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 [25], Stanford Cars [38], CIFAR-10 and CIFAR-100 [39], DTD [14], EuroSAT [31], FGVC Aircraft [50], Flowers [60], ImageNet-R [33], ImageNet-Sketch [90], PCAM [88], 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 [96], 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, 101]. 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 [32] to generate test-time augmentations. For clarity, these baselines can be grouped into those that rely solely on test-time augmentation (e.g., MTA [104] and Ensemble) and those based on prompt tuning (e.g., R-TPT [75], TPT [76], and C-TPT [101]). 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 [37]. The adversarial examples are generated on all the dataset samples using the exact configuration in [75]: **a 100-step Projected Gradient Descent (PGD) attack [49] with a perturbation budget of** $\varepsilon_a = 4.0$. Note that this specific attack setting is

Table 6. **Zero-shot robustness of ET3 across smaller model architectures in defense-unaware setting.** Comparison of clean and robust accuracy for baseline models versus the same models augmented with ET3. Robustness is evaluated against Auto-Attack (AA) at $\epsilon_a = 4/255$.

Model	Method	ImageNet	CalTech	Cars	CIFAR10	CIFAR100	DTD	EuroSAT	FGVC	Flowers	ImageNet-R	ImageNet-S	PCAM	OxfordPets	STL-10	Avg.	Improv.
ViT-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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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)
	+ ET3 (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 [104], TPT [76], and C-TPT [101] 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 [96], 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.

Table 7. **Performance of robust models from RobustBench with and without ET3 in the defense-unaware setting.** We compare several robust models obtained from RobustBench, reporting performance both with and without ET3. All evaluations use the APGD-T attack. For reference, the AutoAttack robust accuracy of each base model is included (in parentheses and shown in gray). All models are trained and evaluated under a standard ℓ_∞ threat model with a perturbation budget of $\epsilon = 4/255$.

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. [45]	Base	76.38	55.60 (55.00)
	+ ET3	75.95	61.40
ConvNeXt-L Liu et al. [45]	Base	77.47	57.70 (57.40)
	+ ET3	76.36	64.50
Swin-B Liu et al. [45]	Base	76.21	55.00 (54.80)
	+ ET3	75.75	61.70
Swin-L Liu et al. [45]	Base	78.18	58.10 (57.80)
	+ ET3	77.21	64.20

Consistent with prior work [72], we use randomly sampled 500 images from each dataset for the adversarial evaluations, and all clean samples for clean evaluations.

A.3. 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.4. 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.5. 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

Table 8. **Evaluating LLaVA 1.5-7B with different vision encoders using *one-step* ET3 in the defense-unaware setting.** 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 [44]				**Flickr30k** [103]				**TextVQA** [77]				**VQAv2** [29]				**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 9. **Zero-shot robustness when using *one-step* ET3 across 14 benchmark datasets in the defense-unaware setting.** Comparison of clean and robust accuracy for baseline models versus same models augmented with ET3. Robustness is evaluated against Auto-Attack (AA) at $\epsilon_a = 4/255$. Across all datasets, ImageNet-21k labels serve as the reference text embeddings for computing the energy $E(\cdot, \theta)$

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)	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)
	+ ET3 (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)
	+ ET3 (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 (FARE)	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)
	+ ET3 (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)
	+ ET3 (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 (TeCoA)	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)
	+ ET3 (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)
	+ ET3 (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 (FARE)	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)
	+ ET3 (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)
	+ ET3 (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	

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.

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. 6 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 $\epsilon_t = 1/255$

Table 10. **Zero-shot robustness of ET3 across 14 benchmark datasets in the defense-unaware setting.** Comparison of clean and robust accuracy for baseline models versus same models augmented with ET3. Robustness is evaluated against Auto-Attack (AA) at $\epsilon_a = 4/255$. Across all datasets, ImageNet-21k labels serve as the reference text embeddings for computing the energy $E(\cdot, \theta)$

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)	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)
	+ ET3 (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	
$\epsilon_t = 4/255$	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	(+8.56)
	+ ET3 (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)	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)
	+ ET3 (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	
$\epsilon_t = 4/255$	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	(+6.29)
	+ ET3 (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)	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)
	+ ET3 (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	
$\epsilon_t = 2/255$	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	(+12.30)
	+ ET3 (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)	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)
	+ ET3 (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	
$\epsilon_t = 2/255$	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	(+11.74)
	+ ET3 (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	

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 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. 7. 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.

