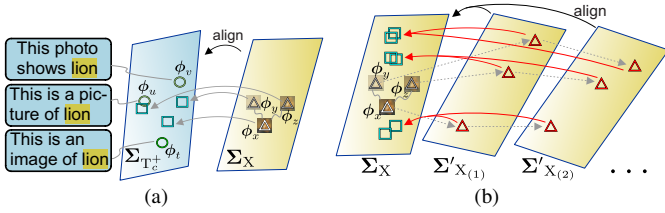


Figure 2. *Understanding subspace alignment.* Fig. 2a shows an image subspace (yellow) obtained from feature vectors (triangles) ϕ_x, ϕ_y and ϕ_z of image sample \mathbf{x} with its two augmentations, \mathbf{y} and \mathbf{z} . The text prompt subspace (blue) is obtained from feature vectors ϕ_t, ϕ_u and ϕ_v of 3 synonymous prompts $\mathbf{t}_x, \mathbf{t}_y$ and \mathbf{t}_z in set \mathcal{T}^+ . As the yellow subspace is aligned with the blue one by Eq. (5), triangles are mapped to squares on the text prompt subspace. This suggests the emergence of auxiliary synonymous prompts, not present in the prompt set \mathcal{T}^+ , benefits the model by capturing distributional “trends” within the subspaces. Following the notations of Fig. 2a, in Fig. 2b, we apply iterative adversarial generation. Three subspaces correspond to iterations $i = 0, 1, 2$. Notice that aligning “adversary” subspaces (middle, right) to the clean subspace (left) results in the emergence of auxiliary augmented images (green squares) of \mathbf{x} . Thus, although green samples are not in the dataset, the model becomes robust to their adversaries.

show that when increasing the perturbation radius, TeCoA exhibits a widening gap, suggesting higher susceptibility of augmented samples to adversaries. In contrast, our subspace alignment method enjoys a smaller distance gap.

Thus, we propose a novel adversarial fine-tuning framework for VLMs where embeddings of an image and its augmentations span a subspace that is aligned with a subspace built from a text prompt (plus its synonymous prompts) embedding. Their adversarial subspace counterparts are also aligned across image and text modalities. Thus, instead of sample-wise adversarial fine-tuning, we use its second-order variant. Fig. 2a shows how subspaces handle points as “groups”. As the “fixed” nature of a text prompt w.r.t. the ground-truth class also affects the trade-off between natural performance and robustness [27, 41], we generate adversaries by perturbing both the vision and text branches simultaneously, yielding challenging joint adversaries that we then use for adversarial fine-tuning. Usage of intermediate adversarial samples (*i.e.*, indices $1, \dots, m-1$) of a clean sample (index 0), derived from iterative adversary generation captures rich information about the decision boundary [29]. Thus, we form a joint “(intermediate) adversarial” subspace based on embeddings of image (plus augmentations) and its matching text prompt (plus synonyms) at each intermediate (or final) adv. generation step, and align it with a joint subspace from the embeddings of corresponding clean samples (Fig. 2b). We list our key contributions:

- i. Based on our observation that adversaries of augmented clean samples tend to deviate more significantly in the feature space from their clean samples than adversaries of non-augmented clean samples with the basic sample-wise adversarial fine-tuning, we propose to align

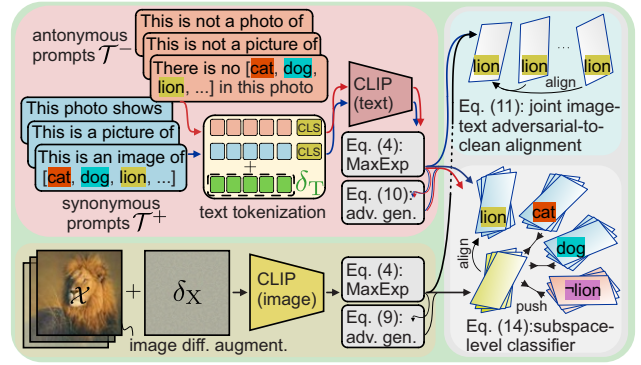


Figure 3. *Our pipeline.* We encode a set \mathcal{X} of *diff. augmentation* [35] augmented images together with the original image by CLIP (image branch). We formulate a set \mathcal{T}^+ of synonymous prompts for each class with a set \mathcal{T}^- of anonymous prompts and encode them with CLIP (text branch). For the entire image set/text set, we learn one Universal Adversarial Set Perturbation (UASP), δ_X/δ_T . The *adv. gen.* (Eq. (9) & (10)) produces $1, \dots, m$ intermediate plus final adversaries for image/text modalities. For the image set \mathcal{X} , we form “clean” and “(intermediate) adversarial” covariances. For each synonymous/anonymous text prompt set for each class c , we form covariances and approximate Grassmann feature maps by *MaxExp* (Eq. (4)) to facilitate subspace alignment by Eq. (6). The *subspace-level classifier* (Eq. (14)) aligns each image subspace (clean, intermediate adversary, final adversary) via SoftMax with the corresponding target class subspace, **lion**, while pushing away from corresponding non-target class subspaces: **cat**, **dog**, **-lion**. Finally, *joint image-text adversarial-to-clean alignment* (Eq. (11)) aligns¹ intermediate adv. subspaces with the clean one.

distributions—we deal with groups of points and match their trends (second-order moments) rather than individual samples between image and text modalities.

- ii. Embeddings of a clean sample plus its augmentations span a subspace that we align with a subspace from embeddings of a class text prompt & synonymous prompts.
- iii. We design a joint adversary generation scheme to find the worst-case adversaries for adversarial fine-tuning. By using intermediate joint image-text adversaries from iterative adversary generation (they capture rich decision boundary information), we form “(intermediate) adversarial” subspaces at each step. Aligning these adversarial subspaces with the corresponding “clean” subspaces in fine-tuning improves adversarial robustness.

Figure 3 is our pipeline detailed in Section 3.

Related works. Numerous defense schemes [2, 4, 16, 17, 49] prevent adversaries. Adversarial training [11, 12, 14, 18, 19, 37, 50] integrates adversaries into training but proves costly in VLMs [43]. Adversarial fine-tuning of VLMs [20, 21, 38, 45] can be performed by parameter-efficient strategies [28, 53]. Mao *et al.* [38] leveraged text-guided contrastive learning. Wang *et al.* [48] prevent robustness

¹Eq. (11) also uses anonymous prompts, skipped in Fig. 3 for clarity.

degradation by a guidance of the pre-trained CLIP. Schlarman *et al.* [45] study the robustness of downstream tasks. In contrast to sample-wise alignment methods, we align image and text subspaces to leverage distributional robustness during fine-tuning. Further details are in Appendix J.

