AFOC: ADAPTIVE FREQUENCY-ORTHOGONAL COUNTERACTION FOR ROBUST CLIP INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot CLIP inference is vulnerable to adversarial perturbations, yet many defenses require training, labels, or prompt optimization. We study the more constrained setting of frozen encoders, unchanged prompts, and no training data. We propose **Adaptive Frequency-Orthogonal Counteraction** (AFOC), a training-free defense that estimates input risk from local embedding instability and sensitivity to frequency attenuation. A unified risk score calibrates orthogonal exploration, trajectory momentum, Fourier-domain update shaping, and a soft final gate, allowing conservative corrections for stable inputs and stronger corrections for risky inputs. Our local analysis formalizes clean-drift control and characterizes the directional and spectral effects of the update; this analysis is mechanistic rather than a robustness certificate. An embedding-geometry diagnostic further shows that PGD can disrupt zero-shot alignment while preserving local visual neighborhoods, whereas AFOC recovers decisions by rearranging embeddings relative to fixed text anchors. Across CIFAR-10, CIFAR-100, and ImageNet, we evaluate both post-hoc defense against attacks on the base classifier and defense-aware EOT-PGD attacks. On CIFAR-100 with CLIP ViT-B/32, AFOC attains 30.18% accuracy under adaptive EOT-PGD and a descriptive average of 39.22% across the four reported attacks, while retaining 56.55% clean accuracy. On ImageNet, it attains 29.53% adaptive accuracy with a 0.11-percentage-point drop in clean accuracy. These results demonstrate a competitive fixed-prompt operating point and motivate broader end-to-end adaptive evaluation.

1 INTRODUCTION

Vision-language models have become a standard foundation for open-world visual recognition. CLIP maps images and natural-language prompts into a shared embedding space and performs zero-shot classification by comparing a visual embedding with textual class anchors (Radford et al., 2021). This inference paradigm removes task-specific classification heads and enables broad transfer across unseen categories. However, because CLIP predicts by ranking image-text similarities, a visually imperceptible perturbation can move the visual embedding away from the correct textual anchor and closer to an incorrect one, disrupting cross-modal alignment and causing zero-shot failure (Mao et al., 2023).

Existing training-based defenses typically improve CLIP robustness through adversarial fine-tuning (Schlarmann et al., 2024) or representation/alignment regularization (Li et al., 2024; Dong et al., 2025b;a), but these methods require updating large pre-trained models and may compromise their general-purpose representations. Prompt-based defenses reduce the number of trainable parameters by optimizing textual or adversarial prompts (Zhang et al., 2024; Wang et al., 2025), yet they still introduce additional optimization and may weaken the semantic transparency of natural-language prompts.

Recent work has investigated whether CLIP can be made more robust entirely at inference time while keeping both encoders and prompts fixed. TTC optimizes a bounded, input-specific counteraction by moving away from the current visual anchor (Xing et al., 2025). DOC adds orthogonal exploration and momentum and uses directional sensitivity to modulate the overall correction strength (Jiang et al., 2026). However, its trajectory controls remain fixed across inputs, and its updates are frequency-agnostic. This limitation exposes a more specific calibration problem: without knowing whether an

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.

input is clean or adversarial, a defense must determine not only *how much* to correct, but also *how broadly* to explore and *where in the spectrum* to allocate its limited budget.

These observations motivate a joint geometry–spectrum view of robust CLIP inference. Geometrically, counteraction must determine how broadly to explore the local embedding landscape: a single local direction may be insufficient for strongly attacked inputs, whereas excessive exploration may damage clean semantics. Spectrally, the defense must determine how to allocate its limited update budget. Prior work suggests that adversarial examples exploit non-robust features (Ilyas et al., 2019), while frequency-domain analyses show that robustness can depend strongly on spectral components (Yin et al., 2019; Guo et al., 2019). Recent CLIP-specific evidence further indicates that adversarial CLIP features can be sensitive to progressive frequency attenuation (Nie et al., 2026). Robust test-time counteraction should therefore be both geometry-adaptive and spectrum-aware.

We therefore formulate robust CLIP inference as *input-conditioned geometry–spectrum regulation* and propose **Adaptive Frequency-Orthogonal Counteraction** (AFOC). Two label-free cues measure local embedding instability and sensitivity to frequency attenuation. The resulting fused risk score controls exploration, momentum, frequency-domain shaping, and final perturbation strength. Low-risk inputs receive mild, stable corrections, whereas high-risk inputs receive broader exploration, lower inertia, stronger high-band shaping, and a larger gate. Both CLIP encoders and the natural-language prompts remain unchanged.

Our contributions are:

- We introduce a label-free, dual-cue risk estimator that combines local geometric instability with sensitivity to frequency attenuation.
- We use one continuous risk signal to adapt trajectory geometry, momentum, Fourier-domain allocation, and final correction strength without modifying either CLIP encoder or the prompts.
- We provide a local mechanism analysis, including an embedding-geometry diagnostic that distinguishes cross-modal misalignment from local visual-neighborhood structure. We also separate base-classifier/post-hoc diagnostics from defense-aware evaluation and report matched ablations across datasets, attacks, prompts, and backbones.

2 PRELIMINARIES

Let $x \in [0,1]^D$ denote an image, where D is the input dimension. CLIP consists of an image encoder f_θ and a text encoder g_ϕ. Given a label set $\mathcal{Y} = \{1, \dots, K\}$, each class i is converted into a prompt t_i, such as "a photo of a {class}." We denote the normalized image and text embeddings by

$$h(x) = \frac{f_\theta(x)}{\|f_\theta(x)\|_2}, \qquad e_i = \frac{g_\phi(t_i)}{\|g_\phi(t_i)\|_2}. \tag{1}$$

The zero-shot prediction is $\hat{y}(x) = \arg\max_{i \in \mathcal{Y}} h(x)^\top e_i$. For a correctly classified image with label y, we define the CLIP alignment margin as

$$m_y(x) = h(x)^\top e_y - \max_{j \neq y} h(x)^\top e_j. \tag{2}$$

Robust zero-shot inference requires this margin to remain positive under small input perturbations.

At inference time, the defender receives an input image x_{in}, which may be clean or adversarial. The encoders f_θ and g_ϕ are frozen, and the defender has no access to the ground-truth label, attack type, or a clean/adversarial indicator. The goal is to construct a counteraction perturbation δ_{ca} in the feasible set

$$\Delta(x_{\text{in}}) = \left\{ \delta \;\middle|\; \|\delta\|_\infty \le \epsilon_{\text{ca}},\ x_{\text{in}} + \delta \in [0,1]^D \right\}, \tag{3}$$

where ϵ_{ca} is the defense budget, and to classify the reconstructed input

$$x_{\text{def}} = \text{clip}_{[0,1]}\left(x_{\text{in}} + \delta_{\text{ca}}\right).$$

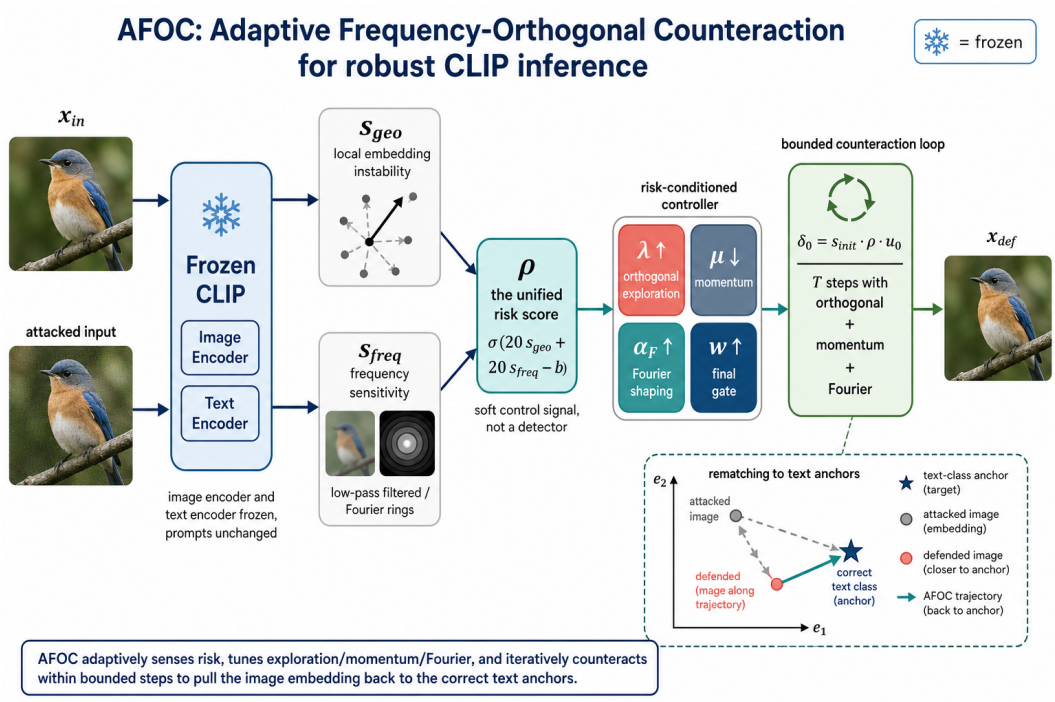


Figure 1: Overview of AFOC. A potentially adversarial input x_{in} is scored by frozen CLIP using two label-free cues: local embedding instability s_{geo} and sensitivity to frequency attenuation s_{freq}. These cues are combined into a continuous risk score $\rho = \sigma(20s_{\text{geo}} + 20s_{\text{freq}} - b)$, which serves as a soft control signal rather than a binary detector. A risk-conditioned controller then jointly sets orthogonal exploration λ, momentum μ, Fourier-domain shaping α_F, and the final gate w. A bounded T-step loop produces the defended image x_{def}; the inset illustrates the corresponding realignment of the visual embedding with a fixed text-class anchor, with the encoders and prompts held fixed.

3 METHOD

3.1 OVERVIEW

AFOC is a label-free inference-time defense for frozen CLIP models. It does not update CLIP encoders, prompts, or any auxiliary defense network. Instead, it performs input-specific counteraction updates to the perturbation δ for each test input. As illustrated in Fig. 1, the procedure comprises three coupled stages. First, the frozen image encoder extracts two complementary, label-free cues—local embedding instability s_{geo} and sensitivity to frequency attenuation s_{freq}—and combines them into a continuous risk score $\rho(x_{\text{in}})$. Second, a risk-conditioned controller maps ρ to the trajectory geometry (λ, μ), Fourier-domain update shaping (α_F), and a soft final gate w. Third, a bounded T-step loop initializes $\delta_0 = s_{\text{init}}\rho\, u_0$ and performs iterative counteraction within the ℓ_∞ budget, producing a defended image x_{def} whose visual embedding can be realigned with a fixed text-class anchor associated with the recovered prediction. Consequently, low-risk inputs receive conservative corrections, whereas high-risk inputs receive broader exploration, lower trajectory inertia, stronger high-band allocation, and a larger gate. For brevity, the main text presents the core mechanism and update equations; the complete inference procedure is provided in Appendix A.

3.2 LABEL-FREE COUNTERACTION OBJECTIVE

Since the ground-truth label is unavailable at test time, the counteraction objective cannot rely on supervised classification loss. We therefore define a label-free surrogate using the frozen CLIP image encoder:

$$\mathcal{J}_{\text{ca}}(\delta; x_{\text{in}}) = 1 - h(x_{\text{in}} + \delta)^\top h(x_{\text{in}}), \tag{4}$$

where $h(\cdot)$ denotes the normalized visual embedding. Maximizing $\mathcal{J}_{\text{ca}}$ moves the updated input away from the current embedding anchor, which may reflect an adversarially corrupted state. However,

this surrogate alone does not distinguish clean and adversarial inputs. Therefore, its update strength, trajectory, and spectral allocation must be calibrated at inference time.

3.3 Input Risk Estimation

The risk score should be label-free, attack-agnostic, and computable using the frozen CLIP image encoder. We estimate this score from two complementary signals.

Geometric risk. We first measure local embedding instability. Given weak probes $\eta_m \sim \mathcal{U}(-\epsilon_p, \epsilon_p)$ for $m = 1, \ldots, M$, we define

$$s_{\mathrm{geo}}(x_{\mathrm{in}}) = 1 - \frac{1}{M} \sum_{m=1}^{M} h(x_{\mathrm{in}})^\top h\big(\mathrm{clip}_{[0,1]}(x_{\mathrm{in}} + \eta_m)\big).$$

A larger s_{geo} indicates that small input changes induce larger directional changes in the CLIP embedding, suggesting a more unstable local embedding neighborhood.

Spectral risk. We then measure the sensitivity of the representation to frequency attenuation. Let $\mathcal{G}_r$ denote a low-pass filtering operator with cutoff radius r. We define

$$s_{\mathrm{freq}}(x_{\mathrm{in}}) = 1 - h(x_{\mathrm{in}})^\top h(\mathcal{G}_r(x_{\mathrm{in}})).$$

A larger s_{freq} indicates a stronger dependence on mid-to-high-frequency components.

