BAPS: BOUNDARY-ANCHORED PROBABILISTIC SEARCH FOR RELIABLE DECISION-BASED ATTACKS AGAINST INPUT-NOISE DEFENSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized input noise causes identical queries to return inconsistent hard labels across repeated calls, making both boundary localization and boundary point selection more difficult. We propose BAPS (Boundary-Anchored Probabilistic Search), which maintains a boundary anchor separately from a set of reliable candidates under the radius constraint. Sequential paired comparisons then allocate raw model queries to resolve uncertain boundary decisions. Once a candidate is selected, it is fixed and tested using 200 fresh draws of defense randomness. Strict attack success requires both $\|x^* - x_0\|_2 \leq \epsilon$ and a one-sided 95% lower confidence bound on the attack probability of at least 0.9. Across nine CIFAR-10 noise conditions, BAPS achieves an overall strict attack success rate (ASR) of 67.7%, compared with 61.4% for a radius-constrained adaptation of PopSkipJump (PSJ). On ImageNet with ResNet50, BAPS achieves macro-average strict ASRs of 37.6%, 54.6%, 75.1%, 87.4%, and 95.1% at $\epsilon \in \{5, 10, 20, 30, 40\}$, respectively. Under the pre-specified analysis plan, all 90 paired comparisons of BAPS against SurFree and CGBA remain significant after multiple-comparison correction. At fixed $\epsilon = 5$, BAPS's ASR increases from 16.7% to 50.1% as the query budget increases from 5K to 50K. Without model-specific tuning, the same ImageNet configuration achieves 31.8% ASR on ConvNeXt-Tiny at $\epsilon = 5$ and $Q = 20{,}000$, compared with 10.8% for CGBA, the strongest baseline evaluated in this setting. Under the evaluated protocols, BAPS improves strict ASR over adapted baselines while maintaining independent final verification.

1 INTRODUCTION

Decision-based black-box attacks can only observe the class labels returned by the model. The attacker needs to infer the local decision boundary through multiple queries and search for adversarial examples under perturbation constraints. Existing methods exploit boundary movement, local direction estimation, or geometric search paths on deterministic hard-label models (Brendel et al., 2018; Chen et al., 2020; Cheng et al., 2020; Rahmati et al., 2020; Li et al., 2020; Chen & Gu, 2020; Wang et al., 2022). However, when the defense resamples input noise before each prediction, the same query point may return different labels. A single label then cannot stably indicate which side of the boundary the point lies on: a hard-label flip caused by random input noise may mislead subsequent geometric updates, and it cannot guarantee that the final sample will still succeed under the next random sampling.

We study the fixed-radius reliable attack when the model has a randomized small-noise defense. Given a clean image x_0 and its original label y_0, let $p_{\mathrm{adv}}(x)$ denote the probability that candidate x is predicted as non-y_0 under the defense randomness. The attacker, with query budget Q and radius constraint $\|x - x_0\|_2 \leq \epsilon$, seeks a candidate with high $p_{\mathrm{adv}}(x)$. Here we need to distinguish two types of information: the region near $p_{\mathrm{adv}} \approx 0.5$ can provide local directional information, but it is not reliable by itself; the final candidate needs a sufficiently high attack probability within the radius, but it may not remain near the boundary. Simply repeating queries cannot determine which uncertain comparisons are worth further sampling, nor can it eliminate the verification bias caused by adaptive selection during the search.

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.

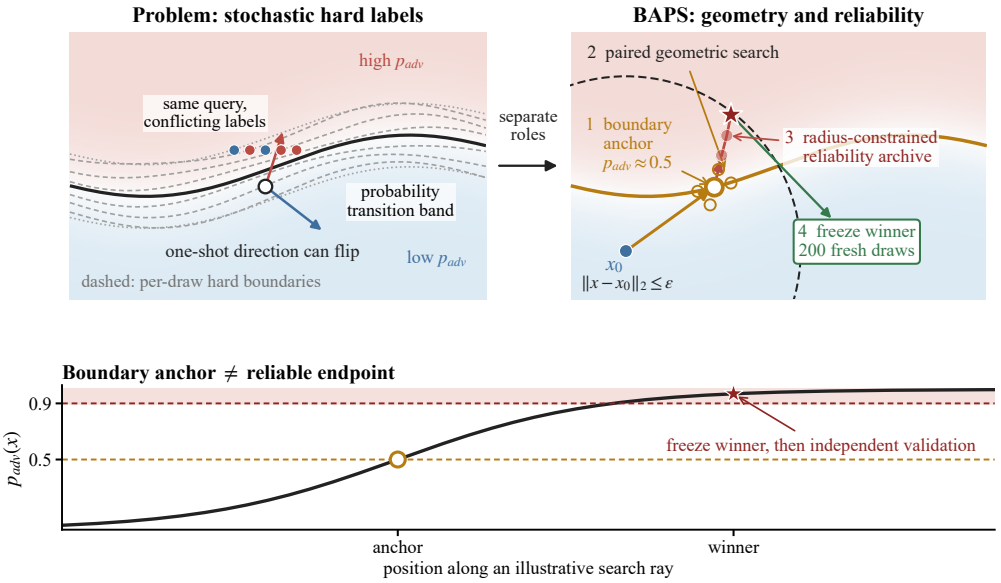

Figure 1: Schematic illustration of Boundary-Anchored Probabilistic Search under stochastic hard labels. Top left: resampling input noise at each query can cause the same location to return opposite labels, forming an observed probability transition band. Top right: BAPS uses probability boundary anchors to obtain local geometric information while maintaining a reliable candidate set within the perturbation radius. Bottom: approaching the probability boundary and achieving high reliability are different goals; the candidate selected by the search is frozen and then subjected to final verification with 200 independent fresh random samples.

Existing research provides two foundations. Sign-OPT (Cheng et al., 2020), SurFree (Maho et al., 2021), and CGBA (Reza et al., 2023) exploit directional or boundary geometry under deterministic hard labels, while PopSkipJump (PSJ) extends decision-based attacks to randomized classifiers. Handling random labels is therefore not new; the open question is how to separate geometry estimation from final candidate verification when labels are unstable. Figure 1 illustrates the design motivation; the paths and probability shapes in the figure are only for explaining the mechanism.

We propose BAPS (Boundary-Anchored Probabilistic Search). Under a limited query budget, BAPS estimates probability boundary anchors by performing sequential pairwise queries along opposite directions around a candidate anchor, and aggregates empirical attack frequency differences into a normal-direction estimate. Updates along effective directions are projected into the allowed radius to update the candidate set. Search-phase evaluations guide anchor updates, construct local directions, and rank candidates, but they are not used as the success criterion. After the search ends, the selected candidate is frozen and verified on 200 fresh defense samples. Strict success requires the radius constraint and a one-sided 95% confidence lower bound on the attack probability of at least 0.9. All conclusions in this paper are based on this independent final verification.

The main contributions of this paper are as follows.

- We formulate stochastic hard-label attacks as a search problem under query budget and perturbation radius constraints, with the objective of maximizing attack success probability, and distinguish local direction estimation during the search phase from independent verification of the final candidate.

- We propose BAPS, which separately maintains probability boundary anchors and a reliable candidate set, and provides query budget allocation and counting rules, decoupling geometric search from statistical verification of the final candidate.

- Under fixed CIFAR-10 and ImageNet evaluation protocols, we compare results across different noise, radius, budget, and network architectures, and report a complete comparison with SurFree, CGBA, and a radius-constrained adapted version of PSJ.

Under the evaluated protocols, BAPS achieves higher strict ASR than the adapted baselines.