2. Background

CLIP [43] contains an image encoder $f_{\theta_X} : \mathcal{X} \rightarrow \mathbb{R}^d$ and a text encoder $f_{\theta_T} : \mathcal{T} \rightarrow \mathbb{R}^d$ with parameters θ_X and θ_T . These encoders generate visual and text feature embeddings for image-text input pairs (\mathbf{x}, \mathbf{t}) . The image-text alignment is achieved by maximizing the cosine similarity between their feature representations. The probability $p_c(\mathbf{x})$ that an input image \mathbf{x} belongs to class $c \in \{1, \dots, C\}$ is defined as:

$$p_c(\mathbf{x}) = \frac{\exp(\cos(f_{\theta_X}(\mathbf{x}), f_{\theta_T}(\mathbf{t}_c)))}{\sum_{c'=1}^C \exp(\cos(f_{\theta_X}(\mathbf{x}), f_{\theta_T}(\mathbf{t}_{c'})))}, \quad (1)$$

where text prompts, $\mathbf{t}_c = g(\text{“[Context] [CLASS}_c\text{]”})$, e.g., “This is a photo of a [CLASS}_c\text{]” tokenized by $g(\cdot)$ and embedded by $f_{\theta_T}(\cdot)$ serve as alignment references. The exponential and cosine similarity functions are denoted as $\exp(\cdot)$ and $\cos(\cdot, \cdot)$. For brevity, we define vector $\mathbf{p} = [p_1, \dots, p_C]^\top \in \mathbb{R}_+^C$. Adversarial training [37] can be applied to VLMs such as CLIP to improve their robustness against (unforeseen) adversaries. Given image-text pairs in the dataset \mathcal{D} , consider the min-max problem [38] below:

$$\min_{\theta_X} \mathbb{E}_{(\mathbf{x}, c) \sim \mathcal{D}} \left[\max_{\|\delta_X\|_\infty \leq \epsilon_X} \mathcal{L}_{\text{CE}}(\mathbf{p}(\mathbf{x} + \delta_X), \mathbf{y}(c)) \right]. \quad (2)$$

The adversarial sample $\mathbf{x}' = \mathbf{x} + \delta_X$ is obtained from perturbation δ_X confined within ℓ_∞ -norm ball with an ϵ_X -radius around the clean sample \mathbf{x} . The one-hot label vector $\mathbf{y}(c) = [\mathbb{1}(c=1), \dots, \mathbb{1}(c=C)]^\top \in \{0, 1\}^C$ encodes category c . The outer minimization of the empirical risk in Eq. (2) over adversarial samples from the inner maximization step, used by studies [38, 48], leads to the iterative update:

$$\mathbf{x}'_{(i+1)} = \Pi_{\mathbb{B}(\mathbf{x}, \epsilon_X)} \left[\mathbf{x}'_{(i)} + \alpha_X \cdot \text{sign} \left(\nabla_{\mathbf{x}^{(i)}} \mathcal{L}_{\text{CE}}(\mathbf{p}(\mathbf{x}'_{(i)}), \mathbf{y}(c)) \right) \right]. \quad (3)$$

The scalar α_X is the gradient ascent step size. $\Pi_{\mathbb{B}(\mathbf{x}, \epsilon_X)}$ projects onto the ℓ_∞ -norm ball of the perturbation radius ϵ_X around clean inputs \mathbf{x} . The initial adversarial example $\mathbf{x}'_{(0)}$ is randomly initialized as $\mathbf{x}'_{(0)} \sim \mathbf{x} + 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$. After m iterations, the final adversary is $\mathbf{x}' = \mathbf{x}'_{(m)}$. The set of all adversarial examples, including intermediate ones, is denoted as $\{\mathbf{x}'_{(i)}\}_{i=1}^m$. Due to the non-linear decision boundaries, these intermediate examples are also adversarial.

3. Methodology

We propose to improve the zero-shot robustness of VLMs via adversarial fine-tuning based on subspace alignment between image and text modalities to achieve distribution-level robustness against adversaries.

Problem Definition. In contrast to conventional robustness evaluations that are restricted to the training data distribution [8], we address a more demanding *zero-shot* adversarial robustness [38]. In this setting, (white-box) attackers have unrestricted access to ground-truth data from novel datasets at the inference stage, while defenders must ensure robustness against these unforeseen adversaries without any prior access to these new datasets. Considering practical defense, text embeddings are not attacked during inference (as text class prompts can be contained within the system). The text-level (and the image-level) adversaries are only used during adversarial fine-tuning, with the objective of constructing subspaces that enhance the alignment of image and text modalities. However, for those arguing otherwise, we evaluate text-only and image-text attacks in Table 6.

Term Definitions. For clarity, we define several terms used in this paper. (i) The worst-case joint adversary is adversarial sample pair $(\mathbf{x} + \delta_X^{(m)}, \mathbf{t} + \delta_T^{(m)})$ of both image and text modalities. The “worst-case adversary” is the final m^{th} -step 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 $\{\mathbf{x} + \delta_X^{(i)}\}_{i=1}^{m-1}$ 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 $i < m$.

3.1. Image and Text Prompt Representations

To capture the distributional information (second-order statistics) of both image and text modalities for adversarially robust alignment, we form covariance matrices on the augmentation sets of each image and the text prompt representing the class label, with synonymous prompts.

Specifically, for a clean image \mathbf{x} , we use an efficient Differentiable Automatic Data Augmentation (DADA) [35] to generate its n augmented variants. We stack them in a set $\mathcal{X} = \{\mathbf{x}\} \cup \{\mathbf{x}_i : 1 \leq i \leq n\}$. Then, feature vectors from CLIP (image branch) of set \mathcal{X} are stacked column-wise into matrix $\Phi_X \in \mathbb{R}^{d \times n'}$ (where $n' = n + 1$) used by our model.

As textual data is fundamentally discrete, for the text prompt $\mathbf{t}_c^+ = g(\text{“[Context] [CLASS}_c\text{]”})$ of the c^{th} class, e.g., “[This is a photo of a][CLASS}_c\text{]”, we form $q = 20$ augmented prompts from synonyms of the context part ordered by ChatGPT-4o [40] in the descending order (from the most similar to the least similar) as in Table 14.

We also form the “opposite meaning” text prompt $\mathbf{t}_c^- =$

g (“[Negating context] [CLASS_{*c*}]” of the c^{th} class, e.g., “[This is not a photo of a] [CLASS_{*c*}]”). We collect $q=20$ antonymous phrases ordered by ChatGPT-4o in descending order from clear negating meaning down to fuzzy negatives (see Table 14). We stack synonymous and antonymous phrases into sets $\mathcal{T}_c^+ = \{\mathbf{t}_c^+\} \cup \{\mathbf{t}_{c,i}^+ : 1 \leq i \leq q\}$ and $\mathcal{T}_c^- = \{\mathbf{t}_c^-\} \cup \{\mathbf{t}_{c,i}^- : 1 \leq i \leq q\}$.

Then feature vectors from CLIP (text branch) of sets \mathcal{T}_c^+ and \mathcal{T}_c^- are stacked column-wise into matrices $\Phi_{\mathcal{T}_c^+} \in \mathbb{R}^{d \times q'}$ and $\Phi_{\mathcal{T}_c^-} \in \mathbb{R}^{d \times q'}$ (for $q' = q + 1$) used by our model.

