WHAT MAKES A GOOD ANCHOR? OFFLINE ANCHOR SELECTION FOR TEST-TIME ADVERSARIAL ROBUSTNESS OF CLIP

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time defense techniques improve the zero-shot adversarial robustness of CLIP without retraining. Among them, rectification-based methods provide a direct and low-overhead strategy by seeking a correction that counteracts the attack by moving the adversarial representation toward an *anchor*, a transformation of the input intended to approximate the unavailable clean representation. Because this anchor serves as the reference for estimating the correction, its selection is fundamental to successful rectification. Yet existing methods typically prescribe an anchor a priori, with no criterion for comparing it against alternatives, and justify it only through the performance of the resulting defense. In this paper, we formulate anchor selection as an offline measurement problem and ask: *what makes a good anchor?* We identify four complementary properties required for effective rectification and quantify them in CLIP feature space to construct an Anchor Quality Score (AQS). We rank representative anchors from 10 transformation families based on their AQS and validate this ordering against the ranking obtained by building the full defense around each anchor. This experiment selects JPEG compression at quality 50 as the strongest anchor. Building on this selection, we introduce **A**nchor-Guided **R**ectification using **C**ompression (**ARC**), a lightweight test-time defense that combines the selected anchor with a soft anchor-consistency scaling of the rectification budget. Across ten zero-shot classification benchmarks, ARC delivers substantial improvements in adversarial robust accuracy over SOTA test-time defenses while maintaining comparable clean accuracy and inference cost, and remains effective under adaptive attacks that explicitly target the complete defense.

1 INTRODUCTION

Pre-trained vision-language models (VLMs) such as CLIP (Radford et al., 2021) embed images and text in a shared space and classify unseen categories by matching an image embedding against text prompts, without task-specific training. This zero-shot ability has made CLIP a standard recognition backbone and a frozen visual encoder inside larger multimodal systems. The same encoder is, however, highly vulnerable to adversarial perturbations that collapse its zero-shot accuracy across downstream tasks (Mao et al., 2022; Zhao et al., 2023; Schlarmann et al., 2024).

The dominant response to such vulnerabilities has been adversarial fine-tuning of the vision encoder or its prompts (Mao et al., 2022; Wang et al., 2024; Schlarmann et al., 2024; Li et al., 2024). These methods require adversarial training on large datasets for every backbone and typically trade clean accuracy and zero-shot generalization for robustness (Mao et al., 2022; Wang et al., 2024). Test-time defenses avoid retraining and instead adapt each input at inference. We group them into three families by how they counter the perturbation (Figure 1). *Multi-view-based* methods (Sheng et al., 2025; Kim & Um, 2026b; Hai et al., 2026) aggregate predictions over many augmented views, which dilutes the perturbation but makes inference cost grow with the number of views. *Counterattack-based* methods (Xing et al., 2025; Jiang et al., 2026) are cheap, but they derive the counter direction from the adversarial input alone, with no reference to the clean representation resulting in sub-optimal performance. *Rectification-based* methods (Tong et al., 2025; Brahma & Biswas, 2026; Nie et al., 2026) estimate a correction that moves the adversarial representation toward an *anchor*, a transfor-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mation of the input that serves as a proxy for the unavailable clean representation. Rectification is attractive because the anchor gives the correction an explicit target, and the procedure needs only a few gradient steps on a single view. We review these defense methods in detail in Appendix A.2.

The anchor defines where rectification moves the representation, but it has never been selected systematically. Existing methods prescribe it a priori, e.g., Gaussian noise (Tong et al., 2025) or a low-pass filter (Nie et al., 2026), and justify the choice only through the accuracy of the full defense. Many other input transformations are known to weaken adversarial perturbations (Xu et al., 2018; Guo et al., 2018; Xie et al., 2018), and comparing them through end-to-end accuracy is expensive and entangles the anchor with the rest of the pipeline.

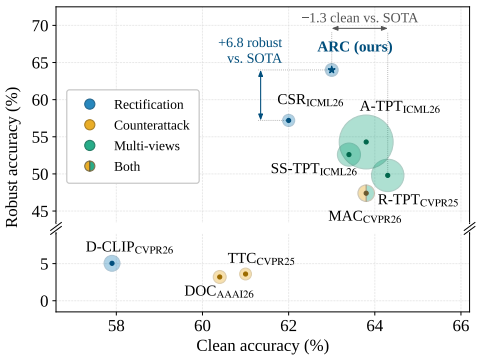

Figure 1: SOTA clean–robust–cost trade-off averaged across ten datasets. Marker color denotes defense family and marker area is proportional to per-image inference time. ARC attains the highest robust accuracy at a fraction of the cost of multi-view tuning methods.

We therefore treat anchor selection as an offline measurement problem and ask: *what makes a good anchor?* Using paired clean and adversarial features, we decompose an anchor's error into a component along the attack displacement and a component orthogonal to it. This yields two geometric properties, *removal* of the attack direction and *avoidance of collateral* distortion. We show that the range of beneficial corrections widens with the first and shrinks with the second. We further add two deployment requirements: *clean fidelity*, since the anchor must not damage benign inputs, and *separability*, since the anchor also supplies the consistency signal that decides which inputs to correct. A weighted harmonic mean combines the four terms into an Anchor Quality Score (AQS), so that an anchor cannot hide a failure on one property behind strength on the others. Ranking representatives of ten transformation families by AQS reproduces the ranking obtained by building the full defense around each anchor, and selects JPEG compression. JPEG has a long history as an input-transformation defense, where it is easily bypassed once the attacker approximates its gradient (Athalye et al., 2018). Here it plays a different role by serving as a reference for the correction and is never the final input to the classifier.

Building on this selection, we propose **A**nchor-Guided **R**ectification using **C**ompression (**ARC**) using JPEG as an anchor for the defense. We further scale the rectification budget by a soft anchor-consistency weight, in place of a fixed budget. Across ten zero-shot classification benchmarks, ARC improves average robust accuracy by 6.8 points over the strongest prior defense while staying within 1.3 points of the best clean accuracy and minimal inference overhead (Figure 1). It retains its advantage under AutoAttack (Croce & Hein, 2020), and under adaptive attacks (Athalye et al., 2018) that target the complete defense. Our contributions are as follows:

- We identify four key properties that characterize a reliable clean proxy and introduce an Anchor Quality Score (AQS), which quantifies these properties to select anchors for test-time rectification without constructing the full defense for each candidate.
- We evaluate anchors within and across ten transformation families, show that AQS predicts the ranking of realized defense performance, and identify JPEG as the strongest anchor, with a frequency-based explanation for its effectiveness.
- We introduce Anchor-Guided Rectification using Compression (ARC) around the selected anchor and show that principled anchor selection improves test-time adversarial robustness on ten zero-shot classification benchmarks under both standard and adaptive attacks.

2 Quantifying Anchor Quality

We formalize anchor quality by comparing how an anchor transforms the adversarial representation relative to its clean counterpart. We decompose the anchor error into the component aligned with the adversarial displacement and the component introduced in a direction orthogonal to the attack. This geometric view lets us distinguish attack removal from arbitrary feature distortion. We then

complement it with two deployment-oriented requirements—preservation of clean semantics and separability of clean and adversarial inputs—yielding the four components of the Anchor Quality Score (AQS).

Setup. Let $f(\cdot)$ be the normalized CLIP image encoder and $\mathcal{A}(z)$ the normalized anchor feature for input z; for an image transform T, $\mathcal{A}(z) = f(T(z))$. For a clean/adversarial (x/x_{adv}) pair,

$$f_c = f(x), \quad f_a = f(x_{\text{adv}}), \quad a = \mathcal{A}(x_{\text{adv}}), \quad a_0 = \mathcal{A}(x).$$

The geometric terms below are computed after projection into the centered, whitened span of the class text embeddings followed by ℓ_2 normalization; we denote this map by $\phi(\cdot)$ and the resulting features by $\bar{f}_c, \bar{f}_a, \bar{a}$ (see Appendix A.3). Clean fidelity and gate consistency are measured in the original CLIP feature space, since they directly correspond to the quantities used by the defense at inference time.

2.1 What Makes an Anchor Good?

Removal. Let $u = \bar{f}_a - \bar{f}_c$ be the attack displacement and $v = \bar{a} - \bar{f}_c$ the anchor error relative to the clean feature. We measure the fraction of the attack direction retained by the anchor and convert it to a $[0, 1]$ goodness score:

$$r = \frac{\langle v, u \rangle}{\|u\|_2^2}, \qquad q_{\text{rem}} = \text{clip}_{[0,1]}(1 - r), \tag{1}$$

Thus $r \approx 1$ describes an identity-like anchor that inherits the attack, while $r \approx 0$ indicates that its attack-aligned component has largely been removed.

Collateral avoidance. Removal alone can be destructive, so we also measure the part of the anchor error orthogonal to the attack direction, normalized by the attack magnitude:

$$c = \frac{\|v - ru\|_2}{\|u\|_2}, \qquad q_{\text{collat}} = \frac{1}{1 + c/s_c}, \tag{2}$$

where s_c is the scale parameter for the collateral score goodness mapping. This distinguishes true purification from simply moving the representation elsewhere. Among anchors with similar removal, the one introducing less unrelated distortion scores higher.

Theorem 1 (When anchor guidance is beneficial)**.** *Write $v = ru + w$ with $w \perp u$ and $\|w\|_2 = c\|u\|_2$, and let $\bar{f}_\beta = (1-\beta)\bar{f}_a + \beta\bar{a}$ for $\beta > 0$. Then*

$$\frac{\|\bar{f}_\beta - \bar{f}_c\|_2^2}{\|\bar{f}_a - \bar{f}_c\|_2^2} = \big[1 - \beta(1-r)\big]^2 + \beta^2 c^2. \tag{3}$$

If $r < 1$, then $\|\bar{f}_\beta - \bar{f}_c\|_2 < \|\bar{f}_a - \bar{f}_c\|_2$ for all $0 < \beta < \frac{2(1-r)}{(1-r)^2 + c^2}$.

Corollary 1 (Optimal correction under a bounded step)**.** *Under the setting of Theorem 1 with $r < 1$, let $d = 1 - r$, let*

$$E(\beta) \;=\; \frac{\|\bar{f}_\beta - \bar{f}_c\|_2^2}{\|\bar{f}_a - \bar{f}_c\|_2^2} \;=\; (1 - \beta d)^2 + \beta^2 c^2 \tag{4}$$

be the normalized error of the anchor-guided update, and let $\beta_{\max} > 0$ be the largest step the rectifier can take. Over $0 < \beta \le \beta_{\max}$, E is uniquely minimized at $\hat{\beta} = \min\{\beta^\star, \beta_{\max}\}$ with $\beta^\star = d/(d^2 + c^2)$, and

$$E(\hat{\beta}) \;=\; \begin{cases} \dfrac{c^2}{d^2 + c^2}, & \beta^\star \le \beta_{\max}, \\ (1 - \beta_{\max} d)^2 + \beta_{\max}^2 c^2, & \beta^\star > \beta_{\max}. \end{cases} \tag{5}$$

Theorem 1 explains why removal and collateral must be measured together, and Corollary 1 shows how the anchor limits the correction a rectifier of finite strength can reach. Removal determines whether the anchor's direction can be exploited within a finite budget, and collateral determines how good that direction is. An anchor that removes the attack with little collateral displacement, $r \to 0$ and $c \to 0$, has $\beta^\star \to 1$ and permits near-complete recovery. When the anchor retains the attack ($d \to 0$), $\beta^\star$ exceeds any fixed budget, and $E(\hat{\beta}) \to 1$ as $d, c \to 0$. Proofs are presented in Appendix A.4.