2 Related Work

Geometric Search for Deterministic Hard-Label Attacks. Boundary Attack (Brendel et al., 2018) starts from an existing adversarial example and gradually reduces the perturbation near the boundary. HopSkipJumpAttack and GeoDA estimate the local normal direction through decision-based queries (Chen et al., 2020; Rahmati et al., 2020). Sign-OPT takes the boundary distance along a direction as the optimization objective and uses label feedback to reduce the number of queries for distance search (Ilyas et al., 2019; Guo et al., 2019; Li et al., 2020; Wang et al., 2022).QEBA reduces gradient-estimation dimension with representative subspaces, RayS performs discrete ray search, and Triangle Attack exploits low-frequency geometric moves (Chen & Gu, 2020). SurFree avoids explicit gradient estimation and performs geometric search in a restricted subspace, while CGBA designs a two-dimensional path search based on boundary curvature (Maho et al., 2021; Reza et al., 2023). These methods provide an effective geometric search framework for deterministic hard-label attacks. However, under random input noise, a single label cannot stably reflect the boundary direction, so boundary direction estimation and final candidate verification need to be handled separately. For fair comparison, we adopt adapted versions of SurFree and CGBA under a fixed evaluation protocol, with unified query counting and independent final verification.

Randomized Defenses and Stochastic Hard-Label Attacks. Small Noise Defense (SND) adds new input noise at inference time, showing that lightweight randomness can already interfere with query processes that rely on stable labels (Byun et al., 2022). More broadly, expectation over transformations attacks randomized transformations, evaluations of obfuscated-gradient defenses emphasize adaptive treatment of test-time randomness, and randomized smoothing formalizes class probabilities under injected noise (Athalye et al., 2018b;a; Cohen et al., 2019).PopSkipJump (PSJ) explicitly attacks probabilistic classifiers and adapts HopSkipJumpAttack's query allocation to uncertain labels (Simon-Gabriel et al., 2021). Thus stochastic hard labels and repeated sampling are not new to this paper. BAPS instead separates a probability-boundary anchor from a radius-constrained candidate archive, followed by independent validation of one frozen output.

Adaptive Sampling and Reliability Verification. Confidence sequences permit optional stopping for a fixed local comparison with fresh independent observations (Howard et al., 2021). However, BAPS uses them only for local pairwise comparisons during the search process; they do not simultaneously certify all adaptively explored candidates. More generally, adaptive data analysis warns against repeatedly reusing evaluation data after data-dependent selection (Dwork et al., 2015). After freezing the final candidate, we therefore compute a one-sided 95% Clopper–Pearson lower bound from independent validation draws (Clopper & Pearson, 1934), denoted by L_{CP}, as the criterion for strict success. This bound applies to that selected candidate, conditional on the search history, rather than to the search path as a whole.

3 Method

3.1 Stochastic Hard-Label Attack Objective

Let $f : [0,1]^d \to \{1, \dots, K\}$ be an undefended deterministic classifier, and let x_0 be a clean image correctly classified by f, with $y_0 = f(x_0)$. The input randomization defense independently samples noise $\eta \sim \mathcal{D}_\sigma$ before each prediction. When the attacker queries x, the observed label is

$$g(x;\eta) = f\big(\Pi_{[0,1]^d}(x+\eta)\big)\,. \tag{1}$$

Here $\Pi_{[0,1]^d}(\cdot)$ denotes element-wise clipping to $[0,1]$. Therefore, the returned label for the same x may change across queries. We consider untargeted attacks, and define $Z(x;\eta) = \mathbb{1}[g(x;\eta) \neq y_0]$.

The attack success probability of candidate x is defined as

$$p_{\mathrm{adv}}(x) = \Pr_{\eta \sim \mathcal{D}_\sigma}\big[g(x;\eta) \neq y_0\big] = \mathbb{E}_\eta\big[Z(x;\eta)\big]. \tag{2}$$

We consider untargeted attacks under query budget Q and allowed radius ϵ: find a candidate with high attack success probability $p_{\mathrm{adv}}(x)$ within $\|x - x_0\|_2 \leq \epsilon$ and $x \in [0,1]^d$. The true optimization objective under a fixed radius is

$$x^*_\epsilon \in \underset{x \in [0,1]^d,\, \|x - x_0\|_2 \leq \epsilon}{\arg\max} p_{\mathrm{adv}}(x). \tag{3}$$

Equation equation 3 gives the true optimization objective; the confidence lower bound obtained from a finite number of queries is used only for candidate ranking and final decision, and does not replace the true attack success probability. The query budget is counted as follows: each call to the defended model g counts as one query and corresponds to one forward pass of the original model f; each sample in a batch, warm-start initialization, failed proposal, repeated sampling, and final verification are all counted into Q.

After the search ends, we adopt a strict success criterion independent of the search phase. The candidate x^* is first fixed, and then the defended model g is called $n_f = 200$ times using defense randomness independent of the search phase. Strict success is declared only when $x^* \in \mathcal{X}_\epsilon$ and the one-sided 95% Clopper–Pearson confidence lower bound $L_{\mathrm{CP}}(x^*) \geq \tau = 0.9$ for its attack success probability, where $\mathcal{X}_\epsilon = \{x \in [0,1]^d : \|x - x_0\|_2 \leq \epsilon\}$. This criterion requires the attack success probability to reach a given confidence lower bound, rather than relying on a single query result. We focus on high-success-probability candidates under a given budget, and do not guarantee global optimality of equation 3.

3.2 Core Design and Algorithm Overview

Random labels lead to two different search goals: locations near $p_{\mathrm{adv}} \approx 0.5$ usually provide the probability differences needed for local directions, while the final candidate must achieve a higher attack success probability within the radius. A point with high attack success probability but nearly constant local observations is not necessarily suitable for estimating the boundary direction; a boundary point with a strong local probability difference does not necessarily satisfy the final verification threshold $\tau = 0.9$. This distinction motivates BAPS to separate boundary direction estimation from final candidate selection, thereby constructing search directions using only local probability comparisons.

BAPS separately maintains probability boundary anchors a_t and a candidate set $\mathcal{A}_t$ within the radius. Anchors are used for sequential pairwise comparisons and geometric updates, while the candidate set stores candidates that satisfy reliability requirements. Each iteration consists of four steps: local direction estimation, tangential proposal generation, ray recalibration, and ϵ-shell candidate evaluation. After the search ends, only one candidate is selected and fixed from the candidate set, and then subjected to independent final verification. Algorithm 1 gives the overall procedure, and the following subsections describe each step and its statistical properties.

3.3 Boundary Anchoring and Sequential Pairwise Geometric Search

Warm-start Initialization and Probability Boundary Calibration. The implementation uses SurFree (Maho et al., 2021) and CGBA (Reza et al., 2023) to generate warm-start initialization directions, and the related queries are counted into the same initialization query budget as the baselines; this budget does not exceed $\max\{300, \lfloor 0.15Q_{\mathrm{search}} \rfloor\}$. For a normalized direction ω, define the ray $x(r;\omega) = \Pi_{[0,1]^d}(x_0 + r\omega)$. The radius r is expanded at most four times, with the update $r \leftarrow \min\{\sqrt{d}, 1.5r + 0.05\}$, until the empirical attack probability exceeds 0.5; then five rounds of binary search are performed between the upper and lower bounds of r, with each midpoint using five queries to estimate the attack probability. The resulting anchor approximates the empirical probability boundary.