3.2. Approximate Subspace Construction

Distribution alignment of an image to its text target is based on minimizing the so-called projection distance between subspaces [25], which requires forming Grassmann feature maps from leading singular vectors of Singular Value Decomposition (SVD) [46]. Alas, SVD is computationally costly (typically $\mathcal{O}(\min(d^2 n', n'^2 d))$ for tall matrices; $\mathcal{O}(d^{2.37})$ for covariance matrix), and its gradient is undetermined for so-called non-simple singular values $\lambda_i \approx \lambda_j : i \neq j$, which often occur in typical covariance spectra [31]. Thus, we employ fast spectral expectation of Max-pooling (MaxExp) [30–32] which only uses matrix-matrix multiplications. MaxExp approximates Grassmann feature maps:

$$\psi(\Sigma) = \mathbf{I} - (\mathbf{I} - \Sigma)^\eta \approx \mathbf{U}_{1:r} \mathbf{U}_{1:r}^T, \quad (4)$$

where $\mathbf{U}_{1:r}$ are the r leading singular vectors of SVD decomposition $\mathbf{U} \text{diag}(\Sigma) \mathbf{V}^T$ of trace-normalized covariance matrix Σ , i.e., $\Sigma := \Sigma / (\text{tr}(\Sigma) + \nu)$ where $\nu = 10^{-5}$ prevents division by zero. Parameter $0 < r < \text{rank}(\Sigma)$ represents r -dimensional linear subspace. The integer $\eta \geq 1$ controls the linear subspace dimension in the approximation. \mathbf{I} is an identity matrix. Thus, in Eq. (5) we propose an approximation of the projection distance [25] in Eq. (6):

$$d^2(\Sigma, \Sigma') = \|\psi(\Sigma) - \psi(\Sigma')\|_F^2 \quad (5)$$

$$\approx \|\mathbf{U}_{1:r} \mathbf{U}_{1:r}^T - \mathbf{U}'_{1:r} \mathbf{U}'_{1:r}{}^T\|_F^2, \quad (6)$$

where Eq. (5) approximates the projection distance between r -dimensional linear subspaces in Eq. (6). Moreover, $\mathbf{U}_{1:r}$ and $\mathbf{U}'_{1:r}$ are the r leading singular vectors of covariance matrices Σ and Σ' between which we measure the distance.

The MaxExp distance enjoys the following properties, which we leverage (proofs are in Appendix C):

1. *Low Computational Complexity:* For a $d \times d$ covariance, forming a Grassmann feature map costs $\mathcal{O}(d^3)$ due to SVD, whose internal operations cannot be easily paralleled on GPU. In contrast, MaxExp requires $\log \eta$ matrix-matrix multiplications with non-parallel and parallel compute costs $\mathcal{O}(d^{2.37})$ [3] and $\mathcal{O}(\log d)$. Thus, MaxExp best complexity is $\mathcal{O}(\log \eta \cdot \log d)$.
2. *Numerical Stability:* Unlike for SVD, the derivative of MaxExp is determined for non-simple singular values.

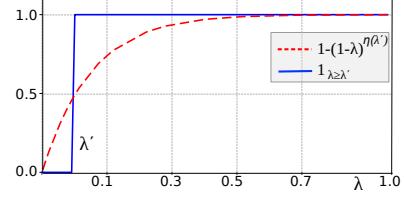


Figure 4. The push-forward functions (given a cut-off λ') map the spectrum of a covariance matrix to the spectrum of the Grassmann feature map and the spectrum of the MaxExp map, respectively.

3. *Spectrum Whitening:* For $\eta \rightarrow \infty$, $\{\lambda_i = 0 : i < j\}$ and $\{0 < \lambda_j \leq 1 \text{ (the spectrum is trace-normalized)}\}$, $1 - (1 - \lambda_j)^\eta \rightarrow 1$ while $1 - (1 - \lambda_i)^\eta = 0$, $\forall i < j$. Fig. 4: for a sufficiently big $\eta > 1$, Eq. (4) whitens the spectrum as Grassmann feature maps, except around the spectral cut-off $\lambda' = \lambda_i$ at index $i < j$. The spectrum of the Grassmann feature map is $1_{\lambda \geq \lambda'}$, i.e., 1 if $\lambda \geq \lambda'$, else 0. By analogy, for $\eta(\lambda') = \log(0.5) / \log(1 - \lambda')$, one gets a soft spectral cut-off in Eq. (4) for singular values $\lambda < \lambda'$ as $1 - (1 - \lambda')^{\eta(\lambda')} = 0.5$, $1 - (1 - \lambda)^{\eta(\lambda')} \rightarrow 0$ if $\lambda \rightarrow 0$, and it rapidly tends to 1 if $\lambda \gg \lambda'$.
4. *Robust Estimator Property:* Point 3 above suggests that any number of points $\phi \in \mathbb{R}^d$ lying on the subspace will not change their orientation if included in the estimation of covariance matrix Σ from which the subspace was obtained as long as $\|\phi\|_2 \geq \sqrt{\lambda'}$ for the cut-off singular value λ' . The point-subspace distance, given as $\bar{d}(\phi, \Sigma) = \|\phi - \mathbf{U}_{1:r} \mathbf{U}_{1:r}^T \phi\|_2$, is equal zero if ϕ is lying anywhere on the subspace, i.e., $\phi \in \text{Span}(\mathbf{U}_{1:r})$.

Point 4 above lists the properties of subspaces [25], helping us treat them as robust statistics, a hyperplane “trend” that represents all possible points that might sit on it, not merely points used in the covariance estimation. Theorem 1 formalizes this for the MaxExp approximation we use.

Theorem 1. *While the projection distance $\bar{d}(\phi, \Sigma) = 0$ for the Grassmann feature map if $\phi \in \text{Span}(\mathbf{U}_{1:r})$, the MaxExp approximation yields error $\epsilon = (1 - \lambda_i)^\eta \|\phi\|_2$ when $\cos(\phi, \mathbf{u}_i) = 1$. Thus, one may set $\eta \geq \log(\epsilon / \|\phi\|_2) / \log(1 - \lambda_i)$ to obtain at most error ϵ for the MaxExp approximation. For any $\phi \in \text{Span}(\mathbf{u}_i, \dots, \mathbf{u}_{max})$ the error is less/equal ϵ . Proof. See Appendix C. \square*

3.3. Image and Text Prompt Alignment

For an image with its n augmentations (set \mathcal{X}), its embedding matrix $\Phi_{\mathcal{X}} \in \mathbb{R}^{d \times n'}$ (for $n' = n + 1$) is used to form the corresponding positive-definite covariance matrix (second-order moment) $\Sigma_{\mathcal{X}} \in \mathbb{S}_+^{d \times d}$. See Appendix B.1 for details. Moreover, from text embedding matrices $\Phi_{\mathcal{T}_c^+} \in \mathbb{R}^{d \times q'}$ and $\Phi_{\mathcal{T}_c^-} \in \mathbb{R}^{d \times q'}$ (for $q' = q + 1$), we compute the corresponding positive-definite covariance matrices $\Sigma_{\mathcal{T}_c^+} \in \mathbb{S}_+^{d \times d}$ and $\Sigma_{\mathcal{T}_c^-} \in \mathbb{S}_+^{d \times d}$. See Appendix B.2 for details.