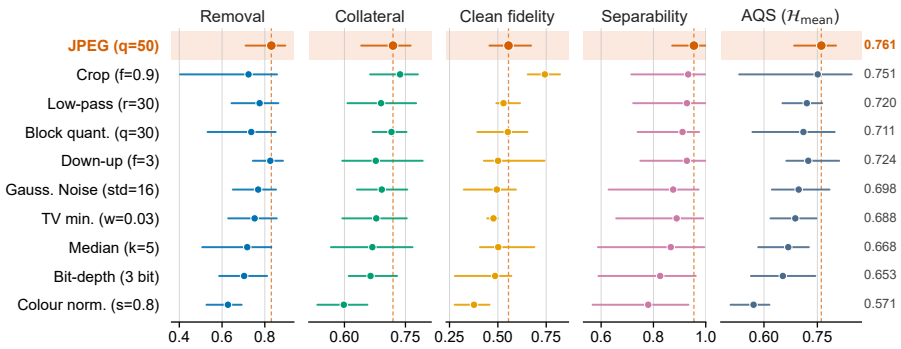

Figure 2: **Inter-family anchor selection.** Component decomposition and resulting harmonic-mean AQS for the family representatives. JPEG50 is highlighted in orange.

Clean fidelity. A useful anchor should not obtain robustness by corrupting benign inputs. Applying the same anchor to the clean image, we measure its cosine cost in the original CLIP space:

$$\kappa = 1 - \langle a_0, f_c \rangle, \qquad q_{\text{clean}} = \frac{1}{1 + \kappa / s_\kappa}, \tag{6}$$

where s_κ is the scale parameter for the clean fidelity goodness mapping. This penalizes transformations that remove the attack only by sacrificing the clean semantic representation.

Separability. The anchor also supplies the consistency signal $s(z)$ used by the test-time gate. With $s(z) = \langle f(z), \mathcal{A}(z) \rangle$, we evaluate clean–adversarial separation without fixing a threshold:

$$\text{AUC}_{\text{gate}} = \Pr[s(x) > s(x_{\text{adv}})], \qquad q_{\text{sep}} = \text{clip}_{[0,1]}(2(\text{AUC}_{\text{gate}} - 0.5)). \tag{7}$$

Hence $q_{\text{sep}} = 0$ is uninformative ordering and $q_{\text{sep}} = 1$ is perfect separation.

2.2 Anchor Quality Score

The four components capture complementary failure modes. We combine them with a weighted harmonic mean,

$$\text{AQS}(\mathcal{A}) = \frac{\sum_i w_i}{\sum_i w_i / \max(q_i, \epsilon_h)}, \quad i \in \{\text{rem}, \text{collat}, \text{clean}, \text{sep}\}. \tag{8}$$

The harmonic mean makes a weak requirement difficult to hide behind strong performance elsewhere. We assign a larger weight to removal, $w_{\text{rem}} = 3$, because it directly reflects adversarial robustness, while using unit weight for the remaining terms, with $s_c = 1$, $s_\kappa = 0.10$, and $\epsilon_h = 10^{-3}$.

AQS is an *offline selection criterion* as removal and collateral require the paired clean feature. We therefore compute it on a development set and restrict analysis to examples correctly classified by clean CLIP, avoiding confounding anchor quality with base-model errors. The necessity of each component is further evaluated in Section 4.3.

2.3 Inter-Family Selection and Validation

We first perform an intra-family strength sweep and retain the best operating point from each transformation family; sweep details are deferred to the Appendix A.6 and A.7. The resulting representatives are then compared on Pets (Parkhi et al., 2012), Flowers102 (Nilsback & Zisserman, 2008), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), and Caltech101 (Fei-Fei et al., 2004) under an l_∞ Projected Gradient Descent (PGD) Madry et al. (2017) attack with 10 steps and attack strength $\epsilon \in \{1, 4\}/255$, giving ten dataset–attack cells. Anchors are ranked by AQS within each cell and the ranks are averaged, emphasizing consistency across conditions.

Figure 2 shows the inter-family AQS scores and their corresponding component decomposition. JPEG at quality 50 (JPEG50) ranks first overall and achieves the highest mean AQS (0.761). The

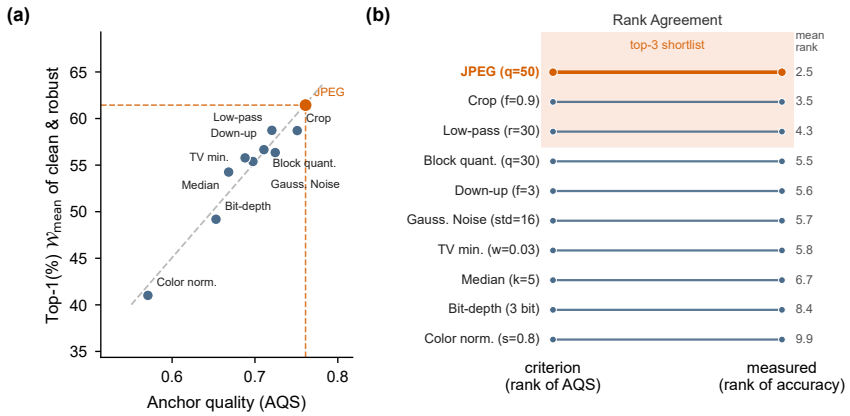

Figure 3: **Does AQS predict defense performance?** **(a)** AQS versus realized defense performance, measured by the weighted mean of clean and robust accuracy. **(b)** Rank agreement between the AQS criterion and the measured defense-performance ranking.

component view shows how anchors differ substantially in how they trade removal against collateral avoidance, clean fidelity, and separability, with JPEG50 providing the strongest overall balance.

We then ask whether this criterion ranking predicts what happens after the full defense is built. Holding the rectification pipeline (see Section 3.2) fixed, we rank each candidate anchor per evaluation cell using a weighted combination of clean- and robust-accuracy ranks, with greater weight on robustness because it is the primary objective of the defense. As shown in Fig. 3(a), higher AQS tracks higher realized performance. The induced rankings are in complete agreement (Fig. 3(b)), and JPEG50 remains the strongest measured anchor with mean performance rank 2.5.

3 Method

3.1 Preliminaries

Given a clean image x with label y, an adversarial example is

$$x_{\text{adv}} = x + \delta^\star, \qquad \delta^\star = \arg\max_{\|\delta\|_p \le \epsilon} \mathcal{L}_{\text{cls}}(x + \delta, y), \tag{9}$$

where ϵ bounds the attack magnitude (Madry et al., 2017). Rectification-based defenses counter this perturbation at test time by seeking a small correction whose feature-space effect moves the adversarial representation back toward its clean counterpart. Because the clean representation is unavailable at inference, these methods use an *anchor*—a transformed view intended to approximate it—as the correction target. The quality of this proxy therefore directly determines whether rectification cancels the adversarial shift or follows a distorted reference.

3.2 Anchor-Guided Rectification using Compression

Section 2 evaluates candidate anchors independently of the defense and selects JPEG50. We instantiate this choice in *Anchor-Guided Rectification using Compression* (ARC), a lightweight test-time pipeline that uses JPEG50 both as the reference for rectification and as the consistency signal controlling how strongly rectification is applied. The complete pipeline is summarized in Fig. 4.

JPEG50 as an anchor. JPEG first converts the image into a luminance–chrominance representation, subsamples chrominance, partitions the image into 8×8 blocks, and expresses each block in a discrete cosine transform (DCT) basis. Its central operation is frequency-dependent quantization:

$$\widehat{G}^{(b)}_{u,v} = Q_{u,v}\ \text{round}\left(\frac{G^{(b)}_{u,v}}{Q_{u,v}}\right), \tag{10}$$

where $G^{(b)}_{u,v}$ is a DCT coefficient and $Q_{u,v}$ its quantization step. Higher-frequency coefficients are quantized more coarsely, so small mid-to-high-frequency components can be removed *entirely*,

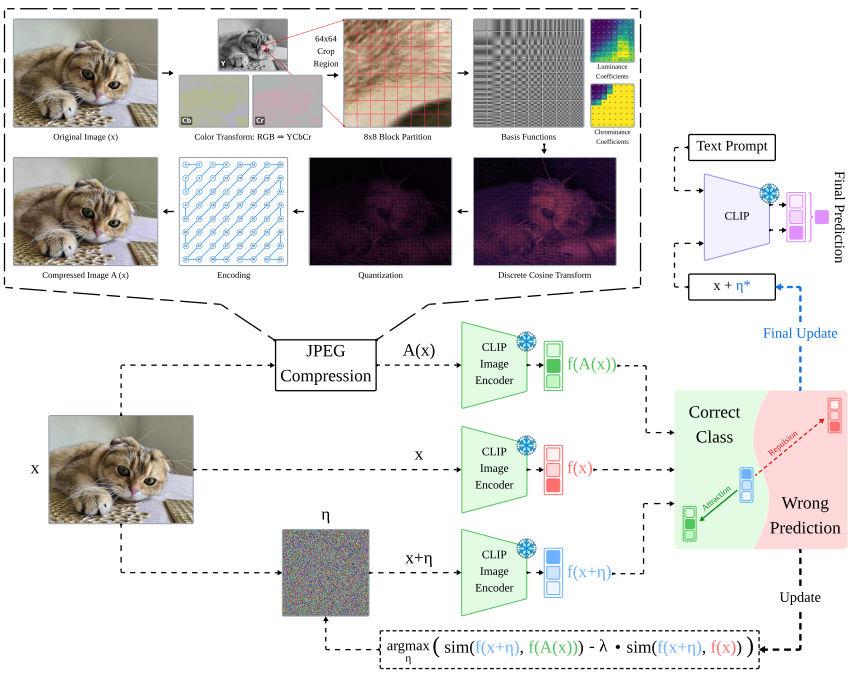

Figure 4: **Overview of ARC.** The upper panel illustrates how JPEG compression produces the selected anchor $A(x)$: the input is transformed to YCbCr, partitioned into 8×8 blocks, mapped to the DCT basis, quantized, and decoded to obtain the compressed image. In the lower panel, ARC extracts features from the original input x, its JPEG anchor $A(x)$, and the rectified input $x + \eta$; the anchor guides the contrastive rectification objective toward the correct representation while the original feature serves as the repulsive reference. A soft consistency weight scales the budget of the optimized correction $\eta^\star$, and the resulting image $x + \eta^\star$ is passed to CLIP together with the text prompts for the final zero-shot prediction.

while dominant low-frequency structure is retained at finer precision. In particular, this is relevant for CLIP, for which prior work identifies much of the adversarial effect in the mid-to-high-frequency range (Nie et al., 2026). Unlike linear smoothing, which continuously attenuates these components and increasingly blurs useful structure as filtering becomes stronger, JPEG combines localized blockwise processing with nonlinear coefficient thresholding. This mechanism also explains JPEG50 through the four AQS components. Removing small attack-sensitive frequency components improves *removal*; preserving dominant image structure limits attack-orthogonal movement and therefore improves *collateral* avoidance; retaining the low-frequency semantic content supports *clean fidelity*; and the larger consistency change induced on attacked inputs than on benign inputs provides the *separability* required by the gate.