Local Probability Comparison. At iteration t, let $r_t = (a_t - x_0)/\|a_t - x_0\|_2$. Let $u_{t,1}, \ldots, u_{t,m}$ be the set of directions used in this iteration, where $m = 16$, including r_t and 15 low-frequency DCT

Algorithm 1 BAPS at native radius ϵ

Require: $x_0, y_0, Q, \epsilon, \tau = 0.9, n_f = 200$
Set $Q_{\text{search}} \leftarrow Q - n_f$; initialize query counter $q \leftarrow 0$
Obtain SurFree/CGBA warm directions within the shared allowance
Calibrate ray anchors to $p^{\star}_{\text{adv}} \approx 0.5$; initialize shell archive $\mathcal{A}$
while $q < Q_{\text{search}}$ and the next atomic operation fits the budget **do**
  Build one radial and 15 multiscale DCT directions
  Estimate paired contrasts with 3–7 fresh draws per side
  Form momentum normal $\hat{n}_t$ and tangent v_t
  for $\lambda \in \{0.5, 1, 2\}$ while budget permits **do**
    Propose $a_t + \lambda s_t v_t$; recalibrate its probability-boundary ray
  end for
  Apply probability/radius gates and update anchor
  Map accepted directions to the ϵ-shell; screen and update $\mathcal{A}$
end while
Freeze $x^* \leftarrow \arg\max_{x \in \mathcal{A}} L_{\text{CP}}(x)$
Query x^* on 200 fresh draws and compute Eq. equation 7

return strict success iff $\|x^* - x_0\|_2 \le \epsilon$ and $L_{\text{CP}} \ge \tau$

basis directions whose spatial scales cycle through $\{8, 16, 32\}$. Low-frequency and representative-subspace priors have previously improved query efficiency under score and deterministic hard-label feedback (Ilyas et al., 2019; Guo et al., 2019; Li et al., 2020; Wang et al., 2022); BAPS uses them only as direction proposals inside paired probability comparisons. The offset is fixed at $h = 0.5$, and we draw independent defense random samples at $a_t \pm h u_{t,j}$. For each direction j, each side performs between 3 and 7 queries. For a fixed query location, let the number of queries at stopping be n, with S_n successful attacks. A Jeffreys mixture prior is used to construct a confidence sequence:

$$M_n(q) = \frac{B(S_n + \frac{1}{2}, n - S_n + \frac{1}{2})}{B(\frac{1}{2}, \frac{1}{2}) q^{S_n} (1-q)^{n - S_n}}, \qquad \mathcal{C}_n = \{q : M_n(q) < 1/\alpha\}. \tag{4}$$

Here $B(\cdot, \cdot)$ is the Beta function, and $\alpha = 0.025$ is the one-sided significance level. Confidence sets $\mathcal{C}_n^+$ and $\mathcal{C}_n^-$ are computed for the two sides; if they are disjoint, the pairwise sampling is stopped early; otherwise it stops at the query cap. After stopping, let $\widehat{p}(x)$ denote the empirical estimate of $p_{\text{adv}}(x)$ obtained from the independent queries. Then

$$\begin{aligned} \widehat{\Delta}_{t,j} &= \widehat{p}(a_t + h u_{t,j}) - \widehat{p}(a_t - h u_{t,j}), \\ \widehat{n}_t &= \text{norm}\left(\sum_{j=1}^{m} \widehat{\Delta}_{t,j} u_{t,j} \right). \end{aligned} \tag{5}$$

Here $\text{norm}(v) = v / \|v\|_2$ ($v \neq 0$). Equation equation 5 gives an empirical quantity used to construct search directions. The normal estimate uses momentum updates with coefficient $\beta = 0.6$; if the estimate degenerates, it falls back to the radial direction r_t. Because new independent observations are used after stopping, the sequential stopping rule preserves the coverage probability of a single local comparison at any stopping time.

Tangential Proposals and Anchor Updates. The tangential direction pointing to the source is

$$v_t = \text{norm}(-r_t - \langle -r_t, \widehat{n}_t \rangle \widehat{n}_t). \tag{6}$$

The algorithm tests candidate points $a_t + \lambda s_t v_t$ for $\lambda \in \{0.5, 1, 2\}$, where $s_t = \max\{\epsilon, \|a_t - x_0\|_2\}/\sqrt{t+1}$ is the step size. For each tested candidate, its ray is recalibrated to the empirical probability boundary; if the empirical attack probability is below 0.35, the candidate is discarded. Anchor updates also require the candidate radius not to exceed 1.02 times the current anchor radius. If the normal estimate degenerates twice consecutively, it falls back to a random direction orthogonal to the current direction. These rules are a concrete implementation of the search strategy in this paper.

3.4 Reliable Candidate Set and Independent Verification

Each acceptable candidate direction is mapped to the sphere of radius ϵ, yielding $x_t^{\text{shell}} = \Pi_{[0,1]^d}(x_0 + \epsilon d_t)$, where d_t is the normalized candidate direction. Candidate points on this sphere are first screened using 7 queries to estimate the empirical attack probability; if the empirical attack probability is at least $\max\{0.5, \widehat{p}_{\text{best}} - 0.1\}$, where $\widehat{p}_{\text{best}}$ is the best empirical attack probability in the current candidate set, then 31 more queries are used to estimate the empirical attack probability, and the one-sided 95% Clopper–Pearson confidence lower bound is used as the ranking criterion. This ranking criterion is used for candidate ranking during the search phase; final verification is performed independently after the candidate is fixed. Only candidates satisfying the radius constraint can become strict success candidates.

After the selection in the search phase ends, one candidate x^* is fixed and no longer modified based on final verification samples. Let S_f be the number of successes in $n_f = 200$ independent fresh random samples. The one-sided 95% Clopper–Pearson confidence lower bound is defined as

$$L_{\text{CP}}(x^*) = \begin{cases} 0, & S_f = 0, \\ \text{Beta}^{-1}(0.05; S_f, n_f - S_f + 1), & S_f > 0, \end{cases} \tag{7}$$

where $\text{Beta}^{-1}(p; \alpha, \beta)$ denotes the p-quantile of the $\text{Beta}(\alpha, \beta)$ distribution. When $\tau = 0.9$ and $n_f = 200$, at least 188 successes are required to satisfy $L_{\text{CP}}(x^*) \geq \tau$. Since x^* is fixed before final verification and the verification samples are independent of all search observations, this lower bound has at least 95% one-sided coverage probability for $p_{\text{adv}}(x^*)$. The total number of queries satisfies $q_{\text{used}} = q_{\text{search}} + n_f \leq Q$; the algorithm may terminate early if the budget is not exhausted.

The above coverage guarantee applies to local direction estimation and independent verification of the final candidate.

4 Experiments

This section reports the complete-method comparison results under the fixed evaluation protocol, focusing on the difference between BAPS and the adapted baselines under the unified strict success criterion, as well as how this difference varies with radius and query budget. Additional network architecture validation and the scope of the current conclusions are discussed at the end.

4.1 Fixed Evaluation Protocol and Comparison Settings

Data, Randomized Defense, and Query Budget. Experiments restrict clean image pixels to $[0, 1]$. At each query, the defense independently applies Gaussian noise $\mathcal{N}(0, \sigma^2)$, uniform noise $\mathcal{U}[-\sqrt{3}\sigma, \sqrt{3}\sigma]$, or Laplace noise with scale parameter $\sigma/\sqrt{2}$, where $\sigma \in \{0.001, 0.005, 0.01\}$. The per-pixel standard deviation is σ for all three distributions. Development and final evaluation image lists are separated; all final evaluation conditions are fixed before formal evaluation, and final evaluation results are not used for tuning.

The CIFAR-10 (Krizhevsky, 2009) main protocol uses 100 test images, with 5 attack seeds per image, $\epsilon = 1$, and $Q = 10{,}000$. The ImageNet (Deng et al., 2009) main protocol uses 100 fixed images correctly classified by ResNet50-V2 (He et al., 2016), with 1 attack seed per image; BAPS runs independently for $\epsilon \in \{5, 10, 20, 30, 40\}$, with $Q = 20{,}000$. The query budget supplementary experiment fixes $\epsilon = 5$ and runs independently for $Q \in \{5{,}000, 10{,}000, 20{,}000, 50{,}000\}$. The additional architecture ConvNeXt-Tiny (Liu et al., 2022) uses 100 fixed images that it correctly classifies, follows the ImageNet fixed configuration, and is evaluated at $\epsilon = 5, Q = 20{,}000$; since the two models use different image lists, they do not form cross-architecture paired samples. All queries, including final verification, are counted into Q.