Subsequently, we align a subspace for the entire image set \mathcal{X} represented by the MaxExp feature map $\psi(\Sigma_{\mathcal{X}})$ with

the corresponding ‘‘positive’’ subspace of the text prompt of class c with its synonymous prompts, captured by the set \mathcal{T}_c^+ , and represented by $\psi(\Sigma_{\mathcal{T}_c^+})$. We also push $\psi(\Sigma_X)$ away from (i) maps $\{\psi(\Sigma_{\mathcal{T}_c^+}) : c' \neq c\}$ and (ii) the map of antonymous phrases $\psi(\Sigma_{\mathcal{T}_c^-})$. To this end, we define:

$$p_c(\mathcal{X} | \mathcal{T}^+, \mathcal{T}^-) = \frac{\exp(-d^2(\Sigma_X, \Sigma_{\mathcal{T}_c^+})/\rho)}{\exp(-d^2(\Sigma_X, \Sigma_{\mathcal{T}_c^-})/\rho) + \sum_{c'=1}^C \exp(-d^2(\Sigma_X, \Sigma_{\mathcal{T}_{c'}^+})/\rho)}, \quad (7)$$

where $d^2(\cdot, \cdot)$ follows Eq. (5), ρ is the temperature factor, and $p_c(\mathcal{X})$ is the probability of set \mathcal{X} belonging to class c . We stack $\mathbf{p} = [p_1(\mathcal{X}), \dots, p_C(\mathcal{X})]^\top \in \mathbb{R}_+^C$. The adversarial generation of joint image-text perturbations is defined as:

$$\min_{\theta_X} \mathbb{E}_{(\mathcal{X}, c) \sim \mathcal{D}} \max_{\substack{\|\delta_X\|_\infty \leq \epsilon_X \\ \|\delta_T\|_\infty \leq \epsilon_T}} \left[\mathcal{L}_{\text{CE}}(\mathbf{p}(\mathcal{X}' | \mathcal{T}'^+, \mathcal{T}'^-), \mathbf{y}(c)) + \beta \Omega(\mathcal{X}', \mathcal{T}'^+) \right], \quad (8)$$

$$\text{where } \mathcal{X}' = \{\mathbf{x} + \delta_X : \mathbf{x} \in \mathcal{X}\}, \mathcal{T}'^- = \{\mathbf{t} + \delta_T : \mathbf{t} \in \mathcal{T}_c^-\} \\ \text{and } \mathcal{T}'^+ = \{\{\mathbf{t} + \delta_T : \mathbf{t} \in \mathcal{T}_{c'}^+\}_{c' \in \{1, \dots, C\}}\}.$$

Moreover, δ_X is a learnable perturbation shared across the original image \mathbf{x} and its augmentations (sample set \mathcal{X}), and δ_T is a learnable perturbation that only affects the context vectors of all text prompts under set \mathcal{X} . We call such a scheme a *Universal Adversarial Set Perturbation (UASP)* as it lowers the computational cost as general universal adversarial perturbations [39] and effectively misaligns image and text modalities. $\Omega(\cdot, \cdot)$ is an additional alignment-based regularization loss whose role we will explain shortly.

Firstly, as we forego individual perturbation per augmentation (and per synonymous/antonymous phrase), we opt to generate/use intermediate UASPs in addition to the final UASP from Eq. (8). Intermediate adversaries can capture a non-linear gradient ascent path, gaining extra information about the decision boundary. For image modality, we have:

$$\delta_X^{(i+1)} = \vartheta_{\alpha_X}(\delta_X^{(i)}, \mathcal{X}'_{(i)}, \mathcal{T}'_{(i)}^+, \mathcal{T}'_{(i)}^-) = \Pi_{\mathbb{B}(\epsilon_X)} \left[\delta_X^{(i)} + \alpha_X \cdot \text{sign} \left(\nabla_{\delta_X^{(i)}} \mathcal{L}_{\text{CE}}(\mathbf{p}(\mathcal{X}'_{(i)} | \mathcal{T}'_{(i)}^+, \mathcal{T}'_{(i)}^-), \mathbf{y}(c)) \right) + \beta \Omega(\mathcal{X}'_{(i)}, \mathcal{T}'_{(i)}^+) \right], \quad (9)$$

$$\text{where } \mathcal{X}'_{(i)} = \{\mathbf{x} + \delta_X^{(i)} : \mathbf{x} \in \mathcal{X}\}, \mathcal{T}'_{(i)}^- = \{\mathbf{t} + \delta_T^{(i)} : \mathbf{t} \in \mathcal{T}_c^-\} \\ \text{and } \mathcal{T}'_{(i)}^+ = \{\{\mathbf{t} + \delta_T^{(i)} : \mathbf{t} \in \mathcal{T}_{c'}^+\}_{c' \in \{1, \dots, C\}}\}.$$

$\Pi_{\mathbb{B}(\epsilon)}(\cdot)$ denotes the projection into the ℓ_∞ -norm ball with radius ϵ . By analogy, for text modality we have:

$$\delta_T^{(i+1)} = \vartheta_{\alpha_T}(\delta_T^{(i)}, \mathcal{X}'_i, \mathcal{T}'_i^+, \mathcal{T}'_i^-) = \Pi_{\mathbb{B}(\epsilon_T)} \left[\delta_T^{(i)} + \alpha_T \cdot \text{sign} \left(\nabla_{\delta_T^{(i)}} \mathcal{L}_{\text{CE}}(\mathbf{p}(\mathcal{X}'_i | \mathcal{T}'_i^+, \mathcal{T}'_i^-), \mathbf{y}(c)) \right) + \beta \Omega(\mathcal{X}'_i, \mathcal{T}'_i^+) \right]. \quad (10)$$

Instead of the inner maximization in Eq. (8), we run Eq. (9) & (10) in the inner loop $i = 0, \dots, m$ to generate/use clean, intermediate and final adversary sets for image and text modalities, *i.e.*, $\mathcal{X}_* = \{\mathcal{X}'_{(i)}\}_{i=0}^m$, $\mathcal{T}_*^+ = \{\mathcal{T}'_{(i)}^+\}_{i=0}^m$ and $\mathcal{T}_*^- = \{\mathcal{T}'_{(i)}^-\}_{i=0}^m$. Note that $\mathcal{X}'_{(0)} = \mathcal{X}$, $\mathcal{T}'_{(0)}^+ = \mathcal{T}^+$ and $\mathcal{T}'_{(0)}^- = \mathcal{T}^-$. Similarly, $\mathcal{X}'_{(m)} = \mathcal{X}'$, $\mathcal{T}'_{(m)}^+ = \mathcal{T}'^+$ and $\mathcal{T}'_{(m)}^- = \mathcal{T}'^-$. Thus, we have $(m-1)$ adversarial perturbation patterns per original image, and $2(m-1)$ per original text prompt and its ‘‘negating meaning’’ text prompt.