C.3. Defense-aware attacks for LVLM

To rigorously assess the reliability of ET3 in the context of Large Vision-Language Models (LVLMs), we evaluate its performance against defense-aware adaptive attacks. Specifically, we employ Backward Pass Differentiable Approximation (BPDA) combined with AutoAttack, maintaining the exact same LLaVA evaluation settings as those established in Table 3.

In Tab. 13 we report the average robust accuracy across the four evaluation datasets, utilizing a total of 200 samples. When using a standard, non-robust CLIP model as the underlying image encoder, we observe a sharp drop in robustness against these adaptive attacks. Conversely, when integrated with robustly trained CLIP encoders, ET3 consistently boosts the model's overall robustness. Crucially, the results of this adaptive evaluation explicitly validate the theoretical analysis presented in Sec. 4. This empirical evidence confirms that ET3 effectively leverages and amplifies the inherent robustness of the foundation encoder, successfully mitigating stronger defense-aware attacks. Under the same settings, we further evaluate ET3 when it uses ℓ_∞ projection instead the default ℓ_2 projection, more details are provided in Sec. D.2.

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.

Table 11. **ET3 improves robustness across increasing attack strengths in defense-unaware setting.** We report clean and robust accuracy on 14 datasets as the attack strength increases, comparing the baseline model to its ET3-augmented variant. ϵ_a indicates the strength of the attack. Across all datasets, ImageNet-21k labels serve as the reference text embeddings for computing the energy $E(\cdot, \theta)$.

(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.
Clean Data	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	
	+ET3	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)
2/255	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	
	+ET3	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)
4/255	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	
	+ ET3	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	(+12.28)
6/255	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	
	ET3	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)
8/255	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	
	ET3	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)
10/255	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	
	ET3	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.
Clean Data	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	
	+ET3	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)
2/255	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	
	+ ET3	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)
4/255	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	
	+ ET3	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	(+8.56)
6/255	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	
	ET3	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)
8/255	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	
	ET3	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)
10/255	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	
	ET3	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)

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. 8 for the LVLM experiments and in Tabs. 9 and 12 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 to 5 for TeCoA and 4 for FARE.

D.2. Budget fairness

Evaluating the robustness of Test-Time Training (TTT) methods introduces unique considerations compared to adversarial training defenses. In Adversarial Training (AT), fair evaluation strictly requires matching the defense budget

Table 12. ***One-step*** **ET3 improves robustness across increasing attack strengths in defense-unaware setting.** We report clean and robust accuracy on 14 datasets as the attack strength increases, comparing the baseline model to its ET3-augmented variant. ϵ_a indicates the strength of attack. Across all datasets, ImageNet-21k labels serve as the reference text embeddings for computing the energy $E(\cdot, \theta)$.

(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.
Clean Data	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	
	+ET3	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)
$2/255$	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	
	+ET3	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)
$4/255$	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	
	+ ET3	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)
$6/255$	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	
	ET3	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)
$8/255$	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	
	ET3	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)
$10/255$	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	
	ET3	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.
Clean Data	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	
	+ET3	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)
$2/255$	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	
	+ ET3	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)
$4/255$	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	
	+ ET3	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)
$6/255$	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	
	ET3	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)
$8/255$	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	
	ET3	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)
$10/255$	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	
	ET3	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)

to the threat model's attack budget. Our base robust models adhere to this standard: they are trained with ℓ_∞ constraints and evaluated against ℓ_∞ attacks that equal or exceed the training budget. However, because ET3 is a TTT method operating actively within the *inference pipeline*, its parameters serve a fundamentally different purpose.

The primary goal of ET3 is to amplify the features of the ground-truth concept rather than to simply mask adversarial noise. Consequently, ET3 utilizes an ℓ_2 projection, even when evaluating against ℓ_∞ attacks. The ℓ_2 constraint regulates the extent of the TTT transformation to ensure clean accuracy is preserved; it is not designed to match the adversarial threat constraint. Thus, the relatively "large" ℓ_2 radius utilized by ET3 should be understood as a hyperparameter of the defense pipeline rather than an unfair budget advantage.