Baseline Adaptation and Comparison Settings. SurFree, CGBA, and PSJ in this paper refer to adapted versions under fixed radius constraints and random queries, and are not direct reproductions of the original implementations. SurFree and CGBA use the same randomized hard-label interface, query counting rules, and 200 independent final verifications. In the multi-radius experiment, BAPS searches independently at each given radius; SurFree and CGBA first perform distance minimization search, and then determine success based on whether the L2 distance of the resulting candidate does

not exceed ϵ. Therefore, this experiment compares results under the unified strict success criterion; the internal optimization objectives differ, but the final criterion is the same.

PSJ uses a fixed commit version of the authors' code, maps the boundary direction obtained at the end of search to the ϵ-shell of the allowed radius, and then performs unified final verification. Its initialization, sampling parameters, iteration cap (32 rounds), and final output criterion all differ from the original paper's settings. This adaptation is documented under a fixed protocol and is not a reproduction of the original implementation. PSJ does not participate in the additional model protocol (ConvNeXt-Tiny) in this paper.

Statistical inference uses images as the unit of analysis; the nine defense conditions and the attack seeds in CIFAR-10 are not treated as mutually independent samples. On ImageNet, differences between each pair of methods are computed using 20,000 image-level paired bootstrap resamples to obtain a 95% empirical confidence interval for each comparison, and a two-sided exact paired test is used for discordant image pairs. The 90 pre-fixed comparisons against SurFree/CGBA on ResNet50 use the same multiple comparison correction family, while the 18 comparisons on ConvNeXt use a separate family. The nine PSJ-related comparisons on CIFAR-10 retain the cluster structure of five attack seeds within the same image, using image-cluster-level bootstrap resampling and an image-level sign-flip test, with a separate multiple comparison correction. Multiple comparison correction only adjusts p-values; the per-comparison resampling intervals do not thereby become joint confidence intervals holding simultaneously for all comparisons. The macro-averages and curves below are descriptive statistics only.

4.2 Defended Clean Accuracy

On the fixed 1,000 ImageNet clean images, the three noise distributions and three noise scales form nine conditions. The defended average top-1 accuracy is 79.20%–80.27%, and the undefended accuracy is 80.40%. Each condition is independently repeated three times, with a maximum drop of 1.20 percentage points relative to the undefended accuracy. These results only reflect overall clean accuracy. Per-image label agreement rates and the natural random error rate on the clean-correct subset require per-query predictions.

4.3 CIFAR-10 Main Comparison

Under the nine defense conditions on CIFAR-10, BAPS achieves an overall strict ASR of 67.7%, PSJ 61.4%, SurFree 4.7%, and CGBA 9.6% (Table 1). The aggregate BAPS–PSJ difference is $+6.3$ percentage points and is reported descriptively only. Table 1 reports each of the nine noise conditions before their aggregate: BAPS has higher observed strict ASR than PSJ in all six conditions with $\sigma \in \{0.001, 0.005\}$, whereas PSJ is higher in all three conditions with $\sigma = 0.01$, by 7.2, 6.2, and 3.2 percentage points, respectively.

Among the nine BAPS–PSJ paired comparisons, only four remain significant after Holm correction (Holm, 1979). The three $\sigma = 0.01$ differences are all non-significant within the same Holm correction family. Therefore, BAPS does not universally outperform PSJ in per-condition comparisons; the non-significant differences do not imply equivalence.

A separate CIFAR-10 budget experiment shows that the relative performance of the two methods varies with query budget. Under Gaussian noise $\sigma = 0.01$, $\epsilon = 1$, BAPS achieves strict ASR of 8.6%, 25.4%, 42.0% at $Q = 2$K, 5K, and 10K, respectively, while PSJ achieves 2.2%, 18.0%, 49.2%. At $Q = 10$K, PSJ's strict ASR is higher than BAPS, indicating that the ranking is not consistent across all budgets. Each data point is measured independently at its respective budget; the 10K endpoint reuses the corresponding main-comparison cell, and no interpolation is applied at other budgets.

4.4 ImageNet Radius and Query-Budget Sweeps

The radius-sweep block of Table 2 reports nine-condition macro strict ASR at $Q = 20$K. All 45 BAPS native-radius cells were evaluated under the fixed protocol. From $\epsilon = 5$ to 40, BAPS rises from 37.6% to 95.1%; the strongest baseline (CGBA) reaches 8.7% at $\epsilon = 5$ and 75.6% at $\epsilon = 40$, while the adapted PSJ reaches 0.3% and 5.1%. Within each noise condition the same 100 images

Table 1: CIFAR-10 strict ASR (%, higher is better) at $\epsilon = 1$ and $Q = 10$K (100 images, five attack seeds per image). Δ is BAPS minus the radius-constrained PSJ adaptation in percentage points; the overall value is descriptive. All methods use the same 200-draw final verification. Bold marks the largest observed ASR in each row, not statistical significance.

Noise	σ	BAPS	PSJ	SurFree	CGBA	Δ (pp)
Gaussian	0.001	**85.2**	71.8	8.8	16.6	+13.4
	0.005	**76.6**	65.4	3.4	7.8	+11.2
	0.010	42.0	**49.2**	2.2	5.8	−7.2
Uniform	0.001	**83.8**	69.2	8.2	17.2	+14.6
	0.005	**74.8**	64.8	2.6	8.0	+10.0
	0.010	41.6	**47.8**	2.2	4.2	−6.2
Laplace	0.001	**85.6**	72.0	8.6	16.0	+13.6
	0.005	**76.2**	65.6	3.6	6.8	+10.6
	0.010	43.6	**46.8**	2.4	3.8	−3.2
Overall		**67.7**	61.4	4.7	9.6	+6.3

Table 2: ImageNet–ResNet50-V2 strict ASR (%) across radius and raw-query-budget sweeps, macro-averaged over nine noise conditions with the same 100 paired images per condition. Bold indicates the largest observed value in each column. †PSJ denotes the fixed radius-constrained adaptation; see text for protocol details.

	Radius sweep at $Q = 20$K					Budget sweep at $\epsilon = 5$			
Method	$\epsilon = 5$	10	20	30	40	$Q = 5$K	10K	20K	50K
BAPS	**37.6**	**54.6**	**75.1**	**87.4**	**95.1**	**16.7**	**26.4**	**37.6**	**50.1**
SurFree	6.2	14.8	38.3	56.9	75.2	2.6	3.1	6.2	10.6
CGBA	8.7	19.1	40.7	59.1	75.6	2.2	4.4	8.7	21.4
PSJ†	0.3	2.0	2.4	4.0	5.1	0.8	0.6	0.3	0.6

are paired. Across the pre-specified family of 90 image-level BAPS–SurFree/CGBA comparisons, paired differences range from 11 to 50 percentage points; every pointwise 95% paired-bootstrap interval excludes zero, and every exact paired test survives Holm correction (maximum adjusted $p = 0.00836$). These comparisons show an endpoint advantage under the stated objectives.

The budget-sweep block fixes $\epsilon = 5$ and varies the allowed raw-call cap. BAPS increases from 16.7% at 5K to 50.1% at 50K, versus 21.4% for CGBA, 10.6% for SurFree, and 0.6% for the adapted PSJ at 50K. The 20K point reuses the main ImageNet evaluation run; the other budget points are separate runs. The raw-call cap is shared, but realized query counts and wall-clock time may differ. Appendix Figure 3 connects measured radius and budget endpoints only.