3.4. Regularization: Joint Image-Text Adversarial-to-Clean Subspace Alignment

Note that \mathcal{L}_{CE} in Eq. (8) is label-driven, *e.g.*, the minimization step strives for an adversarial sample to be correctly classified. However, to achieve further distributional robustness against unforeseen adversaries, we propose intermediate sets of image augmentations and text-synonymous phrases (given index $i > 0$) to form a joint intermediate adversarial image-text subspace and align it with a joint ‘‘clean’’ image-text subspace. Such a step ignores classification *per se* but helps reduce variability, *e.g.*, intermediate adversarial subspaces roughly align with the ‘‘clean’’ subspace to lower the complexity of decision boundaries.

Let $\{\Phi'_{X_{(i)}}\}_{i \in \{1, \dots, m\}}$ and $\{\Phi'_{T_{(i)}^+}\}_{i \in \{1, \dots, m\}}$ be intermediate (and final if $i = m$) adversarial feature embedding matrices for an image with its augmentations, and a text prompt of class c with the synonymous prompts. For each given adversarial level $1 \leq i \leq m$ (and separately for ‘‘clean’’ level), we concatenate image and text feature embedding matrices along the sample mode, and obtain joint image-text covariance matrices $\{\Sigma'_{\text{XT}_{(i)}} \in \mathbb{S}_+^{d \times d}\}_{i \in \{1, \dots, m\}}$ and ‘‘clean’’ $\Sigma_{\text{XT}^+} \in \mathbb{S}_+^{d \times d}$. Then, we propose the following term:

$$\Omega_*(\mathcal{X}_*, \mathcal{T}_*^+, \mathcal{T}_*^-) = \sum_{i=1}^m w_i d^2(\Sigma'_{\text{XT}_{(i)}}, \Sigma_{\text{XT}^+}), \quad (11) \\ \text{where } \delta_i = |p_c(\mathcal{X} | \mathcal{T}^+, \mathcal{T}^-) - p_c(\mathcal{X}'_{(i)} | \mathcal{T}'_{(i)}^+, \mathcal{T}'_{(i)}^-)| \\ \text{and } w_i = \delta_i / \max_{1 \leq j \leq m} \delta_j. \quad (12)$$

As the adversarial strength of each intermediate adversarial subspace varies due to the non-linear gradient ascent path, weight w_i puts more attention when an adversarial subspace leads to larger category perturbation (larger security threat).

Theorem 2. *The weighting mechanism in Eq. (12) captures the localized κ -Lipschitz smoothness of image/text modality branches. For $i = 1, \dots, m$, the following inequalities hold:*

$$\frac{|\delta_i|}{\max(\epsilon_X, \epsilon_T)} \leq \frac{|\delta_i|}{\max \left(\max_{\mathbf{x}' \in \mathcal{X}'_{(i)}} \|\mathbf{x} - \mathbf{x}'\|_\infty, \max_{\mathbf{t}' \in \mathcal{T}'_{(i)}} \|\mathbf{t} - \mathbf{t}'\|_\infty \right)} \\ \leq \kappa \leq \frac{|\delta_i|}{\min(\alpha_X, \alpha_T)}, \forall i = 1, \dots, m. \quad (13)$$

Algorithm 1 Alignment of Subspaces (AoS).

Input: CLIP ($f_{\theta_X}, f_{\theta_T}$); dataset \mathcal{D} of c classes; hyper-parameter m, β, ρ .

```
1: while  $\neg$  converged do
  Form image augmentations  $\mathcal{X}'_{(0)} = \mathcal{X}$ ,
  synonym & antonym augmentations  $\mathcal{T}'_{(0)} = \mathcal{T}^+ \& \mathcal{T}'_{(0)} = \mathcal{T}^-$ 
2:
3: for  $i = 0, 1, \dots, m - 1$  do
4:   Draw  $\delta_X^{(0)} \sim 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $\delta_T^{(0)} \sim 0.001 \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$ 
   Generate image and text adversarial perturbations:
5:    $\delta_X^{(i+1)} = \vartheta_{\alpha_X}(\delta_X^{(i)}, \mathcal{X}'_{(i)}, \mathcal{T}'_{(i)}^+, \mathcal{T}'_{(i)}^-)$ , Eq. (9)
    $\delta_T^{(i+1)} = \vartheta_{\alpha_T}(\delta_T^{(i)}, \mathcal{X}'_{(i)}, \mathcal{T}'_{(i)}^+, \mathcal{T}'_{(i)}^-)$ , Eq. (10)
   Set  $\mathcal{X}'_{(i+1)} = \{\mathbf{x} + \delta_X^{(i+1)} : \mathbf{x} \in \mathcal{X}\}$ ,
6:    $\mathcal{T}'_{(i+1)} = \{\mathbf{t} + \delta_T^{(i+1)} : \mathbf{t} \in \mathcal{T}_c^-\}$ ,
    $\mathcal{T}'_{(i+1)} = \{\{\mathbf{t} + \delta_T^{(i+1)} : \mathbf{t} \in \mathcal{T}_{c'}^+\}_{c' \in \{1, \dots, c\}}\}$ 
7: end for
8:    $\mathcal{L} = \left[ \sum_{i=0}^m \mathcal{L}_{\text{CE}}(\mathbf{p}(\mathcal{X}'_{(i)} | \mathcal{T}'_{(i)}^+, \mathcal{T}'_{(i)}^-), \mathbf{y}(c)) \right]$  (14)
    $+\beta \Omega_*(\mathcal{X}_*, \mathcal{T}_*^+, \mathcal{T}_*^-)$ , where  $\mathcal{X}_* = \{\mathcal{X}'_{(i)}\}_{i=0}^m$ ,
    $\mathcal{T}_*^+ = \{\mathcal{T}'_{(i)}^+\}_{i=0}^m$  and  $\mathcal{T}_*^- = \{\mathcal{T}'_{(i)}^-\}_{i=0}^m$ 
9:   Update student network parameters:
    $\theta_X \leftarrow \theta_X - \rho \nabla_{\theta_X} \mathcal{L}$ 
10: end while
11: return  $\theta_X$ .
```

Proof. Eq. (13) arises from the definition of the Lipschitz smoothness. The localized nature of the smoothness is due to finite adversarial sets, *i.e.*, $\mathcal{X}'_{(i)}$ and $\mathcal{T}'_{(i)} = \mathcal{T}'_{(i)}^+ \cup \mathcal{T}'_{(i)}^-$, $\forall_{i=1, \dots, m}$. Samples \mathbf{x}' and \mathbf{t}' in these sets satisfy $\alpha_X \leq \|\mathbf{x} - \mathbf{x}'\|_\infty \leq \epsilon_X$ and $\alpha_T \leq \|\mathbf{t} - \mathbf{t}'\|_\infty \leq \epsilon_T$, where α_X and α_T are the step-sizes for adversary generation, while ϵ_X and ϵ_T are the maximum perturbation radii. \square

Theorem 2 shows that Eq. (11) improves localized Lipschitz smoothness of our adversarially robust VLM. Intuitively, this means a more stable decision boundary between subspaces (groups of points), contributing to the model’s more generalized robustness against adv. perturbations.