Unified risk score. We combine the two signals into a continuous risk score:

$$\rho(x_{\mathrm{in}}) = \sigma(\beta_g s_{\mathrm{geo}}(x_{\mathrm{in}}) + \beta_f s_{\mathrm{freq}}(x_{\mathrm{in}}) - b)\,, \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function, and β_g, β_f, b are fixed scaling hyperparameters. Unless stated otherwise, we set $\beta_g = \beta_f = 20$ and use the normalized low-pass cutoff ratio $r = 0.5$. The base risk bias is $b_0 = 3.0$. With backbone calibration enabled by default, $b = b_0 + \Delta_b$, where $\Delta_b = 0$ for ViT-B/32, 1.25 for RN50, and 1.75 for ViT-B/16; the corresponding biases are 3.0, 4.25, and 4.75. Thus, for the default ViT-B/32 backbone, $\rho(x) = \sigma(20 s_{\mathrm{geo}}(x) + 20 s_{\mathrm{freq}}(x) - 3)$. We use $\rho(x_{\mathrm{in}}) \in (0, 1)$ as a soft control signal, not as a hard clean/adversarial detector or a calibrated attack probability. The backbone-specific offsets alter only this operating point; all subsequent controller mappings remain unchanged. Appendix D.7 quantifies ranking quality, class-level and texture-related confounds (Figure 5), EOT overlap, and the held-out rule used to set b.

3.4 Risk-Conditioned Counteraction Controller

For readability, we write $\rho = \rho(x_{\mathrm{in}})$ in this subsection. As shown in the controller block of Fig. 1, this risk score modulates four quantities: the orthogonal exploration strength λ, the momentum coefficient μ, the frequency-domain shaping strength α_F, and the final gate w:

$$\begin{aligned} \lambda &= \lambda_{\min} + \rho(\lambda_{\max} - \lambda_{\min}), \\ \mu &= \mu_{\max} - \rho(\mu_{\max} - \mu_{\min}), \\ \alpha_F &= \alpha_{\min} + \rho(\alpha_{\max} - \alpha_{\min}), \\ w &= \sigma\big(\gamma(\rho - \tau_\rho)\big). \end{aligned} \tag{6}$$

Thus, a higher risk score increases orthogonal exploration and frequency-domain shaping, reduces trajectory inertia, and enlarges the final correction gate.

Risk-scaled warm initialization. Because h is ℓ_2-normalized, the self-deviation objective is stationary at the zero perturbation: $\nabla_\delta \mathcal{J}_{\mathrm{ca}}(0; x_{\mathrm{in}}) = -\frac{1}{2}\nabla_{x_{\mathrm{in}}} \|h(x_{\mathrm{in}})\|_2^2 = 0$. To avoid initializing at this stationary point while remaining conservative on low-risk inputs, AFOC uses the risk-scaled warm initialization

$$\begin{aligned} u_0 &\sim \mathcal{U}(-\epsilon_{\mathrm{ca}}, \epsilon_{\mathrm{ca}})^D, \\ \delta_0 &= \Pi_\Delta\big(s_{\mathrm{init}} \rho\, u_0\big), \qquad m_0 = 0, \end{aligned} \tag{7}$$

where s_{init} controls the warm-start magnitude and Π_Δ jointly enforces the ℓ_∞ counteraction budget and the valid pixel range. We use $s_{\mathrm{init}} = 0.5$ by default; setting $s_{\mathrm{init}} = 0$ recovers exact zero initialization.

At step t, the normalized primary direction is

$$q_t = \frac{\nabla_{\delta_t} \mathcal{J}_{\mathrm{ca}}(\delta_t; x_{\mathrm{in}})}{\|\nabla_{\delta_t} \mathcal{J}_{\mathrm{ca}}(\delta_t; x_{\mathrm{in}})\|_2 + \xi},$$

where $\xi > 0$ is a small numerical constant. This direction follows the local counteraction objective, but using it alone may lead to a narrow update trajectory. We therefore sample $r_t \sim \mathcal{N}(0, I)$ in the perturbation space and remove its component along q_t:

$$r_t^{\perp} = r_t - \frac{\langle r_t, q_t \rangle}{\|q_t\|_2^2 + \xi} q_t, \qquad \bar{r}_t^{\perp} = \frac{r_t^{\perp}}{\|r_t^{\perp}\|_2 + \xi}.$$

The geometry-adaptive direction and normalized search direction are

$$\begin{aligned} d_t &= q_t + \lambda \bar{r}_t^{\perp}, \\ \widehat{d}_t &= \frac{d_t}{\|d_t\|_1 + \xi}. \end{aligned} \tag{8}$$

We initialize the momentum buffer directly from the first nonzero search direction and apply the risk-conditioned exponential moving average thereafter:

$$m_{t+1} = \begin{cases} \widehat{d}_0, & t = 0, \\ \mu m_t + (1-\mu)\widehat{d}_t, & t \geq 1. \end{cases} \tag{9}$$

Here, λ controls the degree of orthogonal exploration, whereas μ controls trajectory inertia. Direct assignment at $t = 0$ avoids attenuating the first nonzero direction by the factor $1 - \mu$.

Next, AFOC reshapes the momentum direction in the Fourier domain. Let $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the two-dimensional discrete Fourier transform and its inverse, applied channel-wise. Let M_H be a binary high-frequency mask centered at the Fourier origin, constructed using the normalized high-pass cutoff ratio $r_H = 0.5$. We define

$$\widetilde{m}_{t+1} = \mathrm{Re}\left[\mathcal{F}^{-1}\left((1 + \alpha_F M_H) \odot \mathcal{F}(m_{t+1})\right)\right], \tag{10}$$

where $\mathrm{Re}[\cdot]$ extracts the real part, ensuring that the shaped update remains in the real image space. This operation is applied to the defensive update direction rather than to the input image. Therefore, it does not perform low-pass purification; instead, it emphasizes high-band update components when the estimated risk is high.

Finally, the perturbation and defended input are obtained by

$$\begin{aligned} \delta_{t+1} &= \Pi_{\Delta}\big(\delta_t + \alpha\ \mathrm{sign}(\widetilde{m}_{t+1})\big), \\ \delta_{\mathrm{ca}} &= w\delta_T, \\ x_{\mathrm{def}} &= \mathrm{clip}_{[0,1]}\left(x_{\mathrm{in}} + \delta_{\mathrm{ca}}\right), \end{aligned} \tag{11}$$

where α is the step size and Δ is the feasible perturbation set defined in Eq. equation 3. In this way, AFOC performs continuous, input-conditioned counteraction rather than applying the same fixed update rule to every input.

4 Local Mechanism Analysis

We analyze three effects induced by AFOC—risk-gated correction, orthogonal exploration, and frequency-selective update shaping—under local smoothness or linearization assumptions. These results characterize what each operator controls; they do not certify robustness. Complete proofs are deferred to Appendix B.

Proposition 1 (Local mechanism of risk-conditioned counteraction)**.** *Assume text embeddings are normalized and h is locally L_h-Lipschitz. Let $m_y(x)$ be the CLIP margin defined in Eq.* (2)*. Then:*

(i) ***Soft gating limits clean distortion:*** *if $\|\delta_T\|_\infty \leq \epsilon_{\mathrm{ca}}$, then $\|h(x_{\mathrm{def}}) - h(x)\|_2 \leq L_h w(x) \sqrt{D} \epsilon_{\mathrm{ca}}$; consequently, $w(x) < m_y(x)/(2L_h\sqrt{D}\epsilon_{\mathrm{ca}})$ preserves the clean prediction.*

Table 1: Zero-shot accuracy (%). PGD, CW, and AA attack the base classifier before post-hoc defense and are base-classifier/post-hoc diagnostics; EOT-PGD attacks the stochastic defense using EOT and BPDA. "Mixed Avg." is a descriptive mean across these different protocols and excludes Clean; it is not an end-to-end robustness metric. Bold marks the best method in each attack column.

Dataset	Method	Clean	PGD	CW	AA	EOT-PGD	Mixed Avg.
CIFAR-100	Vanilla CLIP	58.53	0.00	0.00	0.08	0.00	0.02
	Low-pass	55.44	0.06	0.05	5.47	0.01	1.40
	JPEG	**58.60**	0.00	0.00	7.05	0.00	1.76
	TTC	55.33	0.08	0.03	8.88	0.00	2.25
	DOC	55.25	1.48	0.27	8.89	0.49	2.78
	AFOC	56.55	**44.56**	**36.74**	**45.41**	**30.18**	**39.22**
ImageNet	Vanilla CLIP	**59.26**	0.00	0.00	0.12	0.00	0.03
	Low-pass	54.76	1.66	2.20	8.93	1.16	3.49
	JPEG	58.53	0.50	0.34	4.50	0.20	1.39
	TTC	49.37	3.40	3.78	7.27	0.02	3.62
	DOC	43.28	20.38	15.40	20.08	2.72	14.65
	AFOC	59.15	**23.60**	**31.41**	**41.18**	**29.53**	**31.43**

(ii) ***Orthogonal exploration covers missing directions:*** *for a primary direction q and a random unit $r^{\perp}$ orthogonal to q, define $\bar{d}_\lambda = (q + \lambda r^{\perp})/\sqrt{1+\lambda^2}$. Then for any unit v,*

$$\mathbb{E}[(v^\top \bar{d}_\lambda)^2] = \frac{(v^\top q)^2 + \frac{\lambda^2}{D-1}(1-(v^\top q)^2)}{1+\lambda^2},$$

so that missing directional components receive nonzero expected coverage.

(iii) ***Frequency shaping amplifies high-band aligned updates:*** *with $P_F = P_L + (1+\alpha_F)P_H$, for any unit embedding-space direction a,*

$$a^\top J P_F(m) = a^\top J P_L(m) + (1+\alpha_F) a^\top J P_H(m).$$

Thus, when $a^\top J P_H(m) > 0$, increasing α_F strengthens the high-band aligned component while leaving the low-band contribution unchanged.

Interpretation. The soft gate controls clean embedding drift, orthogonal exploration broadens directional coverage when the primary gradient is insufficient, and frequency-domain shaping conditionally amplifies high-band update components. These are operator-level motivations; the empirical evaluation (Sec. 5) establishes whether the resulting trajectory improves classification. The embedding-geometry diagnostic in Sec. 5.4 further separates cross-modal misalignment from local visual structure.

5 Experiments

We evaluate AFOC as a training-free defense for zero-shot classification using frozen CLIP encoders and unchanged natural-language prompts. Our evaluation addresses five questions: **(Q1)** whether AFOC can recover predictions from adversarial examples generated against the base classifier; **(Q2)** whether the improvement persists under a defense-aware adaptive attack; **(Q3)** how the individual controller components affect accuracy; **(Q4)** how the frozen embedding geometry changes under attack and counteraction; and **(Q5)** the sensitivity and computational cost of the inference-time procedure.

5.1 Experimental Setup

Datasets. We evaluate CIFAR-10, CIFAR-100, and ImageNet in the zero-shot setting. CIFAR-10 and CIFAR-100 are evaluated on their complete test sets. For ImageNet, clean accuracy is measured on the validation set, whereas adversarial accuracy is measured on a fixed class-balanced validation subset because iterative defenses and adaptive attacks are computationally intensive. The same ImageNet subset is used for every method, attack, and random seed.

Backbones and prompts. CLIP ViT-B/32 is the default backbone. RN50 and ViT-B/16 results are reported in Appendix D.2. All compared methods use the same frozen image encoder, text encoder, and image preprocessing for a given backbone. The default class prompt is "a photo of a {class}." Prompt-ensemble results are reported separately in Appendix D.3, because changing the text-anchor construction can alter both clean and adversarial behavior.

Threat model and evaluation protocols. We consider untargeted ℓ_∞ attacks. Given a labeled clean image x, the attacker constructs $x_{\mathrm{adv}} = x + \delta_{\mathrm{adv}}$ with $\|\delta_{\mathrm{adv}}\|_\infty \le \epsilon_{\mathrm{atk}}$ by optimizing the cross-entropy loss of the CLIP zero-shot classifier. Unless otherwise specified, $\epsilon_{\mathrm{atk}} = 8/255$ for CIFAR-10/100 and $4/255$ for ImageNet. We distinguish two evaluation protocols. PGD (Madry et al., 2018), CW (Carlini & Wagner, 2017), and AutoAttack (Croce & Hein, 2020) are generated against the frozen zero-shot classifier and then passed through each defense; these attacks are white-box with respect to the base classifier but do not backpropagate through the defense. We therefore treat them as base-classifier/post-hoc diagnostics, not as white-box attacks on the defended classifier. Adaptive EOT-PGD is defense-aware and optimizes a differentiable surrogate of the complete defended prediction pipeline, using Expectation over Transformation (EOT) (Athalye et al., 2018b) together with a Backward Pass Differentiable Approximation (BPDA) (Athalye et al., 2018a).

Attacks. PGD uses random initialization, 40 iterations, and step size $\epsilon_{\mathrm{atk}}/10$. CW uses a logit-margin objective and 100 iterations. AutoAttack uses its standard ensemble configuration. For adaptive EOT-PGD, the attacker solves

$$\max_{\|\delta\|_\infty \le \epsilon_{\mathrm{atk}}} \mathbb{E}_\omega \left[\ell_{\mathrm{ce}}\big(\hat{p}(\mathcal{D}_\omega(x+\delta)), y\big) \right],$$

where $\mathcal{D}_\omega$ denotes the stochastic defense and $\hat{p}(\cdot)$ denotes the defended zero-shot prediction. The main adaptive evaluation uses $K_{\mathrm{EOT}} = 4$ stochastic samples per gradient step. Projection, clipping, and sign-based inner updates use straight-through BPDA surrogates. Consequently, this attack is defense-aware but remains dependent on the quality of the surrogate gradient; it should not be interpreted as an exhaustive adaptive evaluation. Appendix D.5 increases K_{EOT} to 8 and 16 to assess sensitivity to gradient averaging.