This inference-time transformation also changes how we evaluate perceptual quality. In adversarial purification, where external generative models reconstruct images independently of the classifier, pixel-level fidelity is a primary constraint to avoid introducing new perceptual artifacts. In contrast, ET3 performs its transformation directly

Table 13. **Evaluating LLaVA 1.5-7B with ET3 across different vision encoders in the *defense-aware setting*.** Robust scores is reported under $\epsilon_a = 4/255$ using standard CLIP and the TeCoA/FARE backbones adversarially trained with $\epsilon_t = 2/255$ and $\epsilon_t = 4/255$. Across all tasks, ImageNet-21k labels serve as the reference text embeddings for computing the energy $E(\cdot, \theta)$. The column labeled $(4/255)$ reports results without any test-time defense. **+ET3** denotes evaluation under a non-adaptive attack, while **+ET3*** denotes evaluation under an defense-aware adaptive attack. COCO and Flickr30k are evaluated using CIDEr for captioning, while TextVQA and VQAv2 report VQA accuracy.

	COCO [44]			**Flickr30k** [103]			**TextVQA** [77]			**VQAv2** [29]			**Average**		
	4/255	+ET3	+ET3*	4/255	+ET3	+ET3*	4/255	+ET3	+ET3*	4/255	+ET3	+ET3*	4/255	+ET3	+ET3*
CLIP	2.8	19.2	3.5	0.9	16.0	5.0	0.0	12.0	2.0	0.0	19.6	7.4	0.9	16.7 (+15.8)	4.5 (+3.6)
TeCoA^2	24.9	43.8	34.9	20.9	30.3	29.4	5.8	13.8	11.8	21.8	35.8	25.8	18.3	30.9 (+12.6)	25.5 (+7.2)
FARE^2	21.8	37.0	33.4	22.5	29.5	28.4	5.8	15.8	11.8	20.4	33.2	26.4	17.6	28.9 (+11.3)	25.0 (+7.4)
TeCoA^4	25.4	45.0	41.4	22.4	31.6	29.6	7.8	9.8	9.8	23.0	37.4	33.4	19.7	30.9 (+11.2)	28.6 (+8.9)
FARE^4	28.1	36.2	34.3	28.0	39.5	39.9	13.8	19.8	17.8	26.6	35.8	31.8	24.1	32.8 (+8.7)	31.0 (+6.9)

Table 14. **Evaluating LLaVA 1.5-7B with ET3 with ℓ_∞ *projection* across different vision encoders in the *defense-aware setting*.** Robust scores is reported under $\epsilon_a = 4/255$ using standard CLIP and the TeCoA/FARE backbones adversarially trained with $\epsilon_t = 2/255$ and $\epsilon_t = 4/255$. The ET3 update is projected into the same ℓ_∞ ball as the attack with radius $4/255$. Across all tasks, ImageNet-21k labels serve as the reference text embeddings for computing the energy $E(\cdot, \theta)$. The column labeled $(4/255)$ reports results without any test-time defense. **+ET3** denotes evaluation under a non-adaptive attack, while **+ET3*** denotes evaluation under an defense-aware adaptive attack. COCO and Flickr30k are evaluated using CIDEr for captioning, while TextVQA and VQAv2 report VQA accuracy.

	COCO [44]			**Flickr30k** [103]			**TextVQA** [77]			**VQAv2** [29]			**Average**		
	4/255	+ET3	+ET3*	4/255	+ET3	+ET3*	4/255	+ET3	+ET3*	4/255	+ET3	+ET3*	4/255	+ET3	+ET3*
CLIP	2.8	16.1	4.0	0.9	12.8	2.9	0.0	6.0	0.0	0.0	14.6	6.2	0.9	12.4 (+11.5)	3.3 (+2.4)
TeCoA^2	24.9	42.1	36.1	20.9	31.6	29.8	5.8	13.8	11.8	21.8	37.8	29.8	18.3	31.3 (+13.0)	26.9 (+8.6)
FARE^2	21.8	31.4	30.0	22.5	33.2	30.2	5.8	13.8	10.6	20.4	29.8	23.8	17.6	27.0 (+9.4)	23.7 (+6.1)
TeCoA^4	25.4	45.3	39.9	22.4	32.2	31.1	7.8	9.8	7.8	23.0	37.4	33.4	19.7	31.1 (+11.4)	28.1 (+8.4)
FARE^4	28.1	34.2	35.1	28.0	40.2	37.9	13.8	19.8	19.8	26.6	35.8	31.8	24.1	32.5 (+8.4)	31.1 (+7.0)