Inference. For a test input x_t, ARC computes the original feature $z_0 = f(x_t)$ and JPEG50 anchor $a = f(\mathcal{J}_{50}(x_t))$. Their cosine consistency $C(x_t) = z_0^\top a$ is converted into a soft rectification weight

$$g(x_t) = \sigma\left(\frac{\tau - C(x_t)}{T}\right), \qquad \tilde{x}_t = \begin{cases} x_t, & g(x_t) \le g_\tau, \\ x_t + \eta^\star, & g(x_t) > g_\tau, \end{cases} \tag{11}$$

where T controls the transition around τ, g_τ is the gate threshold, and $\eta^\star$ is the corrective perturbation returned by the rectification procedure below, whose budget is scaled by $g(x_t)$. Thus, highly consistent inputs are left unchanged or receive little correction, whereas low-consistency inputs are allowed a larger correction without an abrupt hard decision on its size. The resulting $\tilde{x}_t$ is passed to the frozen CLIP classifier to obtain the final prediction.

For the corrective step, we adopt the contrastive rectification objective of CSR (Nie et al., 2026),

$$\underset{\|\eta\|_\infty \le g(x_t)\,\epsilon_\eta}{\arg\max} \; \big(\,\mathrm{sim}\,(f(x_t+\eta),\, a) - \lambda \cdot \mathrm{sim}\,(f(x_t+\eta),\, z_0)\,\big), \tag{12}$$

Table 1: Clean (Acc) and Robust (Rob) accuracy (%) of test-time defenses on ten zero-shot classification benchmarks using CLIP ViT-B/16 and PGD-10 attack at $\epsilon = 1/255$ and step size $\epsilon/4$. ΔRob is the average robust-accuracy gain of ARC over each method. **Bold** (Acc) and **bold orange** (Rob) mark the best and underline the second-best result per column.

Method	Caltech101		Pets		Cars		Flowers102		Aircraft		DTD		EuroSAT		UCF101		Food101		ImageNet		Average		
	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	ΔRob
R-TPT$_{\text{CVPR25}}$	93.5	88.9	86.8	44.5	**66.8**	47.0	69.1	55.5	24.1	16.4	46.1	39.4	35.7	27.6	67.2	56.2	84.3	67.9	**69.4**	54.5	**64.3**	49.8	+14.2
SS-TPT$_{\text{ICML26}}$	93.5	88.5	86.1	73.9	65.4	48.0	67.9	54.5	23.4	17.6	45.3	38.8	35.2	27.6	65.9	55.9	82.2	66.6	68.8	54.9	63.4	52.6	+11.4
A-TPT$_{\text{ICML26}}$	**93.6**	**89.6**	89.5	78.7	65.5	49.4	69.9	59.1	24.1	17.9	**47.9**	42.2	33.2	25.0	**67.5**	58.5	82.6	68.6	**69.4**	54.5	**64.3**	54.3	+9.7
TTC$_{\text{CVPR25}}$	87.9	16.5	82.4	6.24	54.8	0.20	69.4	1.46	22.7	0.03	41.3	5.32	**47.5**	0.55	66.6	1.30	87.0	0.82	50.8	4.00	61.0	3.6	+60.4
DOC$_{\text{AAAI26}}$	87.8	16.7	76.2	3.98	55.9	0.14	69.8	0.86	23.3	0.03	41.1	4.95	47.3	0.55	66.8	1.11	87.8	0.57	47.8	2.72	60.4	3.2	+60.8
D-CLIP$_{\text{CVPR26}}$	87.1	4.15	82.9	7.96	53.5	2.62	65.6	6.81	22.2	0.84	35.7	6.12	40.5	1.10	65.1	1.67	70.0	9.42	56.3	9.71	57.9	5.0	+59.0
CSR$_{\text{ICML26}}$	83.2	75.3	88.9	65.9	62.7	60.9	65.8	53.5	22.7	**31.2**	41.9	39.1	42.9	**40.4**	67.1	66.3	81.9	80.7	62.5	58.9	62.0	57.2	+6.8
MAC$_{\text{CVPR26}}$	88.1	75.6	89.0	46.9	57.8	48.9	**70.9**	34.9	**24.5**	23.5	42.5	31.8	45.3	33.0	67.2	58.6	**88.0**	72.8	64.3	48.4	63.8	47.4	+16.6
ARC (ours)	85.9	86.2	**89.9**	**82.9**	62.2	**69.4**	70.6	**67.9**	23.5	30.9	41.5	**42.4**	41.9	35.0	67.3	**71.7**	83.3	**85.0**	63.7	**68.7**	63.0	**64.0**	–

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity. This attracts the corrected representation toward the JPEG50 anchor while discouraging it from remaining at the incoming representation. Unlike CSR, which uses a fixed budget ϵ_η, we scale the budget by $g(x_t)$. We optimize this objective with a small fixed number of projected sign-gradient steps, keeping $x_t + \eta$ in $[0, 1]$, and retain the iterate $\eta^\star$ with the highest objective value along the trajectory.

4 Experiments

4.1 Experimental Setup

Datasets. We follow the evaluation protocol of prior test-time defenses for CLIP (Sheng et al., 2025; Xing et al., 2025) and report results on ten zero-shot classification benchmarks: Caltech101 (Fei-Fei et al., 2004), OxfordPets (Parkhi et al., 2012), StanfordCars (Krause et al., 2013), Flowers102 (Nilsback & Zisserman, 2008), FGVC-Aircraft (Maji et al., 2013), DTD (Cimpoi et al., 2014), EuroSAT (Helber et al., 2019), UCF101 (Soomro et al., 2012), Food101 (Bossard et al., 2014), and ImageNet (Deng et al., 2009). These cover generic objects, fine-grained categories, textures, satellite imagery, and human actions. All evaluation is performed on the test split of each dataset.

Implementation Details. We evaluate on pre-trained CLIP with ViT-B/16 and ViT-B/32 backbones using each dataset's default prompt template. Adversarial examples are generated with PGD (Madry et al., 2017) under an ℓ_∞ constraint using 10 steps, $\epsilon = 1/255$, and step size $\alpha = \epsilon/4$, unless otherwise specified. We use JPEG compression at quality 50 as the anchor for all main results. For the rectification stage we follow the setup of CSR (Nie et al., 2026): consistency threshold $\tau = 0.85$, corrective perturbation budget $\epsilon_\eta = 4/255$, step size $\alpha_\eta = 2/255$, repulsion weight $\lambda = 1$, and $N = 3$ rectification steps. Our soft anchor-consistency gate uses temperature $\tau_{\text{temp}} = 0.02$ and gate threshold $g_\tau = 0.25$. All experiments run on a single RTX 6000 GPU.

Baselines. We compare ARC against state-of-the-art test-time defenses for CLIP, including rectification-based methods: D-CLIP (Brahma & Biswas, 2026) and CSR (Nie et al., 2026), counterattack-based methods: TTC (Xing et al., 2025) and DOC (Jiang et al., 2026), prompt-tuning based methods with multi-augmented views: R-TPT (Sheng et al., 2025), SS-TPT (Kim & Um, 2026b), and A-TPT (Hai et al., 2026), and MAC (Kim & Um, 2026a), which combines counterattack with multi-augmented views. All methods are evaluated under the same attack scenario using their officially released implementations and original hyperparameters for a fair comparison.

4.2 Main Results

Comarison with SOTA. Tables 1 and 2 compare ARC with test-time defenses from all three categories on ViT-B/16 and ViT-B/32. ARC achieves the highest average robust accuracy on both backbones and improves over every baseline. On ViT-B/16, it reaches 64.0%, which is 6.8 points above CSR, the strongest baseline, and 9.7 points above the best multi-view method, A-TPT. It is best on seven of the ten datasets, with the largest gains on fine-grained benchmarks (Cars, Flowers102) and on ImageNet (+9.8 over CSR). Its clean accuracy (63.0%) is higher than that of every counterattack

Table 2: Clean (Acc) and Robust (Rob) accuracy (%) of test-time defenses on ten zero-shot classification benchmarks using CLIP ViT-B/32 and PGD-10 attack at $\epsilon = 1/255$ and step size $\epsilon/4$. ΔRob is the average robust-accuracy gain of ARC over each method. **Bold** (Acc) and **bold orange** (Rob) mark the best and underline the second-best result per column.

Method	Caltech101		Pets		Cars		Flowers102		Aircraft		DTD		EuroSAT		UCF101		Food101		ImageNet		Average		
	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	Acc	Rob	ΔRob
R-TPT$_{\text{CVPR25}}$	**90.8**	85.9	85.2	72.9	**62.8**	44.6	62.5	52.9	19.2	13.4	**42.6**	38.1	31.9	22.0	62.3	53.7	78.6	64.3	63.1	46.8	59.9	49.5	+2.3
SS-TPT$_{\text{ICML26}}$	89.9	84.9	83.6	73.9	60.4	43.7	61.8	52.7	18.0	13.7	42.3	36.4	30.5	20.1	61.0	53.2	76.6	62.9	**63.8**	51.7	58.8	49.3	+2.5
A-TPT$_{\text{ICML26}}$	89.6	85.0	**87.8**	77.6	60.2	43.6	64.5	54.4	19.3	15.0	42.4	34.2	35.6	25.2	62.1	55.0	76.8	63.6	63.2	46.6	60.2	50.0	+1.8
TTC$_{\text{CVPR25}}$	86.3	24.5	82.8	28.2	48.2	13.9	64.2	18.6	17.8	2.67	37.2	12.9	52.9	0.56	63.1	5.29	82.4	20.0	49.2	16.5	58.4	14.3	+37.5
DOC$_{\text{AAAI26}}$	86.6	22.6	74.9	20.8	49.0	6.69	64.5	12.0	18.1	1.50	36.5	11.7	**53.2**	0.25	**63.2**	3.62	82.8	11.2	43.2	11.0	57.2	10.1	+41.7
D-CLIP$_{\text{CVPR26}}$	85.3	21.9	81.2	36.2	47.2	21.0	61.7	21.6	16.0	4.98	35.6	13.7	37.2	6.27	60.8	16.4	74.6	37.0	51.9	26.8	55.1	20.6	+31.2
CSR$_{\text{ICML26}}$	81.2	59.0	83.9	49.4	59.6	25.4	62.6	32.0	17.2	16.7	42.4	25.8	34.2	26.1	63.0	45.7	79.4	57.7	56.8	37.1	58.0	37.5	+14.3
MAC$_{\text{CVPR26}}$	85.8	58.0	87.7	27.3	51.8	17.0	64.9	17.5	**19.4**	9.33	40.1	21.9	41.8	17.8	62.3	37.7	**83.1**	34.3	58.9	29.8	59.6	27.1	+24.7
ARC (ours)	86.1	75.0	87.6	67.1	58.6	**50.7**	**66.4**	42.4	19.0	**21.8**	41.3	35.2	42.9	**35.5**	**63.2**	**61.7**	80.9	**74.0**	60.4	**54.8**	**60.6**	**51.8**	–

Table 3: Robust accuracy (%) of test-time adaptation methods on three fine-grained datasets. Adaptive = EOT+BPDA PGD-10 attack and AA = AutoAttack, both with $\epsilon = 1/255$; PGD = PGD-10 with $\epsilon = 4/255$. Best results per column are highlighted.