PSJ† in Table 2 is the documented, fixed radius-constrained adaptation rather than an optimized reproduction of the original method. All nine BAPS–PSJ paired tests at $\epsilon = 5$, $Q = 20$K survive a separate Holm family (maximum adjusted $p = 3.81 \times 10^{-6}$). However, 691/900 PSJ runs reached the 32-iteration cap at 20K, with mean raw use 14.7K among capped runs; initialization and the mapping of the official direction to an ϵ-shell candidate also affect comparability. We therefore report this comparator as an adapted baseline and do not interpret its ASR as the performance of the original PSJ protocol.

4.5 Additional Architecture Validation

Without model-specific tuning for ConvNeXt-Tiny, the fixed ImageNet configuration achieves an average strict ASR of 31.8% at $\epsilon = 5$, $Q = 20$K, while SurFree and CGBA achieve 7.2% and 10.8%, respectively. The additional CIFAR-10 evaluation also includes the canonical VGG16 family (Simonyan & Zisserman, 2015). The 18 within-model image-level paired comparisons against these two baselines are all significant after an independent Holm correction, with a maximum adjusted $p = 0.0009766$. This result shows that the fixed configuration can be reused on one additional architecture. This paper examines the reuse of the same attack configuration across different

Table 3: Strict ASR (%) on additional models. The three CIFAR-10 models use 100 images, 3 attack seeds per image, $\epsilon = 1$, $Q = 10$K, and the values are average strict ASR over the nine noise conditions; the ImageNet value is for ConvNeXt-Tiny on 100 of its own correctly classified clean images, 1 attack seed per image, $\epsilon = 5$, $Q = 20$K, averaged over the nine noise conditions.

Dataset	Model	BAPS	SurFree	CGBA
CIFAR-10	ResNet-20	48.3	4.7	6.3
CIFAR-10	AttackBox CNN	54.7	7.7	12.7
CIFAR-10	VGG16	40.0	2.7	5.3
ImageNet	ConvNeXt-Tiny	31.8	7.2	10.8

architectures, rather than cross-model transfer of adversarial examples. Because ResNet50 and ConvNeXt use different clean image lists, their absolute ASR difference reflects both architecture and image-list effects.

Mechanism ablation on the completed Gaussian/Uniform subset. We further analyze the Gaussian and uniform subsets of the ablation protocol (six cells, 100 images per cell, $\epsilon = 5$, $Q = 20$K). Full BAPS obtains 37.5% macro strict ASR, compared with 31.0% for geometry-only, 21.2% for equal-budget repeat-only, 20.8% for anchor-only, 19.8% with neither component, and 17.5% for warm-start-only. At the image level, full BAPS improves over repeat-only by 16.3 percentage points (pointwise paired-bootstrap 95% CI: [12.0, 20.7]), over anchor-only by 16.7 points ([12.5, 21.0]), and over geometry-only by 6.5 points ([3.8, 9.3]). Full BAPS and repeat-only both use approximately 19.97K raw calls on average, so this contrast does not reduce to making more repeated queries. Geometry-only is consistently the second-best variant, while adding the boundary anchor provides a smaller but complementary gain. Because this six-cell subset covers only two noise distributions and differs from the nine-condition confirmatory family, we report these intervals and any subset tests as exploratory rather than preregistered confirmatory evidence. Appendix C gives all cellwise values, query counts, and paired contrasts.

Scope of Interpretation. Our conclusions are limited to the tested noise distributions, image lists, adapted baselines, and strict endpoint criterion. Perceptual quality and robustness against other randomization mechanisms are not evaluated. PSJ results do not rank the unadapted original method. The statistical interpretation assumes fresh independent noise; caching or correlated draws require a new analysis. Clean-label agreement and natural random-error rates are not evaluated.

5 Conclusion

For decision-based attacks against input noise, BAPS separates probability-boundary geometry from radius-constrained candidate reliability and verifies one frozen output on fresh draws. In the evaluated CIFAR-10 and ImageNet protocols, it outperformed adapted SurFree and CGBA on strict ASR, including 90 Holm-significant ResNet50 comparisons and an additional ConvNeXt-Tiny evaluation. PSJ was stronger in three high-noise CIFAR cells, and its ImageNet adaptation has important fidelity limits. On the completed Gaussian/Uniform ablation subset, the structured probability-geometry search accounts for the largest isolated gain, the boundary anchor adds a complementary improvement, and an equal-budget repeat-only control performs substantially worse. This attribution is limited to the six-cell subset and does not extend to the full nine-condition protocol.

AI Use Statement

In preparing this work, the authors used generative AI tools for language editing, polishing the English expression of the manuscript, and LaTeX formatting assistance. The authors did not use generative AI tools to generate research ideas, design experiments, analyze data, or produce scientific claims. All AI-assisted text was reviewed and edited by the authors, who take full responsibility for the final content of this work.

Ethics Statement

This work studies adversarial attacks on image classifiers under randomized defenses. All experiments use publicly available datasets and do not involve human subjects or personally identifiable information. The proposed method is intended for evaluating and improving model robustness; we do not condone its use for malicious purposes.

Reproducibility Statement

We describe the evaluation protocol, noise distributions, query budgets, and strict success criterion in Section 4. Algorithm 1 summarizes the proposed method. Implementation details, hyperparameters, and additional results are provided in the appendix. The code and evaluation scripts will be released upon publication.

References

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*, volume 80 of *Proceedings of Machine Learning Research*, 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*, volume 80 of *Proceedings of Machine Learning Research*, pp. 284–293, 2018b.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018.

Junyoung Byun, Hyojun Go, and Changick Kim. On the effectiveness of small input noise for defending against query-based black-box attacks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3051–3060, 2022.

Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy*, pp. 1277–1294, 2020. doi: 10.1109/SP40000.2020.00045.

Jinghui Chen and Quanquan Gu. RayS: A ray searching method for hard-label adversarial attack. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1739–1747, 2020. doi: 10.1145/3394486.3403225.

Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2020.

C. J. Clopper and Egon S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375.

Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q. Weinberger. Simple black-box adversarial attacks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2484–2493, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BkMiWhR5K7.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. QEBA: Query-efficient boundary-based blackbox attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1221–1230, 2020.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11976–11986, 2022.

Thibault Maho, Teddy Furon, and Erwan Le Merrer. SurFree: A fast surrogate-free black-box attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10430–10439, 2021.

Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. GeoDA: A geometric framework for black-box adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8446–8455, 2020.

Md Farhamdur Reza, Ali Rahmati, Tianfu Wu, and Huaiyu Dai. CGBA: Curvature-aware geometric black-box attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 124–133, 2023.

Carl-Johann Simon-Gabriel, Noman Ahmed Sheikh, and Andreas Krause. PopSkipJump: Decision-based attack for probabilistic classifiers. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9712–9721, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Xiaosen Wang, Zeliang Zhang, Kangheng Tong, Dihong Gong, Kun He, Zhifeng Li, and Wei Liu. Triangle attack: A query-efficient decision-based adversarial attack. In *European Conference on Computer Vision*, pp. 156–174, 2022. doi: 10.1007/978-3-031-20065-6_10.

A Appendix

A.1 Evaluation and Implementation Details

Randomized oracle and endpoint. All perturbations are applied in clipped $[0, 1]$ pixel space before model normalization. Gaussian noise has per-pixel standard deviation σ; uniform noise is sampled from $[-\sqrt{3}\sigma, \sqrt{3}\sigma]$; Laplace noise has scale $\sigma/\sqrt{2}$. Thus all three distributions have the same per-pixel standard deviation. Each defended-model call resamples noise independently. One raw query is charged for every prediction, including warm starts, rejected proposals, repetitions, batched examples, and final validation. A budget Q reserves 200 calls before search, so $Q_{\text{search}} = Q - 200$; unused calls are not added to later budgets. Strict success requires both the stated ℓ_2 radius and a one-sided 95% Clopper–Pearson lower bound of at least 0.9 from 200 fresh draws of the frozen candidate. At least 188 of the 200 draws must return an attack label.