The final objective (Algorithm 1). Expressing our model as a simple min-max problem in Eq. (8) may be somewhat challenging as we alternate between (i) generating intermediate/final adversaries by Eq. (9) & (10) (equiv. of maximization in the inner loop of Eq. (8)), and (ii) minimizing the loss \mathcal{L} as defined in Algorithm 1 where $\beta \geq 0$ controls the localized Lipschitz smoothness discussed in Theorem 2. During the inference stage, we directly use the adversarially fine-tuned CLIP model for robustness evaluations.

All steps in our method are differentiable (including image augmentation and subspace construction in Eq. (4)). We use PGD for adversary generation and the SGD optimizer for VLM parameter optimization. Table 12 shows training times. Eq. (14) has the complexity $\mathcal{O}(m \cdot d^{2.371} \cdot \log \eta)$ (m : steps, d : feature size, η : hyper-parameter in Eq. (4)) but the GPU parallelized cost is $\mathcal{O}(\log m \cdot \log d \cdot \log \eta)$.

4. Experiments

Below, we compare our *Alignment of Subspaces (AoS)* model with other adversarial fine-tuning approaches.

Datasets. Following Mao *et al.* [38], we adversarially fine-tune CLIP (vision branch) on the ImageNet training set [9]. We evaluate the robustness of the ImageNet test set and 14 other zero-shot test sets. (See Appendix A.1 for details.)

Implementation details. As in prior studies [38, 48], we use CLIP [43] with the ViT-Base/32 architecture [22], unless specified otherwise. We set the image augmentation and text prompt synonym/antonym numbers $n = q = 20$ for subspace construction. During fine-tuning, we generate universal adversarial perturbations using 3-step Projected Gradient Descent (PGD) [37] under the ℓ_∞ -norm threat model for both the image and text levels. For image- and text-level perturbations, we adopt a maximum perturbation radii $\epsilon_X = 1/255$ and $\epsilon_T = 2 \times 10^{-4}$ with step sizes $\alpha_X = 1/255$ and $\alpha_T = 1 \times 10^{-4}$, respectively. Following [38, 48], we assess robustness against three strong white-box attacks: PGD [37] with 20 steps, CW attack [5], and Auto-Attack (AA) [7]. All evaluations are conducted under adaptive attack schemes for fairness. Further details are in Appendix A.3.

4.1. Main Results

Zero-shot performance of our method. Tables 1 and 2 compare our method (AoS) with state-of-the-art adversarial fine-tuning models on vision encoder of CLIP. Zero-shot inference on non-ImageNet datasets assesses generalization capabilities. We report accuracy on clean samples and their adversarial counterparts generated by 20-step PGD across 15 datasets. Table 1 shows that AoS achieves a 2% gain on average clean accuracy over other models, approaching vanilla CLIP. Table 2: vanilla CLIP achieves a mere 7.2% adv. accuracy. Our AoS scores 43.8%, beating the best competitor, FARE [45], by 4% in robustness on 15 datasets.

Robustness across various perturbation radii. Table 3 assesses zero-shot robustness under $\epsilon_X \geq 1/255$. Under stronger attacks, our AoS maintains consistent adv. robustness, outperforming other robust CLIP models.

Adversarial fine-tuning with larger perturbation radii. Prior works [38, 48] use a default radius $\epsilon_X = 1/255$ during fine-tuning. We investigate larger ℓ_∞ -norm perturbation radii $\epsilon_X = 2/255, 3/255, \text{ and } 4/255$. For consistency, we evaluate the robustness of each model using adversaries generated with the identical perturbation radius. Table 4 shows our AoS consistently outperforms other methods.

Robustness with different vision backbones. We apply stronger adversarial attacks, CW [5] and AA [7], on ViT-L and ResNet-50 in addition to ViT-B. Table 5 shows our AoS surpasses other adversarial fine-tuning methods in both natural performance and adversarial robustness on 15 datasets.

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

Method	ImageNet	STL10	CIFAR-10	CIFAR-100	SUN397	Stanf.Cars	Food101	OxfordPet	Flower102	DTD	EuroSAT	FGVC	PCAM	Caltech101	Caltech256	Average
CLIP [43]	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 [38]	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 [48]	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 [45]	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 (Ours)	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

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

Method	ImageNet	STL10	CIFAR-10	CIFAR-100	SUN397	Stanf.Cars	Food101	OxfordPet	Flower102	DTD	EuroSAT	FGVC	PCAM	Caltech101	Caltech256	Average
CLIP [43]	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 [38]	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 [48]	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 [45]	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 (Ours)	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

Table 3. Average robust accuracy (%) under PGD-20 attacks on 15 datasets across diverse perturbation radii set in **evaluation only**.

Method	Robust Accuracy			
	1/255	2/255	3/255	4/255
TeCoA [38]	38.91	25.43	14.72	8.38
PMG-FT [48]	39.72	23.38	12.70	6.58
FARE [45]	37.93	24.87	13.37	7.74
AoS	43.88	26.70	15.83	9.51

Table 4. Average performance (%) of diverse adversary generation setups for **adversarial fine-tuning and robustness evaluations**.

Radius ϵ	Method	Clean	PGD	CW	AA
2/255	TeCoA [38]	49.10	26.86	26.07	25.33
	PMG-FT [48]	50.72	29.38	28.41	27.66
	FARE [45]	51.09	28.55	27.79	27.18
	AoS	51.86	30.48	29.45	28.70
3/255	TeCoA [38]	42.90	19.59	18.69	17.94
	PMG-FT [48]	43.09	22.56	20.61	18.12
	FARE [45]	43.24	22.30	20.85	18.32
	AoS	43.98	23.64	21.78	19.16
4/255	TeCoA [38]	37.67	14.80	13.75	12.96
	PMG-FT [48]	37.84	17.03	15.18	13.29
	FARE [45]	37.95	16.57	14.21	13.43
	AoS	38.49	18.38	16.40	14.05

Robustness against text- & bi-level attacks. Table 6 shows AoS outperforms other adv. fine-tuning models by $\sim 5\%$ given *text-level attacks*: BERT-Attack [34] & Gradient-Based Distributional Attack (GBDA) [24], and *bi-level attacks* using Collaborative Multimodal Adversarial Attack (Co-Attack) [51] & Set-level Guidance Attack (SGA) [36].

