A Provable Energy-Guided Test-Time Defense
Boosting Adversarial Robustness of Large Vision-Language Models

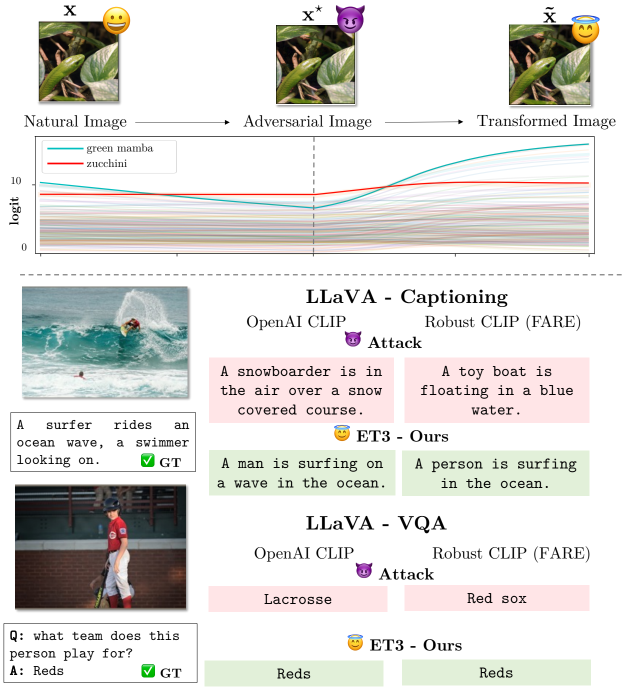
Mujtaba Hussain Mirza[1] Antonio D'Orazio[1] Odelia Melamed[2] Iacopo Masi[1]

[1]OmnAI Lab, Computer Science Department, Sapienza University of Rome, Italy
[2]Weizmann Institute of Science, Israel

{mirza, dorazio, masi}@di.uniroma1.it, odelia.melamed@weizmann.ac.il

Abstract

Despite the rapid progress in multimodal models and Large Visual-Language Models (LVLM), they remain highly susceptible to adversarial perturbations, raising serious concerns about their reliability in real-world use. While adversarial training has become the leading paradigm for building models that are robust to adversarial attacks, Test-Time Transformations (TTT) have emerged as a promising strategy to boost robustness at inference. In light of this, we propose ***Energy-Guided Test-Time Transformation*** *(*ET3*), a lightweight, training-free defense that enhances the robustness by minimizing the energy of the input samples. Our method is grounded in a theory that proves our transformation succeeds in classification under reasonable assumptions. We present extensive experiments demonstrating that* ET3 *provides a strong defense for classifiers, zero-shot classification with CLIP, and also for boosting the robustness of LVLMs in tasks such as Image Captioning and Visual Question Answering. Code will be released upon acceptance of the paper.*



A surfer rides an ocean wave, a swimmer looking on.

A snowboarder is in the air over a snow covered course.

A toy boat is floating in a blue water.

A man is surfing on a wave in the ocean.

A person is surfing in the ocean.

Q: what team does this person play for?
A: Reds

Lacrosse

Red sox

Reds

Reds

Figure 1. *(top)* Presenting a natural image green mamba $\mathbf{x}$, and its adversarial image $\mathbf{x}^\star$ mistakenly classified as a zucchini by a robust classifier f_θ. Given only $\mathbf{x}^\star$ and f_θ, our ET3 test-time defense produces the correctly classified $\tilde{\mathbf{x}}$, thereby boosting adversarial robustness. The plot illustrates the logits' change: even though the ground-truth class is *not* the second best, ET3 still recovers it. *(bottom)* ET3 boosts robust accuracy of Large VLMs like LLaVa [45] using standard or even Robust CLIP [72] on both image captioning and Visual QA. *Note in VQA example, the Reds team is a different team from the Red Sox, and refers to the Cincinnati Reds team.*

1. Introduction

Large vision–language models (LVLMs) have achieved remarkable progress, demonstrating strong multimodal reasoning and zero-shot generalization across tasks such as image captioning, visual question answering, and open-world recognition. Recent LVLMs such as LLaVA [45], OpenFlamingo [5], and Qwen-VL [6] present excellent performance; however, despite their broad generalization, they are still vulnerable to attacks to the visual modality [11, 64], potentially compromising downstream LVLM tasks.

Crucially, many LVLMs rely on the visual encoder from CLIP [65], which provides exceptional flexibility and generalization across diverse visual inputs, but also forms the primary source of vulnerability to image-based adversarial attacks: unnoticeable, carefully crafted perturbations to input images can induce incorrect predictions [27, 82].

During the past decade, a plethora of defenses have been proposed to improve robustness, with adversarial training (AT) [48] emerging as a leading efficient yet time-consuming paradigm. Despite this progress, AT-trained models still exhibit a substantial gap between their performance on clean and perturbed inputs, particularly under strong or previously unseen attacks. In contrast, another line of work focuses on test-time defenses that achieve robustness at inference. Within this class, adversarial purification methods [79] attempt to denoise and remove adversarial perturbations using auxiliary generative models, while Randomized Smoothing (RS) [16] certifies robustness by averaging predictions over noise-perturbed inputs.

More recently, test-time transformations [75, 94] have been proposed to improve classification without explicitly removing perturbations, instead adaptively transforming inputs dynamically at test time. Compared to adversarial training, these approaches defend against previously unseen threats in a plug-and-play manner without requiring model retraining. However, they often incur significant inference overhead, may rely on additional models, and sometimes struggle to scale to stronger attacks.

While test-time transformations (TTT) are promising, they can be computationally expensive, as they typically require optimizing additional tokens or parameters or generating augmented views for each sample. In this work, we present our novel and efficient test-time defense **ET3**, which demonstrates superior robustness achieved without such costly adaptation. The ET3 defense transforms the image at inference to boost robustness of CLIP while remaining lightweight and computationally efficient. Furthermore, our approach can be easily incorporated with existing methods to further improve their robustness significantly, incurring only a slight higher inference cost.

ET3 is inspired by energy-based models (EBMs) that were previously linked to robustness [28] and aims to lower the energy of a given input. Requiring extremely short optimization through the vision encoder, the defense is fast and lightweight, having a very minor overhead over the Visual LVM inference time. Fig. 1 presents a few qualitative examples of the ET3 transformation improving the robustness of classification, image captioning, and VQA. The ET3 boosts robustness for a broad range of models and tasks. In our experiments, we demonstrate increased robustness for zero-shot classification using CLIP-based vision language models and for robust image classifiers on the ImageNet dataset. We also show a significant boost in robustness in downstream tasks for LVLMs, even with a single optimization step through the CLIP encoder alone, improving robustness without noticeably impacting inference speed. Finally, we show our approach is robust against adaptive attacks.

To further support the ET3 defense, we theoretically prove its effectiveness on binary classifiers. For a binary classifiers satisfying few assumptions, we theoretically prove that an input transformed by ET3 will be correctly classified. We later provide experimental evidence of a robust model, along with examples of classifiers that have been theoretically proven to be robust in several different settings [26, 53, 54, 85], satisfying these assumptions.

2. Related Work

Adversarial Attacks. Adversarial examples are slightly perturbed inputs to induce erroneous predictions or behaviors in deep neural networks. As various methods exist to create such perturbations [3, 27, 48, 58], within the domain of LVLMs, a growing body of work has started examining their susceptibility to adversarial attacks [11, 64, 71].

Adversarial Defense. Among many defense strategies, adversarial training (AT)—which incorporates adversarial examples during training—remains the most effective empirical approach. AT has also been extended to vision language models (VLM) [51, 63, 67, 72, 89] and diffusion models [68]. In VLM, AT addresses vulnerabilities arising from multi-modal inputs by improving robustness to perturbations in either the visual or textual components. Beyond training-based defenses, offline prompt tuning methods learn a set of input tokens using a training dataset, optimizing them offline to improve accuracy or robustness without updating the backbone [42, 51, 103]. Test-time defenses constitute another direction in which the model itself is not trained to be robust. Instead, robustness is achieved at inference time through additional models or mechanisms. Within this broader class, Randomized Smoothing (RS) comprises a family of methods that certify robustness by averaging predictions over noise-perturbed inputs [16, 41], with various extensions proposed in recent works [1, 47]. Another approach is adversarial purification, which uses an auxiliary generative model to denoise the test image, removing adversarial perturbations while remaining close to the original image [32, 59, 79, 100]. In another line of work, the input to the network is adaptively transformed at the test time [2, 8, 62, 92, 94], aiming for correct classification without explicitly removing adversarial perturbations or constraining a similarity to the original image, as in [8, 94].

Test-Time Transformations for VL tasks. For Vision-Language tasks, adversarial fine-tuning overhead is significant due to the larger size of the models. Thus, test-time transformations provide lightweight adversarial defense directly at inference.

Test-time adaptation methods aim to improve the robustness or generalization by altering the prompt, adjusting the embedding, or leveraging test-time augmentations. [76] introduces Test-Time Prompt Tuning (TPT), which optimizes a text prompt by minimizing prediction entropy across augmented views. While training-free and effective under distribution shift, its multiple augmentations cause significant inference overhead and reduced clean accuracy. C-TPT mitigates this issue by introducing a calibration term that maximizes text-feature dispersion. R-TPT [75] targets adversarial robustness, optimizing the prompt against adversarial perturbations. [102] proposes MTA, which aggregates the embeddings of multiple augmentations using a robust mean-shift procedure. However, all of these methods introduce substantial inference overhead. [94] introduces TTC, which improves zero-shot adversarial robustness by applying a gradient-based perturbation that maximizes the distance between the adversarial input and its clean embedding in CLIP's feature space, yet offers limited robustness gains.

Adversarially Robust Models and EBM. Energy-based models (EBMs) [40] are generative models that learn an energy function $E(\mathbf{x})$ which assigns low energy values to inputs $\mathbf{x}$ in the data distribution and high energy values to other inputs. Intuitively, lower energy corresponds to samples that align well with the learned data distribution, while higher energy indicates atypical or out-of-distribution inputs. Several studies [22, 84, 93] explore the link between generative models such as EBMs and discriminative models. This connection is formalized in the Joint Energy-based Model (JEM) [28], which reformulates the softmax classifier within an energy-based framework. This formulation allows the classifier's output as energy can enable OOD detection [46]. The connection between learning an energy objective and robustness has been demonstrated in both ways, showing that the addition of EBM objectives to training increases robustness [28], and performing adversarial training has been shown to implicitly learn an EBM [97, 104]. Incorporating explicit energy-based objectives has also been shown to further enhance both robustness and generative performance [34, 55, 83, 98]. The generative aspects of robust classifiers have also been explored in other studies [25, 69, 91, 95, 96].

Theoretically Proven Robustness. There has been a great effort in the theoretical research to characterize robustness guarantees and measure the robustness of a trained classifier. Starting in [73], many have tried to theoretically show adversarial vulnerability. It had been shown for random networks with different natural architectures in [7, 9, 10, 20, 57]. Later, in [26, 85] in had been shown that gradient-based training converge to a non-robust networks while robust network do exist, while even offering a concrete robust classifier as an example. In [54] and [53] the author discuss a different activation function, polynomial ReLU, and prove that training such model converges to a robust classifier.

3. Method

We introduce an *energy-based test-time transformation*, ET3, grounded in the energy-based modeling (EBM) perspective of discriminative classifiers. EBMs associate each input $\mathbf{x}$ with a scalar energy $E(\mathbf{x})$, assigning lower energy to in-distribution (on-manifold) samples than to off-manifold ones. ET3 leverages this property by applying a lightweight test-time transformation that decreases the energy of the input. This is particularly relevant for adversarial examples, which are known to deviate from the natural data manifold and typically go off the manifold of natural data [56, 74, 81]. Unlike approaches that train an explicit generative EBM [32] or auxiliary diffusion/score models [79], ET3 requires no additional model training. It operates directly on a pre-trained classifier or visual encoder f_θ that we aim to defend. This is possible because a standard softmax classifier can itself be viewed through the EBM lens, by interpreting its logits as energies [28]. Our focus on robust models stems from the tight connection between energy-based and robustness objectives. For instance, adding an explicit EBM term during training can enhance adversarial robustness [28]. Conversely, adversarial training has been shown to implicitly induce an EBM with smoother local energy landscapes [55, 97, 104]. For clean inputs, ET3 either preserves the original prediction or can even increase the model's confidence. For adversarial inputs, it guides the sample back toward the correct classification, effectively enhancing the model's robustness.

3.1. Energy Definition

Consider a set of d-dimensional labeled images $X = \{(\mathbf{x}, y) \sim \mathcal{D} |\ \mathbf{x} \in \mathbb{R}^d \text{ and } y \in \{1, .., K\}\}$, let $f_\theta(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^K$ be a K-class classifier or an encoder parameterized by θ trained on a dataset X. For a sample $\mathbf{x} \in \mathbb{R}^d$, we denote by $f_\theta(\mathbf{x})_k$ the k-th coordinate of the output logits vector $f_\theta(\mathbf{x}) \in \mathbb{R}^K$. We define the energy at $\mathbf{x}$ as:

$$E(\mathbf{x}) = -\log\Big(\sum_{k=1}^{K} \exp\big(f_\theta(\mathbf{x})_k\big)\Big), \tag{1}$$

which is the negative `LogSumExp` of the output logits.

3.2. The ET3 Defense

Inspired by the connection between the energy and the perception of the data distribution by f_θ, we present a test-time defense based on energy minimization. Given a test image $\mathbf{x}$ and a defense radius ϵ, our goal is to obtain a modified image $\tilde{\mathbf{x}} = \mathbf{x} + \mathbf{z}$ by iteratively minimizing $E(\cdot, \theta)$. Concretely, we solve the constraint minimization $\tilde{\mathbf{x}} = \mathrm{argmin}_{\mathbf{x}^\star \in \mathcal{B}_\epsilon(\mathbf{x})} E(\mathbf{x}^\star)$, where $\mathcal{B}_\epsilon(\mathbf{x}) = \{\mathbf{x}^\star : \big\|\mathbf{x}^\star - \mathbf{x}\big\| \leq \epsilon\}$, using a gradient-based multi-step optimization for T steps. Formally, in each step $t \in [1, ..., T]$ we calculate:

$$\mathbf{x}^{(t)} = \Pi_{\mathcal{B}_\epsilon(\mathbf{x})}\Big(\mathbf{x}^{(t-1)} - \alpha \nabla_{\mathbf{x}} E\big(\mathbf{x}^{(t-1)}\big)\Big), \tag{2}$$

starting from $\mathbf{x}^{(0)} = \mathbf{x}$. The step size α and number of iterations T are hyperparameters, yet we found our method is fast and works for fewer steps. The projection $\Pi_{\mathcal{B}_\epsilon(\mathbf{x})}(\cdot)$ enforces the defense perturbation in a ℓ_2 ball.

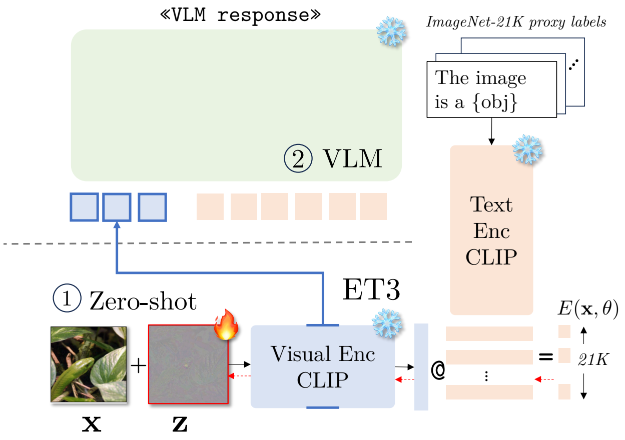


Figure 2. ① ET3 transforms the natural image $\mathbf{x}$ adding a small perturbation $\mathbf{z}$ optimized to lower the energy wrt to ImageNet-$21k$ proxy classes and concepts. This allows robust zero-shot classification; ② the transformed image transfers and protects Large VLM, thereby increasing their robustness. The VLM is *not* used in the optimization, and the optimized image simply transfers to VLM by using the internal representation of the visual encoder.