Method	Flowers102				Pets				Cars			
	Adaptive	AA	PGD	Avg	Adaptive	AA	PGD	Avg	Adaptive	AA	PGD	Avg
SS-TPT$_{\text{ICML26}}$	23.5	58.2	37.1	39.6	34.1	77.6	48.5	53.4	11.3	51.9	25.3	29.5
CSR$_{\text{ICML26}}$	15.3	68.6	45.9	43.3	23.4	87.5	58.2	56.4	7.1	57.3	39.8	34.7
MAC$_{\text{CVPR26}}$	26.0	64.8	**50.9**	47.2	39.0	84.3	**65.2**	62.9	25.1	53.7	**47.1**	42.0
ARC (ours)	**33.9**	**70.4**	**50.9**	**51.7**	**46.0**	**89.0**	64.9	**66.6**	**39.9**	**61.2**	46.5	**49.2**

and rectification baseline and within 1.3 points of the multi-view methods, which aggregate many augmented views per input. Changing the backbone reshuffles the baselines. TTC, DOC, and D-CLIP nearly collapse on ViT-B/16 ($\leq 5\%$ Rob) but recover partially on ViT-B/32 (10–21%), while still trailing far behind. The multi-view methods keep their robust accuracy nearly unchanged and become the strongest baselines, whereas CSR and MAC each lose about 20 points. ARC also degrades on B/32, but it still retains the best average robust accuracy (51.8%) and also attains the best average clean accuracy (60.6%). Its margin over CSR doubles, from 6.8 to 14.3 points.

Robustness under diverse attacks. We further evaluate ARC across additional adversarial attacks including AutoAttack (Croce & Hein, 2020), PGD-10 with a larger budget ($\epsilon = 4/255$), and an end-to-end pipeline-aware adaptive attack that employs both EOT (Expectation Over Transformation) and BPDA (Backward Pass Differentiable Approximation) (Athalye et al., 2018) as a straight-through estimator for non-differentiable components. We evaluate against a representative baseline from each category on three fine-grained datasets as shown in Table 3. ARC achieves the best average robust accuracy on all three datasets, leading MAC by 4.5, 3.7, and 7.2 points on Flowers102, Pets, and Cars, respectively. Detailed attack settings are provided in Appendix A.9.

Table 4: Inference time per image (↓) on Pets, averaged over clean and adversarial inputs, on a single RTX 6000. # Views is the number of augmented views forwarded per input.

Method	# Views	Time (ms)	Rob (%)
R-TPT$_{\text{CVPR25}}$	63	376.4	44.5
SS-TPT$_{\text{ICML26}}$	15	133.6	73.9
A-TPT$_{\text{ICML26}}$	63	1405.9	78.7
TTC$_{\text{CVPR25}}$	–	4.36	6.24
DOC$_{\text{AAAI26}}$	–	9.09	3.98
D-CLIP$_{\text{CVPR26}}$	–	18.9	7.96
CSR$_{\text{ICML26}}$	–	**2.06**	65.9
MAC$_{\text{CVPR26}}$	2	22.1	46.9
ARC (ours)	–	4.38	**82.9**

Inference overhead. Table 4 reports per-image inference time on the Pets dataset with ViT B/16. SS-TPT forwards 15 views and takes 134 ms per image, and A-TPT forwards 63 views and takes over 1.4 s. ARC needs only 4.4 ms, about 30× faster than SS-TPT and 300× faster than A-TPT, while achieving the highest robust accuracy. Among single-view methods, ARC runs at the same cost as TTC and faster than DOC, D-CLIP, and MAC. It is roughly 2 ms slower than CSR, but gains an additional 17.0 points of robust accuracy.

Table 5: **Targeted-degenerate test of AQS component necessity.** Each diagnostic degenerate is constructed to fail the component named in parentheses while surpassing the JPEG50 reference on the other AQS components. Checkmarks indicate components included in the score. Under the full AQS, the reference ranks above every degenerate; removing the targeted component reverses the mean ordering. The final columns report the harmonic mean of clean and robust top-1 accuracy.

		Components in AQS				Mean rank		$H_{\text{clean,rob}}$ (%)	
Variant	Degenerate (target)	Rem.	Collat.	Fid_{cln}	Sep.	Ref.	Degen.	Ref.	Degen.
Full	$\mathcal{D}_{\text{rem}}$ (removal)	✓	✓	✓	✓	1	2	62.43	3.02
	$\mathcal{D}_{\text{col}}$ (collateral)	✓	✓	✓	✓	1	2	62.43	58.33
	$\mathcal{D}_{\text{cln}}$ (clean fidelity)	✓	✓	✓	✓	1	2	62.43	47.70
	$\mathcal{D}_{\text{sep}}$ (separability)	✓	✓	✓	✓	1	2	62.43	61.54
Omit removal	$\mathcal{D}_{\text{rem}}$	×	✓	✓	✓	2	1	62.43	3.02
Omit collateral	$\mathcal{D}_{\text{col}}$	✓	×	✓	✓	2	1	62.43	58.33
Omit clean fidelity	$\mathcal{D}_{\text{cln}}$	✓	✓	×	✓	1.67	1.33	62.43	47.70
Omit separability	$\mathcal{D}_{\text{sep}}$	✓	✓	✓	×	2	1	62.43	61.54

4.3 Ablation Study

Necessity of AQS components. The preceding analysis shows that AQS ranks useful anchors well, but it does not by itself establish that every component contributes non-redundant information. We test this with *targeted degenerate anchors* which are synthetic diagnostic anchors constructed to fail one AQS component while surpassing the reference anchor on the remaining three. Their purpose is to create controlled counterexamples where one property is selectively removed. Construction details and geometric justification are given in Appendix A.11. For each component, we compare the reference anchor (JPEG50) with its corresponding degenerate under the full AQS and a leave-one-component-out variant. A component is necessary if the full score correctly prefers the reference, but removing that component reverses the ordering even though the degenerate performs worse when inserted into the same rectification pipeline. Table 5 shows that with all four components present, JPEG50 outranks every degenerate. Omitting the targeted component reverses the mean ranking in each case; the inversion is weaker for clean fidelity because clean-image distortion is geometrically coupled to the gate score. Full-defense evaluation resolves the ordering consistently where every degenerate underperforms JPEG50. Thus, each AQS component prevents a distinct failure mode that the remaining components alone can miss.

Table 6: Effect of budget scaling with the JPEG50 anchor. Average robust accuracy over ten datasets.

Budget	Rob (%)
Fixed	62.8
Scaled (ARC)	64.0

Impact of budget scaling. Table 6 compares ARC with a variant that keeps the JPEG50 anchor but uses CSR's hard gate with a fixed rectification budget. Scaling the budget by the soft consistency weight adds 1.2 points of robust accuracy. The fixed-budget variant already reaches 62.8%, well above CSR (57.2%, Table 1), so most of ARC's gain over CSR comes from the choice of anchor.

5 Conclusion and Future Work

Conclusion. We explore anchor selection as a central design problem in test-time adversarial rectification for CLIP. We identify four complementary anchor properties: attack removal, collateral avoidance, clean fidelity, and separability, and combine them into the Anchor Quality Score (AQS). AQS consistently selects JPEG50 as the strongest candidate, motivating ARC, which couples this anchor with soft consistency weighting-based rectification. Across ten zero-shot benchmarks and two CLIP backbones, ARC provides strong, robust accuracy while preserving competitive clean performance and low inference cost, including under stronger and adaptive attacks.

Future work. Anchor quality analysis can be extended to other vision-language architectures and multimodal tasks, explore learned or composed anchors beyond fixed transformations, and investigate weaker-supervision strategies for estimating anchor quality. A further direction is to tighten the connection between anchor geometry, adaptive gating, and nonlinear rectification dynamics.

REFERENCES

Motasem Alfarra, Juan C Pérez, Ali Thabet, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Combating adversaries with anti-adversaries. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 5992–6000, 2022.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Roy Betser, Meir Yossef Levi, and Guy Gilboa. Whitened CLIP as a likelihood surrogate of images and captions. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Debarshi Brahma and Soma Biswas. Defending clip via noise-induced feature dynamics for training-free, zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 656–665, 2026.

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

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pp. 1310–1320. PMLR, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Yubo Cui, Xianchao Guan, Zijun Xiong, and Zheng Zhang. Agft: Alignment-guided fine-tuning for zero-shot adversarial robustness of vision-language models. *arXiv preprint arXiv:2603.29410*, 2026.

Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jia Deng, Jin Li, Zhenhua Zhao, and Shaowei Wang. Fpt-noise: Dynamic scene-aware counterattack for test-time adversarial defense in vision-language models. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Junhao Dong, Piotr Koniusz, Xinghua Qu, and Yew-Soon Ong. Stabilizing modality gap & lowering gradient norms improve zero-shot adversarial robustness of vlms. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 236–247, 2025a.

Junhao Dong, Piotr Koniusz, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Improving zero-shot adversarial robustness in vision-language models by closed-form alignment of adversarial path simplices. In *Forty-second International Conference on Machine Learning*, 2025b.

Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.

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

Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *ICLR (Poster)*. OpenReview.net, 2018.

Jia-Wei Hai, Yijun Wang, and Xiu-Shen Wei. Towards fine-grained robustness: Attention-guided test-time prompt tuning for vision-language models. *arXiv preprint arXiv:2605.19956*, 2026.

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

Md Zarif Hossain and Ahmed Imteaj. Slade: Shielding against dual exploits in large vision-language models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24244–24254. IEEE, 2025.

Chengze Jiang, Minjing Dong, Xinli Shi, and Jie Gui. Diversifying counterattacks: Orthogonal exploration for robust cllp inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5359–5368, 2026.

Borna Khodabandeh, Amirabbas Afzali, Amirhossein Afsharrad, Shahab Mousavi, Sanjay Lall, Sajjad Amini, and Seyed-Mohsen Moosavi-Dezfooli. Lore: Lagrangian-optimized robust embeddings for visual encoders. *Advances in Neural Information Processing Systems*, 38:82217–82255, 2025.

Sunoh Kim and Daeho Um. When clip sees more, it fights back harder: Multi-view guided adaptive counterattacks for test-time adversarial robustness. *arXiv preprint arXiv:2606.06938*, 2026a.

Sunoh Kim and Daeho Um. Ss-tpt: Stability and suitability-guided test-time prompt tuning for adversarially robust vision-language models. *arXiv preprint arXiv:2606.06943*, 2026b.

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

Meir Yossef Levi and Guy Gilboa. The double-ellipsoid geometry of CLIP. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24408–24419. IEEE, 2024.

Zhiwei Li, Yitian Pang, Weining Wang, Zhenan Sun, and Qi Li. Ttp: Test-time padding for adversarial detection and robust adaptation on vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1513–1522, 2026.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in neural information processing systems*, 35:17612–17625, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

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

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. *arXiv preprint arXiv:2212.07016*, 2022.