Efficient adversarial fine-tuning with VPT. Visual Prompt Tuning (VPT) [28] enjoys fast fine-tuning due to few learnable parameters at the visual token level. Table 7 shows zero-shot accuracy on clean/adv. samples for robust VPTs, trained/evaluated on the same adv. perturbation radius ϵ and parameter reduction: ViT-B/32 (86M \rightarrow 0.092M), ViT-L/14 (307M \rightarrow 0.246M), ResNet-50 (25.6M \rightarrow 0.11M), each amounting to $< 1\%$ of total parameters.

Extension to BLIP. We apply AoS+BLIP [33] for image-text retrieval and image captioning. Table 8 (setup in Ap-

Table 5. Average clean and robust accuracy (%) of various vision backbones of CLIP with the perturbation radius of $\epsilon_X = 1/255$.

Architecture	Method	Clean	PGD	CW	AA
ViT-B	TeCoA [38]	52.62	38.91	37.85	37.62
	PMG-FT [48]	57.36	39.72	38.70	38.09
	FARE [45]	59.67	37.93	37.56	37.18
	AoS	61.70	43.88	42.94	42.18
ViT-L	TeCoA [38]	66.39	42.86	39.08	38.43
	PMG-FT [48]	67.11	43.64	39.56	38.91
	FARE [45]	67.71	43.18	40.23	39.62
	AoS	68.41	45.76	44.13	43.49
ResNet-50	TeCoA [38]	42.12	28.45	27.72	27.13
	PMG-FT [48]	46.03	30.66	29.20	28.36
	FARE [45]	48.53	29.16	28.41	27.83
	AoS	49.60	32.95	32.14	31.57

Table 6. Average robust accuracy (%) of various 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 [38]	36.22	34.97	25.87	25.14
PMG-FT [48]	37.25	36.73	26.95	26.76
FARE [45]	35.76	35.08	25.10	24.87
AoS	40.83	40.37	30.24	29.51

Table 7. Average robust accuracy (%) under various adversary generation setups for both fine-tuning and evaluations using VPT.

Radius ϵ	Method	Clean	PGD	CW	AA
1/255	TeCoA [38]	51.00	32.27	31.11	30.26
	PMG-FT [48]	52.64	33.09	32.10	30.83
	FARE [45]	52.75	32.69	31.58	30.64
	AoS	54.43	34.38	33.27	32.05
2/255	TeCoA [38]	42.61	18.12	16.88	15.39
	PMG-FT [48]	42.11	19.26	17.68	16.47
	FARE [45]	42.81	18.98	17.46	16.35
	AoS	43.84	20.47	18.60	17.29
3/255	TeCoA [38]	33.86	12.32	10.78	8.89
	PMG-FT [48]	32.52	12.87	11.36	9.38
	FARE [45]	33.70	12.47	10.92	9.04
	AoS	35.13	13.79	12.42	10.14
4/255	TeCoA [38]	26.78	11.04	9.87	7.19
	PMG-FT [48]	23.57	11.73	10.01	7.26
	FARE [45]	26.17	11.49	10.32	7.53
	AoS	27.92	13.17	11.50	8.87

pendix A.3) shows AoS enjoys the highest zero-shot performance in Flickr [42] for image-text retrieval and Nocaps [1]

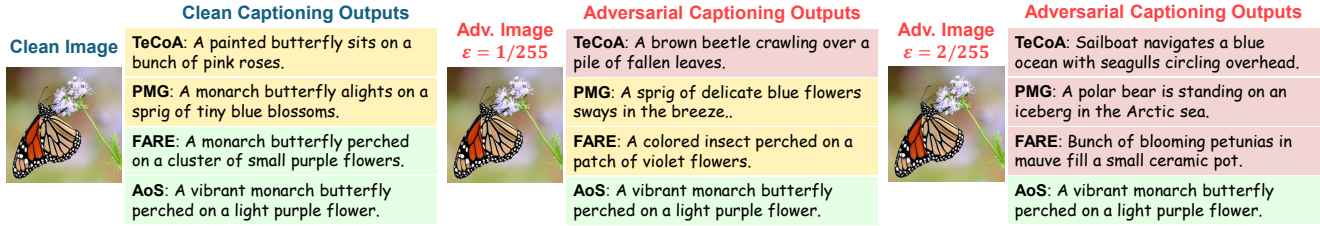


Figure 5. Captioning using robust BLIP. We compare our AoS with other adversarial fine-tuning methods with different perturbation radii.

Table 8. Extension in the context of BLIP on clean and PGD-20 adversarial samples for image-text retrieval and image captioning.

Method	Image-Text Retrieval				Image Captioning	
	Clean		Robust		Clean	Robust
	TR	TR	IR	IR	CIDEr	CIDEr
TeCoA [38]	87.5	54.4	77.0	47.5	96.9	57.8
PMG-FT [48]	87.8	55.6	77.9	48.2	97.5	58.2
FARE [45]	88.2	55.9	78.4	49.0	98.1	58.7
AoS	91.3	58.6	80.8	51.7	101.5	62.1

Table 9. Extensions in the context of 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 [38]	0.674	0.526	0.857	0.685	0.602
PMG-FT [48]	0.692	0.538	0.850	0.688	0.619	0.495
FARE [45]	0.687	0.533	0.845	0.679	0.615	0.490
AoS	0.718	0.572	0.883	0.719	0.643	0.531

Table 10. Ablation study of main AoS components (average clean & robust accuracy (%) on 15 datasets).

	Eq. (8) w/o Ω , \mathcal{T}^-	Eq. (14) w/o Ω , \mathcal{T}^-	with \mathcal{T}^-	with Ω	Clean	PGD	AA
TeCoA [38]					52.62	38.91	37.62
1	✓				57.11	41.20	39.29
2		✓			58.85	42.75	40.82
3			✓		58.64	42.54	40.73
4				✓	61.09	42.86	40.97
5			✓	✓	60.25	43.63	41.85
7		✓	✓	✓	61.70	43.88	42.18

for image captioning. The qualitative evaluations for image captioning are in Figure 5. Our AoS with BLIP remains adv. robust, generating high-quality captions. TeCoA, PMG, and FARE fail quicker even for weak adv. attacks of $\epsilon = 1/255$.

Medical CLIP [52]. Table 9 (chest X-ray imaging) shows that AoS on medical CLIP (setting in Appendix A.3) attains superior AUC on clean/adversarial samples.

4.2. Analysis

Module ablations. We ablate: (i) “Eq. (8) w/o Ω , \mathcal{T}^- ” means no intermediate adversaries were used, no *Joint Image-Text Adversarial-to-Clean Subspace Alignment* (Ω) in Eq. (11), and no antonymous text prompts \mathcal{T}^- , (ii) Eq. (14) that incorporates intermediate adv. samples. Then we (iii) add antonymous text prompts “with \mathcal{T}^- ” and (iv) Ω . Table 10 (average on 15 datasets) shows AoS (with “negative” prompts & Ω) significantly boosts clean/robust zero-shot accuracies over TeCoA [38] (sample-wise alignment).