Baselines. We report vanilla CLIP; low-pass filtering and JPEG compression as simple input-preprocessing baselines; and TTC and DOC as the most directly comparable inference-time counteraction methods. TTC, DOC, and AFOC keep both CLIP encoders fixed and construct an input-specific bounded counteraction. Our quantitative claims are restricted to methods with frozen encoders and unchanged prompts.

AFOC configuration. Unless otherwise specified, $\epsilon_{\mathrm{ca}} = \epsilon_{\mathrm{atk}}$, $T = 10$, $M = 4$, $\epsilon_p = \epsilon_{\mathrm{atk}}/4$, $\alpha = \epsilon_{\mathrm{ca}}/T$, and $s_{\mathrm{init}} = 0.5$. Risk weights, controller ranges, and spectral cutoffs are fixed across datasets and attacks; only the risk bias uses the fixed backbone calibration described in Sec. 3. The complete settings are listed in Appendix C.4. No attack-specific hyperparameter tuning is performed.

Metrics. Clean accuracy is measured on unperturbed inputs after applying the defense. Attack-specific accuracy is measured after defense under each protocol. Unless a table states otherwise, Mixed Avg. is the arithmetic mean of PGD, CW, AutoAttack, and adaptive EOT-PGD accuracy and excludes the Clean column. Because this mean combines post-hoc and defense-aware protocols, we use it only as a compact descriptive summary and base end-to-end robustness claims on EOT-PGD. Stochastic methods are run with three random seeds and the reported entries are seed averages.

5.2 Main Results

Table 1 reports clean and adversarial accuracy on CIFAR-100 and ImageNet; the full three-dataset table is in Appendix D.1.

AFOC outperforms TTC and DOC across all evaluated attack protocols. On CIFAR-100, its mixed-protocol average of 39.22% exceeds both baselines by substantial margins (29.69–44.48 percentage points across the individual attacks). On ImageNet, it achieves 31.43%, surpassing TTC by 27.81 percentage points and DOC by 16.78 percentage points. Under defense-aware adaptive EOT-PGD, AFOC attains 30.18% on CIFAR-100 and 29.53% on ImageNet, whereas TTC and DOC are near zero on CIFAR-100 and reach at most 2.72% on ImageNet. Additional evaluations, including Square Attack (Andriushchenko et al., 2020) (Appendix D.6), provide complementary evidence of the defense behavior.

Clean accuracy is largely preserved: AFOC drops by only 1.98 percentage points on CIFAR-100 and 0.11 percentage points on ImageNet relative to vanilla CLIP, while TTC and DOC lose 9.89 and 15.98 percentage points, respectively, on ImageNet. Among methods that keep both encoders and natural-language prompts frozen, AFOC combines substantially higher adversarial accuracy with a comparatively small clean-accuracy change in these evaluations.

5.3 Ablation Study

We ablate the four risk-conditioned controls on CIFAR-100 with CLIP ViT-B/32 (Table 2). Each variant changes one component and retains the remaining pipeline. The full-model row corresponds to the complete AFOC configuration used throughout this section. Because CW is not evaluated in this ablation, Mixed Avg. is computed over PGD, AutoAttack, and adaptive EOT-PGD.

Table 2: Matched ablations on CIFAR-100 (%). Mixed Avg. is the descriptive mean of PGD, AA, and EOT-PGD. Fixed momentum uses $\mu = 0.6$. The fixed controller uses $(\lambda, \mu, \alpha_F, w) = (0.25, 0.6, 0.25, 0.5)$, i.e., Eq. (6) evaluated at $\rho = 0.5$ (Appendix F.1).

Variant	Clean	PGD	AA	EOT	Mixed Avg.
Full AFOC	56.55	**44.56**	**45.41**	**30.18**	**40.05**
No soft gate	5.79	27.53	38.75	16.29	27.52
No orthogonal branch	**58.54**	0.15	0.40	19.83	6.79
Fixed momentum	57.30	27.91	37.67	7.11	24.23
No frequency shaping	57.32	28.29	38.61	16.75	27.88
Fixed controller	9.67	11.10	21.28	0.20	10.86

Input-conditioned calibration limits destructive updates. Removing the soft gate reduces the three-attack average from 40.05% to 27.52% and collapses clean accuracy from 56.55% to 5.79%, a 50.76-percentage-point loss. The fixed controller performs poorly on both objectives: relative to full AFOC, it loses 46.88 percentage points in clean accuracy and 29.19 percentage points in average robustness. Input-conditioned gating is therefore essential for avoiding destructive updates without sacrificing robustness.

Orthogonal exploration is the primary source of broad attack coverage. Setting $\lambda = 0$ changes clean accuracy by only 1.99 percentage points but reduces PGD, AutoAttack, and adaptive EOT-PGD accuracy by 44.41, 45.01, and 10.35 percentage points, respectively. The three-attack average falls by 33.26 percentage points. Orthogonal exploration is therefore the dominant contributor to broad attack coverage in this ablation, rather than a mechanism that benefits only one attack family.

Adaptive momentum is especially important under the adaptive attack. Replacing the risk-conditioned momentum with the fixed coefficient $\mu = 0.6$ lowers PGD, AutoAttack, and adaptive EOT-PGD accuracy by 16.65, 7.74, and 23.07 percentage points, respectively. The three-attack average decreases from 40.05% to 24.23%. Risk-conditioned momentum therefore contributes most strongly under the adaptive attack, while also improving performance in the post-hoc evaluations.

Frequency shaping is a complementary allocation mechanism. Removing Fourier-domain shaping lowers PGD, AutoAttack, and adaptive EOT-PGD accuracy by 16.27, 6.80, and 13.43 percentage points, respectively, and reduces the three-attack average by 12.17 percentage points, while recovering 0.77 percentage points in clean accuracy. Frequency shaping therefore increases measured robustness at a modest clean-accuracy cost. Together with the spectral diagnostics in Appendices E and F.4, this supports its interpretation as a complementary allocation mechanism within the joint controller rather than as input purification.

5.4 Embedding Geometry under Attack and Counteraction

The quantitative gains above leave open a mechanistic question: does post-hoc recovery restore the clean visual neighborhood geometry, or does it primarily rematch image embeddings to textual anchors under a different local organization? To probe this distinction, we construct a class-balanced CIFAR-10 subset (50 images per class), extract frozen CLIP ViT-B/32 image embeddings for four matched states of the *same* images—clean inputs, base-classifier PGD examples, AFOC without orthogonal exploration, and full AFOC—and report zero-shot accuracy together with cosine-silhouette

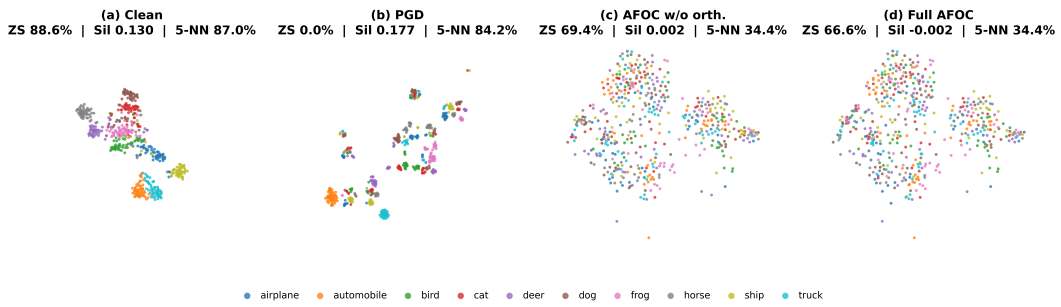


Figure 2: Frozen CLIP embedding geometry on a class-balanced CIFAR-10 subset (ViT-B/32). Subfigures (a)–(d) share one PCA+t-SNE map for qualitative comparison; zero-shot (ZS), cosine silhouette (Sil), and 5-NN are measured in the original embedding space. **(a) Clean:** class neighborhoods are compact and zero-shot alignment is strong. **(b) PGD:** zero-shot accuracy collapses while silhouette and 5-NN remain high, indicating cross-modal misalignment without destroying local visual neighborhood structure. **(c)–(d) AFOC variants:** a large fraction of zero-shot accuracy is recovered while silhouette and 5-NN drop sharply, consistent with rematching to text anchors rather than restoring the attacked cluster geometry. On this diagnostic subset the two AFOC variants induce similar redistributions, so the figure is a geometry diagnostic rather than a component ranking.

and 5-NN accuracy computed in the original embedding space. A shared PCA+t-SNE projection is used only for visualization; all numerical claims refer to the unprojected embeddings.

PGD disrupts image–text alignment while preserving local visual structure. Figure 2(a)–(b) shows that clean embeddings form coherent class neighborhoods (ZS 88.6%, Sil 0.130, 5-NN 87.0%). After PGD, zero-shot accuracy falls to 0.0%, yet silhouette *increases* to 0.177 and 5-NN remains 84.2%. Thus, the attack can sever ranking against frozen text anchors without erasing within-class neighborhood consistency in the visual embedding space. This separation cautions against treating local clustering alone as a proxy for zero-shot robustness.

AFOC recovers zero-shot decisions by rearranging embeddings, rather than by reconstructing the clean cluster geometry. Subfigures 2(c)–(d) show that both AFOC without orthogonal exploration and full AFOC restore a large fraction of zero-shot accuracy (69.4% and 66.6%) while collapsing silhouette to near zero and reducing 5-NN to 34.4%. The recovered predictions are therefore accompanied by a qualitatively different embedding organization: counteraction relocates attacked points relative to the fixed text anchors rather than reconstituting the compact clean/PGD neighborhoods. We treat the two-dimensional map as qualitative evidence of this rematching effect; the accompanying silhouette and 5-NN statistics provide the corresponding geometry summary in the original space. The near-parity of Figure 2(c) and (d) on this CIFAR-10 diagnostic therefore concerns local embedding organization rather than full-protocol robustness: the two variants can rematch text anchors similarly while still differing in attack recovery under the matched CIFAR-100 evaluation. **Additional analyses.** Risk-score distributions, supervised-margin trajectories, and update spectra that further support the geometry–spectrum controller are reported in Appendix E. Detailed analyses of the risk-scaled warm initialization, hyperparameter sensitivity, and inference overhead are provided in Appendices F.2 and G. These diagnostics complement the main quantitative comparison under the same evaluation protocol.

6 Conclusion

We introduced AFOC, a training-free defense for frozen, fixed-prompt CLIP inference. A dual-cue risk score jointly calibrates orthogonal exploration, momentum, frequency-domain shaping, and final correction strength. This joint geometry–spectrum perspective provides a basis for improving inference-time robustness. Extending the label-free risk principle to multimodal and generative foundation models remains an important avenue for future work.

REPRODUCIBILITY STATEMENT

The final artifact will release split manifests, prompt identifiers, attack budgets, generation and controller seeds, raw and controlled trajectory summaries, evaluation outputs, and scripts that regenerate every table and figure.

ETHICS STATEMENT

This paper studies inference-time adversarial robustness of frozen CLIP on public classification benchmarks (CIFAR-10/100 and ImageNet). The attacks and defenses are used only to evaluate zero-shot image classification; we do not target people, private data, or deployed systems.

AI USE STATEMENT

Generative AI tools assisted with language editing, code drafting, and consistency checks. The authors remain responsible for all claims, citations, experimental results, and released artifacts, and will verify every AI-assisted component before submission.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 2020.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 274–283, 2018a.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 284–293, 2018b.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pp. 39–57, 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 2206–2216, 2020.

Junhao Dong, Piotr Koniusz, Liaoyuan Feng, Yifei Zhang, Hao Zhu, Weiming Liu, Xinghua Qu, and Yew-Soon Ong. Robustifying zero-shot vision language models by subspaces alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21037–21047, 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 *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14061–14078, 2025b.

Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In *Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, pp. 1127–1137, 2019.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32, pp. 125–136, 2019.

Chengze Jiang, Minjing Dong, Xinli Shi, and Jie Gui. Diversifying counterattacks: Orthogonal exploration for robust clip inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026.

Xiao Li, Wei Zhang, Yining Liu, Zhanhao Hu, Bo Zhang, and Xiaolin Hu. Language-driven anchors for zero-shot adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24686–24695, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *International Conference on Learning Representations*, 2023.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust clip: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *International Conference on Machine Learning*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19910–19920, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15172–15182, 2025.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin D. Cubuk, and Justin Gilmer. A fourier perspective on model robustness in computer vision. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

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.