3.3. Extending ET3 to Vision-Language Tasks

We extend ET3 to zero-shot classification and vision-language tasks. The core principle remains the same: minimize the energy function defined in Eq. (1) by refining the input image at test time. Fig. 2 depicts how we apply ET3 to both zero-shot classification and VLMs.

Zero-Shot Classification with CLIP. For zero-shot classification, we apply ET3 to CLIP models, where classification is performed by computing similarity scores between image and text embeddings and serve as logits for ET3. To compute energy, we consider two distinct label configurations: a refined subset of labels, or a vast set of labels, such as the full ImageNet-21K label set [66] which yields slightly improved robustness. We primarily adopt the latter, with additional analysis presented in Appendix D.

Large Vision-Language Models (LVLMs). For multimodal tasks such as image captioning and visual question answering, we apply ET3 to the LLaVA model. Similar to the zero-shot setting, ET3 optimizes the visual input at test time by minimizing its energy through the CLIP vision encoder. After refinement, the visual embeddings from the vision encoder are passed to LLaVA's projection layer, and the subsequent language generation process remains unchanged. ET3 enhances robustness specifically in the vision modality, even without modifying or fine-tuning the vision encoder to improve its robustness. To further support ET3, next we prove that the ET3 defense method boosts robustness for a binary classifier, where it provably transforms a clean or adversarial sample into a correctly classified input.

4. ET3 defense provably boosts Robust Nets

In this section, we theoretically prove that for a binary classifier $f_\theta(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^2$, under some assumptions, minimizing the energy with an ϵ budget will transform a given sample to be classified as its original ground truth class for both natural and adversarial samples. First, for simplicity, we assume local linearity in an ϵ-ball around the sample. We note that approximate local linearity has been proved and demonstrated before in the context of adversarial attacks and robust classifier [52]. Second, we assume that the gradient of the energy derived from the ground truth label is larger than the one derived from the opposite label. This assumption is particularly intriguing and arises from the vicinity of the adversarial sample and its corresponding source sample. Details about these assumptions in Remark 4.1.

Data and Model. We consider d dimensional data inputs $\mathbf{x} \in \mathbb{R}^d$, with unknown binary ground truth labels $y_t \in \{-1, 1\}$. The incorrect label is $\hat{y_t} = -y_t$. Our model is a binary classifier with two output logits denoted by $f_\theta(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}^2$. We use a two-logit output to adapt to the multiclass classifier regime, for which each possible class relates to a logit. Similarly, for a given input $\mathbf{x} \in \mathbb{R}^d$ and its ground truth label $y_t \in \{-1, 1\}$, if $f_\theta(\mathbf{x})_{y_t} > f_\theta(\mathbf{x})_{\hat{y_t}}$, then $\mathbf{x}$ is correctly classified.

Theorem 4.1. *Let $\mathbf{x} \in \mathbb{R}^d$ be a data sample, and y_t be its ground truth label. Let $f_\theta : \mathbb{R}^d \to \mathbb{R}^2$ be a binary classifier such that it's locally linear in $\mathcal{B}_\epsilon(\mathbf{x})$. Denote $r_x = f_\theta(\mathbf{x})_{y_t} - f_\theta(\mathbf{x})_{\hat{y_t}}$, $\mathbf{g}_i = \nabla_\mathbf{x} f_\theta(\mathbf{x})_i$ and $e_i = \mathit{SoftMax}\left(f\left(\mathbf{x}\right)\right)_i$. Let $\epsilon > 0$ a defense budget such that $\epsilon > \frac{-2r_x}{\left\|\mathbf{g}_{y_t}\right\|}$. Then, if*

$$C\left\|e_{\hat{y_t}}\mathbf{g}_{\hat{y_t}}\right\| < \left\|e_{y_t}\mathbf{g}_{y_t}\right\|$$

for

$$C > \max\left\{\frac{\exp(|r_x|)\epsilon\left\|\mathbf{g}_{y_t}\right\|}{\epsilon\left\|\mathbf{g}_{y_t}\right\| + 2r_x}, 1\right\},$$

the ET3 *defense transformation $\mathbf{z}$, parametrized by $T = 1$ and $\alpha = \frac{\epsilon}{e_{y_t}\left(1+\frac{1}{C}\right)\left\|\mathbf{g}_{y_t}\right\|}$, will satisfy $\left\|\mathbf{z}\right\| \leq \epsilon$ and*

$$f_\theta(\mathbf{x}+\mathbf{z})_{y_t} > f_\theta(\mathbf{x}+\mathbf{z})_{\hat{y_t}}\ .$$

Proof Idea. Formally, given a data point $\mathbf{x} \in \mathbb{R}^d$, we denote $\mathbf{g}_i$, the gradient of the network's i-th logit w.r.t. the input and e_i the gradient of the energy w.r.t. the i-th logit. Then, the gradient of `LogSumExp` w.r.t. $\mathbf{x}$ is

$$\begin{aligned}\nabla_\mathbf{x} E(\mathbf{x}) &= -\ \mathrm{SoftMax}\left(f_\theta(\mathbf{x})\right)^\top \nabla_\mathbf{x} f_\theta(\mathbf{x}) \\ &= -\ e_{-1}\mathbf{g}_{-1} - e_1\mathbf{g}_1.\end{aligned}$$

Thus, the purification step is of the form $\mathbf{z} = \alpha\left(e_{-1}\mathbf{g}_{-1} + e_1\mathbf{g}_1\right)$. To see the effect of the transformation on each logit, we use the local linearity and look at the

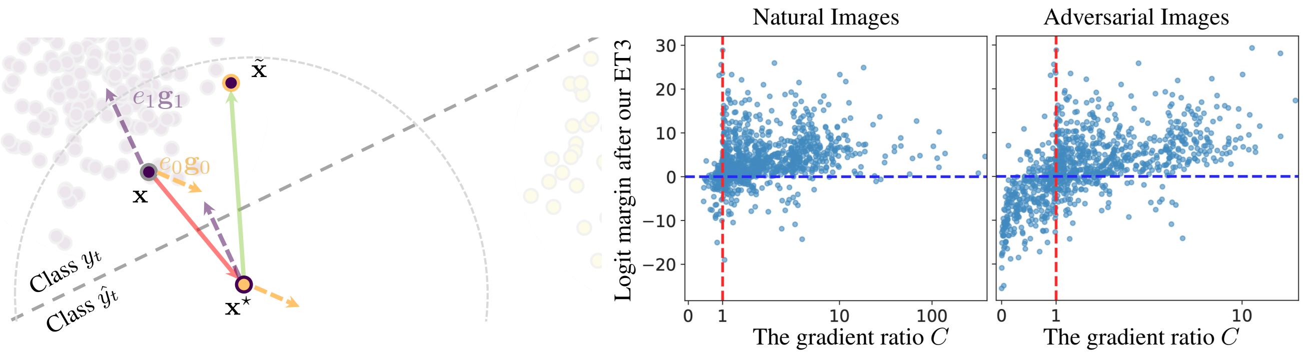


Figure 3. *(left)* The ET3 defense transformation for adversarial examples. Assuming local linearity of the model in the defense neighborhood $\mathcal{B}_\epsilon(\mathbf{x})$, and a large enough ratio C between the norms of the gradients of the energy through each class logit, $e_0\mathbf{g}_0$ and $e_1\mathbf{g}_1$. The adversarial attack, determined to reduce the ground truth logit, follows the negative direction of the larger gradient ($-\mathbf{g}_1$), while our transformation follows its positive direction ($\mathbf{g}_1$), increasing the ground truth logit and pulling the adversarial point back to its ground truth region. Both might also increase the other logit, corresponding to the smaller gradient $\mathbf{g}_0$, that may introduce some smaller deviation. *(right)* Scatter plot of the ratio between the gradients norms C and the logit margin at the transform image $\tilde{\mathbf{x}}$ on ImageNet robust classifier for 1000 randomly sampled images from ImageNet. For most samples for which $C > 1$, we can see that the purified image is correctly classified (logit margin > 0). One can see a correlation between the norms ratio and the logits difference of the transformed image.

classifier as two locally linear functions $f_{-1}, f_1 : \mathbb{R}^d \to \mathbb{R}$, one for each logit. We observe that for $i \in \{-1, 1\}$

$$f_i(\mathbf{x}+\mathbf{z}) = f_i(\mathbf{x}) + \mathbf{z}^\top \mathbf{g}_i.$$

We complete the proof by using the ratio of the gradients' norms, C, to show that multiplying by the larger norm gradient, $\mathbf{z}^\top \mathbf{g}_{y_t}$, will create a larger effect, implying correct classification. Formally, we note that the energy function allows a relaxed assumption on the ratio of the gradients as:

$$\exp\left(f_\theta(\mathbf{x})_{\hat{y}_t} - f_\theta(\mathbf{x})_{y_t}\right) C \left\|\mathbf{g}_{\hat{y}_t}\right\| < \left\|\mathbf{g}_{y_t}\right\|,$$

that is adaptive to the logits of $f_\theta(\mathbf{x})$. When $\mathbf{x}$ is classified correctly, we will have $f_\theta(\mathbf{x})_{\hat{y}_t} - f_\theta(\mathbf{x})_{y_t} \leq 0$ and thus $\exp\left(f_\theta(\mathbf{x})_{\hat{y}_t} - f_\theta(\mathbf{x})_{y_t}\right) \leq 1$ allowing a smaller ratio between the gradients to preserve correct classification. Otherwise, the assumed ratio will be weighted by the difference between the logits. The full proof provided in Appendix E.

Remark 4.1 (The Local Linearity and Gradient Norm Radio Assumptions)**.** ***In Practice.*** In the right part of Fig. 3 we present on the x-axis, for each natural image of ImageNet and its respective adversarial example, the norm ratio of the gradients $e_{y_t}\mathbf{g}_{y_t}$ and $e_{y_{\text{adv}}}\mathbf{g}_{y_{\text{adv}}}$, for which we used the APGD attack [17] and a robust classifier [23]. All inputs are then transformed using our method ET3, and the difference between the resulting output logits is presented in the y-axis. One can see that many natural and adversarial example has high ratio values, i.e., $C \gg 1$. Moreover, one can see that transforming samples with a higher ratio C results in a larger logit difference for the transformed image, indicating successful transformation. For the locally linearity assumption, previous works showed that robust networks are experimentally approximately locally linear [52].

In Theory. When proving robustness of a model, a helpful property is have an approximate locally linear model, allowing a more immediate upper bound on the effectiveness of a worst-case perturbation. To this end, and to also allow correct classification, one idea for a robust network construction will be to assign higher inner products between the data points and the local gradients corresponding to the ground truth label, and smaller inner product with the others, or assigning a sufficiently large bias. In previous works, [26, 85] construct a robust network enforcing both local linearity with a large radius and a larger gradient of the ground truth label in the settings of almost orthogonal data. In [54] and [53], the network trained using polynomial ReLU is approximately locally linear and with a similar larger ground truth gradient. Sec. 4.1 gives more details on the robust construction from [26] for which both assumptions hold, adapting the classic binary classifier to our two-logit regime.

We note that as there are no assumptions of $\mathbf{x}$, the ET3 defense will be successful for both clean sampled and adversarial data input, yet the assumption are milder on a clean data sample. First, since $r_x > 0$ for a clean sample, the assumption on ϵ will hold for any $\epsilon > 0$. Second, $\exp\left(f_\theta(\mathbf{x})_{\hat{y}_t} - f_\theta(\mathbf{x})_{y_t}\right)$ will be smaller than 1, inducing a relaxed lower bound for the ratio C.

This theorem extends to the case of multi-step transformation within the same budget. Since the classifier is locally linear, each step to minimize the energy, will still increase the ground truth logit and the gap between it and the other logit. In other words, the theorem shows that

$$f_\theta(\mathbf{x}+\mathbf{z})_{y_t} - f_\theta(\mathbf{x}+\mathbf{z})_{\hat{y}_t} > f_\theta(\mathbf{x})_{y_t} - f_\theta(\mathbf{x})_{\hat{y}_t}$$

for a transformation $\mathbf{z}$ of any size. Therefore, the new input $\mathbf{x}+\mathbf{z}$ will preserve the assumptions on C and ϵ, allowing a multi-step attack with the same budget ϵ to take multiple small gradient steps till accumulated to a sufficient size.

4.1. Theoretical Application

We look at the construction for a robust two-layer ReLU network, stated in [26].

Data. We use the following data distribution $\mathcal{D}_{\text{clusters}}$ on $\mathbb{R}^d \times \{-1, 1\}$ in which we have k clusters with means $\mu^{(1)}, ..., \mu^{(k)} \in \mathbb{R}^d$ and covariance $\sigma^2 I_d$, where examples in the j-th cluster are labeled $y^{(j)} \in \{-1, 1\}$.

Model. We adapt the two-layer binary classification network into our multi-class network. For a given binary classifier $f_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{j \in J} v_j \sigma(\mathbf{w}_j^T \mathbf{x} + b_j)$, we define a two-logits output classifier as following. We define $J_+ = \{j : v_j \geq 0\}$ and $J_- = \{j : v_j < 0\}$, and denote for $l \in \{-1, 1\}$

$$\overline{f}_{\boldsymbol{\theta}}(\mathbf{x})_l = l \sum_{j \in J_i} v_j \sigma(\mathbf{w}_j^T \mathbf{x} + b_j) \ ,$$

We note that for any $\mathbf{x}$, $f_{\boldsymbol{\theta}}(\mathbf{x}) \equiv \overline{f}_{\boldsymbol{\theta}}(\mathbf{x})_1 - \overline{f}_{\boldsymbol{\theta}}(\mathbf{x})_{-1}$.

In [26], the authors present a robust network, proving its robustness to a large $O(\sqrt{d})$ adversarial perturbation.

For the robust network of width k, they take for any $j \in [k]$, $v_j = y^{(j)}$ and $\mathbf{w}_j = \frac{4\mu^{(j)}}{d}$ and $b_j = -2$ (for wider networks, i.e. $J > k$, we set the extra weights to be zero).

Let $(\mathbf{x}, y) \in \mathcal{D}_{\text{clusters}}$, meaning that $\mathbf{x} = \mu^{(q)} + \xi$ for some $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ and $y = y^{(q)}$. For the gradient ratio and local linearity assumptions, we look at the analysis for $\mathbf{w}_j^T \mathbf{x} + b_j$ for different $j \in [k]$, and get that for $j = q$ $\mathbf{w}_q^T \mathbf{x} + b_q > 1$, while for $j \neq q$, $\mathbf{w}_j^T \mathbf{x} + b_q < -1$. Thus, denoting $\mathbb{1}_{\{A\}}$ as the indicator function for an event A, we have w.h.p.

$$\mathbf{g}_{y_t} = \frac{\partial \overline{f}_{\theta}(\mathbf{x})_{y_t}}{\partial \mathbf{x}} = \sum_{j \in J_+} \mathbf{w}_j \mathbb{1}_{\{\mathbf{w}_j^T \mathbf{x} + b_j \geq 0\}} = y \mathbf{w}_q \ ,$$

and $\mathbf{g}_{\hat{y}_t} = 0$, satisfying the ratio assumption. In Theorem 4.1 the authors show that w.h.p. over $\mathbf{x}$, the network is locally linear in $\mathcal{B}_\epsilon(\mathbf{x})$ for the large $\epsilon \leq \frac{\sqrt{d}}{8}$. Applying the ET3 defense, we can see that the transformation will be $\mathbf{z} = \alpha \mathbf{w}_q$ for some $\alpha > 0$, which will ensure correct classification for the transformed input.

5. Experiments

We evaluate our defense in multiple settings to demonstrate its ability to boost the robustness of adversarially trained models. We show that our test-time defense in zero-shot classification can boost robustness alone and when combined with test-time augmentation and test-time prompt tuning approaches. Additionally, we demonstrate that CLIP vision encoder's embeddings obtained after ET3, when used in downstream LVLMs, enhance their resilience to attacks. Finally, we present the defense success under adaptive attacks for classifiers.