Data and models. The CIFAR-10 headline matrix evaluates 100 frozen test images with five attack seeds per image; the additional-model study is a separate 100-image, three-seed protocol. ImageNet–ResNet50 uses 100 clean-correct final images disjoint from development and confirmation lists. Its torchvision `IMAGENET1K_V2` weights use a 232-pixel resize and 224-pixel crop. ConvNeXt-Tiny uses `IMAGENET1K_V1` weights, a 236-pixel resize and 224-pixel crop, and its own deterministically selected clean-correct list of 100 images excluding the prior ImageNet split indices. Noise is injected before ImageNet channel normalization. Cross-architecture comparisons use different image sets. We use a fixed ImageNet attack seed (20260828). Final lists and configurations were locked before the corresponding formal runs.

Table 4: Headline evaluation settings. All Q values are maximum raw defended-model calls including the 200 independent final calls.

Dataset/model	Images	Seeds/image	ℓ_2 radius	Raw-call cap
CIFAR-10 headline	100	5	1	10K
CIFAR-10 budget	100	5	1	2K, 5K, 10K
ImageNet ResNet50 radius	100	1	5, 10, 20, 30, 40	20K
ImageNet ResNet50 budget	100	1	5	5K, 10K, 20K, 50K
ImageNet ConvNeXt-Tiny	100	1	5	20K

BAPS implementation. For ImageNet at $Q = 20$K, the shared warm-start allowance is $\max\{300, \lfloor 0.15(Q - 200) \rfloor\} = 2{,}970$ calls, divided between SurFree and CGBA warm directions and charged to BAPS. Each warm ray is expanded at most four times and then calibrated toward empirical attack probability 0.5 by five bisection steps, using five calls at each tested radius. Starting from an anchor, an iteration compares $a_t \pm hu_j$ with $h = 0.5$ along one radial and 15 multiscale low-frequency DCT directions; the initial DCT sizes cycle through 8, 16, and 32. Each side uses 3–7 calls. The paired differences form a normal estimate with momentum 0.6. Tangent proposals use multipliers 0.5, 1, and 2 and are recalibrated on their rays. A proposal can replace the anchor when its empirical boundary probability is at least 0.35 and its radius is less than 1.02 times the current anchor radius. The proposal direction is separately mapped to the allowed ϵ shell, screened with seven calls, and, if it passes the archive gate, evaluated with 31 further calls. A search-phase lower bound ranks archived candidates but does not certify final success. ImageNet smoothing starts at 0.5 within the configured range 0.1–1.5; the low-frequency size starts at 32 with configured maximum 64. These rules describe the evaluated implementation.

Baseline adaptations. SurFree and CGBA use a repeated randomized hard-label interface and the same raw-call counting and independent 200-draw endpoint. In the ResNet50 radius sweep, their distance-minimizing outputs are thresholded at each radius, while BAPS is rerun independently at each native radius; the final criterion is shared but their internal objectives differ. PSJ denotes the fixed radius-constrained adaptation of the authors' code at commit `b789e476696c2660e04688c0b52254e00850ed8b`. Its internal probability boundary is 0.5; its final direction is mapped to the allowed shell and independently verified. The adapter uses 32 iterations, 100 initial evaluations, a 5,000-evaluation cap, grid size 100, batch size 32, nine start repetitions, and at most 100 start attempts. Initialization and shell output differ from the original study. At ResNet50 $\epsilon = 5$, $Q = 20$K, 691/900 adapted PSJ runs reached the 32-iteration cap; capped runs used 14.7K raw calls on average. These results pertain to the adapted PSJ implementation.

Statistical units and correction families. ResNet50 inference pairs the same 100 images within each noise condition. Each pointwise 95% paired percentile interval uses 20,000 image-level bootstrap resamples. Two-sided exact discordant-pair tests receive Holm correction over the pre-specified 90 BAPS–SurFree/CGBA comparisons. ConvNeXt's 18 comparisons and the nine fixed-$\epsilon = 5$ ResNet50 PSJ comparisons use separate Holm families. CIFAR-10 keeps each image and its five attack seeds as one cluster; BAPS–PSJ inference uses image-cluster bootstrap and image-level sign flips with a separate nine-test Holm family. Holm correction adjusts the specified p-values, not the pointwise intervals. Macro averages and plotted segments are descriptive.

A.2 Additional Measured Results

A.2.1 Defended Clean Accuracy and CIFAR-10 Budget Endpoints

Table 5: ImageNet–ResNet50 defended clean top-1 accuracy (%), mean of three noise repeats on the fixed 1,000-image pool. Undefended accuracy is 80.40%.

σ	Gaussian	Uniform	Laplace
0.001	80.27	80.27	80.17
0.005	80.23	80.00	80.13
0.010	79.37	79.20	79.60

Table 6: Measured CIFAR-10 strict ASR (%) under Gaussian noise with $\sigma = 0.01$ and $\epsilon = 1$. The 10K endpoint reuses the headline cell; smaller budgets are separate runs.

Maximum Q	BAPS	Adapted PSJ
2K	8.6	2.2
5K	25.4	18.0
10K	42.0	49.2

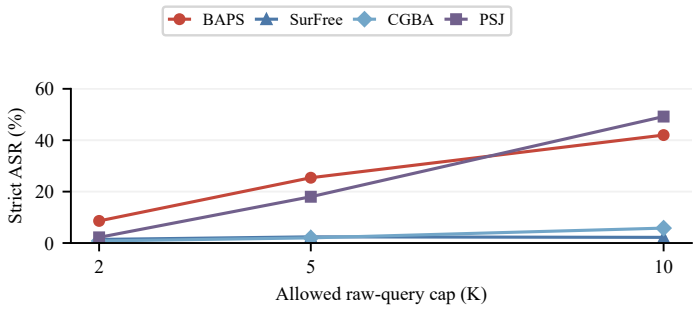

Figure 2: CIFAR-10 strict-ASR endpoints from Table 6. Segments connect independently measured budget points.

A.3 IMAGENET RADIUS, BUDGET, AND ADDITIONAL-MODEL CELLS

Table 7: BAPS strict ASR (%) in all 45 independently run native-radius ImageNet–ResNet50 cells at $Q = 20$K. Each cell has 100 images.

Noise	σ	$\epsilon = 5$	10	20	30	40
Gaussian	0.001	53	67	87	91	99
	0.005	37	57	76	88	97
	0.010	22	38	63	81	94
Uniform	0.001	54	69	86	90	97
	0.005	39	53	76	90	95
	0.010	20	41	66	82	91
Laplace	0.001	53	68	88	93	99
	0.005	40	56	75	90	95
	0.010	20	42	59	82	89

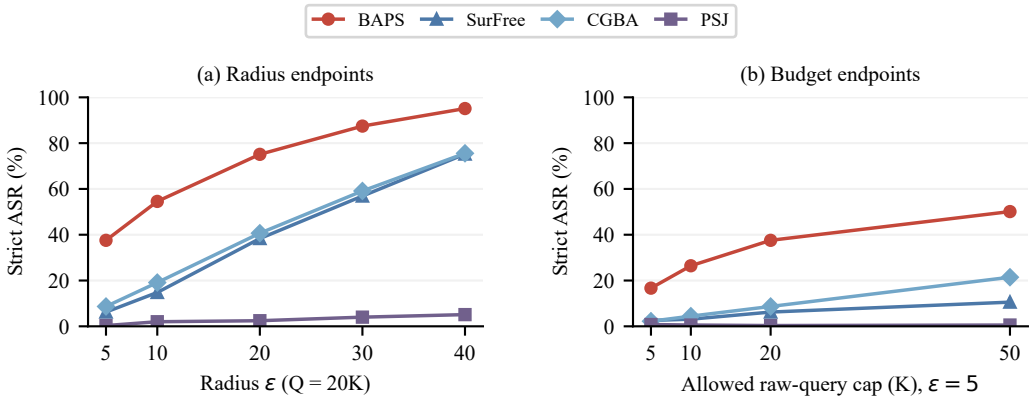

Figure 3: Measured ImageNet–ResNet50 strict-ASR endpoints. Left: BAPS was independently run at each native radius ($Q = 20$K), whereas SurFree and CGBA distance-minimizing outputs were radius-thresholded; PSJ is the fixed radius-constrained adaptation. Right: separate raw-call-cap experiments at fixed $\epsilon = 5$. The 20K point reuses the formal main run; 5K, 10K, and 50K are separate runs. Segments guide the eye only.