Sen Nie, Jie Zhang, Zhuo Wang, Shiguang Shan, and Xilin Chen. Contrastive spectral rectification: Test-time defense towards zero-shot adversarial robustness of clip. *arXiv preprint arXiv:2601.19210*, 2026.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

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*, pp. 722–729. IEEE, 2008.

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

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 *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 81–91. IEEE, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Ivan Reyes-Amezcua, Gilberto Ochoa-Ruiz, and Andres Mendez-Vazquez. Enhancing image classification robustness through adversarial sampling with delta data augmentation (dda). In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 274–283. IEEE, 2024.

Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. *arXiv preprint arXiv:2402.12336*, 2024.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29958–29967, 2025.

Richard Shin, Dawn Song, et al. Jpeg-resistant adversarial images. In *NIPS 2017 workshop on machine learning and computer security*, volume 1, 2017.

Khurram Soomro, Amir Roshan Zamir, Mubarak Shah, et al. A dataset of 101 human action classes from videos in the wild. *Center for Research in Computer Vision*, 2(11):1–7, 2012.

Linxiang Su and András Balogh. Atac: Augmentation-based test-time adversarial correction for clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 541–551, 2026.

Baoshun Tong, Hanjiang Lai, Yan Pan, and Jian Yin. On the zero-shot adversarial robustness of vision-language models: A truly zero-shot and training-free approach. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19921–19930. IEEE, 2025.

Sibo Wang, Jie Zhang, Zheng Yuan, and Shiguang Shan. Pre-trained model guided fine-tuning for zero-shot adversarial robustness. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24502–24511. IEEE, 2024.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19910–19920. IEEE, 2025.

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.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects through randomization. In *ICLR (Poster)*. OpenReview.net, 2018.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15172–15182. IEEE, 2025.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *NDSS*. The Internet Society, 2018.

Cheng Zhang and Pan Gao. Countering adversarial examples: Combining input transformation and noisy training. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 102–111. IEEE, 2021.

Jiaming Zhang, Xingjun Ma, Xin Wang, Lingyu Qiu, Jiaqi Wang, Yu-Gang Jiang, and Jitao Sang. Adversarial prompt tuning for vision-language models. In *European conference on computer vision*, pp. 56–72. Springer, 2024.

Mingkun Zhang, Keping Bi, Wei Chen, Jiafeng Guo, and Xueqi Cheng. Clipure: Purification in latent space via clip for adversarially robust zero-shot classification. In *International conference on learning representations*, volume 2025, pp. 17360–17384, 2025.

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

Xingyu Zhu, Beier Zhu, Shuo Wang, Kesen Zhao, and Hanwang Zhang. Enhancing clip robustness via cross-modality alignment. *Advances in Neural Information Processing Systems*, 38:17553–17576, 2025.

Xingyu Zhu, Huanshen Wu, Shuo Wang, Beier Zhu, Jiannan Ge, Jiaheng Zhang, and Long Chen. Robustifying vision-language models via test-time prompt adaptation. *arXiv preprint arXiv:2607.09450*, 2026.

A Appendix

A.1 Limitations

Our study focuses on anchor-based rectification for zero-shot CLIP classification in order to isolate the role of anchor quality. AQS is intentionally an offline selection criterion and uses paired clean/adversarial examples only during anchor analysis; the deployed ARC defense does not require clean counterparts or model updates. The theoretical results characterize the geometry of anchor-guided correction in feature space, while the complete nonlinear pixel-space procedure is validated empirically.

A.2 Related Work

Adversarial fine-tuning and purification. TeCoA (Mao et al., 2022) first studied zero-shot adversarial robustness of CLIP, aligning adversarial image features with text embeddings through a contrastive min-max objective. Follow-up work regularizes toward the pre-trained model to preserve generalization (Wang et al., 2024), removes the need for labels (Schlarmann et al., 2024), or refines the alignment objective (Dong et al., 2025b;a; Cui et al., 2026; Khodabandeh et al., 2025; Hossain & Imteaj, 2025). Other approaches tune text prompts adversarially instead of the encoder (Li et al., 2024; Zhang et al., 2024). All of these require adversarial training for each backbone. Purification leaves the classifier untouched and removes the perturbation before classification, either with a diffusion model in pixel space (Nie et al., 2022) or by iterative optimization in CLIP's latent space (Zhang et al., 2025).

Multi-view-based defenses. Transformation ensembles average predictions over augmented copies of the input (Pérez et al., 2021; Reyes-Amezcua et al., 2024). Test-time prompt tuning methods optimize prompts over augmented views of each test sample (Sheng et al., 2025; Wang et al., 2025; Kim & Um, 2026b; Hai et al., 2026; Zhu et al., 2026). COLA (Zhu et al., 2025) projects adversarial embeddings onto the subspace spanned by class text features and aligns image and text distributions over views with optimal transport. Our AQS geometry is computed in a similar text-induced subspace. TTP (Li et al., 2026) detects adversarial inputs from the feature shift under spatial padding and adapts detected inputs with trainable padding and a view ensemble. MAC (Kim & Um, 2026a) combines multi-view guidance with adaptive counterattacks.

Counterattack-based defenses. Early counterattack methods perturb the input to increase the classifier's confidence but assume an adversarially trained model (Wu et al., 2021; Alfarra et al., 2022). TTC (Xing et al., 2025) observed that adversarial images are "falsely stable" under CLIP's own encoder and counterattacks them at inference without training. DOC (Jiang et al., 2026) diversifies the counterattack through orthogonal exploration, and FPT-Noise (Deng et al., 2026) adapts the counterattack strength to the scene.

Rectification-based defenses. AOM (Tong et al., 2025) treats Gaussian-noised adversarial inputs as anchors and follows a path in embedding space toward them, adapting its step size to the anchor distance. D-CLIP (Brahma & Biswas, 2026) builds on the same feature-correction operator using noise-induced feature dynamics. ATAC (Su & Balogh, 2026) corrects the embedding along a recovery direction inferred from augmentation-induced drift. CSR (Nie et al., 2026) gates inputs by spectral consistency and optimizes an input perturbation that pulls features toward a low-pass filtered anchor and away from the adversarial features. Each method fixes its anchor a priori. Our work instead measures anchor quality directly and selects the anchor before the defense is built.

Input transformations as defenses. Before CLIP, many input transformations were proposed to remove adversarial perturbations: bit-depth reduction and median filtering (Xu et al., 2018), cropping, rescaling, and total variation minimization (Guo et al., 2018; Rudin et al., 1992), random resizing and padding (Xie et al., 2018), and JPEG compression (Dziugaite et al., 2016; Das et al., 2017; Shin et al., 2017; Zhang & Gao, 2021). Gaussian noise underlies randomized smoothing (Cohen et al., 2019). Used as the classifier's input, these transformations are largely bypassed by adaptive attacks that approximate their gradients (Athalye et al., 2018). We revisit them in a different role, as candidate anchors that guide the correction.

A.3 Working space for removal and collateral

For a dataset with C classes, let $W \in \mathbb{R}^{C\times d}$ stack the ℓ_2-normalized text embeddings w_c of the class prompts, one prompt per class using the dataset's default template. Let $\mu_W = \frac{1}{C}\sum_c w_c$ and $W_c = W - \mathbf{1}\mu_W^\top$. The rows of W_c sum to zero, so its rank is at most $C-1$. We take the thin singular value decomposition $W_c = USV^\top$ and retain its $k = C-1$ leading components, which gives $V_k \in \mathbb{R}^{d\times k}$ and $S_k = \mathrm{diag}(s_1, \ldots, s_k)$. Each ℓ_2-normalized image-side feature $z \in \{f_c, f_a, a\}$ is mapped to

$$\phi(z) = \frac{S_k^{-1}V_k^\top(z-\mu_W)}{\left\|S_k^{-1}V_k^\top(z-\mu_W)\right\|_2}, \tag{13}$$

and removal and collateral are computed from $\bar{f}_c = \phi(f_c)$, $\bar{f}_a = \phi(f_a)$ and $\bar{a} = \phi(a)$.

Related projections have been used for CLIP. Zhu et al. (2025) project adversarial image features onto the span of the leading singular vectors of an uncentered matrix of LLM-generated class descriptions and use the projection itself as a defense. Betser et al. (2025) whiten the full CLIP space using the covariance of image and caption embeddings. Our map differs from both in centering the class prototypes and serves only as a space in which to measure anchor quality. The centering is motivated by the fact that it removes the component shared by all class prompts. That component adds the same amount to every logit, and CLIP embeddings occupy a narrow cone (Liang et al., 2022) and lie on ellipsoid shells that are not centered at the origin (Levi & Gilboa, 2025).

A.4 Additional Theoretical Insights

Proof of Theorem 1. Assume $u \neq 0$. Since $\bar{f}_a - \bar{f}_c = u$ and $\bar{a} - \bar{f}_c = v = ru + w$,

$$\bar{f}_\beta - \bar{f}_c = (1-\beta)u + \beta v = \big[1 - \beta(1-r)\big]u + \beta w.$$

Because $w \perp u$ and $\|w\|_2 = c\|u\|_2$, dividing the squared norm by $\|u\|_2^2$ gives equation 3. Let $d = 1 - r$. The ratio is below one iff $-2\beta d + \beta^2(d^2 + c^2) < 0$. For $\beta > 0$, this is equivalent to $\beta < 2d/(d^2 + c^2)$, and the interval is nonempty iff $d > 0$, that is, $r < 1$. □

Proof of Corollary 1. Expanding equation 4 gives $E(\beta) = 1 - 2\beta d + \beta^2(d^2 + c^2)$, a quadratic with second derivative $2(d^2 + c^2) > 0$ since $d > 0$. Setting $E'(\beta) = -2d + 2\beta(d^2 + c^2) = 0$ gives the unique stationary point $\beta^\star = d/(d^2 + c^2)$, which is positive. By strict convexity, E decreases on

$(0, \beta^\star]$ and increases on $[\beta^\star, \infty)$, so on $(0, \beta_{\max}]$ its unique minimizer is $\hat{\beta} = \min\{\beta^\star, \beta_{\max}\}$. If $\beta^\star \le \beta_{\max}$, substituting gives $E(\beta^\star) = 1 - d^2/(d^2 + c^2) = c^2/(d^2 + c^2)$; otherwise the minimum is $E(\beta_{\max}) = (1 - \beta_{\max} d)^2 + \beta_{\max}^2 c^2$. □

A.5 Additional Experimental Insights

Figure 2 in the main paper reports the AQS of each anchor family representative averaged over cells. Averages conceal how an anchor behaves on individual datasets and attack budgets, and our selection rule ranks anchors within each cell before averaging. Figure 5 therefore reports the score of every representative in every cell, together with the mean rank on which the inter-family selection is based.