5.1. Evaluating Zero-Shot Classification

We perform zero-shot classification using CLIP models with the robust vision encoders TeCoA [50] and FARE [72], and evaluate across 15 benchmark datasets—full implementation details can be found in Appendix A.

Attack setup. Unless stated otherwise, following standard practice [72], we use the two attacks from AutoAttack [17]: APGD-CE and APGD-DLR with 100 iterations each.

Results. In Tab. 1 we show our test-time defense ET3 consistently enhances robust accuracy on adversarially robust models (TeCoA and FARE) across 14 datasets under attacks with a perturbation magnitude of $\epsilon_a = 4/255$. These models trained against attacks with either $\epsilon_t = 2/255$ or $4/255$, where ET3 show substantial improvements even with the "weaker" models, which were trained against the weaker attacks. This trend is further illustrated in Fig. 4, where ET3 improves robustness even for larger magnitude attacks—see Appendix C for per-dataset results and additional details. Next, in Tab. 2 we show that ET3 surpasses similar defense methods in diverse defense scopes. Following the comparison standardization of the previous methods, we report results on a subset of 8 datasets. In the **lightweight defense scope**, ET3 consistently improves robust accuracy over baseline CLIP model and TTC [94], and surpasses even the slower TPT and C-TPT methods [76, 99], despite requiring no training. Within the **multiple augmentation scope**, we show that an easy incorporation of the lightweight ET3 yields further gains to the ensemble that averages predictions across all augmented views. It surpassing MTA [102] and matching the state-of-the-art robustness of R-TPT [75]. In the last **augmentation-based methods with prompt-tuning scope**, we outperform existing methods by similarly combining ET3 with the former state-of-the-art R-TPT. Unless otherwise stated, the ϵ for ET3 is set to 5 for TeCoA and 4 for FARE, with the number of steps fixed at $T = 2$. Further details—including results using a single-step variant of the defense and an ablation on the label sets used in ET3 —are provided in Appendix C and Appendix D.

5.2. Evaluating Large Vision-Language Models

We evaluate ET3 on LLaVA 1.5-7B using as vision encoder (i) standard CLIP, (ii) TeCoA-robust CLIP [50], and (iii) FARE-robust CLIP [72]. For both robust methods, we employ the two variants trained with $\epsilon = 2/255$ and $\epsilon = 4/255$, marked as ^{2} or ^{4} respectively. We consider image captioning (COCO, Flickr30k) and visual question answering tasks (TextVQA, VQAv2); evaluations on clean data use the full datasets, while adversarial results are computed over 500 perturbed inputs generated with an attack budget of $\epsilon_a = 4/255$, following the evaluation of [72]. Across all tasks, we employ ImageNet-21k text labels [66], to compute the energy. Further details in Appendix A.

Table 1. ET3 boosts Robustness Zero-shot robustness across 14 benchmark datasets. Comparison of clean and robust accuracy for baseline models versus the same models augmented with our defense. Robustness is evaluated against Auto-Attack (AA) at ϵ_a (4/255).

Model	Method	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	**Improv.**
ViT-L/14 (TeCoA) $\epsilon_t = 4/255$	Base (Clean)	74.91	78.36	37.83	79.61	50.26	38.03	22.48	11.76	38.41	74.35	54.22	49.95	76.07	93.44	55.69	
	+ **ET3** (Clean)	75.20	78.16	37.21	81.27	49.97	38.09	23.31	11.61	39.31	76.66	55.22	50.03	76.18	93.08	56.09	(+0.4)
	Base (Robust)	44.50	60.90	8.50	37.10	21.50	16.50	6.40	2.20	12.60	41.90	32.80	45.70	55.00	74.30	32.85	
	+ **ET3** (Robust)	53.00	65.10	13.50	56.40	34.50	22.70	15.80	4.40	22.50	51.90	40.00	51.50	60.90	80.80	40.93	(+8.08)
ViT-L/14 (FARE) $\epsilon_t = 4/255$	Base (Clean)	70.78	84.70	63.84	77.67	56.53	43.83	18.28	21.96	58.07	80.24	56.74	50.02	87.14	96.04	61.85	
	+ **ET3** (Clean)	70.97	84.73	62.96	80.58	55.91	43.94	18.93	21.96	58.86	82.49	57.25	50.02	86.59	96.17	62.24	(+0.39)
	Base (Robust)	34.80	64.20	12.70	34.80	20.20	17.50	11.10	3.00	12.20	40.50	30.60	52.30	50.60	74.30	32.77	
	+ **ET3** (Robust)	41.20	69.40	18.40	53.50	33.10	26.80	14.70	8.30	24.20	50.80	37.40	52.30	58.50	79.80	40.60	(+7.83)
ViT-L/14 (TeCoA) $\epsilon_t = 2/255$	Base (Clean)	80.11	80.67	50.08	87.53	60.69	44.36	26.06	14.04	51.80	80.12	58.43	49.89	80.02	96.08	61.42	
	+ **ET3** (Clean)	79.82	79.62	46.33	86.34	59.90	44.15	31.52	13.35	50.02	82.05	58.99	49.99	80.43	94.97	61.25	(–0.17)
	Base (Robust)	37.00	57.40	6.40	31.00	17.90	14.70	7.80	1.00	9.60	36.60	30.90	17.40	50.40	69.10	27.66	
	+ **ET3** (Robust)	44.40	63.10	14.10	49.20	32.20	23.40	24.00	4.70	20.50	46.00	38.40	47.10	58.40	75.40	38.64	(+10.98)
ViT-L/14 (FARE) $\epsilon_t = 2/255$	Base (Clean)	74.48	84.77	70.53	89.52	69.13	50.05	25.39	26.70	70.60	85.52	59.72	50.01	91.06	98.47	67.57	
	+ **ET3** (Clean)	74.07	84.60	68.31	89.64	66.64	48.03	32.98	24.90	69.47	86.94	59.86	50.03	90.38	98.00	67.42	(–0.15)
	Base (Robust)	17.80	46.40	5.00	25.70	14.20	11.60	0.40	0.90	7.10	25.60	22.10	19.10	28.10	61.50	20.39	
	+ **ET3** (Robust)	25.20	56.40	12.20	43.90	28.70	21.90	25.30	6.50	16.00	35.10	29.40	35.90	38.90	69.10	31.75	(+11.36)

Table 2. Robustness of ET3 on fine-grained classification. We compare against other test-time adaptation techniques on eight fine-grained datasets. All defenses are applied to a TeCoA pre-trained CLIP-ViT-B/32 model and evaluated against PGD-100 ($\epsilon = 4/255$). Results show that ET3 consistently improves robustness, both as a standalone method and when combined with other defenses.

Method	**Caltech101**		**Pets**		**Cars**		**Flower102**		**Aircraft**		**DTD**		**EuroSAT**		**UCF101**		**Avg.**		**Improv.**
	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Acc.	Rob.	Rob.
CLIP (Robust)	78.82	43.45	66.88	15.94	10.20	0.99	30.82	9.09	6.60	0.45	24.53	10.70	14.53	10.78	34.58	6.71	33.37	12.26	
Lightweight Defense using Image Transformation (Vision Input Only)																			
TTC [94]	71.26	44.90	65.00	21.10	10.62	1.39	26.02	8.93	6.93	0.69	23.99	11.38	20.84	**12.34**	32.91	10.45	32.20	13.90	(+1.64)
ET3 (ours)	79.07	50.59	66.86	27.15	10.32	2.04	28.91	12.10	5.28	0.87	24.00	13.18	13.37	11.17	37.35	16.02	33.15	16.51	(+4.25)
Defenses using Multiple Augmentations (Vision Input Only)																			
Ensemble	73.02	55.66	59.96	38.35	5.55	2.80	26.35	16.12	4.29	1.89	23.82	15.96	12.51	11.01	26.35	14.12	28.98	19.4	(+7.14)
MTA [102] *	79.70	55.70	66.20	31.20	9.00	2.50	29.10	14.00	6.50	1.60	24.40	13.50	13.30	11.20	34.60	12.50	32.90	17.80	(+5.54)
Ensemble + **ET3**	76.23	61.38	62.03	43.17	5.77	3.03	26.07	18.11	3.93	2.22	22.64	17.55	11.88	11.14	32.62	21.44	30.14	22.26	(+10)
Defenses using Multiple Augmentations + Prompt Tuning (Vision + Text Input)																			
TPT [76] *	79.30	52.70	65.20	27.40	9.60	2.00	27.90	12.30	6.70	1.70	25.50	14.60	12.20	11.20	34.90	10.20	32.70	16.50	(+4.24)
C-TPT [99] *	79.80	47.30	66.10	19.50	10.60	1.30	29.40	10.70	6.40	0.70	26.20	12.40	13.00	11.10	36.40	8.10	33.50	13.90	(+1.64)
R-TPT [75]	75.58	61.01	63.75	41.43	5.57	2.79	26.29	16.57	5.73	2.46	25.59	18.09	11.46	11.30	31.48	17.63	30.68	21.41	(+9.15)
R-TPT+ **ET3**	79.27	**65.07**	63.86	**45.71**	6.21	**3.27**	26.51	**18.72**	5.85	**3.51**	26.42	**19.92**	11.37	11.16	36.93	**24.72**	32.05	**23.88**	(+11.62)

Results. In Tab. 7, one can see that across all settings, ET3 consistently improves the adversarial robustness of the base models while preserving performance on clean data. Both TeCoA and FARE, as well as standard CLIP encoder, show improved performance with ET3. Moreover, ET3 introduces no significant overhead in inference time. Further details in Appendix A.

5.3. Evaluation with Robust Classifiers

We evaluate ET3 on robust classifiers from `RobustBench` [18] using the ImageNet dataset. We compare our method against other test-time transformation defenses as well as popular strong defenses, including adversarial training and adversarial purification

Adaptive Attacks. Unlike evaluating VLMs, where the common threat model is where the attacker is unaware of the test-time defense [75, 90, 94, 103], with classifiers, we assess ET3 under *adaptive attacks* too, where the attacker has full knowledge of the defense and actively attempts to circumvent it as suggested by [19]. Following [19], we adopt a strong attack configuration, targeted APGD-DLR with gradients approximated through the defense via Backward Pass

Table 3. **Evaluation of ET3 on LLaVA 1.5-7B with different vision encoders.** Clean and ℓ_∞-robust performance ($\epsilon_a = 4/255$) using standard CLIP and the TeCoA/FARE backbones adversarially trained with $\epsilon_t = 2/255$ and $\epsilon_t = 4/255$. Clean results use full test sets, while adversarial scores are computed on 500 APGD perturbations following the ensemble protocol of [72]. Across all tasks, ImageNet-21k labels serve as the reference text embeddings for computing the energy $E(\cdot, \theta)$. ET3 performs two gradient descent iterations. COCO and Flickr30k are evaluated with CIDEr for captioning, while TextVQA and VQAv2 report VQA accuracy.

	COCO [43]				**Flickr30k** [101]				**TextVQA** [77]				**VQAv2** [33]				**Average**			
	Clean	**+ET3**	4/255	**+ET3**	Clean	**+ET3**	4/255	**+ET3**	Clean	**+ET3**	4/255	**+ET3**	Clean	**+ET3**	4/255	**+ET3**	Clean	**+ET3**	4/255	**+ET3**
CLIP	**115.5**	111.3	2.7	**61.1**	**77.5**	76.1	1.1	**37.1**	**37.1**	34.9	0.2	**17.1**	**74.5**	73.1	0.0	**40.0**	**76.2**	73.9 (-2.3)	1.0	**38.8** (+37.8)
TECOA^2	98.4	**99.0**	30.0	**56.0**	57.1	**58.1**	14.8	**32.1**	24.1	**24.2**	7.8	**13.2**	66.9	**67.7**	25.1	**40.7**	61.6	**62.3** (+0.7)	19.4	**35.5** (+16.1)
FARE^2	109.9	**110.2**	32.5	**52.7**	71.1	**71.5**	17.5	**31.9**	31.9	31.9	7.2	**15.3**	71.7	71.7	24.5	**32.6**	71.2	**71.3** (+0.1)	20.4	**33.1** (+12.7)
TECOA^4	**88.3**	88.2	34.4	**53.7**	**48.6**	48.2	19.5	**30.2**	20.7	**21.0**	9.5	**12.6**	63.2	63.2	31.1	**42.6**	55.2	55.2 (+0.0)	23.6	**34.8** (+11.2)
FARE^4	102.4	**103.0**	42.2	**56.4**	61.6	**61.8**	23.1	**33.5**	**27.6**	27.4	10.2	**17.2**	68.3	68.3	29.5	**42.2**	65.0	**65.1** (+0.1)	26.3	**37.3** (+11.0)

Table 4. Worst case robustness comparison on ImageNet under ℓ_∞ threat model ($\epsilon = 4/255$). All the TTT methods build upon the same base robust model.

Method (Architecture)	**Adaptive Attack**	**Clean**	**Robust**
Adversarial Purification			
DiffPure [59] (WRN-50-2)	DiffBreak [35]	74.22	12.11
DiffPure [59] (DeiT-S)	DiffBreak [35]	73.63	25.00
GDMP [88] (DeiT-S)	DiffBreak [35]	69.14	20.70
***TTT** — Base: Salman et al. (RN-50) [70]*			
Base Model Only	AutoAttack [17]	64.02	34.96
+ Singh et al. [78]	APGD-DLR [17]	63.91	34.68 (-0.28)
+ Kulkarni et al. [39]	IW-WC [39]	64.10	35.48 (+0.52)
+ ET3 (Ours)	Tr. APGD-T + BPDA [19]	63.12	37.70 (+2.74)

Differentiable Approximation (BPDA) [4], effectively accounting for the entire test-time optimization process. We also perform a transfer attack from the static base model and report the worst-case accuracy. More details in Appendix A. Due to the computational cost of adaptive attacks, we restrict this evaluation to ResNet-50 classifiers. For other competing defenses, we show robustness using the strongest attacks appropriate for each method—for example, DiffBreaker [35] for diffusion-based purification defenses, and AutoAttack for robust classifiers.

Results. Tab. 4 reports the worst-case clean and robust accuracies—i.e., the lowest accuracies achieved under the strongest attacks reported in the original papers, including both defense and attack evaluations where applicable—on ImageNet. Methods are organized by defense family—adversarial purification, and test-time transformation (TTT)—and worst-case robustness is reported using the strongest adaptive attack available for each method. Consistent with prior work, DiffBreak significantly reduces the robustness of diffusion-based purification methods, whereas adversarially trained models maintain high robustness. In the TTT setting, where all methods build on the same robust base classifier [70], ET3 performs the best and improves worst-case robust accuracy from 34.96 to 37.70 while preserving clean accuracy. To ensure meaningful comparisons, we focus on worst-case robustness under the relevant adaptive attack for each method and omit defenses previously shown to fail under such evaluations [19]. Additional results on adversarially trained models, along with improvements achieved by ET3 across other robust classifiers, are provided in Appendix A.

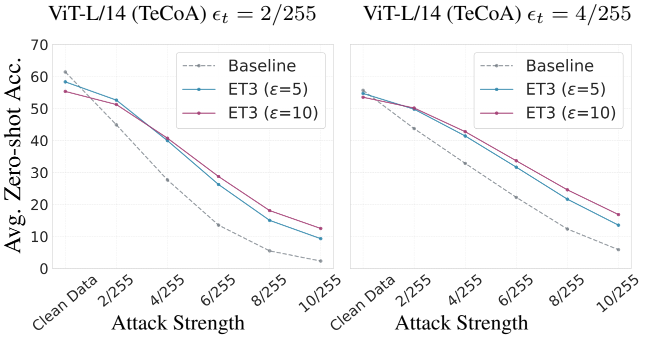


Figure 4. ET3 enhances models across different defense strengths. The plot shows that ET3 consistently boosts the robust accuracy of TeCoA models as the attack strength (ϵ_a) increases. We also compare the performance of the various ϵ for ET3

6. Conclusion

In this paper, we presented the ET3 defense, a novel test-time transformation method based on energy minimization. We presented extensive experiments of boosting robustness for image classification, zero-shot classification using CLIP, and with Large Vision Language Models, demonstrating superiority over many datasets and downstream tasks. We also presented theoretical justification for which our ET3 will provably transform a clean or adversarial image into correctly classified images. For future directions, we note that our lightweight ET3 defense can be incorporated into any image-related ML future defenses, to further increase robustness. Another interesting follow-up work is to optimize the network to support conditions that enable the ET3 defense, increasing its local linearity radius or the energy gradient ratio.