Table 15. **Zero-shot robustness of ET3 with ℓ_∞ *projection* across 14 benchmark datasets in the defense-unaware setting.** Comparison of clean and robust accuracy for baseline models versus the same models augmented with ET3. Robustness is evaluated against AutoAttack (AA) at $\epsilon_a = 4/255$. The ET3 update is projected into the same ℓ_∞ ball as the attack, with radius $4/255$.

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)	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)	74.78	78.09	36.81	80.77	50.26	38.09	21.72	12.27	39.05	73.66	54.95	49.98	75.93	94.49	55.78	(+0.09)
$\epsilon_t = 4/255$	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)	52.00	64.40	10.70	51.30	31.60	21.70	13.40	4.60	20.10	48.30	39.80	51.00	60.60	82.20	39.41	(+6.56)

within the predictive model itself. Therefore, the relevant criterion is the preservation of *clean accuracy* rather than the raw visual fidelity of the image (visualizations of ET3-transformed images are provided in the Fig. 7).

While our ℓ_2 formulation is by design, we also demonstrate that ET3's effectiveness is not merely an artifact of utilizing a larger or different norm. To confirm this, we evaluate ET3 using an ℓ_∞ projection strictly equal to the attack budget ($\epsilon = 4/255$). Under the same settings as Table 1, the average performance across 14 datasets for TeCoA (ViT-L/14, $\epsilon = 4/255$) demonstrates consistent gains: clean accuracy improves from 55.69% to 55.78%, and robust accuracy improves significantly from 32.85% to 39.41% as shown in Tab. 15. Finally, we also evaluate ET3 with this strict ℓ_∞ projection on Large Vision-Language Models (LVLMs) against adaptive attacks as shown in Tab. 14. Across all

benchmark datasets, the improvement persists when using a robust vision encoder, further confirming that genuine feature induction drives the observed robustness gains.

D.3. 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 fine-tuned 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. Detailed numerical results are provided in Tab. 11. 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. 12.

D.4. 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, 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).

[1] https://github.com/mosjel/ImageNet_21k_Original_OK

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. 3 and 9 to 12. In contrast, Tables 1 and 2 of the main paper, along with Tab. 6 and Tab. 15, 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 comparable 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.