A Detailed Algorithm

Algorithm 1: AFOC: Adaptive Frequency-Orthogonal Counteraction

Input: f_θ, x_{in}, budgets $\epsilon_{\text{ca}}, \epsilon_p$, cutoffs r, r_H, number of probes M, warm-start scale s_{init}, step size α, number of steps T, and fixed risk and controller parameters.
Output: Defended image x_{def}.

```
// Risk estimation and controller
```
Estimate s_{geo} with M weak probes and s_{freq} by low-pass attenuation
Compute ρ by Eq. (5)
Compute $\lambda, \mu, \alpha_F, w$ by Eq. (6)
```
// Risk-scaled initialization
```
Sample $u_0 \sim \mathcal{U}(-\epsilon_{\text{ca}}, \epsilon_{\text{ca}})^D$ and set $\delta_0 = \Pi_\Delta(s_{\text{init}}\rho u_0)$, $m_0 = 0$
```
// Geometry--spectrum counteraction
```
for $t = 0$ **to** $T-1$ **do**
    Compute the normalized primary direction q_t
    Sample $r_t \sim \mathcal{N}(0, I)$ and compute $\bar{r}_t^{\perp}$
    Set $d_t = q_t + \lambda \bar{r}_t^{\perp}$ and $\widehat{d}_t = d_t/(\|d_t\|_1 + \xi)$
    if $t = 0$ **then**
        Set $m_1 = \widehat{d}_0$
    else
        Set $m_{t+1} = \mu m_t + (1-\mu)\widehat{d}_t$
    end
    Compute $\widetilde{m}_{t+1}$ by Fourier-domain shaping
    Update δ_{t+1} by Eq. (11)
end
Set $\delta_{\text{ca}} = w\delta_T$
Return $x_{\text{def}} = \text{clip}_{[0,1]}(x_{\text{in}} + \delta_{\text{ca}})$

B Proofs

This section provides the complete proofs for the local mechanism analysis in Sec. 4. The analysis is local and does not claim certified robustness or guaranteed recovery of the correct class. Its purpose is to clarify how risk-gated correction, orthogonal exploration, and frequency-selective shaping affect the CLIP embedding under standard smoothness and first-order assumptions.

B.1 Margin Preservation under Embedding Perturbation

We first show that keeping the CLIP embedding within a margin-dependent radius is sufficient to preserve the zero-shot prediction.

Lemma 1 (CLIP margin preservation)**.** *Assume all text embeddings are normalized, i.e.,* $\|e_i\|_2 = 1$ *for all* i*. Let* x *be a correctly classified image with label* y *and CLIP margin*

$$m_y(x) = h(x)^\top e_y - \max_{j \neq y} h(x)^\top e_j > 0.$$

If another input $\tilde{x}$ *satisfies*

$$\|h(\tilde{x}) - h(x)\|_2 < \frac{m_y(x)}{2},$$

then $\hat{y}(\tilde{x}) = y$*.*

Proof. For any competing class $j \neq y$, consider the logit difference between classes y and j at $\tilde{x}$:

$$h(\tilde{x})^\top e_y - h(\tilde{x})^\top e_j = h(\tilde{x})^\top (e_y - e_j).$$

Add and subtract $h(x)$:

$$h(\tilde{x})^\top (e_y - e_j) = h(x)^\top (e_y - e_j) + \big(h(\tilde{x}) - h(x)\big)^\top (e_y - e_j).$$

By the definition of the margin,

$$h(x)^\top (e_y - e_j) = h(x)^\top e_y - h(x)^\top e_j \geq m_y(x).$$

By the Cauchy–Schwarz inequality,

$$\left|\big(h(\tilde{x}) - h(x)\big)^\top (e_y - e_j)\right| \le \|h(\tilde{x}) - h(x)\|_2 \cdot \|e_y - e_j\|_2.$$

Since $\|e_y\|_2 = \|e_j\|_2 = 1$, we have

$$\|e_y - e_j\|_2 \le \|e_y\|_2 + \|e_j\|_2 = 2.$$

Therefore,

$$h(\tilde{x})^\top e_y - h(\tilde{x})^\top e_j \ge m_y(x) - 2\|h(\tilde{x}) - h(x)\|_2.$$

If

$$\|h(\tilde{x}) - h(x)\|_2 < \frac{m_y(x)}{2},$$

then

$$h(\tilde{x})^\top e_y - h(\tilde{x})^\top e_j > 0$$

for every $j \neq y$. Hence, class y has a larger zero-shot logit than every other class, and therefore $\hat{y}(\tilde{x}) = y$. □

B.2 Clean-Input Distortion Control by the Soft Gate

We next show that the final soft gate controls clean embedding drift, explaining why the counteraction strength should depend on the estimated input risk.

Proposition 2 (Clean-input distortion control)**.** *Assume h is locally L_h-Lipschitz in ℓ_2 on a neighborhood containing the clean input x and the defended input, i.e.,*

$$\|h(x + \delta) - h(x)\|_2 \le L_h \|\delta\|_2$$

for perturbations in the considered local neighborhood. Suppose AFOC produces $\|\delta_T\|_\infty \le \epsilon_{\mathrm{ca}}$ and applies the final gate

$$\delta_{\mathrm{ca}} = w(x)\delta_T.$$

Then

$$\|h(x_{\mathrm{def}}) - h(x)\|_2 \le L_h w(x) \sqrt{D}\epsilon_{\mathrm{ca}},$$

where D is the input dimension. Consequently, if

$$w(x) < \frac{m_y(x)}{2L_h\sqrt{D}\epsilon_{\mathrm{ca}}},$$

then the clean prediction is preserved.

Proof. For a clean input, $x_{\mathrm{in}} = x$. The final defended input is

$$x_{\mathrm{def}} = \mathrm{clip}_{[0,1]}(x + \delta_{\mathrm{ca}}), \qquad \delta_{\mathrm{ca}} = w(x)\delta_T.$$

Because $x \in [0,1]^D$ and coordinate-wise clipping is non-expansive relative to x, $\|\,\mathrm{clip}_{[0,1]}(x + \delta_{\mathrm{ca}}) - x\|_2 \le \|\delta_{\mathrm{ca}}\|_2$. It therefore suffices to bound the perturbation norm before clipping. Since $\|\delta_T\|_\infty \le \epsilon_{\mathrm{ca}}$, we have

$$\|\delta_T\|_2 \le \sqrt{D}\|\delta_T\|_\infty \le \sqrt{D}\epsilon_{\mathrm{ca}}.$$

Therefore,

$$\|\delta_{\mathrm{ca}}\|_2 = w(x)\|\delta_T\|_2 \le w(x)\sqrt{D}\epsilon_{\mathrm{ca}}.$$

By the local Lipschitz continuity of h,

$$\begin{aligned}\|h(x_{\mathrm{def}}) - h(x)\|_2 &\le L_h\|\delta_{\mathrm{ca}}\|_2 \\ &\le L_h w(x)\sqrt{D}\epsilon_{\mathrm{ca}}.\end{aligned}$$

If

$$w(x) < \frac{m_y(x)}{2L_h\sqrt{D}\epsilon_{\mathrm{ca}}},$$

then

$$\|h(x_{\mathrm{def}}) - h(x)\|_2 < \frac{m_y(x)}{2}.$$

By Lemma 1, the prediction remains $\hat{y}(x_{\mathrm{def}}) = y$. □

B.3 Coverage of Orthogonal Components

We now characterize the expected coverage provided by orthogonal exploration for components not captured by the primary gradient direction. This result motivates using an orthogonal direction when the true corrective direction is unknown.

Proposition 3 (Coverage of orthogonal components)**.** *Let $q \in \mathbb{R}^D$ be a unit primary direction. Let $r^\perp$ be a random unit vector uniformly distributed on the unit sphere in the subspace orthogonal to q. For any unknown unit corrective direction $v \in \mathbb{R}^D$, define*

$$\bar{d}_\lambda = \frac{q + \lambda r^\perp}{\sqrt{1+\lambda^2}}, \qquad \lambda \geq 0.$$

Then

$$\mathbb{E}\big[(v^\top \bar{d}_\lambda)^2\big] = \frac{(v^\top q)^2 + \frac{\lambda^2}{D-1}\big(1 - (v^\top q)^2\big)}{1+\lambda^2}.$$

Proof. Since q is a unit vector, decompose v into the component parallel to q and the component orthogonal to q:

$$v = (v^\top q)q + v_\perp, \qquad v_\perp \perp q.$$

Because $\|v\|_2 = 1$ and $\|q\|_2 = 1$,

$$\|v_\perp\|_2^2 = 1 - (v^\top q)^2.$$

We now compute the inner product between v and $\bar{d}_\lambda$:

$$\begin{aligned} v^\top \bar{d}_\lambda &= v^\top \frac{q + \lambda r^\perp}{\sqrt{1+\lambda^2}} \\ &= \frac{v^\top q + \lambda v^\top r^\perp}{\sqrt{1+\lambda^2}}. \end{aligned}$$

Since $r^\perp \perp q$, we have

$$v^\top r^\perp = \big((v^\top q)q + v_\perp\big)^\top r^\perp = v_\perp^\top r^\perp.$$

Thus,

$$v^\top \bar{d}_\lambda = \frac{v^\top q + \lambda v_\perp^\top r^\perp}{\sqrt{1+\lambda^2}}.$$

Squaring and taking expectation gives

$$\mathbb{E}\big[(v^\top \bar{d}_\lambda)^2\big] = \frac{1}{1+\lambda^2}\mathbb{E}\left[(v^\top q)^2 + 2\lambda (v^\top q)(v_\perp^\top r^\perp) + \lambda^2 (v_\perp^\top r^\perp)^2\right].$$

Because $r^\perp$ is uniformly distributed on the unit sphere in the $(D-1)$-dimensional subspace orthogonal to q, its distribution is symmetric around the origin in that subspace. Hence

$$\mathbb{E}[v_\perp^\top r^\perp] = 0,$$

and the cross term vanishes:

$$\mathbb{E}\left[2\lambda (v^\top q)(v_\perp^\top r^\perp)\right] = 0.$$

It remains to compute

$$\mathbb{E}\big[(v_\perp^\top r^\perp)^2\big].$$

For a random unit vector uniformly distributed on the unit sphere in a d-dimensional space, the second moment in any fixed direction a is

$$\mathbb{E}\big[(a^\top r)^2\big] = \frac{\|a\|_2^2}{d}.$$

Here the orthogonal subspace has dimension $d = D - 1$, so

$$\mathbb{E}\big[(v_\perp^\top r^\perp)^2\big] = \frac{\|v_\perp\|_2^2}{D-1} = \frac{1 - (v^\top q)^2}{D-1}.$$

Substituting this into the expectation yields

$$\mathbb{E}\big[(v^\top \bar{d}_\lambda)^2\big] = \frac{(v^\top q)^2 + \lambda^2 \frac{1-(v^\top q)^2}{D-1}}{1+\lambda^2}.$$

This proves the claim. □

Remark. In the implementation, $r^{\perp}$ is obtained by projecting a Gaussian random vector onto the subspace orthogonal to q and normalizing it. This produces a direction uniformly distributed on the unit sphere of the orthogonal subspace, so the above analysis applies. The proposition isolates the additional orthogonal second-moment term. Because the combined direction is normalized by $\sqrt{1+\lambda^2}$, it does not assert that every $\lambda > 0$ increases total alignment, and the squared inner product does not determine the sign of a correction.

B.4 Effect of Frequency-Selective Update Shaping

Finally, we show that, under a local linearization of the frozen image encoder, the frequency-shaping operator amplifies high-band components of the defensive update.

Proposition 4 (High-band amplification under local linearization)**.** *Let P_H and $P_L = I - P_H$ denote the high- and low-frequency projections, respectively. Define the frequency-shaped update operator*

$$P_F(m) = P_L(m) + (1+\alpha_F)P_H(m), \qquad \alpha_F \geq 0.$$

Assume that the CLIP image embedding is locally linearized around x_{in}:

$$h(x_{\mathrm{in}} + \delta) \approx h(x_{\mathrm{in}}) + J\delta,$$

where J is the Jacobian of h at x_{in}. For any unit embedding direction a, the first-order target-aligned effect of $P_F(m)$ is

$$a^{\top} J P_F(m) = a^{\top} J P_L(m) + (1+\alpha_F) a^{\top} J P_H(m).$$

Therefore, if $a^{\top} J P_H(m) > 0$, increasing α_F amplifies the high-band aligned correction while leaving the low-band contribution unchanged.

Proof. By definition,

$$P_F(m) = P_L(m) + (1+\alpha_F)P_H(m).$$

Under the local linearization,

$$h(x_{\mathrm{in}} + P_F(m)) - h(x_{\mathrm{in}}) \approx J P_F(m).$$

The first-order effect of this update along an embedding-space direction a is

$$a^{\top} J P_F(m).$$

Substituting the decomposition of $P_F(m)$ gives

$$\begin{aligned} a^{\top} J P_F(m) &= a^{\top} J \left[P_L(m) + (1+\alpha_F) P_H(m) \right] \\ &= a^{\top} J P_L(m) + (1+\alpha_F) a^{\top} J P_H(m). \end{aligned}$$

The low-band term $a^{\top} J P_L(m)$ is independent of α_F, while the high-band term is multiplied by $(1+\alpha_F)$. Hence, if

$$a^{\top} J P_H(m) > 0,$$

then increasing α_F strictly increases the target-aligned first-order effect contributed by the high-frequency component. This proves the claim. □

Interpretation. The positivity condition is essential: the proposition characterizes how the operator reallocates an already aligned component, but does not establish that the high-frequency component is aligned for every input. The empirical ablation and spectrum analyses test whether this conditional mechanism is useful in the evaluated setting.

B.5 Proof of Proposition 1

Proof. Part (i) follows directly from Proposition 2. Specifically, the soft gate gives

$$\|h(x_{\mathrm{def}}) - h(x)\|_2 \leq L_h w(x) \sqrt{D} \epsilon_{\mathrm{ca}},$$

and combining this inequality with Lemma 1 yields the stated sufficient condition for preserving the clean prediction.

Part (ii) is exactly Proposition 3. It shows that adding a normalized orthogonal component to the primary direction provides expected coverage for the component of any unknown corrective direction outside the span of the primary gradient.

Part (iii) is exactly Proposition 4. It follows from decomposing the shaped update into low- and high-frequency projections and applying a first-order linearization of the frozen image encoder.