References

[1] Motasem Alfarra, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Data dependent randomized smoothing. In *Uncertainty in Artificial Intelligence*, pages 64–74. PMLR, 2022. 2

[2] Motasem Alfarra, Juan C Pérez, Ali Thabet, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Combating adversaries with anti-adversaries. In *AAAI Conference on Artificial Intelligence*, 2022. 2

[3] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *ECCV*, 2020. 2

[4] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 8, 15

[5] Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023. 1

[6] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1

[7] Peter Bartlett, Sébastien Bubeck, and Yeshwanth Cherapanamjeri. Adversarial examples in multi-layer random relu networks. In *NeurIPS*, 2021. 3

[8] Tsachi Blau, Roy Ganz, Chaim Baskin, Michael Elad, and Alex M. Bronstein. Class-conditioned transformation for enhanced robust image classification. In *WACV*, 2025. 2

[9] Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *ICML*, 2019. 3

[10] Sébastien Bubeck, Yeshwanth Cherapanamjeri, Gauthier Gidel, and Rémi Tachet des Combes. A single gradient step finds adversarial examples on random two-layers neural networks. In *NeurIPS*, 2021. 3

[11] Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *NeurIPS*, 2023. 1, 2

[12] Mehdi Cherti and Romain Beaumont. Clip benchmark, 2025. 13

[13] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, 2023. 13

[14] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, 2014. 13

[15] 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*. JMLR Workshop and Conference Proceedings, 2011. 13

[16] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *ICML*, pages 1310–1320. PMLR, 2019. 2

[17] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 5, 6, 8, 13

[18] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *ICLR Workshops 2021 - Workshop on Security and Safety in Machine Learning Systems*, 2021. 7, 15

[19] Francesco Croce, Sven Gowal, Thomas Brunner, Evan Shelhamer, Matthias Hein, and Taylan Cemgil. Evaluating the adversarial robustness of adaptive test-time defenses. In *ICML*. PMLR, 2022. 7, 8, 15, 17

[20] Amit Daniely and Hadas Shacham. Most relu networks suffer from ℓ_2 adversarial perturbations. In *NeurIPS*, 2020. 3

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 13

[22] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *NeurIPS*, 2019. 3

[23] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Brandon Tran, and Aleksander Madry. Adversarial robustness as a prior for learned representations. *arXiv preprint arXiv:1906.00945*, 2019. 5

[24] 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 *CVPR Workshops*, 2004. 13

[25] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021. 3

[26] Spencer Frei, Gal Vardi, Peter Bartlett, and Nati Srebro. The double-edged sword of implicit bias: Generalization vs. robustness in relu networks. *NeurIPS*, 2023. 2, 3, 5, 6

[27] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1, 2

[28] Will Grathwohl, Kuan-Chieh Wang, Jörn-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, 2020. 2, 3

[29] 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*, 2019. 13

[30] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 13

[31] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 13

[32] Mitch Hill, Jonathan Craig Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *ICLR*, 2021. 2, 3

[33] Ziheng Jia, Zicheng Zhang, Jiaying Qian, Haoning Wu, Wei Sun, Chunyi Li, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. Vqa^2: Visual question answering for video quality assessment, 2024. 8, 16

[34] Kaichao Jiang, He Wang, Xiaoshuai Hao, Xiulong Yang, Ajian Liu, Qi Chu, and Yunfeng Diao. Your classifier can do more: Towards bridging the gaps in classification, robustness, and generation. *arXiv preprint arXiv:2505.19459*, 2025. 3

[35] A. Kassis, U. Hengartner, and Y. Yu. DiffBreak: Breaking diffusion-based purification with adaptive attacks. In *NeurIPS*, 2025. 8

[36] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 13

[37] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *CVPR Workshops*, 2013. 13

[38] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 13

[39] Akshay Kulkarni and Tsui-Wei Weng. Interpretability-guided test-time adversarial defense. In *ECCV*, 2024. 8

[40] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 3

[41] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*, 2019. 2

[42] 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 *CVPR*, 2024. 2

[43] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. 8, 16

[44] Chang Liu, Yinpeng Dong, Wenzhao Xiang, Xiao Yang, Hang Su, Jun Zhu, Yuefeng Chen, Yuan He, Hui Xue, and Shibao Zheng. A comprehensive study on robustness of image classification models: Benchmarking and rethinking. *IJCV*, 2025. 15

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023. 1

[46] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *NeurIPS*, 2020. 3

[47] Saiyue Lyu, Shadab Shaikh, Frederick Shpilevskiy, Evan Shelhamer, and Mathias Lécuyer. Adaptive randomized smoothing: Certified adversarial robustness for multi-step defences. In *NeurIPS*, 2024. 2

[48] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1, 2, 13

[49] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 13

[50] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023. 6

[51] Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *ICLR*, 2023. 2, 13, 17

[52] Odelia Melamed, Gilad Yehudai, and Adi Shamir. Malt powers up adversarial attacks. In *NeurIPS*, 2024. 4, 5

[53] Hancheng Min and René Vidal. Can implicit bias imply adversarial robustness? *arXiv preprint arXiv:2405.15942*, 2024. 2, 3, 5

[54] Hancheng Min and Rene Vidal. Gradient flow provably learns robust classifiers for orthonormal GMMs. 2025. 2, 3, 5

[55] Mujtaba Hussain Mirza, Maria Rosaria Briglia, Senad Beadini, and Iacopo Masi. Shedding more light on robust classifiers under the lens of energy-based models. In *ECCV*, 2024. 3

[56] Mujtaba Hussain Mirza, Maria Rosaria Briglia, Filippo Bartolucci, Senad Beadini, Giuseppe Lisanti, and Iacopo Masi. Understanding adversarial training with energy-based models, 2025. 3

[57] Andrea Montanari and Yuchen Wu. Adversarial examples in random neural networks with general activations. *arXiv preprint arXiv:2203.17209*, 2022. 3

[58] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *CVPR*, 2016. 2

[59] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *ICML*, 2022. 2, 8

[60] 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*, 2008. 13

[61] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, 2012. 13

[62] Juan C. Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via test-time transformation ensembling. In *ICCV*, 2021. 2

[63] Brian Pulfer, Yury Belousov, and Slava Voloshynovskiy. Robustness tokens: Towards adversarial robustness of transformers. In *ECCV*, 2024. 2

[64] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI Conference on Artificial Intelligence*, 2024. 1, 2

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[66] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses, 2021. 4, 6

[67] Elias Abad Rocamora, Christian Schlarmann, Naman Deep Singh, Yongtao Wu, Matthias Hein, and Volkan Cevher. Robustness in both domains: Clip needs a robust text encoder, 2025. 2

[68] Briglia Maria Rosaria, Mujtaba Hussain Mirza, Giuseppe Lisanti, and Iacopo Masi. What is adversarial training for diffusion models? *arXiv preprint arXiv:2505.21742*, 2025. 2

[69] Mozhdeh Rouhsedaghat, Masoud Monajatipoor, C-C Jay Kuo, and Iacopo Masi. MAGIC: Mask-guided image synthesis by inverting a quasi-robust classifier. In *AAAI Conference on Artificial Intelligence*, 2023. 3

[70] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In *NeurIPS*, 2020. 8, 15

[71] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023. 2

[72] 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. In *ICML*, 2024. 1, 2, 6, 8, 13, 14, 16

[73] Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance. *arXiv preprint arXiv:1901.10861*, 2019. 3

[74] Adi Shamir, Odelia Melamed, and Oriel BenShmuel. The dimpled manifold model of adversarial examples in machine learning. *arXiv preprint arXiv:2106.10151*, 2021. 3

[75] Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *CVPR*, 2025. 2, 6, 7, 13, 14

[76] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *NeurIPS*, 2022. 2, 6, 7, 13, 14

[77] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326, 2019. 8, 16

[78] Anurag Singh, Mahalakshmi Sabanayagam, Krikamol Muandet, and Debarghya Ghoshdastidar. Robust feature inference: A test-time defense strategy using spectral projections. 2024. 8

[79] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *ICLR*, 2018. 2, 3

[80] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 13

[81] David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *CVPR*, 2019. 3

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

[83] Keke Tang, Tianrui Lou, Weilong Peng, Nenglun Chen, Yawen Shi, and Wenping Wang. Effective single-step adversarial training with energy-based models. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024. 3

[84] Zhuowen Tu. Learning generative models via discriminative approaches. In *CVPR*, 2007. 3

[85] Gal Vardi, Gilad Yehudai, and Ohad Shamir. Gradient methods provably converge to non-robust networks. *Advances in Neural Information Processing Systems*, 35: 20921–20932, 2022. 2, 3, 5

[86] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2018. 13

[87] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 2019. 13

[88] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022. 8

[89] Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *CVPR*, 2024. 2

[90] Xin Wang, Kai Chen, Jiaming Zhang, Jingjing Chen, and Xingjun Ma. Tapt: Test-time adversarial prompt tuning for robust inference in vision-language models. In *CVPR*, 2025. 7

[91] Yifei Wang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. A unified contrastive energy-based model for understanding the generative ability of adversarial training. In *ICLR*, 2022. 3

[92] Boxi Wu, Heng Pan, Li Shen, Jindong Gu, Shuai Zhao, Zhifeng Li, Deng Cai, Xiaofei He, and Wei Liu. Attacking adversarial attacks as a defense. *arXiv preprint arXiv:2106.04938*, 2021. 2

[93] Jianwen Xie, Yang Lu, Song-Chun Zhu, and Yingnian Wu. A theory of generative convnet. In *ICML*. PMLR, 2016. 3

[94] Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *CVPR*, 2025. 2, 3, 6, 7, 13, 14

[95] Xiulong Yang and Shihao Ji. M-ebm: Towards understanding the manifolds of energy-based models. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 291–302. Springer, 2023. 3

[96] Xiulong Yang, Qing Su, and Shihao Ji. Towards bridging the performance gaps of joint energy-based models. In *CVPR*, pages 15732–15741, 2023. 3

[97] Xuwang Yin, Shiying Li, and Gustavo K Rohde. Learning energy-based models with adversarial training. In *ECCV*. Springer, 2022. 3

[98] Xuwang Yin, Claire Zhang, Julie Steele, Nir Shavit, and Tony T Wang. Joint discriminative-generative modeling via dual adversarial training. *arXiv preprint arXiv:2510.13872*, 2025. 3

[99] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark A. Hasegawa-Johnson, Yingzhen Li, and Chang D. Yoo. C-TPT: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *ICLR*, 2024. 6, 7, 13, 14

[100] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pages 12062–12072. PMLR, 2021. 2

[101] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014. 8, 16

[102] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *CVPR*, 2024. 3, 6, 7, 13, 14

[103] Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *ECCV*, 2024. 2, 7

[104] Yao Zhu, Jiacheng Ma, Jiacheng Sun, Zewei Chen, Rongxin Jiang, Yaowu Chen, and Zhenguo Li. Towards understanding the generative capability of adversarially robust classifiers. In *ICCV*, 2021. 3

A Provable Energy-Guided Test-Time Defense Boosting Adversarial Robustness of Large Vision-Language Models

Supplementary Material

A. Implementation details

This section provides the implementation details for all experiments presented in the main paper.

A.1. Details on Zero-shot evaluation

Unless stated otherwise, all zero-shot experiments, consistent with the setup adopted in [72], follow the standard CLIP evaluation protocol used in the CLIP Benchmark [12] and OpenCLIP [13].

For each dataset, every class name is paired with a set of prompt templates, producing multiple natural-language descriptions per class. These prompts are encoded with the CLIP text encoder to obtain their corresponding textual embeddings. For each class, we average all template-derived embeddings to form a single class-level representation. Zero-shot predictions are then computed by taking the cosine similarity between the CLIP image embedding and all class embeddings, assigning the label with the highest similarity score.

Throughout the experiments, we report results from the following datasets: Caltech101 [24], Stanford Cars [37], CIFAR-10 and CIFAR-100 [38], DTD [14], EuroSAT [29], FGVC Aircraft [49], Flowers [60], ImageNet-R [31], ImageNet-Sketch [87], PCAM [86], Oxford Pets [61], and STL-10 [15] and UCF-101[80]. We also report results on the validation set of ImageNet-1k [21] consistent with prior work[51, 72].

Attack setup. Consistent with the established evaluation setup from [72], we measure adversarial robustness on a subset of 1000 randomly selected samples from each dataset, while clean accuracy is computed over all clean samples. Adversarial examples are generated using the first two attacks from the AutoAttack suite [17]: APGD with cross-entropy loss (APGD-CE) and APGD with the DLR loss (APGD-DLR), each executed for 100 iterations. For binary datasets such as PCAM, where the DLR loss is not applicable, only APGD-CE is used. All evaluations assume an ℓ_∞ threat model with perturbation magnitudes of $\varepsilon_a = 4/255$. Unless noted otherwise, all robustness experiments are conducted at 224×224 resolution, while CIFAR-10, CIFAR-100, and STL-10 are evaluated at their native image sizes.

A.1.1. Settings for Comparisons with Prior Work

For the comparisons reported in **Table 2** of the main paper, we use the exact same model checkpoints of the robust CLIP model and strictly follow the experimental settings established by [75], ensuring a fair and consistent comparison across all test-time adaptation and test-time augmentation defenses.

ET3 is applied under the identical settings used by each respective baseline, enabling a direct and principled evaluation. For the base Robust CLIP model and the state-of-the-art image–transformation defense for CLIP, TTC [94], ET3 is used in a zero-shot setting as exactly described in the method section of our main paper, with no modifications to their original inference pipelines.

The evaluation also includes the standard set of test-time prompt-tuning and augmentation baselines commonly used in CLIP robustness research [75, 99]. Following [75], we also include the simple *Ensemble* baseline, which aggregates predictions across multiple augmented views. All methods operate under identical constraints: they use CLIP as the underlying vision–language model and rely exclusively on AugMix [30] to generate test-time augmentations.

For clarity, these baselines can be grouped into those that rely solely on test-time augmentation (e.g., MTA [102] and Ensemble) and those based on prompt tuning (e.g., R-TPT [75], TPT [76], and C-TPT [99]). For these baseline methods, the generation of multiple augmented views is an integral and necessary part of their method. To evaluate ET3 under the same input conditions, it must therefore operate on the identical augmented input distribution used by each baseline.

Consequently, we apply ET3 directly on top of their existing mechanisms. Because ET3 is orthogonal to both test-time augmentation and prompt tuning, it can be integrated without altering the underlying baseline methods. Specifically, ET3 operates exclusively on the visual input space, leaving textual parameters and prompt embeddings unmodified. In all these settings, multiple augmented views are generated per input, and ET3 is applied to these augmented images before it is processed by the baseline method such as Ensemble or R-TPT.

All experiments share a common set of hyperparameters following the implementation of [75]. The text prompt template is initialized as “a photo of a”. For prompt-tuning methods, the learnable component consisted of a four-token prompt, updated via a single step with learning rate: 5×10^{-3} using the Adam optimizer [36]. The adversarial examples are generated on all the dataset samples using the exact configuration in [75]: **a 100-step Projected Gradient Descent (PGD) attack [48] with a perturbation budget of** $\varepsilon_a = 4.0$. Note that this specific attack setting is

Table 5. ET3 enhances zero-shot robustness across diverse benchmarks for various robust models. We report clean and robust accuracy on 14 datasets, comparing baseline models with versions augmented with ET3.