Condition-level ImageNet budget endpoints. Tables 8–11 give the measured strict-ASR percentage for every noise cell at fixed $\epsilon = 5$. Each entry is a success count out of 100 images. The 20K endpoints reuse the headline runs; the other caps are separate runs. All caps include final validation.

Table 8: ImageNet–ResNet50 strict ASR (%) at $\epsilon = 5$ and maximum raw-call cap $Q = 5$K. PSJ is the fixed radius-constrained adaptation.

Noise	σ	BAPS	SurFree	CGBA	PSJ
Gaussian	0.001	23	4	4	0
Gaussian	0.005	18	3	3	1
Gaussian	0.010	13	2	0	1
Uniform	0.001	21	4	4	0
Uniform	0.005	16	1	1	0
Uniform	0.010	9	1	2	2
Laplace	0.001	26	4	3	0
Laplace	0.005	16	2	2	1
Laplace	0.010	8	2	1	2

Table 9: ImageNet–ResNet50 strict ASR (%) at $\epsilon = 5$ and maximum raw-call cap $Q = 10$K. PSJ is the fixed radius-constrained adaptation.

Noise	σ	BAPS	SurFree	CGBA	PSJ
Gaussian	0.001	37	3	9	0
Gaussian	0.005	28	2	3	1
Gaussian	0.010	17	2	1	1
Uniform	0.001	38	6	7	0
Uniform	0.005	22	4	4	0
Uniform	0.010	14	2	1	1
Laplace	0.001	40	6	10	1
Laplace	0.005	27	2	3	0
Laplace	0.010	15	1	2	1

Table 10: ImageNet–ResNet50 strict ASR (%) at $\epsilon = 5$ and maximum raw-call cap $Q = 20$K. PSJ is the fixed radius-constrained adaptation.

Noise	σ	BAPS	SurFree	CGBA	PSJ
Gaussian	0.001	53	9	12	0
Gaussian	0.005	37	5	7	0
Gaussian	0.010	22	3	4	1
Uniform	0.001	54	11	17	0
Uniform	0.005	39	8	8	0
Uniform	0.010	20	3	4	0
Laplace	0.001	53	8	11	0
Laplace	0.005	40	4	10	1
Laplace	0.010	20	5	5	1

Table 11: ImageNet–ResNet50 strict ASR (%) at $\epsilon = 5$ and maximum raw-call cap $Q = 50$K. PSJ is the fixed radius-constrained adaptation.

Noise	σ	BAPS	SurFree	CGBA	PSJ
Gaussian	0.001	65	18	36	0
Gaussian	0.005	57	10	16	0
Gaussian	0.010	30	8	10	2
Uniform	0.001	62	16	36	0
Uniform	0.005	56	10	19	0
Uniform	0.010	28	3	12	2
Laplace	0.001	64	17	34	0
Laplace	0.005	57	9	17	0
Laplace	0.010	32	4	13	1

ConvNeXt-Tiny condition-level results. The frozen ResNet50-selected BAPS settings were reused without ConvNeXt-specific tuning. The ConvNeXt image list differs from the ResNet50 list; these are within-model comparisons.

Table 12: ImageNet–ConvNeXt-Tiny strict ASR (%) at $\epsilon = 5$, $Q = 20$K on 100 clean-correct images per condition. BAPS, SurFree, and CGBA all use the same model-specific images and 200-draw endpoint.

Noise	σ	BAPS	SurFree	CGBA
Gaussian	0.001	44	10	16
Gaussian	0.005	29	6	9
Gaussian	0.010	21	5	7
Uniform	0.001	45	10	16
Uniform	0.005	31	8	12
Uniform	0.010	21	6	8
Laplace	0.001	46	9	15
Laplace	0.005	29	5	9
Laplace	0.010	20	6	5
Macro ASR (%)		31.8	7.2	10.8

B Complete Paired Comparison Tables

The following tables report all rows of the three named ImageNet correction families and the nine CIFAR-10 BAPS–PSJ tests. Entries are BAPS minus the indicated adapted baseline in percentage points. Each ImageNet row has 100 paired images; a CIFAR row has 100 image clusters with five seeds each. Intervals are pointwise; adjusted p-values apply to their named Holm families.

ResNet50: pre-specified family of 90 SurFree/CGBA comparisons. Tables 13–17 retain all five radii, three noise families, three scales, and two adapted baselines. Every row's Holm-adjusted p-value is below 0.05; the largest is 0.00836. The five native BAPS radii were independent runs, whereas baseline outputs were distance-thresholded.

Table 13: ImageNet–ResNet50 BAPS minus adapted baselines at $\epsilon = 5$, $Q = 20$K (100 image pairs per row). Differences and pointwise intervals are percentage points; p-values are Holm-adjusted within the pre-specified 90-test family.

Noise	σ	Baseline	Δ	95% CI	Holm p
Gaussian	0.001	SurFree	44.0	[34.0, 54.0]	9.78×10^{-12}
Gaussian	0.001	CGBA	41.0	[32.0, 51.0]	7.19×10^{-11}
Gaussian	0.005	SurFree	32.0	[23.0, 41.0]	2.56×10^{-8}
Gaussian	0.005	CGBA	30.0	[20.0, 40.0]	2.43×10^{-6}
Gaussian	0.010	SurFree	19.0	[12.0, 27.0]	7.63×10^{-5}
Gaussian	0.010	CGBA	18.0	[11.0, 26.0]	1.22×10^{-4}
Uniform	0.001	SurFree	43.0	[33.0, 53.0]	1.93×10^{-11}
Uniform	0.001	CGBA	37.0	[28.0, 47.0]	1.06×10^{-9}
Uniform	0.005	SurFree	31.0	[22.0, 40.0]	4.84×10^{-8}
Uniform	0.005	CGBA	31.0	[22.0, 40.0]	4.84×10^{-8}
Uniform	0.010	SurFree	17.0	[10.0, 25.0]	1.68×10^{-4}
Uniform	0.010	CGBA	16.0	[9.0, 24.0]	7.25×10^{-4}
Laplace	0.001	SurFree	45.0	[35.0, 55.0]	4.95×10^{-12}
Laplace	0.001	CGBA	42.0	[32.0, 52.0]	3.73×10^{-11}
Laplace	0.005	SurFree	36.0	[27.0, 45.0]	2.01×10^{-9}
Laplace	0.005	CGBA	30.0	[21.0, 39.0]	9.31×10^{-8}
Laplace	0.010	SurFree	15.0	[8.0, 22.0]	4.27×10^{-4}
Laplace	0.010	CGBA	15.0	[8.0, 22.0]	4.27×10^{-4}

Table 14: ImageNet–ResNet50 BAPS minus adapted baselines at $\epsilon = 10$, $Q = 20$K (100 image pairs per row). Differences and pointwise intervals are percentage points; p-values are Holm-adjusted within the pre-specified 90-test family.