Thus, the three component-wise statements in Proposition 1 follow under their respective assumptions. □

C Additional Experimental Details

C.1 Dataset and Evaluation Details

CIFAR-10 and CIFAR-100 are evaluated on their complete test sets. ImageNet clean accuracy is evaluated on the validation set, while adversarial accuracy is measured on one fixed class-balanced validation subset to control the cost of iterative defenses and adaptive attacks. The subset, attack examples, and random seeds are matched across methods. Images follow the official CLIP preprocessing for each backbone. Unless otherwise stated, the class prompt is "a photo of a {class}." For prompt ensembling, class embeddings are obtained by averaging the template-specific text embeddings and normalizing the result.

C.2 Attack Details

For PGD (Madry et al., 2018) and CW (Carlini & Wagner, 2017), the attack variable is initialized inside the permitted ℓ_∞ ball and optimized against CLIP zero-shot logits. PGD uses 40 iterations and step size $\epsilon_{\text{atk}}/10$. CW uses 100 iterations with the margin loss

$$\ell_{\text{cw}}(x, y) = \max_{j \neq y} s_j(x) - s_y(x),$$

where $s_j(x)$ is the zero-shot logit of class j. AutoAttack (Croce & Hein, 2020) uses its standard ensemble configuration. The default budgets are $\epsilon_{\text{atk}} = 8/255$ for CIFAR-10/100 and $4/255$ for ImageNet. These three attacks are generated against the frozen base classifier and then evaluated after defense.

Adaptive EOT-PGD optimizes a BPDA surrogate (Athalye et al., 2018a) of the complete defended pipeline, taking the Monte Carlo expectation under EOT (Athalye et al., 2018b). For a stochastic defense $\mathcal{D}_\omega$, its Monte Carlo objective is

$$\max_{\|\delta\|_\infty \leq \epsilon_{\text{atk}}} \frac{1}{K_{\text{EOT}}} \sum_{k=1}^{K_{\text{EOT}}} \ell_{\text{ce}}\big(\hat{p}(\mathcal{D}_{\omega_k}(x + \delta)), y\big).$$

The main tables use $K_{\text{EOT}} = 4$, consistent with Table 3. Projection, clipping, and sign-based inner updates use a straight-through BPDA approximation. This produces a surrogate gradient through the non-smooth defense rather than an exact derivative. Appendix D.5 reports $K_{\text{EOT}} \in \{4, 8, 16\}$; those rows are a sensitivity analysis and do not replace the $K_{\text{EOT}} = 4$ values in the main tables. Appendix D.6 further queries the defended logits with the ℓ_∞ Square Attack (Andriushchenko et al., 2020). This experiment assesses gradient-free search rather than replacing the main-table BPDA protocol.

C.3 Baselines

Vanilla CLIP. The frozen zero-shot classifier is evaluated without a defense.

Input preprocessing. Low-pass filtering and JPEG compression are included as non-iterative input-level baselines.

Inference-time counteraction. TTC and DOC are the most direct baselines. Both keep CLIP frozen and optimize an input-specific counteraction. Recommended implementation settings are used when available, with the counteraction budget matched to AFOC.

Table 3: Default hyperparameters and backbone bias calibration of AFOC.

Hyperparameter	Value
Counteraction steps T	10
Number of probes M	4
Probe radius ϵ_p	$\epsilon_{\text{atk}}/4$
Risk weights β_g, β_f	20, 20
Base risk bias b_0	3.0
Backbone bias offset Δ_b	RN50: 1.25; ViT-B/16: 1.75; ViT-B/32: 0
Low-pass cutoff ratio r	0.5
High-pass cutoff ratio r_H	0.5
Defense budget ϵ_{ca}	ϵ_{atk}
Warm-start scale s_{init}	0.5
Step size α	ϵ_{ca}/T
$\lambda_{\min}, \lambda_{\max}$	0, 0.5
$\mu_{\min}, \mu_{\max}$	0.3, 0.9
$\alpha_{\min}, \alpha_{\max}$	0, 0.5
Gate sharpness γ	8
Risk threshold τ_ρ	0.5
EOT samples K_{EOT}	4

C.4 AFOC Hyperparameters

Table 3 lists the default hyperparameters used in the main experiments. The same risk weights, controller ranges, and frequency cutoffs are used across datasets and attacks. The only backbone-dependent quantity is the fixed bias offset Δ_b shown in the table. Appendix D.7 sets b on a held-out clean validation split by matching the median of ρ to $1/2$; ranking metrics do not depend on this offset.

D Additional Results

D.1 Full Dataset Results

Table 4 reports all three datasets. Mixed Avg. is recomputed uniformly as the arithmetic mean of the four attack columns; in particular, it never includes Clean. It is a descriptive cross-protocol summary rather than an end-to-end robustness metric. The CIFAR-100 and ImageNet entries exactly match Table 1.

On CIFAR-10, AFOC attains a 62.83% mixed-protocol average, exceeding TTC and DOC by 61.65 and 59.02 percentage points, respectively, while its clean accuracy is 0.92 percentage points above that of vanilla CLIP. On CIFAR-100, it reaches 39.22% with a 1.98-percentage-point drop in clean accuracy. On ImageNet, AFOC has the highest mixed average among the compared methods and remains 0.11 percentage points below vanilla CLIP. The large variation across datasets also motivates reporting each attack rather than relying only on a single aggregate score.

D.2 Cross-Backbone Results

Table 5 evaluates CIFAR-100 under AutoAttack and adaptive EOT-PGD for RN50, ViT-B/32, and ViT-B/16.

AFOC improves both attacks over DOC for every backbone. The AutoAttack gains are 11.80, 36.52, and 55.76 percentage points for RN50, ViT-B/32, and ViT-B/16, respectively; the corresponding adaptive EOT-PGD gains are 9.90, 29.69, and 38.66 percentage points. Relative to vanilla CLIP, AFOC changes clean accuracy by -0.84, -1.98, and $+1.02$ percentage points on RN50, ViT-B/32, and ViT-B/16, respectively. Thus, the gains transfer across all three visual backbones without a large clean-accuracy penalty in this evaluation. ViT-B/16 yields the strongest AutoAttack and adaptive EOT-PGD accuracies, while RN50 remains substantially weaker in absolute robustness, indicating that the attainable operating point is backbone-dependent.

Table 4: Full results on all datasets. PGD, CW, and AA attack the base classifier before post-hoc defense; EOT-PGD is defense-aware and uses EOT with BPDA. Mixed Avg. averages PGD, CW, AA, and EOT-PGD and excludes Clean; it is a descriptive summary across different protocols.

Dataset	Method	Clean	PGD	CW	AA	EOT-PGD	Mixed Avg.
CIFAR-10	Vanilla CLIP	85.09	0.00	0.00	0.01	0.04	0.01
	Low-pass	84.60	0.01	0.00	2.72	0.00	0.68
	JPEG	83.51	0.00	0.00	3.48	0.00	0.87
	TTC	80.80	0.32	0.10	4.30	0.00	1.18
	DOC	80.47	4.25	1.18	4.84	4.95	3.81
	AFOC	**86.01**	**54.51**	**73.26**	**73.39**	**50.15**	**62.83**
CIFAR-100	Vanilla CLIP	58.53	0.00	0.00	0.08	0.00	0.02
	Low-pass	55.44	0.06	0.05	5.47	0.01	1.40
	JPEG	**58.60**	0.00	0.00	7.05	0.00	1.76
	TTC	55.33	0.08	0.03	8.88	0.00	2.25
	DOC	55.25	1.48	0.27	8.89	0.49	2.78
	AFOC	56.55	**44.56**	**36.74**	**45.41**	**30.18**	**39.22**
ImageNet	Vanilla CLIP	**59.26**	0.00	0.00	0.12	0.00	0.03
	Low-pass	54.76	1.66	2.20	8.93	1.16	3.49
	JPEG	58.53	0.50	0.34	4.50	0.20	1.39
	TTC	49.37	3.40	3.78	7.27	0.02	3.62
	DOC	43.28	20.38	15.40	20.08	2.72	14.65
	AFOC	59.15	**23.60**	**31.41**	**41.18**	**29.53**	**31.43**

Table 5: Cross-backbone results on CIFAR-100.

Backbone	Method	Clean	AA	EOT-PGD
RN50	Vanilla CLIP	**31.48**	0.12	0.06
	DOC	23.65	6.04	0.17
	AFOC	30.64	**17.84**	**10.07**
ViT-B/32	Vanilla CLIP	**58.53**	0.08	0.00
	DOC	55.25	8.89	0.49
	AFOC	56.55	**45.41**	**30.18**
ViT-B/16	Vanilla CLIP	63.58	0.02	0.01
	DOC	52.32	5.25	1.07
	AFOC	**64.60**	**61.01**	**39.73**

D.3 Prompt Ensembling

Table 6 changes the text anchors by averaging multiple prompt templates while retaining the same visual backbone and attack budgets.

With prompt ensembling, AFOC outperforms DOC under all three attacks, with gains of 38.10 percentage points under PGD, 38.16 percentage points under AutoAttack, and 30.52 percentage points under adaptive EOT-PGD. Its clean accuracy is also within 0.25 percentage points of prompt-ensemble vanilla CLIP and exceeds that of DOC by 3.15 percentage points. Relative to the default single-prompt results in Table 1, prompt ensembling increases AFOC's clean, AutoAttack, and adaptive EOT-PGD accuracies by 3.54, 0.31, and 0.60 percentage points, respectively, while reducing PGD accuracy by 6.23 percentage points. These results indicate that the robustness benefit is not tied to a single text template and that the adaptive advantage remains intact after changing the text anchors. At the same time, the attack-dependent shift in performance indicates that prompt design and controller calibration should be considered jointly rather than assuming a uniform gain from ensembling.

D.4 Attack Budget Sensitivity

Table 7 varies the CIFAR-100 ℓ_∞ attack budget for adaptive EOT-PGD.

Table 6: Prompt-ensemble results on CIFAR-100 with CLIP ViT-B/32.

Method	Clean	PGD	AA	EOT-PGD
Vanilla CLIP	**60.34**	0.00	0.11	0.00
DOC	56.94	0.23	7.56	0.26
AFOC	60.09	**38.33**	**45.72**	**30.78**

Table 7: Matched attack/defense-budget sensitivity on CIFAR-100 under adaptive EOT-PGD. For each column, $\epsilon_{\mathrm{ca}} = \epsilon_{\mathrm{atk}}$.

Method	$4/255$	$8/255$	$12/255$
Vanilla CLIP	0.00	0.00	0.00
TTC	0.00	0.00	0.00
DOC	2.61	0.49	0.57
AFOC	**30.45**	**30.18**	**27.05**

AFOC exceeds DOC by 27.84, 29.69, and 26.48 percentage points at $4/255$, $8/255$, and $12/255$, respectively. Because the defense budget is matched to the attack budget in each column, this experiment jointly changes attack strength and counteraction capacity and should not be interpreted as a monotonic attack-only curve. AFOC remains near 30% at $4/255$ and $8/255$ and decreases to 27.05% at $12/255$. This result indicates that increasing the matched counteraction budget does not fully offset the stronger attack at the largest evaluated radius.

D.5 EOT Sample Sensitivity

Table 8 increases the number of EOT samples used by the adaptive attacker. This is a stricter gradient estimate than the default $K_{\mathrm{EOT}} = 4$ evaluation.

Increasing K_{EOT} from 4 to 16 reduces AFOC from 30.18% to 27.17%, corresponding to a 3.01-percentage-point decrease; the intermediate $K_{\mathrm{EOT}} = 8$ evaluation yields 28.79%. DOC remains near zero across all three settings, decreasing from 0.49% to 0.13%. Consequently, AFOC retains margins of 28.66 and 27.04 percentage points over DOC at $K_{\mathrm{EOT}} = 8$ and 16, respectively. The modest monotonic decline indicates some sensitivity to the accuracy of the EOT gradient estimate, while the consistently large margin over DOC shows that the observed advantage is not specific to the default $K_{\mathrm{EOT}} = 4$ setting. This comparison probes only the EOT sample count; it does not rule out failure modes shared by all tested BPDA configurations.

D.6 Gradient-Free Square Attack

The main tables attack AFOC using EOT-PGD with a BPDA surrogate. To test whether that surrogate conceals gradient masking, we query the complete defended prediction pipeline using the ℓ_∞ Square Attack (Andriushchenko et al., 2020), which does not use gradients through the defense. We evaluate Vanilla CLIP, TTC, DOC, and AFOC on the same evaluation sets as Table 1: the complete CIFAR-100 test set (10,000 images) and the same ImageNet class-balanced attack subset (5 images per class). The ℓ_∞ budget is $4/255$ on both datasets, with $\epsilon_{\mathrm{ca}} = \epsilon_{\mathrm{atk}}$. Accuracy is recorded after 100, 1,000, and 2,000 queries (Table 9, Figure 3). On ImageNet, AFOC uses the Table 1 counteraction settings ($T = 10$, $M = 4$, repulsion). On CIFAR-100, Square evaluation uses a hybrid counteraction objective with $w_{\min} = 0.35$ and $b = 2.0$.

Figure 3 and Table 9 separate the four methods along the query axis on both datasets. On CIFAR-100, AFOC retains 27.46% / 25.75% / 25.67% at 100 / 1,000 / 2,000 queries, compared with 13.22% / 3.73% / 2.93% for DOC; vanilla CLIP and TTC fall below 1% by 1,000 queries. On ImageNet, vanilla CLIP is highest at 100 queries (27.47%), but AFOC is the strongest fixed-prompt method at 1,000 and 2,000 queries (8.45% and 6.97%), compared with 3.46% / 1.06% for vanilla CLIP and 2.94% / 2.62% for DOC. TTC remains near 0.02% at every ImageNet budget. Because Square queries the defended logits and does not use BPDA, this evaluation is complementary to the adaptive EOT-PGD column in Table 1 and should not be read as interchangeable with it.