Model	Method	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	Improv.
ViT-B/32 (TeCoA)	Base (Clean)	56.16	73.39	13.77	74.89	40.93	24.57	22.67	5.79	29.31	49.11	29.58	50.01	70.89	87.30	44.88	(-0.6)
	+ E-3T (Clean)	55.15	75.20	11.25	74.74	38.19	23.83	19.26	4.95	28.85	51.65	30.73	50.00	69.66	85.41	44.21	
$\epsilon_t = 4/255$	Base (Robust)	24.05	52.18	2.79	32.38	17.06	11.54	7.35	0.30	7.42	22.04	13.87	49.90	33.22	58.55	23.76	(+8.11)
	+ E-3T (Robust)	34.47	63.02	5.29	51.60	26.85	16.44	13.59	2.76	14.91	33.81	21.59	49.98	44.73	67.09	31.87	
ViT-B/32 (FARE)	Base (Clean)	51.38	78.98	38.52	68.18	45.69	31.17	17.54	10.74	37.68	53.60	32.27	50.02	78.09	89.41	48.80	(-0.75)
	+ E-3T (Clean)	49.98	78.44	36.08	70.87	37.77	29.84	18.02	9.36	38.17	54.94	31.73	50.02	78.52	88.90	48.05	
$\epsilon_t = 4/255$	Base (Robust)	14.62	50.30	2.33	28.10	14.33	13.46	9.63	0.39	5.40	19.05	11.79	49.20	23.55	55.20	21.24	(+7.26)
	+ E-3T (Robust)	21.31	57.22	7.67	46.51	23.63	18.56	13.00	3.78	13.74	28.27	17.63	49.21	36.96	61.50	28.50	
ConvNeXt-B (TeCoA)	Base (Clean)	67.68	79.95	61.32	74.18	49.02	43.14	25.13	12.84	47.88	67.37	50.38	49.24	80.54	90.81	57.11	(-0.72)
	+ E-3T (Clean)	67.05	79.54	60.91	74.24	46.89	45.00	22.26	11.31	45.80	68.64	50.03	48.89	79.42	89.51	56.39	
$\epsilon_t = 4/255$	Base (Robust)	37.10	62.20	22.20	35.90	20.20	22.50	13.50	1.60	17.80	35.30	31.20	34.50	48.90	69.30	32.30	(+8.56)
	+ E-3T (Robust)	48.40	67.30	33.20	52.90	32.40	32.60	15.70	5.40	26.30	47.30	37.10	39.10	58.70	75.60	40.86	
ConvNeXt-B (FARE)	Base (Clean)	63.45	82.53	84.75	74.26	53.33	48.14	23.04	14.52	52.07	74.42	54.55	48.17	81.98	92.17	60.53	(-1.35)
	+ E-3T (Clean)	62.67	81.91	84.19	61.84	45.41	47.50	24.13	14.37	53.03	74.81	53.71	49.93	83.76	91.24	59.18	
$\epsilon_t = 4/255$	Base (Robust)	23.80	63.20	27.20	29.10	17.70	21.50	13.00	1.10	13.10	34.20	27.60	15.00	35.50	67.00	27.79	(+5.36)
	+ E-3T (Robust)	30.40	65.70	32.80	33.60	24.50	27.50	16.40	3.70	20.60	39.40	31.90	23.60	44.00	70.00	33.15	
ViT-B/32 (TeCoA)	Base (Clean)	70.53	77.14	28.88	85.89	54.96	32.82	28.80	12.30	48.41	61.65	41.16	44.19	81.22	93.45	54.39	(-0.87)
	+ E-3T (Clean)	69.83	76.24	26.28	82.33	50.65	31.33	33.20	11.13	47.91	63.81	41.68	42.25	81.25	91.41	53.52	
$\epsilon_t = 1/255$	Base (Robust)	2.83	15.00	0.57	9.99	2.12	5.59	5.24	0.54	1.63	3.03	3.19	34.77	7.39	23.60	8.25	(+2.88)
	+ E-3T (Robust)	4.57	17.44	1.22	13.47	4.99	10.48	10.83	1.11	4.15	5.07	4.71	36.86	12.37	28.52	11.13	
ViT-B/32 (FARE)	Base (Clean)	62.60	82.45	56.29	88.52	64.22	40.85	30.81	16.98	61.83	67.40	41.45	52.06	86.94	96.16	60.61	(-1.37)
	+ E-3T (Clean)	60.33	80.12	53.03	85.29	56.51	37.50	43.50	14.13	59.26	67.81	40.21	52.21	85.20	94.24	59.24	
$\epsilon_t = 1/255$	Base (Robust)	0.14	4.54	0.41	8.18	1.25	3.14	5.41	0.63	0.73	1.27	1.28	35.10	0.90	9.80	5.20	(+3.05)
	+ E-3T (Robust)	1.30	7.23	1.49	12.35	4.41	9.20	13.85	1.68	1.72	2.76	2.61	38.31	2.97	15.61	8.25	

used exclusively for this comparative table; stronger attacks are employed in all other experiments throughout our paper.

The results for MTA [102], TPT [76], and C-TPT [99] are taken directly from [75]. For the methods we re-evaluated, we verified that our reproduced results match those reported in [75]; therefore, we rely on their reported numbers for the remaining baselines. For TTC [94], we use the author's official code base with default hyperparameter and evaluate on the same model checkpoint used for the other baselines to ensure comparability.

A.2. Details on LVLM evaluation

In addition to evaluating robustness on zero-shot classification with CLIP models, we extend our analysis to Large Vision-Language Models (LVLMs) that employ these CLIP models as visual encoders, following the approach of prior work [72]. We specifically examine the susceptibility of the visual modality to adversarial perturbations and seek to enhance robustness against such attacks. Consistent with the procedure described in the Method section of the main paper, ET3 is applied exclusively to the visual encoder, offering a fast and computationally efficient transformation. As shown in Figure 2 of the main paper, embeddings are extracted from the CLIP visual encoder after the ET3 transformation. Following the original LLAVA implementation, we use the feature obtained from the layer before the last layer of the visual encoder.

Attack setup. We adopt the *ensemble* adversarial evaluation procedure introduced in [72]. For each test instance, we run a sequence of APGD attacks (ℓ_∞, $\epsilon = 4/255$, 100 steps) with different initialization conditions. In captioning tasks, we retain the perturbation that yields the lowest CIDEr score; in VQA tasks, we retain the perturbation that yields the lowest answer accuracy. The procedure begins with clean inference, followed by five APGD runs initialized from different ground-truth references, and concludes with a refinement step initialized from the current best perturbation. After each round, if the newly generated output worsens the evaluation metric, the perturbation is kept. If the metric crosses a stopping threshold (low CIDEr or zero VQA accuracy), the attack is terminated early for that sample. For each evaluation setting, we report both the original model performance and the performance obtained when applying ET3 at test time on the same adversarial inputs. Consistent with prior work [72], we use randomly sampled 500 images from each dataset for the adversarial evalua-

Model	Defense	Clean Acc.	Robust Acc.
ResNet-50 Salman et al. [70]	Base	64.02	35.40 (35.20)
	+ ET3	63.12	46.20
ConvNeXt-B Liu et al. [44]	Base	76.38	55.60 (55.00)
	+ ET3	75.95	61.40
ConvNeXt-L Liu et al. [44]	Base	77.47	57.70 (57.40)
	+ ET3	76.36	64.50
Swin-B Liu et al. [44]	Base	76.21	55.00 (54.80)
	+ ET3	75.75	61.70
Swin-L Liu et al. [44]	Base	78.18	58.10 (57.80)
	+ ET3	77.21	64.20

Table 6. We compare several robust models obtained from RobustBench, reporting performance both with and without ET3. All evaluations here use the APGD-T attack. For reference, we also include the AutoAttack robust accuracy of each base model—reported in parentheses and shown in gray.

tions, and all clean samples for clean evaluations.

Computational overhead on LLaVA with ET3. We measure the inference latency of ET3 on an NVIDIA A100 GPU with the LLaVA-1.5 7B model. The baseline inference time is 593.9 ms per sample. Incorporating the ET3 step increases the latency to 607.3 ms per sample (+2.3%) for a single step and 640.0 ms per sample (+7.7%) for two steps. These results are averaged over 500 samples.

A.3. Details on Robust Classifiers

We evaluate the robust ImageNet classifiers obtained from RobustBench [18] under a standard ℓ_∞ threat model with perturbation budget $\varepsilon = 4/255$, in accordance with the RobustBench evaluation protocol [18]. Clean accuracy is reported on the full validation set, while robust accuracy is computed on a subset of 1,000 randomly selected images, following the standard practice in prior works. We also provide additional experiments in Sec. C.2 besides the one presented in the main paper.

A.4. Details on Adaptive attacks

To rigorously assess the robustness of our defense ET3, we evaluate it on robust image classifiers under adaptive attacks tailored to test-time defenses, ensuring a fair and meaningful comparison. We present the results in Table 4 of the main paper and provide additional details here. Specifically, we follow exactly the protocol proposed for test-time defenses by [19], adopting their “Transfer APGD-T + BPDA” attack. We exactly follow their implementation and use APGD-T with the DLR loss, 5 restarts, and 100 iterations per restart. We approximate gradients through our defense using Backward Pass Differentiable Approximation (BPDA), following the standard practice in [19]. This is necessary because directly differentiating through the full unrolled iterative procedure of our defense leads to gradient shattering and memory instabilities. Such effects can artificially impair the attacker and result in gradient obfuscation. By replacing the backward pass with a stable surrogate, BPDA provides a reliable gradient signal. In addition, we also perform **transfer attacks** by generating adversarial perturbations using APGD-T (with DLR loss, 5 restarts, and 100 iterations per restart) on the underlying static base model (without ET3) and applying them to the defended model to evaluate whether the defense provides genuine robustness beyond gradient masking. We report **worst-case** robust accuracy, where a sample is considered misclassified if it is successfully misclassified by either the adaptive attack (APGD-T + BPDA) or the transfer attack. Note that we do not use EOT [4] in evaluations as used in [19], as ET3 does not introduce stochasticity. We specifically adopt the Transfer APGD-T + BPDA attack specifically because it has been demonstrated in [19] to reliably circumvent most iterative test-time defenses similar to ours.

B. Qualitative examples

We provide a series of qualitative examples to illustrate the effect of ET3 across captioning, question answering, and image classification. These examples demonstrate how ET3 mitigates the effect of adversarial perturbations across various tasks discussed in the paper.

We present Fig. 5 which shows qualitative comparisons of generated captions for several sample images. ET3 consistently mitigates the impact of adversarial perturbations on standard CLIP and improves the robustness of both TeCoA and FARE. Green rows indicate semantically correct captions, red rows denote incorrect captions, and yellow rows correspond to partially correct descriptions that still broadly reflect the scene. All adversarial examples are generated with $\epsilon_a = 4/255$.

Fig. 6 shows short question–answer evaluations across various representative images. As with captioning, ET3 corrects adversarially induced errors in standard CLIP and further refines outputs from TeCoA and FARE. Green rows correspond to correct predictions, while red rows indicate incorrect ones. All adversarial samples use $\epsilon_a = 4/255$.

We also analyze the effect of ET3 on classification. To better understand the effect of perturbations, Fig. 7 plots logit as perturbations are smoothly scaled. For each example, we interpolate linearly from 0% to 100% of the perturbation in 100 steps. The top row shows the model’s logit evolution under the adversarial perturbation, while the bottom row shows the corresponding evolution under the ET3 transformation.

Table 7. Evaluating on LLaVA 1.5-7B with different vision encoders and using just **one-step** ET3 defense. Clean and ℓ_∞-robust performance ($\epsilon_a = 4/255$) using standard CLIP and the TeCoA/FARE backbones adversarially trained with $\epsilon_t = 2/255$ and $\epsilon_t = 4/255$. Clean results use full test sets, while adversarial scores are computed on 500 APGD perturbations following the ensemble protocol of [72]. Across all tasks, ImageNet-21k labels serve as the reference text embeddings for computing the energy $E(\cdot, \theta)$. ET3 performs one gradient descent iteration. COCO and Flickr30k are evaluated with CIDEr for captioning, while TextVQA and VQAv2 report VQA accuracy.

	COCO [43]				**Flickr30k** [101]				**TextVQA** [77]				**VQAv2** [33]				**Average**			
	Clean	**+ET3**	4/255	**+ET3**	Clean	**+ET3**	4/255	**+ET3**	Clean	**+ET3**	4/255	**+ET3**	Clean	**+ET3**	4/255	**+ET3**	Clean	**+ET3**	4/255	**+ET3**
CLIP	**115.5**	112.2	2.7	**68.2**	**77.5**	75.3	1.1	**38.9**	**37.1**	34.7	0.2	**18.0**	**74.5**	73.3	0.0	**43.9**	**76.2**	73.9	1.0	**42.3** (+41.3)
TeCoA[2]	98.4	**98.9**	30.0	**57.3**	57.1	**57.3**	14.8	**33.1**	24.1	24.1	7.8	**13.4**	**66.9**	66.8	25.1	**41.9**	61.6	**61.8**	19.4	**36.4** (+17.0)
FARE[2]	109.9	**110.3**	32.5	**57.0**	71.1	**71.3**	17.5	**32.4**	**31.9**	31.8	7.2	**15.0**	71.7	71.7	24.5	**38.1**	71.2	**71.3**	20.4	**35.6** (+15.2)
TeCoA[4]	**88.3**	88.1	34.4	**55.5**	**48.6**	47.8	19.5	**29.8**	20.7	**20.9**	9.5	**12.46**	63.2	63.2	31.1	**42.9**	**55.2**	55.0	23.6	**35.2** (+11.6)
FARE[4]	102.4	**102.6**	42.2	**57.7**	**61.6**	60.9	23.1	**33.6**	**27.6**	27.4	10.2	**16.8**	68.3	68.3	29.5	**43.3**	**65.0**	64.8	26.3	**37.9** (+11.6)

Table 8. ET3 with only **one-step defense** enhances zero-shot robustness across diverse benchmarks for various robust models. We report clean and robust accuracy on 14 datasets, comparing baseline models with versions augmented with ET3..

Model	Defense	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	Improv.
ViT-L/14	None (Clean)	74.91	78.36	37.83	79.61	50.26	38.03	22.48	11.76	38.41	74.35	54.22	49.95	76.07	93.44	55.69	(-0.73)
(TeCoA)	+ E-3T (Clean)	74.50	77.85	36.54	73.55	44.90	37.93	24.91	12.03	39.21	73.41	54.89	49.98	75.63	94.06	54.96	
$\epsilon_t = 4/255$	None (Robust)	44.50	60.90	8.50	37.10	21.50	16.50	6.40	2.20	12.60	41.90	32.80	45.70	55.00	74.30	32.85	(+7.71)
	+ E-3T (Robust)	52.70	64.40	11.20	54.10	32.70	21.70	18.30	5.30	21.00	49.80	40.10	51.00	62.10	83.50	40.56	
ViT-L/14	None (Clean)	70.78	84.70	63.84	77.67	56.53	43.83	18.28	21.96	58.07	80.24	56.74	50.02	87.14	96.04	61.85	(-2.77)
(FARE)	+ E-3T (Clean)	70.11	84.14	61.62	67.43	43.18	43.46	17.04	22.05	55.91	76.88	56.39	50.02	85.53	93.41	59.08	
$\epsilon_t = 4/255$	None (Robust)	34.80	64.20	12.70	34.80	20.20	17.50	11.10	3.00	12.20	40.50	30.60	52.30	50.60	74.30	32.77	(+6.39)
	+ E-3T (Robust)	42.80	68.70	18.10	47.20	30.50	24.90	14.20	6.50	20.50	46.10	38.60	52.30	57.90	80.00	39.16	
ViT-L/14	None (Clean)	80.11	80.67	50.08	87.53	60.69	44.36	26.06	14.04	51.80	80.12	58.43	49.89	80.02	96.08	61.42	(-2.90)
(TeCoA)	+ E-3T (Clean)	78.48	79.34	42.88	76.78	49.71	42.39	29.67	15.36	48.46	76.30	57.80	49.87	76.53	95.73	58.52	
$\epsilon_t = 2/255$	None (Robust)	37.00	57.40	6.40	31.00	17.90	14.70	7.80	1.00	9.60	36.60	30.90	17.40	50.40	69.10	27.66	(+12.28)
	+ E-3T (Robust)	47.80	63.00	13.90	52.10	31.90	21.90	24.30	8.30	22.80	45.50	40.20	46.80	59.70	81.00	39.94	
ViT-L/14	None (Clean)	74.48	84.77	70.53	89.52	69.13	50.05	25.39	26.70	70.60	85.52	59.72	50.01	91.06	98.47	67.57	(-3.48)
(FARE)	+ E-3T (Clean)	73.29	83.94	65.68	80.07	53.46	47.55	28.43	25.47	64.29	81.72	58.55	50.02	88.72	96.03	64.09	
$\epsilon_t = 2/255$	None (Robust)	17.80	46.40	5.00	25.70	14.20	11.60	0.40	0.90	7.10	25.60	22.10	19.10	28.10	61.50	20.39	(+11.61)
	+ E-3T (Robust)	28.20	56.20	12.80	45.30	27.00	19.60	23.00	6.60	14.60	36.20	31.40	37.30	39.10	70.70	32.00	

Finally, Fig. 7b specifically illustrates how ET3 enhances salient, class-relevant features. The left panel shows an Angora bunny image originally misclassified as a Blue Tick. Applying ET3 highlights key attributes—most notably the pinkish eye region—allowing the model to recover the correct prediction. The right panel provides a clean reference image for comparison. Similar behavior is observed throughout the paper, including the teaser example where ET3 makes a snake's eye features more prominent—features absent from its adversarial misclassifications as a zucchini. These examples collectively illustrate that ET3 transformation amplifies discriminative, class-relevant features, enabling recovery from adversarial perturbation.