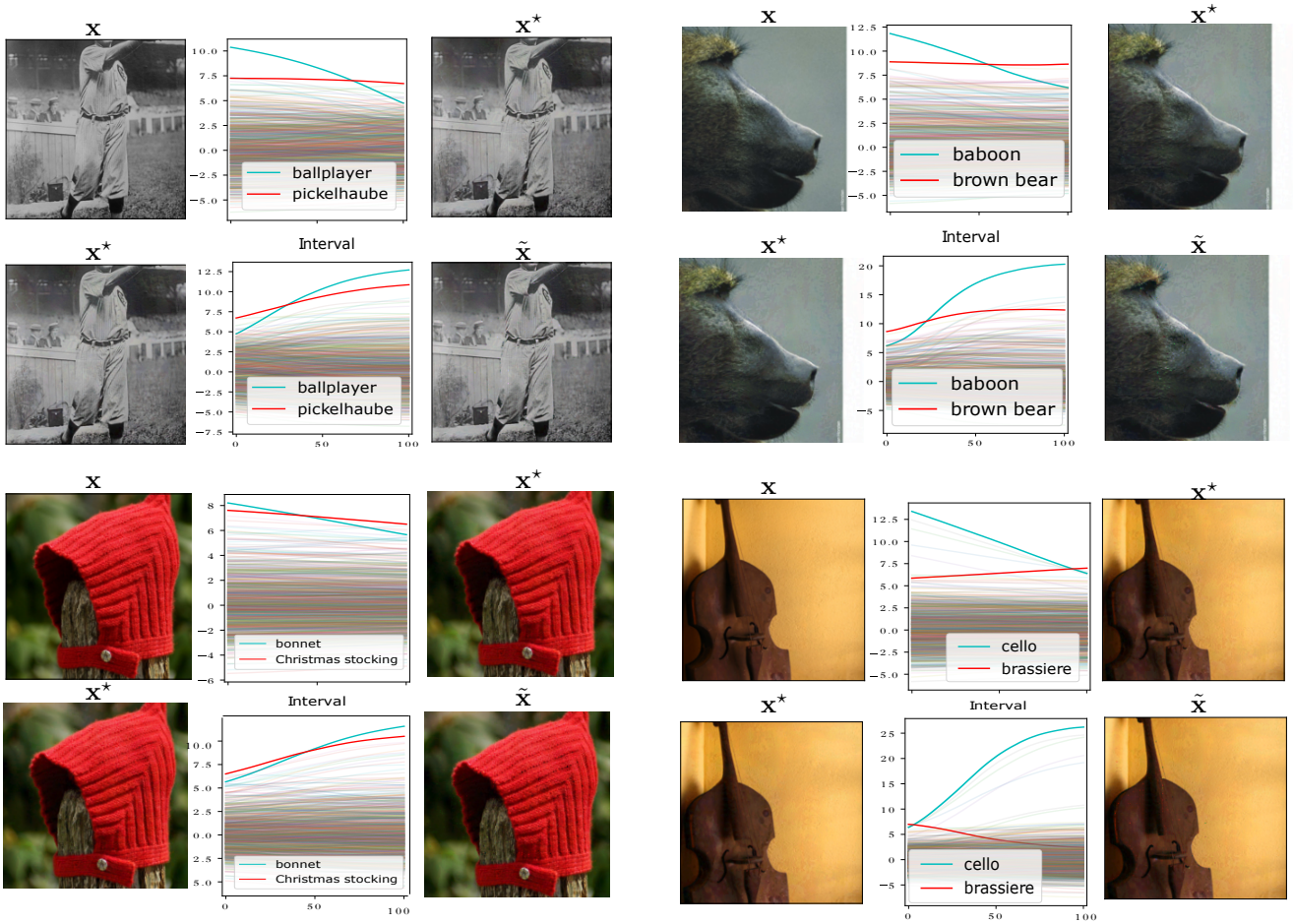
Q: which food is being advertised?

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.

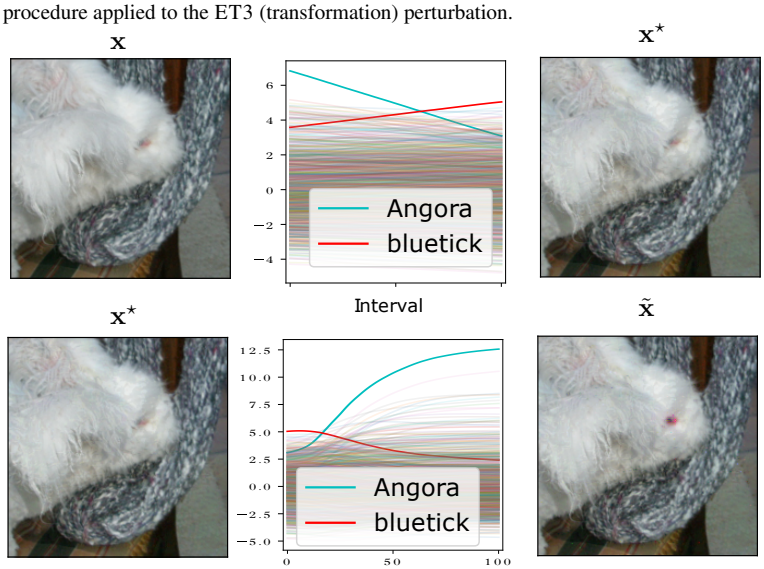
Q: What kind of animal is this?

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.



reference image of an Angora rabbit

(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

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

leading to

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

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

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

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

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

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

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

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

$$2r_x + \epsilon\left\|\mathbf{g}_1\right\| > 0$$
$$\epsilon > \frac{-2r_x}{\left\|\mathbf{g}_1\right\|}$$

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

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

the claim follows.

□