Table 8: Sensitivity to the number of EOT samples on CIFAR-100.

Method	$K_{\mathrm{EOT}} = 4$	$K_{\mathrm{EOT}} = 8$	$K_{\mathrm{EOT}} = 16$
DOC	0.49	0.13	0.13
AFOC	**30.18**	**28.79**	**27.17**

Table 9: Square Attack ($\ell_\infty = 4/255$) accuracy (%) after defense as a function of the query budget. The attack queries defended logits directly and does not use BPDA. CIFAR-100 uses the full 10,000-image test set; ImageNet uses the same 5-image-per-class subset as Table 1. ‡CIFAR-100 AFOC uses a hybrid counteraction objective with $w_{\min} = 0.35$ and $b = 2.0$; ImageNet uses the Table 1 configuration.

Dataset	Method	100	1,000	2,000
CIFAR-100	Vanilla CLIP	11.02	0.59	0.18
	TTC	2.37	0.09	0.05
	DOC	13.22	3.73	2.93
	AFOC‡	**27.46**	**25.75**	**25.67**
ImageNet	Vanilla CLIP	**27.47**	3.46	1.06
	TTC	0.02	0.02	0.02
	DOC	5.63	2.94	2.62
	AFOC	22.02	**8.45**	**6.97**

D.7 Risk-Score Reliability

The main text uses ρ as a continuous controller rather than as an attack detector. Nevertheless, it remains important to determine whether higher ρ ranks more fragile inputs. We therefore evaluate the existing CIFAR-100 Figure 6 scores (10,000 test images; CLIP ViT-B/32) as a ranking problem: clean versus each attacked copy of the same image. Let $z = 20s_{\mathrm{geo}} + 20s_{\mathrm{freq}}$ and $\rho = \sigma(z - b)$ with the paper default $b = 3.0$. Because $\sigma(z - b)$ is strictly monotone in z, AUROC and AUPR are invariant to any global bias b or Δ_b; those offsets change only the operating point of the gate $w(\rho)$.

Table 10 and Figure 4 show that ρ provides a strong ranking for PGD and CW (AUROC ≥ 0.998), a moderate ranking for AutoAttack (0.810), and a weaker ranking of adaptive EOT-PGD (0.755), with an overlap coefficient of 0.61. The EOT overlap observed qualitatively is therefore substantial: 83.9% of EOT-PGD inputs fall below the 95th percentile of clean ρ. This supports the paper's refusal to treat ρ as a detector, while still showing that standard ℓ_∞ attacks shift the score.

The two cues are not interchangeable. s_{freq} alone already separates PGD, CW, and EOT-PGD (AUROC ≥ 0.999). s_{geo} is *anti-aligned* with those attacks: mean geometric instability drops from 0.089 on clean inputs to below 0.001 after PGD, CW, or EOT-PGD, so AUROC of s_{geo} as a "higher-is-more-attacked" score is 0. Fused ρ still ranks PGD/CW well because the s_{freq} term dominates z; under EOT-PGD the collapsed s_{geo} pulls ρ back toward the clean range and produces the overlap in Figure 6. We therefore describe s_{geo} as a clean-side stabilizer of the controller, not as a detector of adversarial perturbations.

A remaining concern is that even a strong aggregate AUROC could be carried by a few easily separated classes, or that ρ could simply track natural image complexity. Figure 5 tests both.

Panel (a) reports PGD-versus-clean AUROC of ρ for every CIFAR-100 class: the minimum is 0.986, the median is 0.999, and all 100 classes exceed 0.90. The ranking is therefore not an average over a handful of separable categories. Panel (b) plots clean ρ against a Fourier high-frequency energy ratio on the same 10,000 unattacked images. If ρ were a texture proxy, the cloud would rise with high-band energy. It does not: the Spearman correlation is $r_s = -0.06$, and a linear fit is essentially flat. Correlation with Laplacian variance is likewise weak and negative (-0.27). Removing a linear texture fit from z leaves the PGD AUROC unchanged at 0.998. High-frequency natural content, including the textured or crowded scenes that could inflate geometric-risk scores, does not explain why attacked inputs rank above their clean counterparts. Together with Table 10, these checks support using ρ as an input-conditioned control signal: it orders standard ℓ_∞ attacks without collapsing onto

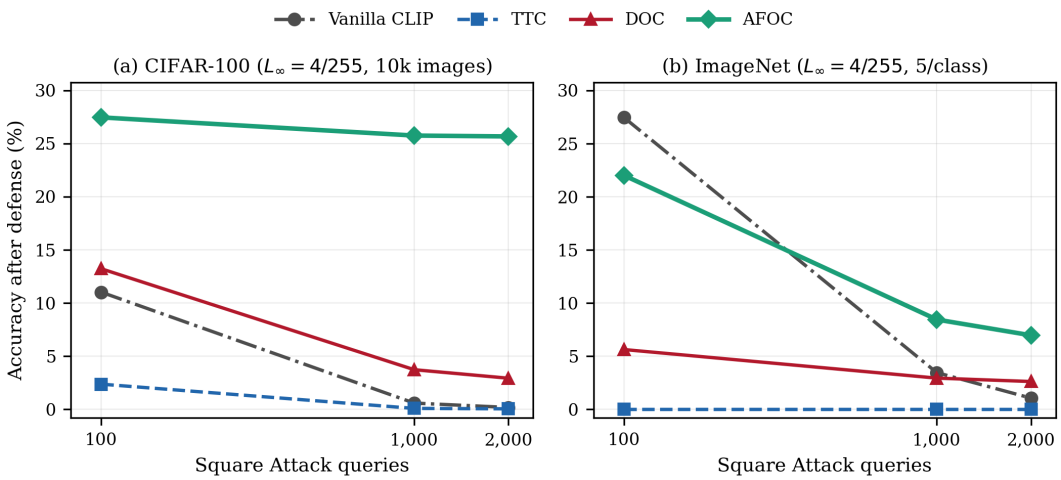


Figure 3: Square Attack ($\ell_\infty = 4/255$) query–accuracy curves on the defended pipeline for Vanilla CLIP, TTC, DOC, and AFOC. **(a) CIFAR-100:** full 10,000-image test set. **(b) ImageNet:** the same 5-image-per-class subset as Table 1. CIFAR-100 AFOC uses a hybrid counteraction objective with $w_{\min} = 0.35$ and $b = 2.0$; ImageNet AFOC uses the Table 1 configuration.

Table 10: CIFAR-100 ranking quality of ρ (and of the two cues) for clean versus attacked copies of the same 10,000 images. Overlap is the histogram overlapping coefficient on $[0, 1]$ at $b = 3.0$.

Attack	AUROC ρ	AUPR ρ	Overlap	Mean ρ	AUROC s_{freq}	AUROC s_{geo}
PGD	0.998	0.998	0.04	0.97	1.000	0.00
CW	1.000	1.000	0.00	1.00	1.000	0.00
AutoAttack	0.810	0.865	0.42	0.77	0.807	0.22
EOT-PGD	0.755	0.711	0.61	0.56	0.999	0.00

class identity or low-level complexity, while the EOT-PGD overlap in Figure 4 remains a reason not to threshold it as a detector.

On a class-stratified 2,000-image validation split used only for calibration, we set b so that the median clean risk is $1/2$, i.e. $b = \text{median}(z^{\text{val}}_{\text{clean}})$. This yields $b = 2.64$ versus the paper default 3.0 ($\Delta_b = -0.36$ relative to ViT-B/32). Held-out PGD AUROC on the remaining 8,000 images is 0.998 at either value, as required by monotonicity. Applying the same clean-only median rule on a 1,000-image subset for three CLIP backbones (Table 11) recovers $\Delta_b \approx -0.32$ for ViT-B/32 and 0.87 for RN50, close in sign and order to the paper offsets 0 and 1.25. The paper ViT-B/16 offset 1.75 is more conservative than median matching (recovered $\Delta_b \approx -0.15$): it lowers mean clean ρ to 0.16 rather than 0.50. PGD ranking AUROC remains ≥ 0.998 for all three backbones at either choice, so Δ_b is an operating-point shift rather than an evaluation-set fit of a detector.

The default $b = 3.0$ is thus a slightly conservative ViT-B/32 operating point (median clean $\rho = 0.41$ rather than 0.50) rather than a detector threshold fit. Thresholding ρ at 0.5 would flag 29% of clean CIFAR-100 images. Sweeping b on a 500-image AFOC evaluation subset (Table 12) changes clean accuracy from 14.8% at $b = 2.0$ to 58.6% at $b = 4.75$, while PGD accuracy stays in a narrow 31–36% band. Lower b over-corrects clean inputs; the paper default sits on the clean-preserving side of this trade-off. A 2,000-image auxiliary analysis finds only a weak association with the vanilla CLIP margin (Spearman -0.11 on clean inputs and -0.27 after PGD); clean errors have slightly higher mean ρ (0.45 versus 0.42). The score tracks attack-induced spectral sensitivity more than it tracks per-sample clean margin.

E Mechanism Analysis

We examine whether the controller exhibits the intended behavior: a continuous input-risk ordering, sustained recovery of the CLIP decision margin, and frequency-selective allocation of the

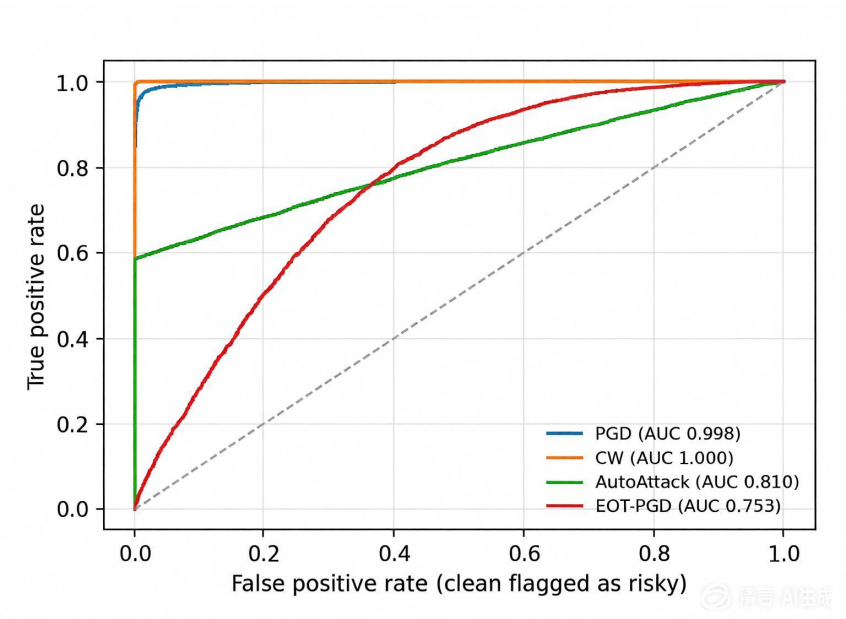


Figure 4: ROC curve for using ρ to distinguish attacked CIFAR-100 inputs from their clean counterparts. PGD and CW are essentially separated, whereas AutoAttack and EOT-PGD retain the overlap visible in Figure 6.

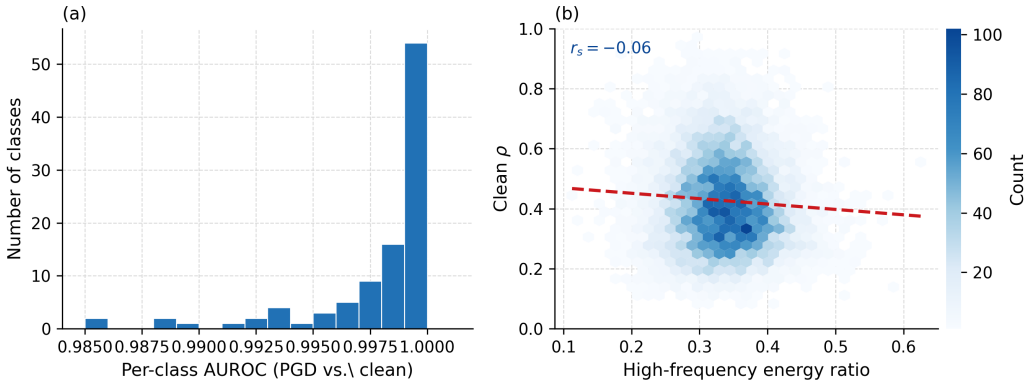


Figure 5: CIFAR-100 diagnostics for two confounds of ρ. **(a)** Per-class AUROC of ρ for PGD versus the matched clean image (100 classes, 100 images each). Separation is uniformly high. **(b)** Clean ρ versus high-frequency energy. The risk score is unrelated to texture ($r_s = -0.06$); the dashed line shows a linear fit. Both panels use CLIP ViT-B/32 and the default bias $b = 3.0$.

counteraction update. Unless otherwise stated, the analysis uses CIFAR-100 with CLIP ViT-B/32. The complementary embedding-geometry diagnostic on CIFAR-10 is reported in the main text (Figure 2, Section 5.4).