C. Additional Experimental Results

C.1. Zero-shot Robustness with additional models

As shown in Table 1 of the main paper, ET3 improves zero-shot robustness. Here, we report analogous results on additional CLIP models under same settings, demonstrating that the observed improvements hold consistently across a broader set of models. The Tab. 5 reports zero-shot performance of ET3 across a diverse set architectures, including transformer-based ViT and ConvNeXt models, as well as models, as well as models trained with a smaller $\epsilon_t = 1/255$ perturbation. Across all configurations, ET3 consistently improves robust accuracy, with minimal or modest impact on clean accuracy. These results demonstrate that the benefits of ET3 are consistent across model architectures and training configurations, further illustrating its effectiveness

Table 9. ET3 enhances zero-shot robustness across diverse benchmarks for various robust models. We report clean and robust accuracy on 14 datasets, comparing baseline models with versions augmented with ET3.

Model	Defense	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	Improv.
ViT-L/14 (TeCoA) $\epsilon_t = 4/255$	None (Clean)	74.91	78.36	37.83	79.61	50.26	38.03	22.48	11.76	38.41	74.35	54.22	49.95	76.07	93.44	55.69	(-0.98)
	+ E-3T (Clean)	74.21	77.95	35.79	73.41	45.09	37.61	23.15	12.54	39.29	72.81	54.85	50.00	75.06	94.15	54.71	
	None (Robust)	44.50	60.90	8.50	37.10	21.50	16.50	6.40	2.20	12.60	41.90	32.80	45.70	55.00	74.30	32.85	(+8.56)
	+ E-3T (Robust)	54.70	66.00	11.50	57.70	32.60	22.40	16.00	6.00	21.40	50.70	40.80	51.40	63.10	85.40	41.41	
ViT-L/14 (FARE) $\epsilon_t = 4/255$	None (Clean)	70.78	84.70	63.84	77.67	56.53	43.83	18.28	21.96	58.07	80.24	56.74	50.02	87.14	96.04	61.85	(-3.26)
	+ E-3T (Clean)	69.86	83.68	60.34	66.23	42.89	42.98	18.20	21.48	54.82	76.00	56.25	50.02	84.93	92.56	58.59	
	None (Robust)	34.80	64.20	12.70	34.80	20.20	17.50	11.10	3.00	12.20	40.50	30.60	52.30	50.60	74.30	32.77	(+6.29)
	+ E-3T (Robust)	42.10	69.00	17.80	47.10	30.30	24.50	13.70	6.50	20.90	46.30	38.30	52.30	57.50	80.60	39.06	
ViT-L/14 (TeCoA) $\epsilon_t = 2/255$	None (Clean)	80.11	80.67	50.08	87.53	60.69	44.36	26.06	14.04	51.80	80.12	58.43	49.89	80.02	96.08	61.42	(-3.09)
	+ E-3T (Clean)	77.79	78.98	40.74	78.50	50.55	42.50	29.94	15.21	47.99	75.40	57.29	49.97	75.91	95.79	58.33	
	None (Robust)	37.00	57.40	6.40	31.00	17.90	14.70	7.80	1.00	9.60	36.60	30.90	17.40	50.40	69.10	27.66	(+12.30)
	+ E-3T (Robust)	47.40	63.50	13.10	51.20	31.40	21.60	22.60	8.80	24.00	47.00	40.50	46.30	59.80	82.20	39.96	
ViT-L/14 (FARE) $\epsilon_t = 2/255$	None (Clean)	74.48	84.77	70.53	89.52	69.13	50.05	25.39	26.70	70.60	85.52	59.72	50.01	91.06	98.47	67.57	(-4.00)
	+ E-3T (Clean)	72.90	83.75	63.93	78.19	53.22	46.70	30.61	24.81	63.05	80.68	58.18	50.02	88.31	95.59	63.57	
	None (Robust)	17.80	46.40	5.00	25.70	14.20	11.60	0.40	0.90	7.10	25.60	22.10	19.10	28.10	61.50	20.39	(+11.74)
	+ E-3T (Robust)	27.10	56.40	12.70	49.10	27.80	20.00	19.20	5.80	15.40	37.00	32.10	36.30	38.80	72.10	32.13	

in enhancing zero-shot robustness.

C.2. Robustness with larger classifier architectures

We further evaluate the effectiveness of ET3 on an extended set of robust ImageNet classifiers obtained from RobustBench, shown in Tab. 6. Specifically, we include larger and architecturally distinct models, such as Swin Transformers and ConvNeXt variants, to assess the generality of ET3. For a fair comparison, we evaluate the base models with the same attack used to assess the ET3, as this protocol is best suited for test-time defenses in classifiers, following [19]. We use APGD-T with DLR loss, 5 restarts, and 100 iterations per restart, following the attack protocol described previously. Clean accuracy is measured on the full ImageNet validation set, while robust accuracy is computed on a 1,000-image subset, consistent with established evaluation practices. We also report robust accuracy obtained with AutoAttack on the same samples.

Across all evaluated classifiers, ET3 consistently enhances robust accuracy, with only minor reductions in clean accuracy.

D. Ablation Study

In this section, we provide additional ablation studies to further analyze the behavior and key design choices of our proposed defense, ET3.

D.1. Single-Step ET3 Defense

Our proposed ET3 method uses a small, fixed number of iterative steps to perform the energy minimization. To demonstrate that ET3 can be made faster at inference if needed, we conduct an ablation in which the defense is restricted to a *single* transformation step. We evaluate this "single-step ET3" on both zero-shot classification and downstream LVLM tasks.

The results—shown in Tab. 7 for the LVLM experiments and in Tabs. 8 and 11 for the zero-shot evaluations—demonstrate that even a single step yields a substantial robustness improvement over the baseline. Although the full multi-step version of ET3 achieves the slightly stronger overall performance. For this ablation, we keep the overall perturbation budget ϵ identical to the multi-step setup, increasing the step size α only.

D.2. Performance under Increased Attack Strength

To further evaluate the resilience of ET3, we conduct an ablation in which we systematically increase the attack strength. For this study, we focus on zero-shot evaluation and use CLIP models whose image encoders are finetuned with TECoA [51], trained with perturbation budgets of $\epsilon_t = 2/255$ and $\epsilon_t = 4/255$, respectively.

We then evaluate two configurations of our defense under attacks of varying strength: **Default Defense:** our standard ET3 configuration with a transformation budget of $\epsilon = 5$ and $\alpha = 2.5$. **Stronger Defense:** a configuration with increased bound for defense transformation, $\epsilon = 10$ and $\alpha = 5$. In both setting, the number of steps is set to 2.

As shown in Figure 4 of the main paper, ET3 maintains a consistent robustness advantage as the attack strength increases, with results averaged across all benchmarks. De-

Table 10. ET3 improves robustness across increasing attack strengths. We report clean accuracy and robust accuracy on 14 datasets as the attack strength increases. Results compare the baseline model to its ET3-augmented variant. ϵ_a indicates the strength of the attack.

(a) ViT-L/14 TeCoA ($\epsilon_t = 2/255$)

ϵ_a	Defense	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	Improv.
	None	80.11	80.67	50.08	87.53	60.69	44.36	26.06	14.04	51.80	80.12	58.43	49.89	80.02	96.08	61.42	
Clean Data	+E-3T	77.79	78.98	40.74	78.50	50.55	42.50	29.94	15.21	47.99	75.40	57.29	49.97	75.91	95.79	58.33	(-3.09)
	None	61.90	70.20	21.90	63.50	34.90	27.10	12.60	6.40	27.50	58.70	43.00	42.60	69.60	88.60	44.89	
$2/255$	+E-3T	69.70	73.70	30.60	72.40	44.70	34.30	27.30	12.10	38.00	64.90	50.50	51.70	73.90	92.90	52.62	(+7.73)
	None	37.00	57.40	6.40	31.00	17.90	14.70	7.80	1.00	9.60	36.60	30.90	17.40	50.40	69.10	27.66	(+12.28)
$4/255$	+ E-3T	47.80	63.00	13.90	52.10	31.90	21.90	24.30	8.30	22.80	45.50	40.20	46.80	59.70	81.00	39.94	
	None	16.30	36.00	1.40	11.90	6.80	7.90	0.00	0.20	2.80	20.60	21.30	1.70	21.60	41.10	13.54	
$6/255$	E-3T	27.40	45.00	7.70	30.20	20.10	14.10	18.90	6.10	13.20	29.40	29.30	32.70	35.80	57.50	26.24	(+12.70)
	Base	4.70	18.40	0.30	2.70	2.20	2.90	0.00	0.00	1.00	10.80	14.20	0.10	4.70	14.90	5.49	
$8/255$	E-3T	13.90	28.30	4.80	16.40	13.20	8.60	10.00	4.10	8.40	18.00	22.20	16.80	15.40	30.30	15.03	(+9.54)
	Base	1.00	8.80	0.00	0.30	0.70	1.10	0.00	0.00	0.00	6.40	9.60	0.00	0.30	4.10	2.31	
$10/255$	E-3T	7.30	15.60	4.10	9.50	9.50	5.00	9.70	3.90	6.00	11.80	16.70	8.50	7.70	15.00	9.31	(+7.00)

(b) ViT-L/14 TeCoA ($\epsilon_t = 4/255$)

ϵ_a	Defense	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	Improv.
	None	74.91	78.36	37.83	79.61	50.26	38.03	22.48	11.76	38.41	74.35	54.22	49.95	76.07	93.44	55.69	
Clean Data	+E-3T	74.21	77.95	35.79	73.41	45.09	37.61	23.15	12.54	39.29	72.81	54.85	50.00	75.06	94.15	54.71	(-0.98)
	None	59.20	69.70	18.10	59.60	33.60	26.50	7.90	5.60	23.90	59.10	42.90	51.10	68.00	86.80	43.71	
$2/255$	+ E-3T	68.30	73.40	23.30	69.10	41.20	30.40	19.60	9.50	30.70	65.40	49.50	52.10	71.70	92.80	49.79	(+6.08)
	None	44.50	60.90	8.50	37.10	21.50	16.50	6.40	2.20	12.60	41.90	32.80	45.70	55.00	74.30	32.85	(+8.56)
$4/255$	+ E-3T	54.70	66.00	11.50	57.70	32.60	22.40	16.00	6.00	21.40	50.70	40.80	51.40	63.10	85.40	41.41	
	None	27.50	49.40	3.40	19.80	11.50	11.30	0.20	0.50	5.80	29.40	25.30	34.00	37.30	55.70	22.22	
$6/255$	E-3T	37.80	56.00	6.20	39.70	24.40	15.40	12.30	3.50	13.20	36.10	32.40	48.30	47.90	70.00	31.66	(+9.44)
	Base	15.40	33.70	0.60	9.20	5.80	6.70	0.00	0.00	2.60	17.30	17.90	11.00	16.90	35.40	12.32	
$8/255$	E-3T	24.50	41.00	3.50	24.60	17.80	10.80	4.20	2.20	9.40	24.40	25.80	36.60	28.30	50.00	21.65	(+9.33)
	Base	6.10	20.90	0.20	2.80	2.40	3.80	0.00	0.00	1.10	10.90	13.00	0.70	5.30	14.90	5.86	
$10/255$	E-3T	14.60	26.30	1.90	15.30	13.60	7.20	3.80	2.00	6.40	17.00	19.70	19.40	13.60	28.60	13.53	(+7.67)

tailed numerical results are provided in Tab. 10. We observe that ET3 improves robustness across all attack strengths in a stable manner. Furthermore, increasing the defense budget to $\epsilon = 10$ yields additional improvements, particularly under the strongest adversarial settings. These findings indicate that ET3 is not only effective under threat levels the robust model has been trained for but also scales gracefully to stronger adversaries without any additional training. We also provide the same analysis when the using only single-step ET3 defense in Tab. 11.

D.3. Impact of Label Set Choice for ET3

As described in the main paper, our energy-based defense, ET3, leverages a set of class labels to guide its energy minimization process. A critical design choice is the composition of this label set. We considered two primary options:

A Vast, General-Purpose Label Set: Using a comprehensive set of labels such as the ∼21,000 classes from the full ImageNet-21k dataset. We use these labels without any further preprocessing, treating each row as one class, for example: {person, individual, someone, somebody, mortal,

Table 11. ET3 with only **one-step defense** improves robustness across increasing attack strengths. We report clean accuracy and robust accuracy on 14 datasets as the attack strength increases. Results compare the baseline model to its ET3-augmented variant. ϵ_a indicates the strength of the attack..