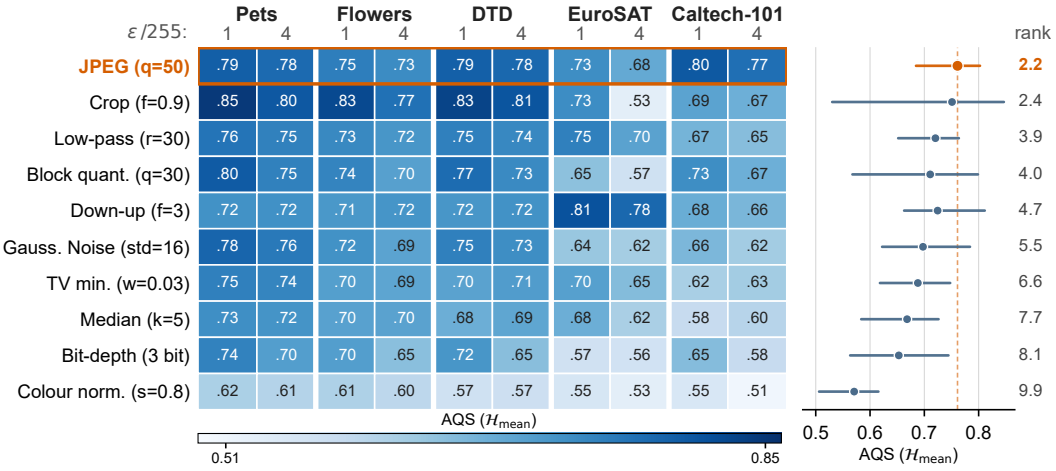

Figure 5: **Cell-wise AQS of the family representatives.** Left, AQS of each representative in each dataset–budget cell. Right, mean AQS across cells (dot), its range across cells (bar), and mean rank across cells (right column), with the dashed line marking the mean of JPEG50. Rows are ordered by mean rank. JPEG50 ranks first although it is the best anchor in only two cells, because it is never ranked below third in any cell.

No anchor is best everywhere. Crop wins all six cells on Pets, Flowers and DTD but falls to tied-last on EuroSAT at $\epsilon = 4/255$, and down-up wins both EuroSAT cells while scoring close to the pool median elsewhere. JPEG50 wins the two Caltech101 cells, yet it places second or third in every other cell, which gives it the best mean rank (2.2) with a range less than half as wide as Crop's. Crop's mean AQS is close to JPEG50's, but its range is the widest in the pool.

The comparison also shows why we select by mean rank rather than by mean score. Down-up has a higher mean AQS than low-pass and block quantization but a worse mean rank, because its average is carried by two strong EuroSAT cells and it trails both in most of the others. Mean rank weights every cell equally and rewards anchors that are consistently near the top, which is the property a general-purpose defense needs. At the other end of the ordering, color normalization is last or near-last in every cell, so its position does not depend on the aggregation rule. Finally, scores at $\epsilon = 4/255$ are at or below those at $\epsilon = 1/255$ for nearly every anchor and dataset, and EuroSAT is the dataset on which the ordering changes most, which is why we evaluate anchors over a grid of cells rather than on any single one.

A.6 Intra-family Sweeps

A single operating point does not represent an anchor mechanism, since every transformation in our study exposes a strength parameter and its quality as an anchor varies with that parameter. Before comparing families, we therefore sweep each family's strength parameter, score every member with AQS over a common grid of cells, and select one representative per family. This section describes the families, the protocol, and the resulting trajectories.

Table 7: Anchor families, the strength parameter swept for each, and the member selected as the family's representative for the inter-family comparison. For every family except down-up, Gaussian noise, TV minimization, median and color normalization, a smaller parameter value corresponds to a stronger transformation.

Family	Parameter	Swept values	Selected
JPEG	quality q	$10, 30, 50, 60, 75, 90$	50
Block quant.	quality q	$10, 20, 30, 40, 50, 60, 75, 90$	30
Low-pass	cutoff radius r	$10, 15, 20, 30, 60, 80$	30
Bit-depth	bits b	$2, 3, 4, 5, 6, 7$	3
Median	kernel size k	$3, 5, 7, 9$	5
Down-up	factor f	$1.5, 2, 3, 4$	3
TV minimization	weight w	$0.03, 0.1, 0.3, 0.6$	0.03
Crop	fraction f	$0.65, 0.75, 0.85, 0.90, 0.95$	0.90
Color norm.	strength s	$0.2, 0.4, 0.6, 0.8, 1.0$	0.8
Gaussian noise	std. σ ($\times 1/255$)	$1, 2, 4, 8, 16, 32, 64$	16

A.6.1 Anchor families

We evaluate ten families, summarized in Table 7. They span the mechanisms proposed for input purification in the literature: compression (JPEG and blockwise DCT quantization), spectral filtering (a Gaussian low-pass filter in the Fourier domain), amplitude quantization (bit-depth reduction), local and spatial smoothing (median filtering, bicubic down- and up-sampling, and total-variation minimization), geometric resampling (random cropping), color normalization, and additive Gaussian noise. All operations act on 224×224 inputs in $[0, 1]$ and are applied to the whole image.

Operations. *JPEG* is a standard libjpeg compression round trip at quality q, including the YCbCr color transform and chroma subsampling (Dziugaite et al., 2016; Guo et al., 2018). *Block quantization* isolates JPEG's lossy stage by applying the 8×8 block DCT and quantizes each coefficient with the JPEG luminance table, scaled by quality with the standard libjpeg law, directly on the RGB channels, with no color transform, chroma subsampling or entropy coding. It therefore shares JPEG's quality parameter, and the gap between the two families at a matched q measures what JPEG's chroma path contributes. *Low-pass* multiplies the image spectrum by $\exp\big(-\|\omega\|^2/(2r^2)\big)$, where ω is the centered frequency and r the cutoff radius. *Bit-depth* rounds each channel to 2^b levels (Xu et al., 2018). *Median* replaces each pixel by the median of its $k \times k$ neighborhood (Xu et al., 2018). *Down-up* downsamples by a factor f with antialiased bicubic interpolation and upsamples back to the input resolution. *TV minimization* reconstructs the image from a random half of its pixels under an anisotropic total-variation prior with weight w, solved with 30 Adam steps at learning rate 0.05 (Guo et al., 2018). *Crop* takes a randomly placed crop covering a fraction f of each side and resizes it to the input resolution (Guo et al., 2018). *Color normalization* converts to YCbCr, equalizes the histograms of the two chroma channels while leaving luminance untouched, and blends the result with the input at strength s. *Gaussian noise* adds independent noise with standard deviation $\sigma/255$. Crop, TV minimization and Gaussian noise are stochastic; we evaluate each with a single draw per image.

A.6.2 Protocol

We use CLIP ViT-B/16 with the pretrained weights and each dataset's default prompt template. A cell is a pair of a dataset and an attack budget. We use five datasets, Oxford Pets, Flowers102, DTD, EuroSAT and Caltech101, and ℓ_∞ adversarial examples at $\epsilon \in \{1/255, 4/255\}$ generated as described in Section 4.1, giving ten cells. In each cell we score anchors on 300 development pairs, restricted to images CLIP classifies correctly without attack. Adversarial examples are generated once per cell and shared by all anchors, so every member of every family is scored on identical inputs.

AQS is computed as in Section 2.2. Removal and collateral are measured in the whitened span of the class text embeddings; semantic cost and the gate score are measured in the original feature space.

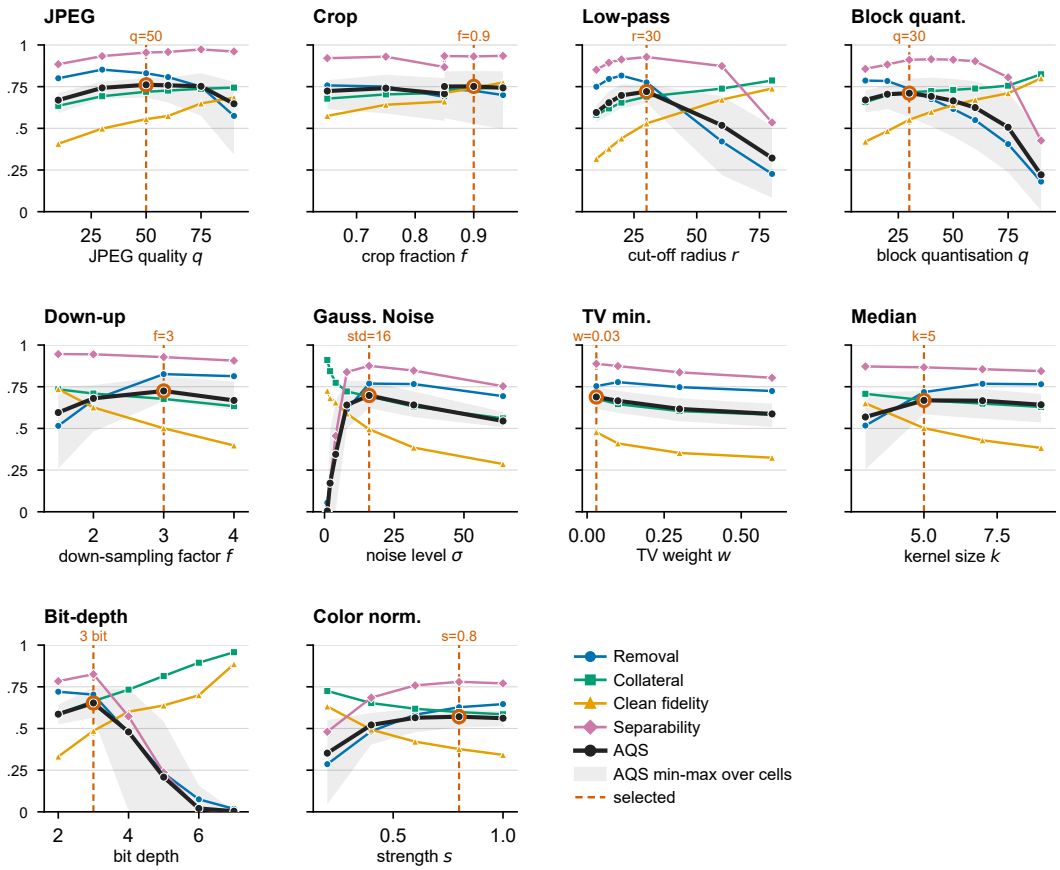

Figure 6: **Intra-family sweeps.** Each panel sweeps the strength parameter of one anchor family. Colored lines show the four AQS components as goodness scores in $[0, 1]$, and higher is better for every curve; the black line shows AQS; all five are averaged over the ten dataset–budget cells. The gray band shows the range of AQS across cells, and the dashed line marks the member selected by mean rank. In most families the removal and clean-fidelity curves move in opposite directions with strength, and AQS peaks between the two extremes, where neither component has collapsed.

We use weight 3 for removal and 1 for the other components, scales $s_c = 1$ and $s_\kappa = 0.1$, and floor $\epsilon_h = 10^{-3}$.

Selection. Within each cell, we rank the members of a family by AQS, and select the member with the best rank averaged over the ten cells. We select by mean rank rather than by mean score because the scale of the components varies across datasets, so a mean score is dominated by the cells in which the members' scores differ most, whereas a mean rank weights every cell equally. Ties are broken by the worst-cell score, which penalizes members that perform well on average but poorly on one dataset or budget.

A.6.3 Results

Figure 6 shows the trajectory of every family.