The risk score provides soft, attack-dependent ordering. Figure 6 shows that PGD, CW, and AutoAttack inputs are concentrated near high risk, whereas clean inputs place most mass at lower scores. Adaptive EOT-PGD has a broader distribution that partially overlaps the clean distribution. Table 10 quantifies this ordering: PGD and CW are nearly separated from clean inputs, AutoAttack is weaker, and the EOT-PGD overlap coefficient is 0.61. This overlap is consistent with using ρ as a soft control variable rather than a binary detector: higher-risk inputs receive stronger exploration and correction, whereas lower-risk inputs are adjusted more conservatively. Additional reliability checks, including class-wise AUROC and a texture-confound test (Figure 5), are provided in Appendix D.7.

Table 11: Clean-only median matching of b on 1,000 CIFAR-100 images. Recovered $\Delta_b = \hat{b} - 3.0$. Paper $b = 3 + \Delta_b$.

Backbone	Paper Δ_b	Recovered Δ_b	Mean clean ρ (paper b)	PGD AUROC of z
ViT-B/32	0.00	-0.32	0.43	0.998
RN50	1.25	0.87	0.44	1.000
ViT-B/16	1.75	-0.15	0.16	1.000

Table 12: AFOC clean and post-hoc PGD accuracy (%) on a 500-image CIFAR-100 subset while varying the risk bias b (ViT-B/32). These subset numbers are not comparable to Table 1.

b	2.0	2.5	3.0	3.5	4.0	4.25	4.75
Clean	14.8	28.4	40.6	51.2	55.8	58.0	58.6
PGD	35.6	30.6	34.6	33.0	34.8	34.6	31.2

Risk-conditioned updates sustain label-free margin recovery. The left panel of Figure 7 uses the supervised margin m_y only as a diagnostic; it is not part of the objective in Eq. equation 4. All methods start from a strongly negative margin. DOC recovers rapidly in the first iterations but plateaus, whereas TTC progresses more slowly. AFOC continues improving after DOC saturates and ends closest to the zero-margin boundary. The margin remains slightly negative, so the result indicates substantial recovery of class separation rather than the complete correction of every attacked sample.

Frequency shaping redistributes update energy rather than filtering the input. The right panel of Figure 7 reports the spectrum of the final counteraction perturbation. TTC is dominated by low frequencies, while DOC retains a broad high-frequency tail. AFOC preserves a low-frequency component, maintains a mid-frequency shoulder, and allocates more mid-to-high-frequency power than TTC while attenuating the extreme tail relative to DOC. Because the spectra are normalized, the plot describes relative allocation rather than perturbation magnitude. Appendix F.4 separately analyzes the spectrum of the update directions, which should not be conflated with the spectrum after iterative accumulation, projection, and risk-dependent gating.

Together, Figures 6 and 7 show that the risk score changes the controller continuously, the resulting label-free trajectory improves an externally evaluated class-margin diagnostic, and the Fourier branch changes the distribution of update energy. Read jointly with Figure 2, these observations suggest that AFOC acts through input-conditioned rematching in the joint embedding space rather than through restoring a unique clean cluster geometry. These diagnostics complement the component-wise evidence in Table 2.

F Additional Ablation and Mechanism Analysis

F.1 Ablation Variant Definitions

w/o soft gate. Set $w = 1$, so the full projected perturbation δ_T is always applied.

w/o orthogonal exploration. Retain the same risk-scaled warm initialization and matched random seeds as the full model, but set $\lambda = 0$ at every counteraction iteration. This variant removes iterative orthogonal exploration without simultaneously removing the mechanism that escapes the stationary point at $\delta = 0$.

w/o adaptive momentum. Replace the risk-conditioned momentum with the constant $\mu = 0.6$ and keep the same trajectory update; all other controller outputs remain input-conditioned. This variant is reported as "Fixed momentum" in Table 2. The constant is obtained by evaluating the default map $\mu = \mu_{\max} - \rho(\mu_{\max} - \mu_{\min})$ at $\rho = 0.5$, using $(\mu_{\min}, \mu_{\max}) = (0.3, 0.9)$ from Table 3. Here $\rho = 0.5$ is the midpoint of $(0, 1)$ and the default value of τ_ρ.

w/o frequency shaping. Set $\alpha_F = 0$, reducing the Fourier-domain shaping operator to the identity.

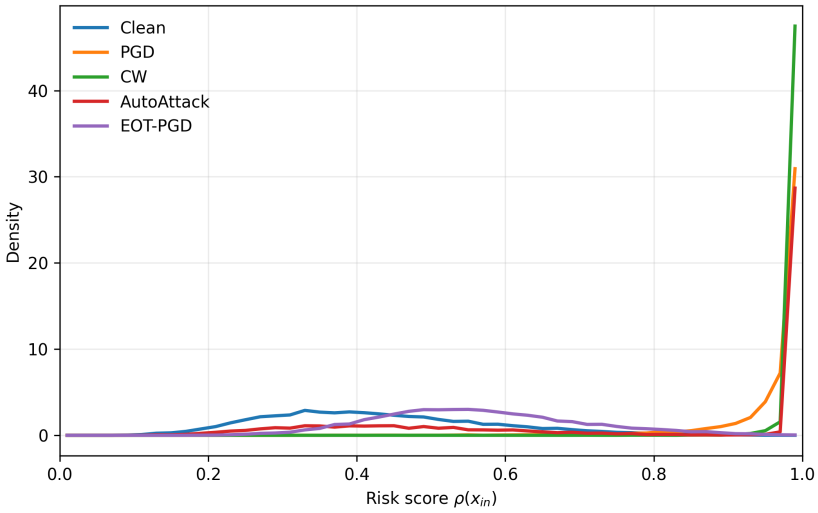


Figure 6: Distribution of the input risk score $\rho(x_{\text{in}})$ on clean and adversarial inputs. Standard attacks concentrate near high risk, whereas adaptive EOT-PGD is broader and partially overlaps the clean distribution. The score is therefore used as a soft controller rather than a binary attack detector.

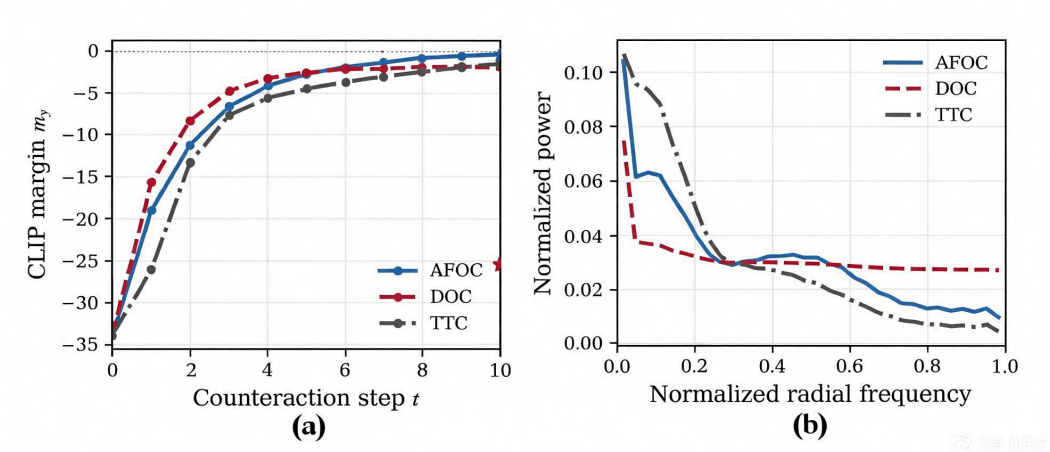


Figure 7: Mechanistic comparison on CIFAR-100 with CLIP ViT-B/32. **Left:** supervised CLIP alignment margin m_y along the label-free counteraction trajectory; the dotted line is the zero-margin decision boundary. **Right:** normalized radial power spectrum of the final counteraction perturbation δ_{ca}. AFOC exhibits sustained margin recovery and a structured mid-to-high-frequency allocation.

fixed controller. Replace $(\lambda, \mu, \alpha_F, w)$ with the constants $(0.25, 0.6, 0.25, 0.5)$, obtained by evaluating Eq. (6) at $\rho = 0.5$ under the default ranges in Table 3. Substituting those ranges yields $\lambda = \lambda_{\min} + \frac{1}{2}(\lambda_{\max} - \lambda_{\min}) = 0.25$, $\mu = \mu_{\max} - \frac{1}{2}(\mu_{\max} - \mu_{\min}) = 0.6$, $\alpha_F = \alpha_{\min} + \frac{1}{2}(\alpha_{\max} - \alpha_{\min}) = 0.25$, and $w = \sigma\big(\gamma(\rho - \tau_\rho)\big) = \sigma(0) = 0.5$. These constants freeze the default controller at a single operating point; they are not obtained by searching over fixed controllers. Risk estimation, the risk-scaled warm start, and the T-step counteraction loop are otherwise unchanged. Table 2 reports the resulting accuracies.

F.2 Warm-Start Ablation

We isolate two roles of the initialization in Eq. (7): escaping the stationary point at zero perturbation and adapting the initial search radius to the estimated input risk. The fixed-scale control uses $\delta_0 = \Pi_\Delta(s_{\text{init}} u_0)$, whereas the risk-scaled variants use $\delta_0 = \Pi_\Delta(s_{\text{init}} \rho u_0)$. All variants retain

Table 13: Warm-start ablation on CIFAR-100 with CLIP ViT-B/32. Mixed Avg. is computed over PGD, AA, and adaptive EOT-PGD. The fixed-scale control removes the factor ρ from Eq. (7); all other components are unchanged. Bold and underlined values denote the best and second-best result in each column, respectively.

Initialization	s_{init}	Clean	PGD	AA	EOT-PGD	Mixed Avg.
Zero initialization	0	<u>56.74</u>	42.97	44.15	**33.39**	<u>40.17</u>
Fixed-scale warm start	0.5	**58.69**	43.29	43.52	20.33	35.71
Risk-scaled warm start	0.25	56.59	42.46	44.54	<u>31.91</u>	39.64
Risk-scaled warm start (default)	0.5	56.55	<u>44.56</u>	<u>45.41</u>	30.18	40.05
Risk-scaled warm start	1.0	56.27	**47.29**	**46.25**	28.28	**40.61**

the remaining AFOC pipeline and use matched random seeds, so the comparison changes only the initialization rule and its scale.

A fixed-scale random start is brittle under adaptive evaluation. The fixed-scale warm start attains the highest clean accuracy (58.69%) and standard-attack performance comparable to that of the weaker initializations, but its adaptive EOT-PGD accuracy drops to 20.33%. This is 13.06 percentage points below zero initialization and 9.85 percentage points below the default risk-scaled variant, resulting in the lowest attack average (35.71%). Thus, a generic nonzero random start is not sufficient to obtain a balanced defense; relative to this fixed-scale start, modulating the warm-start magnitude by ρ improves the measured robustness of the full stochastic defense pipeline, particularly under adaptive EOT-PGD.

The warm-start scale induces a standard–adaptive robustness trade-off. Within the risk-scaled variants, increasing s_{init} from 0.25 to 1.0 improves PGD from 42.46% to 47.29% and AA from 44.54% to 46.25%, while clean accuracy changes by only -0.32 percentage points. The same change reduces EOT-PGD from 31.91% to 28.28%. Consequently, the attack average increases from 39.64% to 40.61%, but the gain is not uniform across attacks. A stronger initial displacement shifts robustness toward PGD and AA at the expense of part of the adaptive-attack performance.

The default is a moderate operating point rather than the maximum-average setting. Relative to $s_{\text{init}} = 0.25$, the default scale of 0.5 gains 2.10 percentage points under PGD and 0.87 percentage points under AA, with only a 0.04-percentage-point change in clean accuracy, although EOT-PGD decreases by 1.73 percentage points. Compared with $s_{\text{init}} = 1.0$, it preserves an additional 1.90 percentage points under EOT-PGD and 0.28 percentage points in clean accuracy, while giving up 0.56 percentage points in Mixed Avg. The competitive zero-initialization result further indicates that the subsequent stochastic search can partially compensate for the stationary primary gradient, but its PGD and AA accuracies remain below those of the stronger risk-scaled warm starts. Overall, Table 13 supports risk-scaled initialization as a calibration mechanism and motivates $s_{\text{init}} = 0.5$ as an intermediate risk-scaled default rather than a universally optimal setting.

F.3 Risk Score and Controller Behavior

With the defaults in Table 3, the controller has the explicit form

$$\lambda(\rho) = 0.5\rho, \quad \mu(\rho) = 0.9 - 0.6\rho,$$
$$\alpha_F(\rho) = 0.5\rho, \quad w(\rho) = \sigma\big(8(\rho - 0.5)\big).$$

Thus, λ and α_F increase linearly from 0 to 0.5, μ decreases from 0.9 to 0.3, and the gate changes smoothly from approximately 0.018 at $\rho = 0$ to 0.5 at $\rho = 0.5$ and 0.982 at $\rho = 1$. Figure 8 visualizes these mappings and shows that the controller has no hard clean/adversarial threshold.

F.4 Update Spectrum Analysis

Figure 9 reports the normalized radial spectrum of the *update directions*, before the full trajectory is accumulated and the final risk gate is applied. DOC and TTC have larger low-frequency peaks, whereas AFOC has a lower low-band peak, comparable mid-band energy, and a consistently higher tail in the high-frequency region. This supports the intended frequency-selective reallocation of the update direction. It does not contradict Figure 7, which measures the *final perturbation*: momentum, projection, repeated updates, and the gate can attenuate the extreme-frequency tail by the end of the trajectory.