(a) ViT-L/14 TeCoA ($\epsilon_t = 2/255$)

ϵ_a	Defense	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	Improv.
	None	80.11	80.67	50.08	87.53	60.69	44.36	26.06	14.04	51.80	80.12	58.43	49.89	80.02	96.08	61.42	
Clean Data	+E-3T	75.79	77.65	31.75	72.03	44.36	39.36	32.78	12.93	42.25	72.44	56.04	50.11	72.39	94.83	55.34	(-6.08)
	None	61.90	70.20	21.90	63.50	34.90	27.10	12.60	6.40	27.50	58.70	43.00	42.60	69.60	88.60	44.89	
$2/255$	+E-3T	67.90	73.20	25.70	66.10	39.40	34.20	31.80	11.70	36.60	64.20	50.80	52.40	71.20	92.10	51.24	(+6.35)
	None	37.00	57.40	6.40	31.00	17.90	14.70	7.80	1.00	9.60	36.60	30.90	17.40	50.40	69.10	27.66	
$4/255$	+ E-3T	48.50	62.90	15.30	48.60	29.50	24.00	28.60	8.90	24.20	46.50	42.10	50.70	59.30	81.00	40.72	(+13.06)
	None	16.30	36.00	1.40	11.90	6.80	7.90	0.00	0.20	2.80	20.60	21.30	1.70	21.60	41.10	13.54	
$6/255$	E-3T	29.40	46.60	8.60	30.10	19.60	16.40	23.90	6.90	15.50	30.60	31.70	45.30	39.10	58.60	28.74	(+15.20)
	Base	4.70	18.40	0.30	2.70	2.20	2.90	0.00	0.00	1.00	10.80	14.20	0.10	4.70	14.90	5.49	
$8/255$	E-3T	15.50	29.70	5.60	16.70	13.50	9.80	14.80	5.80	10.90	18.70	24.90	36.60	18.10	32.60	18.09	(+12,60)
	Base	1.00	8.80	0.00	0.30	0.70	1.10	0.00	0.00	0.00	6.40	9.60	0.00	0.30	4.10	2.31	
$10/255$	E-3T	9.50	17.60	4.60	10.10	9.60	6.50	13.80	4.70	8.00	13.40	19.30	29.90	10.30	17.70	12.50	(+10.19)

(b) ViT-L/14 TeCoA ($\epsilon_t = 4/255$)

ϵ_a	Defense	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	Improv.
	None	74.91	78.36	37.83	79.61	50.26	38.03	22.48	11.76	38.41	74.35	54.22	49.95	76.07	93.44	55.69	
Clean Data	+E-3T	72.75	77.12	32.82	69.83	41.19	36.01	26.07	12.42	38.05	71.13	54.31	50.01	73.81	93.71	53.52	(-2.17)
	None	59.20	69.70	18.10	59.60	33.60	26.50	7.90	5.60	23.90	59.10	42.90	51.10	68.00	86.80	43.71	
$2/255$	+ E-3T	68.00	73.30	24.60	66.30	37.30	31.40	25.20	10.50	33.60	66.30	49.40	52.20	71.70	92.30	50.15	(+6.44)
	None	44.50	60.90	8.50	37.10	21.50	16.50	6.40	2.20	12.60	41.90	32.80	45.70	55.00	74.30	32.85	
$4/255$	+ E-3T	55.30	66.80	13.30	56.10	31.50	24.90	22.40	7.30	23.40	52.20	42.40	52.20	64.70	86.00	42.75	(+9.90)
	None	27.50	49.40	3.40	19.80	11.50	11.30	0.20	0.50	5.80	29.40	25.30	34.00	37.30	55.70	22.22	
$6/255$	E-3T	39.90	56.40	7.30	41.00	23.80	16.90	18.20	5.30	17.40	38.50	33.80	51.10	50.10	71.40	33.65	(+11.43)
	Base	15.40	33.70	0.60	9.20	5.80	6.70	0.00	0.00	2.60	17.30	17.90	11.00	16.90	35.40	12.32	
$8/255$	E-3T	25.70	41.90	4.50	26.90	18.00	13.20	12.90	4.10	12.40	26.70	27.50	46.10	32.10	52.00	24.57	(+12.25)
	Base	6.10	20.90	0.20	2.80	2.40	3.80	0.00	0.00	1.10	10.90	13.00	0.70	5.30	14.90	5.86	
$10/255$	E-3T	15.70	30.20	3.50	16.70	13.70	9.30	8.60	3.50	9.90	18.80	21.00	37.50	16.70	30.70	16.84	(+10.98)

soul} as one class. We obtain the full set from[1].

A Refined, curated set of labels: Manually curating and refining a label set of such magnitude for every potential use case is impractical and outside the scope of this work. Therefore, for our experiments, we adopt the more practical approach of using the refined label set included in the evaluation dataset itself: in this case, we use the set of labels associated with the specific downstream benchmark (e.g., using all the 1,000 class labels of ImageNet-1k when evaluating on it).

Throughout this work, we evaluate both label-set choices, though we predominantly rely on the 21-k proxy ImageNet labels. Specifically, the LVLM experiments shown in Table 3 of the main paper, as well as the zero-shot robustness results presented in Figure 4, use the full 21k ImageNet label set. Additional results using this label set appear in Tabs. 7 to 11. In contrast, Tables 1 and 2 of the main paper, along with Tab. 5, use the label sets associated with their respective evaluation benchmarks.

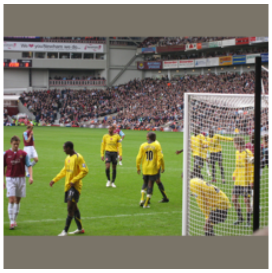
Using the refined, dataset-specific label set has a negligible impact on clean accuracy while still providing com-

[1] `link hidden for reviewing`

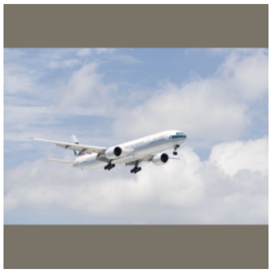
parable improvements in robustness. Overall, we observe that the 21k label set yields slightly higher robustness than the refined label set, albeit at a modest cost in clean accuracy. The extent of this drop varies across models and methods—TeCoA is minimally affected, whereas FARE is impacted more noticeably.

To better understand FARE's clean-accuracy drop, we conducted additional analysis and used the **full 21k labels for evaluation rather than the dataset-specific labels** commonly adopted in standard zero-shot evaluation practices (without applying ET3 or any attacks). We found that classes such as "cat" and "dog" are frequently mapped to semantically related but incorrect labels, including "petfood," "pet-food," or "pet food." When this occurs, the transformed image obtained after ET3 becomes more likely to be misclassified, as the ET3 amplifies features associated with these incorrect labels. Our analysis here is intentionally preliminary and does not constitute a comprehensive study of how zero-shot evaluation should be designed or assessed; a more thorough investigation lies beyond the scope of this work.

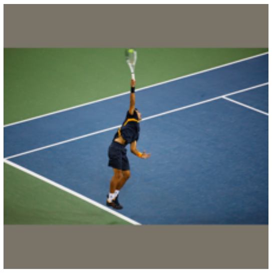
Nevertheless, in realistic deployment scenarios, large and diverse label sets are typically more appropriate for zero-shot classification. Under such conditions, we would not expect to observe the same degree of clean-accuracy degradation that appears in these controlled experimental settings.



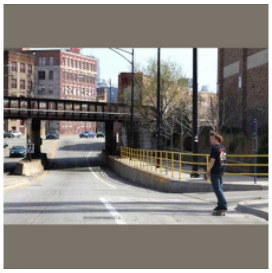
GT:	Sports team on a field wearing yellow jerseys with a goal net to the right.	
CLIP:	A group of people are playing a game of dunking a hot dog in a bun.	🟥
+ ET3:	A group of people playing soccer on a field.	🟩
TeCoA^2:	A group of people are playing with a net full of tennis balls.	🟥
+ ET3:	A group of people are standing on a field with a soccer goal in the background.	🟩
TeCoA^4:	A group of people are standing on a field, with some of them wearing yellow shirts.	🟩
+ ET3:	A group of people are standing on a field, with some of them wearing yellow shirts.	🟩
FARE^2:	A group of young boys playing soccer on a field.	🟩
+ ET3:	A group of soccer players on a field.	🟩
FARE^4:	A group of men are standing on a field, some of them wearing yellow shirts.	🟩
+ ET3:	A group of soccer players standing on a field.	🟩



GT:	Jet flying in the sky among the clouds.	
CLIP:	Angry Angry Birds are angry at the airport.	🟥
+ ET3:	A plane with a bunch of angry looking Sesame Street characters on it.	🟨
TeCoA^2:	A large airplane is on the runway.	🟨
+ ET3:	A large airplane is taking off from a runway.	🟨
TeCoA^4:	A large jetliner is flying through the sky.	🟩
+ ET3:	A large airplane flying in the sky.	🟩
FARE^2:	A large jet airplane is taking off from a runway.	🟨
+ ET3:	A large airplane is taking off into the sky.	🟩
FARE^4:	A large airplane is on the runway.	🟨
+ ET3:	A large airplane is flying through the sky.	🟩



GT:	A man hitting a tennis ball with a racquet.	
CLIP:	A cartoon cat with a football in its mouth.	🟥
+ ET3:	A woman in a purple shirt and black shorts is playing tennis.	🟨
TeCoA^2:	A woman is playing tennis on a court.	🟨
+ ET3:	A woman is playing tennis on a court.	🟨
TeCoA^4:	A tennis player is in the middle of a serve, holding a tennis racket and jumping up.	🟩
+ ET3:	A tennis player is swinging a racket on a tennis court.	🟩
FARE^2:	A woman is playing tennis and is in the middle of a serve.	🟨
+ ET3:	A woman is playing tennis and is about to hit the ball.	🟨
FARE^4:	A woman is playing tennis on a court.	🟨
+ ET3:	A tennis player is in the middle of a serve.	🟩



GT:	A man on a skateboard waits at the corner of a city street, with an overpass in the background.	
CLIP:	Two girls are hugging each other in a parking lot.	🟥
+ ET3:	A woman with long hair is skateboarding in a parking lot.	🟨
TeCoA^2:	A person is snowboarding on a ramp.	🟨
+ ET3:	A person is skateboarding on a sidewalk.	🟩
TeCoA^4:	A person is walking on a sidewalk near a bridge.	🟨
+ ET3:	A person is walking on a sidewalk near a bridge.	🟨
FARE^2:	A person is standing on a sidewalk near a train track.	🟨
+ ET3:	A person is skateboarding on a street.	🟩
FARE^4:	A man is standing on a sidewalk next to a bus.	🟨
+ ET3:	A man is standing on a sidewalk next to a street.	🟨

Figure 5. Qualitative comparison of generated captions for a sample image. ET3 corrects captions affected by adversarial attacks on standard CLIP and further refines captions produced by robust TeCoA and FARE. Green rows indicate semantically correct captions, red rows denote incorrect captions, and yellow rows highlight outputs with partial errors that still broadly reflect the image content. All attacks are generated with $\epsilon_a = 4/255$.

Q: Is this photo taken indoors or outdoors?

Answer:	Outdoors.
CLIP:	maybe.
+ ET3:	Indoors.
TeCoA^2:	Indoors.
+ ET3:	Outdoors.
TeCoA^4:	Indoors.
+ ET3:	Outdoors.
FARE^2:	Indoors.
+ ET3:	Outdoors.
FARE^4:	Indoors.
+ ET3:	Outdoors.

Q: what brewery makes this beer?

Answer:	Asahi.
CLIP:	Asain.
+ ET3:	Asahi.
TeCoA^2:	Asahi.
+ ET3:	Asahi.
TeCoA^4:	Asahi.
+ ET3:	Asahi.
FARE^2:	Pabst blue ribbon.
+ ET3:	Asahi.
FARE^4:	Asahi.
+ ET3:	Asahi.

Q: what does this sign say to do?

Answer:	Stop.
CLIP:	Stop limiting pelicans.
+ ET3:	Stop.
TeCoA^2:	No liquor.
+ ET3:	Stop.
TeCoA^4:	Stop.
+ ET3:	Stop.
FARE^2:	Stop at geyser.
+ ET3:	Stop.
FARE^4:	Stop.
+ ET3:	Stop.

Q: What color is the vehicle?

Answer:	Yellow.
CLIP:	Black.
+ ET3:	Black.
TeCoA^2:	White.
+ ET3:	Yellow.
TeCoA^4:	Yellow.
+ ET3:	Yellow.
FARE^2:	White.
+ ET3:	Yellow.
FARE^4:	White.
+ ET3:	Yellow.

Q: which program is seen on the screen?

Answer:	Office.
CLIP:	Windows xp.
+ ET3:	Office.
TeCoA^2:	Windows.
+ ET3:	Windows.
TeCoA^4:	Windows.
+ ET3:	Windows.
FARE^2:	Flickr.
+ ET3:	Office.
FARE^4:	Windows.
+ ET3:	Windows.

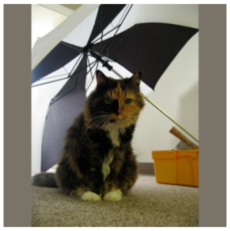
Q: How many people are there?

Answer	2.
CLIP:	5.
+ ET3:	3.
TeCoA^2:	3.
+ ET3:	3.
TeCoA^4:	3.
+ ET3:	2.
FARE^2:	3.
+ ET3:	3.
FARE^4:	3.
+ ET3:	2.

Q: which food is being advertised?

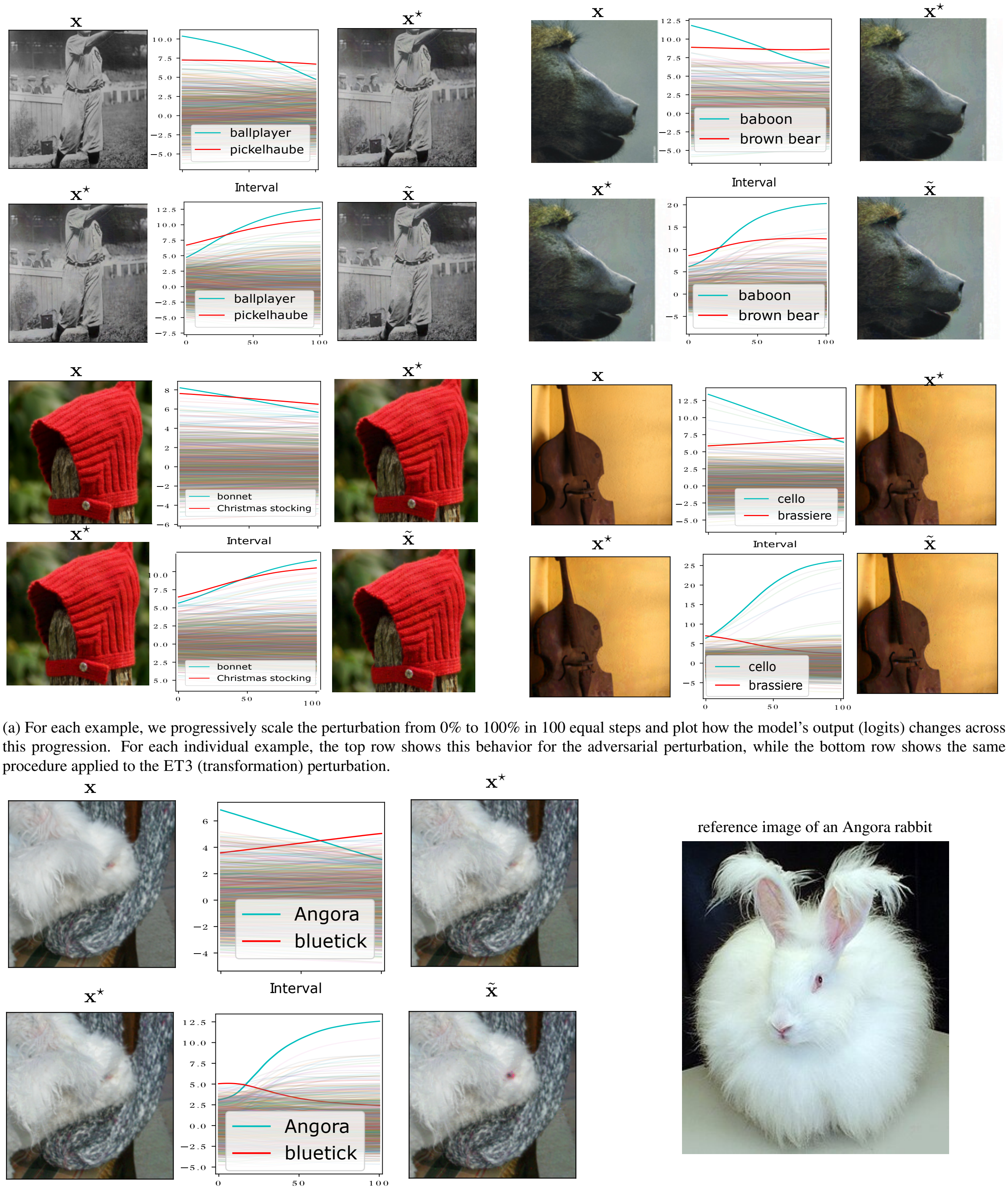
Answer:	Fajita.
CLIP:	Tortilla.
+ ET3:	Tortilla.
TeCoA^2:	Taco.
+ ET3:	Fajita.
TeCoA^4:	Taco.
+ ET3:	Taco.
FARE^2:	Pizza.
+ ET3:	Fajita.
FARE^4:	Fajita.
+ ET3:	Fajita.

Q: What kind of animal is this?