Anchor quality is an interior optimum. In most families, removal and clean fidelity move in opposite directions as strength varies, indicating that strengthening the transform removes more of the attack and damages more of the clean representation. AQS therefore rises and then falls, and its peak lies at an intermediate strength where neither component has collapsed. JPEG, block quantization, low-pass, and bit-depth show this shape most clearly.

The weak end realizes the degenerate anchor. As a transform approaches the identity, it inherits the perturbation. Bit-depth reduction at 6 and 7 bits is the clearest instance where removal and separability fall to nearly zero while collateral avoidance and clean fidelity approach their maxima, and AQS falls to nearly zero with them. Low-pass at large radius and block quantization at high quality follow the same path. These are real transformations reproducing the identity endpoint, and the collapse of AQS despite two near-perfect components is the harmonic mean's bound acting as intended. JPEG degrades more gracefully at high quality than block quantization does at the same q where its removal remains substantial at $q = 90$, whereas that of block quantization has largely vanished. Since the two share the same quantization law, the difference reflects JPEG's chroma path, which continues to discard chroma detail when the luminance steps are small.

Noise does not purify at small magnitude. At small σ, AQS is near zero because the anchor inherits the attack, meaning that adding noise does not remove a perturbation that was optimized to survive small changes. AQS rises only once the noise is large enough to overwrite the perturbation, by which point clean fidelity has already declined, and it falls again as the noise grows.

Some families are insensitive to strength. Crop, median and color normalization at $s \geq 0.6$ produce nearly flat AQS trajectories, so their ranking in the inter-family comparison reflects the mechanism rather than the choice of operating point. For TV minimization, AQS decreases monotonically with the weight, and the selected member lies at the smallest value we swept. Its representative may therefore understate the family's best achievable score; we did not extend the sweep, because each TV anchor requires an inner optimization per image and is substantially more expensive than the other families.

The spread across cells, shown by the gray bands, widens toward the extremes of most sweeps. The selected members lie in the regions where the bands are comparatively narrow. An operating point that is strong on some cells and weak on others ranks below one that is consistently near the top. The ten selected members form the pool compared in Section 2.3.

A.7 Ablation: JPEG Quality

ARC fixes the JPEG quality at $q = 50$, the member that AQS selects within the JPEG family (Appendix A.6). We test this choice against the defense itself. We run the complete ARC pipeline with the anchor set to each quality $q \in \{10, 30, 50, 60, 75, 90\}$, holding every other hyperparameter fixed, on the same ten dataset–attack cells used for anchor selection. Figure 7 reports clean accuracy, robust accuracy and their harmonic mean, each averaged over cells. We rank the qualities by the harmonic mean because, like AQS, it cannot be raised by trading one accuracy for the other.

The defense performs best at $q = 50$. At $q = 10$ both accuracies collapse to nearly the same level. The anchor removes much of the perturbation but discards so much content that clean images disagree with their own anchors, the gate flags them as suspicious, and rectification pulls benign and attacked inputs alike toward a degraded target. At $q = 90$ clean accuracy is near its maximum while robust accuracy falls to its lowest value, because the anchor now retains enough of the perturbation that rectifying toward it no longer undoes the attack. Between these extremes, clean accuracy saturates from $q = 50$ onward while robust accuracy peaks at $q = 50$ and declines steadily beyond it. The ranking by harmonic mean is $q = 50$, 60, 75, 30, 90 and 10.

AQS predicts the ranking. Figure 8 shows the same sweep through AQS. The two rankings agree. AQS orders the first four members as the defense does, with $q = 50$ first at a mean rank of 1.7, and places $q = 10$ and $q = 90$ jointly last, the two members the defense also ranks last. Counting the tie as shared, the Spearman correlation between the two rankings is 0.99. The criterion therefore identifies the operating point the defense would select, without constructing the defense for any member of the sweep.

The components explain the defense curves. Each accuracy curve in Figure 7 has a counterpart among the components. Clean fidelity increases monotonically with quality, as clean accuracy does, and is lowest at $q = 10$, where the defense loses clean accuracy. Removal peaks at $q = 30$ and declines past $q = 60$ to its minimum at $q = 90$, where robust accuracy drops most sharply. Robust accuracy nevertheless peaks at $q = 50$ rather than at the removal peak, and the difference is collat-

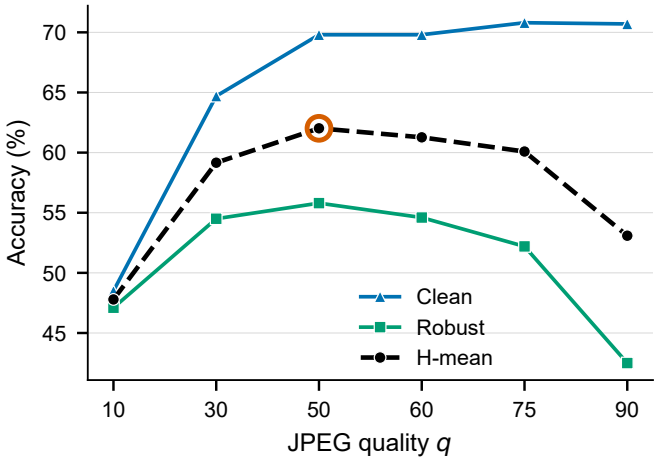

Figure 7: **Defense accuracy as a function of JPEG quality.** Clean accuracy, robust accuracy and their harmonic mean for ARC with the anchor at each JPEG quality q, averaged over the ten dataset–budget cells, with all other hyperparameters fixed. The harmonic mean peaks at $q = 50$ (circled), the quality ARC uses. Strong compression ($q = 10$) lowers clean and robust accuracy alike, whereas weak compression ($q = 90$) preserves clean accuracy and loses robustness.

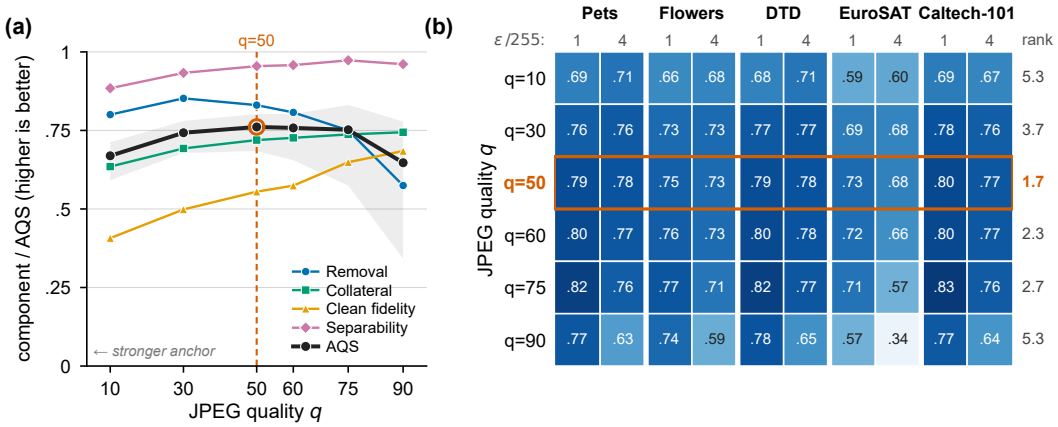

Figure 8: **AQS analysis of the JPEG family. (a)** The four AQS components and the composite as functions of JPEG quality, averaged over the ten cells, with the gray band giving the range of AQS across cells. All curves are goodness scores for which higher is better. **(b)** AQS of each quality in each cell, with the mean rank across cells in the rightmost column. AQS selects $q = 50$ (outlined), and its ranking of the qualities agrees with the ranking by defended accuracy in Figure 7 on every member it distinguishes.

eral. Collateral avoidance is lower at $q = 30$ than at $q = 50$, and by Corollary 1 the correction a rectifier can reach is limited by the collateral as well as by the removal. Neither component alone locates the defense's optimum, whereas their harmonic mean does. Separability remains high at every quality, so for JPEG the gate is not the limiting factor and the relevant trade-off is between removal and preservation.

The per-cell scores explain the selection. The heatmap in Figure 8(b) shows why $q = 50$ wins by mean rank. It is the best or tied-best member in six of the ten cells and within 0.03 of the best in the remaining four. Those four are the $\epsilon = 1/255$ cells of Pets, Flowers, DTD and Caltech101, where the milder $q = 75$ scores highest. The best quality therefore shifts toward weaker compression as the attack weakens, since a small perturbation requires less purification. At $\epsilon = 4/255$ this advantage disappears, and $q = 90$ loses substantially in every cell, falling to 0.34 on EuroSAT, which accounts

for the widening band at high quality in panel (a). No single quality is best everywhere. $q = 50$ is the most consistent, and together with $q = 30$ it attains the highest worst-cell score of 0.68.

A.8 DATASETS

We evaluate on ten zero-shot classification benchmarks, using the official test split of each dataset without using any training data. For ImageNet, the validation set serves as the test set. Table 8 summarizes each dataset.

Table 8: Statistics of the ten evaluation datasets.

Dataset	Description	# Classes	# Test
Caltech101 (Fei-Fei et al., 2004)	Generic objects	101	8,677
OxfordPets (Parkhi et al., 2012)	Fine-grained cat and dog breeds	37	3,669
StanfordCars (Krause et al., 2013)	Fine-grained car models	196	8,041
Flowers102 (Nilsback & Zisserman, 2008)	Fine-grained flower species	102	6,149
FGVC-Aircraft (Maji et al., 2013)	Fine-grained aircraft variants	100	3,333
DTD (Cimpoi et al., 2014)	Texture patterns	47	1,880
EuroSAT (Helber et al., 2019)	Satellite land-use imagery	10	27,000
UCF101 (Soomro et al., 2012)	Human actions (video frames)	101	3,783
Food101 (Bossard et al., 2014)	Fine-grained food dishes	101	25,250
ImageNet (Deng et al., 2009)	Generic objects and scenes	1,000	50,000

A.9 ATTACK SETTINGS

All attacks operate on image tensors in the $[0, 1]$ range, and the perturbation budget is expressed in this domain as $\epsilon/255$. All PGD-based attacks, including the adaptive attacks, use a random initialization within the ϵ-ball and a single run without restarts. Unless stated otherwise, attacks are untargeted and maximize the cross-entropy between the zero-shot logits and the ground-truth label. The configurations are as follows.

- **PGD attack.** We use an ℓ_∞-bounded PGD attack (Madry et al., 2017) with 10 iterations and step size $\epsilon/4$. The main results (Tables 1 and 2) use $\epsilon = 1/255$, and the stronger setting in Table 3 uses $\epsilon = 4/255$. Adversarial examples are generated on the base CLIP model and then passed to each defense.
- **AutoAttack.** We use the standard ℓ_∞ AutoAttack (Croce & Hein, 2020) with $\epsilon = 1/255$, keeping all internal settings at their library defaults. As with PGD, adversarial examples are generated on the base CLIP model and transferred to each defense.
- **Adaptive attack.** To evaluate robustness against a defense-aware adversary, we construct a separate adaptive attack for each defense in Table 3. Each attack runs the complete defense pipeline inside the attack objective. It combines EOT (Athalye et al., 2018) with 10 samples per iteration, averaging the gradient over the random transformations of the defense, and BPDA (Athalye et al., 2018) as a straight-through estimator for its non-differentiable components. The perturbation is updated by ℓ_∞-bounded PGD with $\epsilon = 1/255$, step size $\epsilon/4$, and 10 iterations.