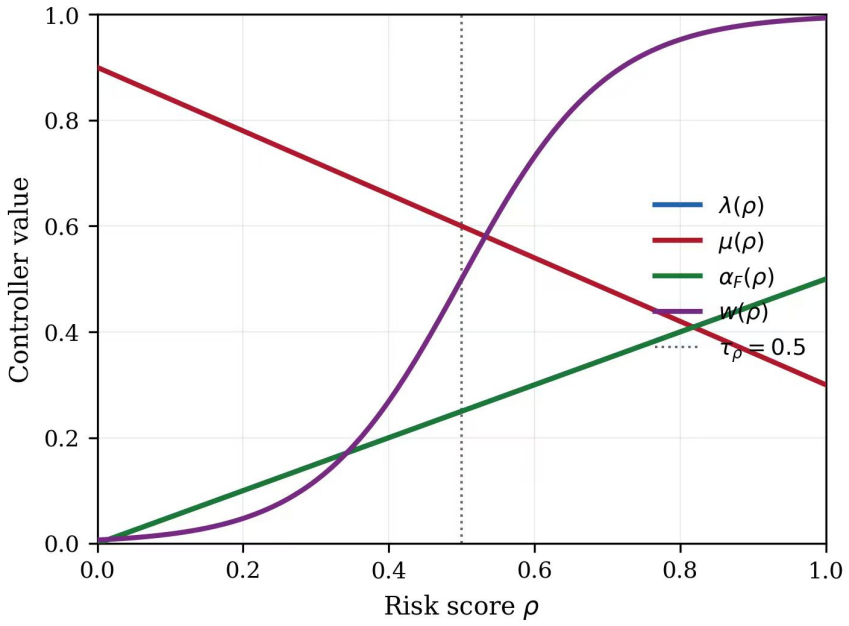


Figure 8: Controller values as functions of the risk score ρ. Increasing risk strengthens orthogonal exploration, frequency-domain shaping, and final gating, while reducing momentum inertia.

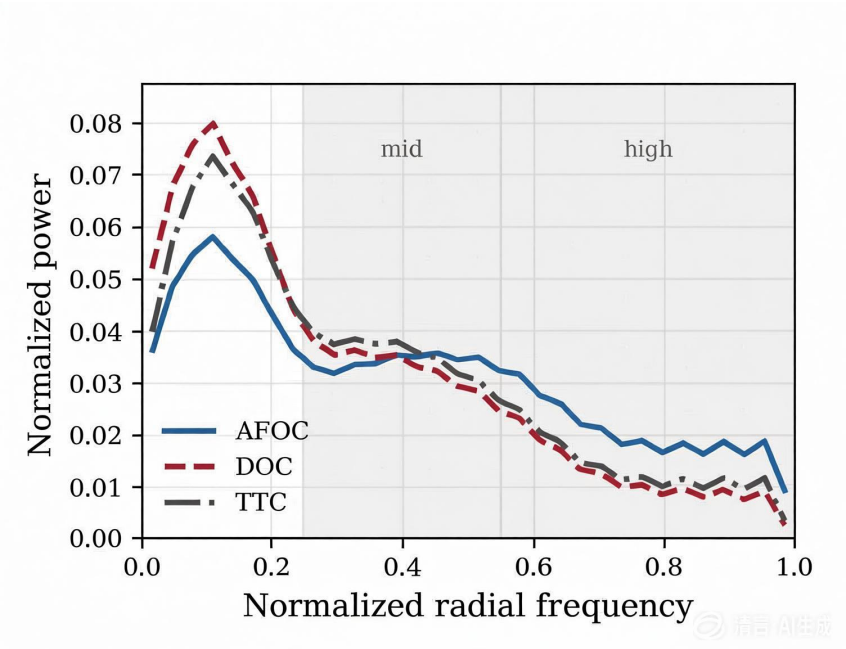


Figure 9: Normalized radial spectrum of defensive update directions. Relative to TTC and DOC, AFOC shifts energy away from the lowest-frequency peak and retains more energy in the high-frequency tail.

G Sensitivity and Efficiency Analysis

Figure 10 reveals two qualitatively different regimes. Optimization depth and the reachable correction set have the largest effect: robust accuracy rises from below the plotted range for $T \leq 3$ to roughly 20%, 36%, and 51% at $T = 5$, 10, and 20, respectively. Most of the corresponding clean-accuracy decrease occurs within the first five steps, after which it stays close to 52%. Varying ϵ_{ca} produces a

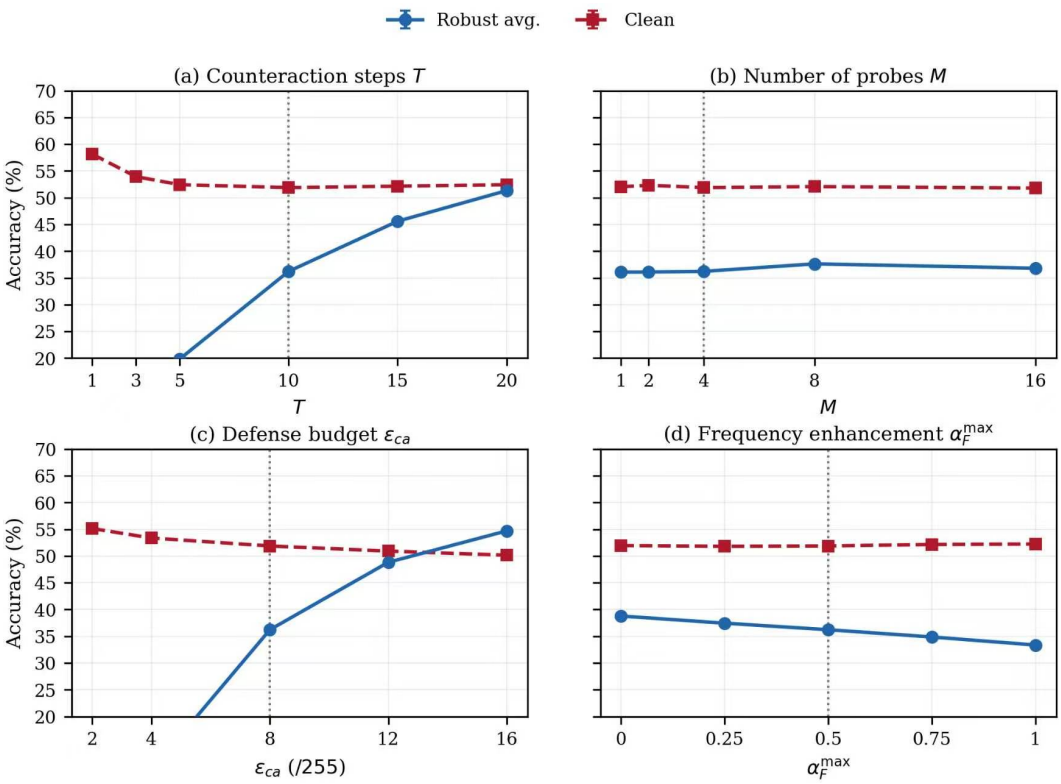


Figure 10: Sensitivity of clean and robust accuracy to the counteraction steps T, number of risk probes M, defense budget ϵ_{ca}, and maximum frequency-enhancement strength $\alpha_F^{\max}$ on CIFAR-100. Each panel varies one hyperparameter while holding the others fixed; dotted vertical lines mark the reported configuration. Robust avg. denotes the attack average used for this sweep and is comparable only within each panel.

similar frontier. The robust average is below 20% at the two smallest budgets, reaches about 36% at $8/255$, and increases to approximately 55% at $16/255$, while clean accuracy decreases gradually from about 55% to 50%. Thus, $T = 10$ and the matched budget $\epsilon_{\text{ca}} = \epsilon_{\text{atk}} = 8/255$ occupy an intermediate accuracy–cost operating point; moving farther along either axis buys robustness through additional computation or a larger admissible correction.

By contrast, performance is insensitive to the number of probes over the tested range: both curves change little for $M \in \{1, 2, 4, 8, 16\}$, making $M = 4$ a reasonable compromise between estimation cost and accuracy. The frequency sweep shows a different trend. Clean accuracy remains nearly constant, whereas the robust average decreases smoothly from about 39% at $\alpha_F^{\max} = 0$ to about 34% at $\alpha_F^{\max} = 1$. The auxiliary sweep therefore does not suggest that stronger spectral amplification is uniformly beneficial. We retain the intermediate value $\alpha_F^{\max} = 0.5$ because the matched three-attack component ablation in Table 2 favors including frequency-domain shaping. Taken together, the two results suggest that its contribution depends on attack composition and that aggressive amplification is unnecessary. We accordingly use this figure to assess local sensitivity, while reserving the four-attack Mixed Avg. in Table 1 for model comparison.

Measurement protocol. We benchmark all methods with the same frozen CLIP ViT-B/32 model, image preprocessing, batch size, and mixed-precision setting. The implementation uses 512 CIFAR-100 test images with batch size 16, performs three warm-up batches, and reports synchronized CUDA wall-clock time normalized by the number of images. Data loading and host-to-device transfer are outside the timed region. Peak allocated memory is measured after resetting CUDA peak-memory statistics. The benchmark specifies the timing dataset, sample count, and batch size above. Because absolute latency also depends on the accelerator and software stack, the relative operation counts provide the more portable efficiency comparison. The dataset, sample count, batch size, precision

Table 14: Inference overhead on CLIP ViT-B/32 under the matched hardware setup. Time is measured per image and includes risk estimation, counteraction, and the final zero-shot classification. All AFOC configurations use the default risk-scaled initialization with $s_{\text{init}} = 0.5$.

Method	Extra updates	Time / image	Memory
Vanilla CLIP	–	0.857 ms	0.745 GB
TTC	10	5.415 ms	1.086 GB
DOC	4	2.891 ms	1.089 GB
AFOC	10	6.801 ms	1.107 GB

Table 15: Detailed inference overhead and exploratory efficiency variants on CLIP ViT-B/32. The first four rows use the matched full configurations reported in Table 14; the remaining rows diagnose implementation and accuracy–efficiency trade-offs. All AFOC rows use risk-scaled warm initialization with $s_{\text{init}} = 0.5$.

Method	Forwards	Backwards	Time / image	Memory
Vanilla CLIP	1.00	0.00	0.857 ms	0.745 GB
TTC	12.00	10.00	5.415 ms	1.086 GB
DOC	6.00	4.00	2.891 ms	1.089 GB
AFOC	17.00	10.00	6.801 ms	1.107 GB
AFOC+FusedRisk	17.00	10.00	6.537 ms	1.107 GB
AFOC-Lite	10.00	5.00	3.866 ms	1.104 GB
AFOC-Optimized	11.83	4.83	6.796 ms	5.748 GB

mode, and number of repetitions are stored with each run because these details materially affect absolute latency and memory.

The overhead is primarily computational. Table 14 uses the same measurements as the detailed accounting in Appendix H. Full AFOC requires 6.801 ms per image, which is 25.6% slower than TTC, $2.35\times$ the latency of DOC, and $7.94\times$ that of vanilla CLIP. The operation counts help explain the difference: AFOC uses 17 image-encoder forward and 10 backward evaluations, compared with 12/10 for TTC and 6/4 for DOC. Its peak memory is 1.107 GB, only 0.021 GB above TTC and 0.018 GB above DOC. Thus, the principal overhead is iterative computation rather than model storage: AFOC introduces no trainable defense network or prompt parameters, but its risk probes and gradient-based counteraction increase the image-encoder workload.

H Efficiency Details

Table 15 gives the operation counts and the measurements used in Table 14. Time includes risk estimation, all counteraction iterations, and final classification. Forward-pass counts represent sample-level image-encoder evaluations; consequently, batching several risk views into one encoder call reduces launch overhead but does not change the sample-level forward count.

Implementation and budget variants. The full AFOC row uses the reported $T = 10$ counteraction steps and $M = 4$ risk probes. AFOC+FusedRisk preserves the same estimator and update budget but encodes risk views in micro-batches. Its sample-level counts therefore remain 17/10, while latency decreases from 6.801 to 6.537 ms (3.9%) with unchanged peak memory. This isolates a modest implementation-level gain without reducing the algorithmic budget.

AFOC-Lite jointly reduces the budget to $T = 5$ and $M = 2$. It requires 10 forward passes and 5 backward passes and lowers latency to 3.866 ms, a 43.2% reduction relative to full AFOC, while peak memory changes only from 1.107 to 1.104 GB. This row is an explicit accuracy–efficiency operating point rather than the configuration used for the main robustness results; its robustness must therefore be reported separately before drawing Pareto conclusions.

AFOC-Optimized is an exploratory dynamic-computation candidate that combines fused risk evaluation with risk-gated $0/5/10$ update budgets. Its average operation counts decrease to 11.83 forward passes and 4.83 backward passes, but the current implementation provides essentially no latency gain (6.796 versus 6.801 ms) and raises peak memory to 5.748 GB. The measurements suggest

that active-sample compaction, indexing, and temporary fused-view tensors offset the savings from skipped updates at batch size 16. We therefore do not use this row as the paper's primary AFOC result. It instead exposes an implementation bottleneck: reducing theoretical operations is insufficient unless dynamic batch construction and temporary allocations are also controlled.

Overall, the first four rows quantify the cost of the reported full-accuracy configuration, whereas the additional rows form an exploratory implementation study. Fused risk batching provides a modest speedup, and the Lite budget provides the largest measured latency reduction; the current dynamic variant requires further memory-aware engineering and matched robustness evaluation.