Answer	Cat.
CLIP:	Snake.
+ ET3:	Cat.
TeCoA^2:	Dog.
+ ET3:	Dog.
TeCoA^4:	Dog.
+ ET3:	Cat.
FARE^2:	Dog.
+ ET3:	Cat.
FARE^4:	Dog.
+ ET3:	Dog.

Figure 6. Qualitative comparison across 8 examples with short Q&A format. ET3 corrects answers affected by adversarial attacks on standard CLIP and further refines the ones produced by robust TeCoA and FARE. Green rows indicate correct captions, while red indicates incorrect ones. All attacks are generated with $\epsilon_a = 4/255$.



(a) For each example, we progressively scale the perturbation from 0% to 100% in 100 equal steps and plot how the model's output (logits) changes across this progression. For each individual example, the top row shows this behavior for the adversarial perturbation, while the bottom row shows the same procedure applied to the ET3 (transformation) perturbation.

(b) The left panel shows the ET3 transformation applied to an Angora bunny image that was originally misclassified as a Blue Tick. ET3 enhances salient features, most notably the pinkish eye region, that are essential for recognizing the correct class. The right panel provides a generic reference image of an Angora bunny.

Figure 7. Presenting a natural image $\mathbf{x}$, and its adversarial image $\mathbf{x}^\star$ wrongly classified by a robust classifier f_θ. Given only $\mathbf{x}^\star$ and f_θ, our ET3 produces $\tilde{\mathbf{x}}$ which is correctly classified.

E. Purification in Robust Networks (Proof for Theorem 4.1)

Proof:

For $\mathbf{v} = f(\mathbf{x})$, we denote the energy loss function we wish to minimize by

$$E(\mathbf{v}) = -\log \sum_{i \in \{-1,1\}} e^{\mathbf{v}_i}$$

thus its gradient with respect to the logits vector $\mathbf{v}$ will be a two dimensional vector

$$\nabla_{\mathbf{v}} E(\mathbf{v}) = -\text{SoftMax}(\mathbf{v}) = \left(-\frac{e^{\mathbf{v}_{-1}}}{\sum_{i \in \{-1,1\}} e^{\mathbf{v}_i}}, -\frac{e^{\mathbf{v}_1}}{\sum_{i \in \{-1,1\}} e^{\mathbf{v}_i}} \right) .$$

Therefore, the gradient of the energy w.r.t. the input calculated during the defense ET3 is

$$\frac{\partial E(f_\theta(\mathbf{x}))}{\partial \mathbf{x}} = -\text{SoftMax}(f_\theta(\mathbf{x}))^T \frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{x}} .$$

We denote

$$\mathbf{g}_0 = \frac{\partial f_\theta(\mathbf{x})_0}{\partial \mathbf{x}},\ \mathbf{g}_1 = \frac{\partial f_\theta(\mathbf{x})_1}{\partial \mathbf{x}} ,$$

and $e_0 = \text{SoftMax}\left(f_\theta\left(\mathbf{x}\right)\right)_0$ and $e_1 = \text{SoftMax}\left(f_\theta\left(\mathbf{x}\right)\right)_1$, the defense is calculating the gradient

$$\begin{aligned}\frac{\partial E(f_\theta(\mathbf{x}))}{\partial \mathbf{x}} &= -\ \text{SoftMax}(f_\theta(\mathbf{x}))^T \frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{x}} \\ &= -\ \left(\text{SoftMax}\left(f_\theta\left(\mathbf{x}\right)\right)_0 \mathbf{g}_0 + \text{SoftMax}\left(f_\theta\left(\mathbf{x}\right)\right)_1 \mathbf{g}_1\right) \\ &= -\ \left(e_0 \mathbf{g}_0 + e_1 \mathbf{g}_1\right)\ .\end{aligned}$$

For the defense optimization we take a gradient descent step of a norm upper bounded by ϵ, and get $\mathbf{x}_p = \mathbf{x} + \mathbf{z}$ for

$$\mathbf{z} = \alpha \left(e_0 \mathbf{g}_0 + e_1 \mathbf{g}_1\right)\ .$$

Since $\left\|\mathbf{z}\right\| \leq \epsilon$, the upper bound for α will be

$$\alpha \leq \frac{\epsilon}{\left\|e_0 \mathbf{g}_0 + e_1 \mathbf{g}_1\right\|} .$$

We remind the reader that $f_\theta(\mathbf{x}) \in \mathbb{R}^2$ by definition, leading to $\frac{\partial f_\theta(\mathbf{x})}{\partial \mathbf{x}} \in \mathbb{R}^2 \times \mathbb{R}^d$, thus for readability, we denote two functions, $f_0(\mathbf{x}) = f_\theta(\mathbf{x})_0$ and $f_1(\mathbf{x}) = f_\theta(\mathbf{x})_1$, concluding that $f_\theta(\mathbf{x}) = [f_0(\mathbf{x}), f_1(\mathbf{x})]$. Following the local linearity assumption of f_θ in $\mathcal{B}_\epsilon(\mathbf{x})$, f_0 and f_1 are linear functions in $\mathcal{B}_\epsilon(\mathbf{x})$, and we note that for any $\mathbf{x}' \in \mathcal{B}_\epsilon(\mathbf{x})$

$$f_0(\mathbf{x}') = \langle \mathbf{g}_0, \mathbf{x}' \rangle + a_0\ ,\ f_1(\mathbf{x}') = \langle \mathbf{g}_1, \mathbf{x}' \rangle + a_1$$

for some $a_0, a_1 \in \mathbb{R}$

We are now ready to show that for an input $\mathbf{x}$ with ground truth label y_t, the defense permutation $\mathbf{z}$ leads to

$$f_\theta(\mathbf{x} + \mathbf{z})_{y_t} > f_\theta(\mathbf{x} + \mathbf{z})_{\hat{y_t}} .$$

We look at $f(\mathbf{x} + \mathbf{z})$, having

$$\begin{aligned}f_0(\mathbf{x} + \mathbf{z}) &= \langle \mathbf{g}_0, \mathbf{x} + \mathbf{z} \rangle + a_0 = f_0(\mathbf{x}) + \langle \mathbf{g}_0, \mathbf{z} \rangle \\ f_1(\mathbf{x} + \mathbf{z}) &= \langle \mathbf{g}_1, \mathbf{x} + \mathbf{z} \rangle + a_1 = f_1(\mathbf{x}) + \langle \mathbf{g}_1, \mathbf{z} \rangle\end{aligned}$$

We denote

$$r_{\mathbf{x}} = f(\mathbf{x})_1 - f(\mathbf{x})_0 \ .$$

We assume W.L.O.G that the truth label is $y_t = 1$. The case $y_t = 0$ is proven similarly. We have $C > 1$ and

$$C\|e_0\mathbf{g}_0\| \leq \|e_1\mathbf{g}_1\| \ ,$$

leading to

$$\|\mathbf{g}_0\| \leq \frac{e_1}{Ce_0}\|\mathbf{g}_1\| \ .$$

We show that $f_1(\mathbf{x}+\mathbf{z}) - f_0(\mathbf{x}+\mathbf{z}) > 0$. We have

$$\begin{aligned} f_1(\mathbf{x}+\mathbf{z}) - f_0(\mathbf{x}+\mathbf{z}) =& f_1(\mathbf{x}) - f_0(\mathbf{x}) + \langle \mathbf{g}_1, \mathbf{z}\rangle - \langle \mathbf{g}_0, \mathbf{z}\rangle \\ =& r_x + \langle \mathbf{g}_1, \alpha\left(e_0\mathbf{g}_0 + e_1\mathbf{g}_1\right)\rangle - \langle \mathbf{g}_0, \alpha\left(e_0\mathbf{g}_0 + e_1\mathbf{g}_1\right)\rangle \\ =& r_x + \alpha\left[e_1\|\mathbf{g}_1\|^2 - e_0\|\mathbf{g}_0\|^2 + \left(e_0 - e_1\right)\langle \mathbf{g}_0, \mathbf{g}_1\rangle\right] \geq \\ \geq& r_x + \alpha\left[e_1\|\mathbf{g}_1\|^2 - \frac{e_1^2}{C^2e_0}\|\mathbf{g}_1\|^2 + \left(e_0 - e_1\right)\langle \mathbf{g}_0, \mathbf{g}_1\rangle\right] \\ \geq& r_x + \alpha\left[\left(e_1 - \frac{e_1^2}{C^2e_0}\right)\|\mathbf{g}_1\|^2 + \left(e_0 - e_1\right)\langle \mathbf{g}_0, \mathbf{g}_1\rangle\right] \ , \end{aligned}$$

for $\alpha \leq \frac{\epsilon}{\|e_0\mathbf{g}_0 + e_1\mathbf{g}_1\|}$. We note that

$$\begin{aligned} \|e_0\mathbf{g}_0 + e_1\mathbf{g}_1\|^2 =& \|e_0\mathbf{g}_0\|^2 + \|e_1\mathbf{g}_1\|^2 + 2\langle e_0\mathbf{g}_0, e_1\mathbf{g}_1\rangle \\ =& e_0^2\|\mathbf{g}_0\|^2 + e_1^2\|\mathbf{g}_1\|^2 + 2e_0e_1\langle \mathbf{g}_0, \mathbf{g}_1\rangle \\ \leq& \left(\frac{1}{C^2} + 1\right)\|e_1\mathbf{g}_1\|^2 + 2\langle e_0\mathbf{g}_0, e_1\mathbf{g}_1\rangle \\ \leq& \|e_1\mathbf{g}_1\|^2\left(\left(\frac{1}{C^2} + 1\right) + \frac{2\langle e_0\mathbf{g}_0, e_1\mathbf{g}_1\rangle}{\|e_1\mathbf{g}_1\|^2}\right) \\ \leq& \|e_1\mathbf{g}_1\|^2\left(\frac{1}{C^2} + 1 + \frac{2}{C}\right) \\ \leq& \|e_1\mathbf{g}_1\|^2\left(1 + \frac{1}{C}\right)^2 \ , \end{aligned}$$

where the last inequality hold since we assumed that $C\|e_0\mathbf{g}_0\| \leq \|e_1\mathbf{g}_1\|$, and for any two vectors $\mathbf{u}_1, \mathbf{u}_2$ we have that $\frac{\langle \mathbf{u}_1, \mathbf{u}_1\rangle}{\|\mathbf{u}_1\|\|\mathbf{u}_2\|} \leq 1$. Therefore, if we take

$$\alpha = \frac{\epsilon}{e_1\left(1 + \frac{1}{C}\right)\|\mathbf{g}_1\|} \leq \frac{\epsilon}{\|e_0\mathbf{g}_0 + e_1\mathbf{g}_1\|}$$

we get

$$
\begin{aligned}
f_1(\mathbf{x}+\mathbf{z}) - f_0(\mathbf{x}+\mathbf{z}) \geq & r_x + \alpha \left[\left(e_1 - \frac{e_1^2}{C^2 e_0} \right) \left\| \mathbf{g}_1 \right\|^2 + (e_0 - e_1) \left\langle \mathbf{g}_0, \mathbf{g}_1 \right\rangle \right] \\
\geq & r_x + \frac{\epsilon}{e_1 \left(1 + \frac{1}{C}\right) \left\| \mathbf{g}_1 \right\|} \left[\left(e_1 - \frac{e_1^2}{C^2 e_0} \right) \left\| \mathbf{g}_1 \right\|^2 + (e_0 - e_1) \left\langle \mathbf{g}_0, \mathbf{g}_1 \right\rangle \right] \\
\geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \left[\frac{e_1 - \frac{e_1^2}{C^2 e_0}}{e_1 \left(1 + \frac{1}{C}\right)} + \frac{(e_0 - e_1) \left\langle \mathbf{g}_0, \mathbf{g}_1 \right\rangle}{e_1 \left(1 + \frac{1}{C}\right) \left\| \mathbf{g}_1 \right\|^2} \right] \\
\geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \left[\frac{e_1 - \frac{e_1^2}{C^2 e_0}}{e_1 \left(1 + \frac{1}{C}\right)} - \frac{e_0 - e_1}{e_1 \left(1 + \frac{1}{C}\right) C} \right] \\
\geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \left[\frac{1 - \frac{e_1}{C^2 e_0}}{1 + \frac{1}{C}} - \frac{e_0 - e_1}{e_1 \left(1 + \frac{1}{C}\right) C} \right] \\
\geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \left[\left(\frac{1}{1 + \frac{1}{C}} \right) \left(1 - \frac{e_1}{C^2 e_0} - \frac{e_0 - e_1}{e_1 C} \right) \right] \\
\geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \left[\left(\frac{1}{1 + \frac{1}{C}} \right) \left(1 - \frac{e_1}{C^2 e_0} - \frac{e_0}{e_1 C} + \frac{1}{C} \right) \right] \ .
\end{aligned}
$$

We note that

$$
\frac{e_0}{e_1} = \frac{\text{SoftMax}(f(\mathbf{x}))_0}{\text{SoftMax}(f(\mathbf{x}))_1} = \frac{\frac{\exp(f(\mathbf{x})_0)}{\exp(f(\mathbf{x})_0) + \exp(f(\mathbf{x})_1)}}{\frac{\exp(f(\mathbf{x})_1)}{\exp(f(\mathbf{x})_0) + \exp(f(\mathbf{x})_1)}} = \exp(f(\mathbf{x})_0 - f(\mathbf{x})_1) = \exp(-r_x) \ ,
$$

and similarly $\frac{e_1}{e_0} \leq \exp(r_x)$, having

$$
\begin{aligned}
f_1(\mathbf{x}+\mathbf{z}) - f_0(\mathbf{x}+\mathbf{z}) \geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \left[\left(\frac{1}{1 + \frac{1}{C}} \right) \left(1 - \frac{e_1}{C^2 e_0} - \frac{e_0}{e_1 C} + \frac{1}{C} \right) \right] \\
\geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \frac{1}{2} \left(1 - \frac{\exp(r_x)}{C^2} - \frac{\exp(-r_x)}{C} + \frac{1}{C} \right) \\
\geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \frac{1}{2} \left(1 - \frac{1}{C^2} - \frac{\exp(|r_x|)}{C} + \frac{1}{C} \right) \ ,
\end{aligned}
$$

where the last inequality holds since

$$
-\frac{\exp(r_x)}{C^2} - \frac{\exp(-r_x)}{C} = -\frac{\exp(r_x) + C\exp(-r_x)}{C^2} \geq -\frac{\exp(-|r_x|) + C\exp(|r_x|)}{C^2} \geq -\frac{1 + C\exp(|r_x|)}{C^2} \ .
$$

Therefore we have

$$
\begin{aligned}
f_1(\mathbf{x}+\mathbf{z}) - f_0(\mathbf{x}+\mathbf{z}) \geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \frac{1}{2} \left(1 - \frac{1}{C^2} - \frac{\exp(|r_x|)}{C} + \frac{1}{C} \right) \\
\geq & r_x + \epsilon \left\| \mathbf{g}_1 \right\| \frac{1}{2} \left(1 - \frac{\exp(|r_x|)}{C} \right) \\
= & r_x + \frac{\epsilon \left\| \mathbf{g}_1 \right\|}{2} - \frac{\exp(|r_x|) \epsilon \left\| \mathbf{g}_1 \right\|}{2C} \\
= & \frac{1}{2} + \frac{r_x}{\epsilon \left\| \mathbf{g}_1 \right\|} - \frac{\exp(|r_x|)}{2C} \ .
\end{aligned}
$$

We note that ϵ should satisfy

$$\begin{aligned} 2r_x + \epsilon\|\mathbf{g}_1\| &> 0 \\ \epsilon &> \frac{-2r_x}{\|\mathbf{g}_1\|} \end{aligned}$$

for the correct classification to be possible in $\mathcal{B}_\epsilon(\mathbf{x})$. We note that this condition adds a necessary constraint only for adversarial samples, and applies directly where $\mathbf{x}$ is already correctly classified.

Finally, for

$$C > \frac{\exp(|r_x|)\epsilon\|\mathbf{g}_1\|}{\epsilon\|\mathbf{g}_1\| + 2r_x}$$

the claim follows.

□