ARC-adaptive attack. At each iteration, the attacker forms the perturbed image and runs the full ARC pipeline on it. This pipeline computes the JPEG50 anchor, evaluates the soft anchor-consistency gate, performs the gate-scaled rectification (Algorithm 1), and classifies the rectified image. The attack maximizes the cross-entropy between the resulting logits and the ground-truth label. The forward pass is exact. In the backward pass, the non-differentiable operations are replaced by the identity: JPEG quantization, the sign update and greedy selection inside the rectification loop, and the gate threshold test. Gradients flow through all remaining operations, including the CLIP encoder and the sigmoid gate, so the attacker accounts for both the anchor and the rectification it induces. ARC is deterministic, so all EOT samples yield the same gradient. We apply the same procedure to CSR, SS-TPT, and MAC, running each method's own pipeline in the forward pass and applying BPDA to its non-differentiable operations. From the multi-view category we evaluate SS-

TPT instead of A-TPT, because A-TPT's per-sample prompt tuning over 63 views makes adaptive evaluation prohibitively expensive.

A.10 ALGORITHM

We provide the detailed pseudocode for ARC in Algorithm 1.

Algorithm 1 Anchor-Guided Rectification using Compression (ARC)

Input: Input image $x \in [0,1]^{3\times H\times W}$, normalized CLIP image encoder f, anchor transform $T = \mathrm{JPEG}_{50}$, consistency threshold τ, temperature τ_{temp}, gate threshold g_τ, rectification budget ϵ_η, step size α_η, steps N, repulsion weight λ
Output: Rectified image $x^\star$
Initialize: $\eta \leftarrow \mathbf{0}$, $\mathcal{L}^\star \leftarrow -\infty$, $x^\star \leftarrow x$
// Anchor and soft anchor-consistency gate
$a \leftarrow f(T(x)), \quad f_o \leftarrow f(x)$
$s \leftarrow \cos(f_o, a)$
$g \leftarrow \sigma\big((\tau - s)/\tau_{\mathrm{temp}}\big)$
if $g \le g_\tau$ **then**
  return x ▷ consistent with anchor: treated as clean
end if
$\epsilon_g \leftarrow g \cdot \epsilon_\eta$ ▷ gate-scaled budget
// Anchor-guided rectification
for $t = 1$ **to** N **do**
  $\tilde{x} \leftarrow \mathrm{clip}_{[0,1]}(x + \eta)$
  $\mathcal{L} \leftarrow \cos\big(f(\tilde{x}), a\big) - \lambda \cos\big(f(\tilde{x}), f_o\big)$
  if $\mathcal{L} > \mathcal{L}^\star$ **then** ▷ greedy selection
    $\mathcal{L}^\star \leftarrow \mathcal{L}, \quad x^\star \leftarrow \tilde{x}$
  end if
  $\eta \leftarrow \eta + \alpha_\eta \cdot \mathrm{sign}(\nabla_\eta \mathcal{L})$
  $\eta \leftarrow \mathrm{clip}_{[-\epsilon_g, \epsilon_g]}(\eta)$
  $\eta \leftarrow \mathrm{clip}_{[0,1]}(x + \eta) - x$
end for
return $x^\star$

A.11 ABLATION: TARGETED DEGENERATE ANCHORS

The ablation in Section 4.3 uses synthetic *degenerate anchors* to test whether each AQS component contributes non-redundant information. Each degenerate is constructed to fail one target component while matching, or slightly surpassing, the JPEG50 reference on the remaining components. These constructions are diagnostic controls rather than deployable transformations.

Let $\bar{f}_c$ and $\bar{f}_a$ denote the normalized clean and adversarial features, and let

$$u = \bar{f}_a - \bar{f}_c$$

be the attack displacement. For an attacked input, the anchor error relative to the clean feature is $v = A(x_{\mathrm{adv}}) - \bar{f}_c$. Recall that removal is determined by the attack-aligned residual

$$r = \frac{\langle v, u\rangle}{\|u\|_2^2},$$

while collateral distortion measures the component of v orthogonal to u. For readability, let

$$\mathcal{N}(z) = \frac{z}{\|z\|_2}$$

denote ℓ_2 normalization.

A.11.1 DEGENERATE $\mathcal{D}_{rem}$: INSUFFICIENT REMOVAL

$\mathcal{D}_{rem}$ isolates the case in which the anchor moves in the correct direction but not far enough. For attacked inputs, it is constructed by interpolating between the clean and adversarial features,

$$A_{rem}(x_{\mathrm{adv}}) = \bar{f}_c + r_{rem}u, \qquad 0 < r_{rem} < 1,$$

with r_{rem} chosen so that removal is worse than the reference while the other components remain favorable. Because the anchor error is exactly collinear with the attack displacement,

$$v = r_{rem}u,$$

its collateral distortion is zero. The construction therefore fails only in how much of the original attack remains.

The clean-side anchor is kept very close to $\bar{f}_c$, giving very high clean fidelity. Since clean and attacked consistency scores remain perfectly ordered in this construction, separability is also maximal. Thus $\mathcal{D}_{rem}$ demonstrates that zero collateral distortion, excellent clean preservation, and perfect gate ordering do not imply sufficient purification.

A.11.2 DEGENERATE $\mathcal{D}_{col}$: OFF-AXIS ERROR

$\mathcal{D}_{col}$ targets collateral distortion. Unlike $\mathcal{D}_{rem}$, it retains an explicit attack-aligned component so that removal can be matched to the reference while a separate perpendicular error is increased. We construct the pre-normalized attacked anchor as

$$z = \bar{f}_c + \alpha u + \lambda\|u\|_2 n, \qquad A_{col}(x_{\mathrm{adv}}) = \mathcal{N}(z),$$

where $\|n\|_2 = 1$, $n \perp u$, and $n \perp \bar{f}_c$. The term αu preserves a controlled residual of the original attack, while $\lambda\|u\|_2 n$ injects the targeted off-axis error. Requiring $n \perp u$ prevents this error from being counted as residual attack displacement; requiring $n \perp \bar{f}_c$ makes it tangential to the clean feature rather than a radial rescaling.

The normalization couples the two coefficients, so they must be chosen jointly. Using

$$\langle \bar{f}_c, u\rangle = \langle \bar{f}_c, \bar{f}_a - \bar{f}_c\rangle = -\frac{\|u\|_2^2}{2},$$

the normalization factor is

$$\begin{aligned}\|z\|_2^2 &= \|\bar{f}_c\|_2^2 + \alpha^2\|u\|_2^2 + \lambda^2\|u\|_2^2 + 2\alpha\langle \bar{f}_c, u\rangle \\ &= 1 + \|u\|_2^2\left(\alpha^2 - \alpha + \lambda^2\right).\end{aligned}$$

Hence, with

$$s = \sqrt{\|z\|_2^2} = \sqrt{1 + \|u\|_2^2(\alpha^2 - \alpha + \lambda^2)},$$

the post-normalization attack residual and collateral term depend on the same scale factor. In the construction used here,

$$r = \frac{\alpha - \frac{1}{2}}{s} + \frac{1}{2}, \qquad c = \frac{\lambda}{s}.$$

We therefore solve for α and λ so that removal matches the reference while collateral distortion is deliberately worse. Clean fidelity and separability are tuned to match or slightly exceed the reference.

A.11.3 DEGENERATE $\mathcal{D}_{cln}$: CLEAN-SIDE DAMAGE

$\mathcal{D}_{cln}$ isolates clean fidelity. Its attacked-side construction is made reference-like, so the attack-oriented components—removal and collateral distortion—are matched to the reference. The degenerate is therefore not defined by a different adversarial geometry; instead, it deliberately damages the anchor produced for clean inputs.

The clean-side distortion is increased only enough to make clean fidelity worse than the reference while retaining competitive separability. These attack-oriented components depend on $A(x_{\mathrm{adv}})$,

whereas clean fidelity depends on $A(x)$. Good adversarial-side geometry therefore places no guarantee on how benign inputs are treated. Under the soft gate used by ARC, even moderate clean-side damage can receive nonzero rectification weight and degrade an already-correct representation.

Because clean-anchor consistency also contributes to the gate score, clean fidelity and separability are partially coupled. The construction is consequently tuned within the regime where separability remains matched to, or slightly better than, the reference, so the targeted failure remains clean fidelity alone.

A.11.4 DEGENERATE $\mathcal{D}_{sep}$: ERRATIC GATE

$\mathcal{D}_{sep}$ targets separability by preserving good *average* attack geometry while making the gate unreliable on a subset of samples. Clean anchors are kept reference-like. For attacked inputs,

$$A_{sep}(x_{\mathrm{adv}}) = \mathcal{N}\big(\bar{f}_c + r_i u\big)\,, \qquad r_i = \begin{cases} r_{\mathrm{big}}, & \text{with probability } p, \\ r_{\mathrm{small}}, & \text{with probability } 1-p. \end{cases}$$

The large-residual branch leaves some attacked anchors close to the incoming adversarial representation, producing clean-like consistency scores and weak rectification. The remaining samples compensate with a smaller residual.

This heterogeneity must not improve or degrade mean removal relative to the reference. Let r_{mean} denote the target mean residual. Because removal is linear in r, preserving mean removal is equivalent to preserving $\mathbb{E}[r_i]$:

$$\mathbb{E}[r_i] = p\, r_{\mathrm{big}} + (1-p)\, r_{\mathrm{small}} = r_{\mathrm{mean}}.$$

Solving for the compensating branch gives

$$r_{\mathrm{small}} = \frac{r_{\mathrm{mean}} - p\, r_{\mathrm{big}}}{1-p}.$$

Thus the pathological subgroup can be made substantially less purified while the population-average removal remains matched to the reference. The remaining non-target components are likewise calibrated to match or slightly surpass the reference.

Separability nevertheless falls because it depends on *per-image ordering*, not only on mean geometry. Samples from the large-residual branch produce attacked consistency scores that overlap the clean distribution; the gate can therefore assign too little rectification where it is needed.

A.11.5 INTERPRETATION

The four constructions isolate complementary failure modes. $\mathcal{D}_{rem}$ shows that correct direction with zero collateral distortion can still leave too much of the attack. $\mathcal{D}_{col}$ shows that good removal does not prevent a large off-axis error. $\mathcal{D}_{cln}$ shows that strong attacked-side behavior says nothing about preserving benign inputs. $\mathcal{D}_{sep}$ shows that good mean geometry can hide a subgroup on which the gate fails.

For each target component X, the necessity test compares JPEG50 with $\mathcal{D}_X$ under the full AQS and under an ablated score that omits X. The degenerate is calibrated so that its non-target components match or slightly exceed the reference. If removing X causes the degenerate to outrank JPEG50 while full-defense evaluation still favors JPEG50, then X carries information that the remaining components cannot recover. The results in Table 5 exhibit this behavior for all four AQS components.

A.12 USE OF LLMS

We use Claude only for the purpose of assisting in writing in terms of improving the grammar, clarity, and tone. We do not use it for any technical content or idea generation. All concepts, diagrams, analyses, experimental design choices, and conclusions are entirely our own.