Adversary generation strategies. Table 11 shows that our *universal perturbations* are almost as good as *instance-wise perturbations* (each augmentation instance receives its own perturbation) but can be generated 2.5 times faster.

Table 11. Performance (%) of various adversary generation strategies for adv. fine-tuning with the average training time per epoch.

Adversary Generation Strategy	Clean	PGD	AA	Time (min)
Instance-wise perturbation	60.16	44.12	42.26	242.6
Original samples perturbation	58.42	41.43	39.85	72.9
Universal Perturbation	61.70	43.88	42.18	96.0

Table 12. Performance (%) of various augmentation numbers for subspace construction with the average training time per epoch.

Augmentation Number	Clean	PGD	AA	Time (min)
5	60.55	42.95	40.89	64.6
10	61.34	43.48	41.75	82.5
20	61.70	43.88	42.18	96.0
30	61.86	43.96	42.28	137.3
50	61.98	44.07	42.32	175.6

Table 13. Performance (%) of various covariance metrics for adv. subspace learning with the average training time per epoch.

Metric	Clean	PGD	AA	Time (min)
Frobenius norm	60.16	40.86	39.05	94.7
SVD	60.89	42.35	40.47	171.5
MaxExp	61.70	43.88	42.18	96.0

Image/text augmentation numbers. Table 12 shows how varying the number of image augmentations and synonymous/antonymous text prompts, (for brevity $n = q$), impacts zero-shot robustness. As text prompts are ordered from most meaningful to fuzzy, including $n = q = 20$ augmentations seems optimal for performance and speed.

Different covariance metrics. Table 13 shows the Frobenius dist. instead of Eq. (5) leads to large drop in robust accuracy. Using SVD to form subspaces (Eq. (6)) is costly and recovers only some accuracy due to backpropagation instability. Approximate subspaces by fast/numerically stable MaxExp yield 1.2% & 3% gain on clean/robust acc.

Appendices. A & B discuss experimental details. **C & D** discuss MaxExp/other dist. replacing Eq. (5). **E & F** study hyper-parameters & text prompt aug. **G** analyzes weighting. **H** studies EMA in covariances. **I** discusses the efficacy.

5. Conclusions

We have shown that aligning individual adversarial samples to clean samples is suboptimal compared to forming “adversarial” subspaces and aligning them with “clean” subspaces, which capture trends of groups of points. Thus, we showed subspaces formed by samples of differential image augmentation and synonymous/antonymous text prompts of class labels. Subspace alignment provides localized Lipschitz smoothness, improving decision smoothness.

Acknowledgments

This research is supported in part by National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative, partly supported through the AI Singapore Programme under the project titled “AI-based Urban Cooling Technology Development”(Award No. AISG3-TC-2024-014-SGKR), the Centre for Frontier Artificial Intelligence Research, Institute of High Performance Computing, A*Star, and the College of Computing and Data Science at Nanyang Technological University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, and Infocomm Media Development Authority. Piotr Koniusz and Hao Zhu are supported by CSIRO’s Science Digital.

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] Ahmed Aldahdooh, Wassim Hamidouche, Sid Ahmed Fezza, and Olivier Déforges. Adversarial example detection for dnn models: A review and experimental comparison. *Artificial Intelligence Review*, 55(6):4403–4462, 2022. 2
- [3] Josh Alman, Ran Duan, Virginia Vassilevska Williams, Yinzhao Xu, Zixuan Xu, and Renfei Zhou. More asymmetry yields faster matrix multiplication, 2024. 4
- [4] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI*, pages 4312–4321, 2021. 2
- [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. 6
- [6] Junxi Chen, Junhao Dong, and Xiaohua Xie. Mind the trojan horse: Image prompt adapter enabling scalable and deceptive jailbreaking. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23785–23794, 2025. 1
- [7] 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
- [8] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo DeBenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021. 3
- [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. 6
- [10] 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
- [11] Junhao Dong, Yuan Wang, Jian-Huang Lai, and Xiaohua Xie. Improving adversarially robust few-shot image classification with generalizable representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9025–9034, 2022. 2
- [12] 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
- [13] Junhao Dong, Yuan Wang, Jianhuang Lai, and Xiaohua Xie. Restricted black-box adversarial attack against deepfake face swapping. *IEEE Transactions on Information Forensics and Security*, 18:2596–2608, 2023. 1
- [14] 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. 2
- [15] 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. 1
- [16] 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. 2
- [17] 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 (CVPR)*, pages 28432–28442, 2024. 2
- [18] 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
- [19] Junhao Dong, Yuan Wang, Xiaohua Xie, Jianhuang Lai, and Yew-Soon Ong. Generalizable and discriminative representations for adversarially robust few-shot learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(3):5480–5493, 2024. 2
- [20] 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*, pages 236–247, 2025. 2

- [21] 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 *Forty-second International Conference on Machine Learning*, 2025. [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] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. [1](#)
- [24] 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](#)
- [25] Jihun Hamm and Daniel D Lee. Grassmann discriminant analysis: a unifying view on subspace-based learning. In *Proceedings of the 25th international conference on Machine learning*, pages 376–383, 2008. [4](#)
- [26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. [1](#)
- [27] Bo Huang, Mingyang Chen, Yi Wang, Junda Lu, Minhao Cheng, and Wei Wang. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24668–24677, 2023. [2](#)
- [28] 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. [2](#), [7](#)
- [29] Hamid Karimi, Tyler Derr, and Jiliang Tang. Characterizing the decision boundary of deep neural networks. *arXiv preprint arXiv:1912.11460*, 2019. [2](#)
- [30] Piotr Koniusz and Anoop Cherian. Sparse coding for third-order super-symmetric tensor descriptors with application to texture recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5395–5403, 2016. [4](#)
- [31] 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. [4](#)
- [32] Piotr Koniusz, Fei Yan, Philippe-Henri Gosselin, and Krystian Mikolajczyk. Higher-order occurrence pooling on mid- and low-level features: Visual concept detection. 2013. [4](#)
- [33] 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](#)
- [34] 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](#)
- [35] 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. [1](#), [2](#), [3](#)
- [36] 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](#)
- [37] 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](#)
- [38] 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](#), [6](#), [7](#), [8](#)
- [39] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1765–1773, 2017. [5](#)
- [40] OpenAI. Chatgpt [large language model]. <https://chatgpt.com>, 2024. [3](#)
- [41] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, pages 17258–17277. PMLR, 2022. [2](#)
- [42] 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](#)
- [43] 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](#), [6](#), [7](#)
- [44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. [1](#)
- [45] 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](#), [3](#), [6](#), [7](#), [8](#)
- [46] Christian Simon, Piotr Koniusz, Richard Nock, and Mehrtash Harandi. Adaptive subspaces for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4136–4145, 2020. [4](#)
- [47] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. [1](#)
- [48] Sibow 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](#), [3](#), [6](#), [7](#), [8](#)
- [49] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations, ICLR*, 2018. [2](#)
- [50] 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](#)
- [51] 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](#)
- [52] 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. [8](#)
- [53] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. [2](#)