Noise	σ	Baseline	Δ	95% CI	Holm p
Gaussian	0.001	SurFree	50.0	[40.0, 60.0]	1.60×10^{-13}
Gaussian	0.001	CGBA	39.0	[29.0, 49.0]	2.80×10^{-10}
Gaussian	0.005	SurFree	43.0	[33.0, 53.0]	1.93×10^{-11}
Gaussian	0.005	CGBA	40.0	[31.0, 50.0]	1.42×10^{-10}
Gaussian	0.010	SurFree	26.0	[17.0, 35.0]	7.13×10^{-6}
Gaussian	0.010	CGBA	29.0	[19.0, 39.0]	4.45×10^{-6}
Uniform	0.001	SurFree	47.0	[37.0, 57.0]	1.25×10^{-12}
Uniform	0.001	CGBA	36.0	[27.0, 46.0]	2.01×10^{-9}
Uniform	0.005	SurFree	38.0	[29.0, 48.0]	5.53×10^{-10}
Uniform	0.005	CGBA	37.0	[28.0, 47.0]	1.06×10^{-9}
Uniform	0.010	SurFree	33.0	[24.0, 42.0]	1.37×10^{-8}
Uniform	0.010	CGBA	32.0	[23.0, 42.0]	1.87×10^{-7}
Laplace	0.001	SurFree	49.0	[39.0, 59.0]	3.16×10^{-13}
Laplace	0.001	CGBA	36.0	[27.0, 46.0]	2.01×10^{-9}
Laplace	0.005	SurFree	43.0	[33.0, 53.0]	1.93×10^{-11}
Laplace	0.005	CGBA	36.0	[26.0, 46.0]	1.65×10^{-8}
Laplace	0.010	SurFree	29.0	[20.0, 38.0]	1.75×10^{-7}
Laplace	0.010	CGBA	34.0	[25.0, 43.0]	7.10×10^{-9}

Table 15: ImageNet–ResNet50 BAPS minus adapted baselines at $\epsilon = 20$, $Q = 20$K (100 image pairs per row). Differences and pointwise intervals are percentage points; p-values are Holm-adjusted within the pre-specified 90-test family.

Noise	σ	Baseline	Δ	95% CI	Holm p
Gaussian	0.001	SurFree	36.0	[26.0, 46.0]	1.65×10^{-8}
Gaussian	0.001	CGBA	38.0	[28.0, 48.0]	4.77×10^{-9}
Gaussian	0.005	SurFree	38.0	[28.0, 48.0]	4.77×10^{-9}
Gaussian	0.005	CGBA	40.0	[30.0, 50.0]	1.37×10^{-9}
Gaussian	0.010	SurFree	30.0	[20.0, 40.0]	7.27×10^{-6}
Gaussian	0.010	CGBA	34.0	[24.0, 45.0]	7.76×10^{-7}
Uniform	0.001	SurFree	38.0	[29.0, 48.0]	5.53×10^{-10}
Uniform	0.001	CGBA	30.0	[21.0, 39.0]	9.31×10^{-8}
Uniform	0.005	SurFree	42.0	[32.0, 52.0]	3.73×10^{-11}
Uniform	0.005	CGBA	38.0	[29.0, 48.0]	5.53×10^{-10}
Uniform	0.010	SurFree	36.0	[26.0, 46.0]	1.65×10^{-8}
Uniform	0.010	CGBA	35.0	[26.0, 44.0]	3.78×10^{-9}
Laplace	0.001	SurFree	42.0	[33.0, 52.0]	3.73×10^{-11}
Laplace	0.001	CGBA	32.0	[23.0, 41.0]	2.56×10^{-8}
Laplace	0.005	SurFree	43.0	[33.0, 53.0]	1.14×10^{-9}
Laplace	0.005	CGBA	35.0	[25.0, 45.0]	1.36×10^{-7}
Laplace	0.010	SurFree	26.0	[17.0, 36.0]	2.26×10^{-5}
Laplace	0.010	CGBA	28.0	[18.0, 38.0]	7.39×10^{-6}

Table 16: ImageNet–ResNet50 BAPS minus adapted baselines at $\epsilon = 30$, $Q = 20$K (100 image pairs per row). Differences and pointwise intervals are percentage points; p-values are Holm-adjusted within the pre-specified 90-test family.

Noise	σ	Baseline	Δ	95% CI	Holm p
Gaussian	0.001	SurFree	28.0	[19.0, 37.0]	2.14×10^{-6}
Gaussian	0.001	CGBA	20.0	[12.0, 28.0]	1.42×10^{-4}
Gaussian	0.005	SurFree	35.0	[25.0, 45.0]	2.93×10^{-8}
Gaussian	0.005	CGBA	31.0	[22.0, 41.0]	3.32×10^{-7}
Gaussian	0.010	SurFree	26.0	[16.0, 36.0]	1.09×10^{-4}
Gaussian	0.010	CGBA	31.0	[20.0, 42.0]	9.72×10^{-6}
Uniform	0.001	SurFree	25.0	[15.0, 35.0]	1.42×10^{-4}
Uniform	0.001	CGBA	18.0	[9.0, 27.0]	0.00160
Uniform	0.005	SurFree	34.0	[25.0, 43.0]	7.10×10^{-9}
Uniform	0.005	CGBA	36.0	[27.0, 45.0]	2.01×10^{-9}
Uniform	0.010	SurFree	32.0	[22.0, 42.0]	7.76×10^{-7}
Uniform	0.010	CGBA	32.0	[23.0, 42.0]	1.87×10^{-7}
Laplace	0.001	SurFree	29.0	[20.0, 39.0]	1.13×10^{-6}
Laplace	0.001	CGBA	24.0	[15.0, 33.0]	2.25×10^{-5}
Laplace	0.005	SurFree	38.0	[28.0, 48.0]	4.77×10^{-9}
Laplace	0.005	CGBA	32.0	[23.0, 42.0]	1.87×10^{-7}
Laplace	0.010	SurFree	28.0	[19.0, 37.0]	2.14×10^{-6}
Laplace	0.010	CGBA	31.0	[22.0, 40.0]	3.32×10^{-7}

Table 17: ImageNet–ResNet50 BAPS minus adapted baselines at $\epsilon = 40$, $Q = 20$K (100 image pairs per row). Differences and pointwise intervals are percentage points; p-values are Holm-adjusted within the pre-specified 90-test family.

Noise	σ	Baseline	Δ	95% CI	Holm p
Gaussian	0.001	SurFree	21.0	[13.0, 29.0]	2.38×10^{-5}
Gaussian	0.001	CGBA	17.0	[10.0, 25.0]	1.68×10^{-4}
Gaussian	0.005	SurFree	21.0	[13.0, 30.0]	1.09×10^{-4}
Gaussian	0.005	CGBA	24.0	[15.0, 33.0]	2.25×10^{-5}
Gaussian	0.010	SurFree	23.0	[15.0, 31.0]	7.39×10^{-6}
Gaussian	0.010	CGBA	28.0	[19.0, 37.0]	3.20×10^{-7}
Uniform	0.001	SurFree	12.0	[5.0, 20.0]	0.00836
Uniform	0.001	CGBA	11.0	[4.0, 18.0]	0.00836
Uniform	0.005	SurFree	18.0	[11.0, 26.0]	1.22×10^{-4}
Uniform	0.005	CGBA	21.0	[13.0, 29.0]	2.38×10^{-5}
Uniform	0.010	SurFree	24.0	[15.0, 33.0]	6.37×10^{-5}
Uniform	0.010	CGBA	23.0	[14.0, 32.0]	3.56×10^{-5}
Laplace	0.001	SurFree	18.0	[11.0, 26.0]	1.22×10^{-4}
Laplace	0.001	CGBA	13.0	[7.0, 20.0]	9.77×10^{-4}
Laplace	0.005	SurFree	23.0	[14.0, 32.0]	3.56×10^{-5}
Laplace	0.005	CGBA	18.0	[10.0, 26.0]	3.20×10^{-4}
Laplace	0.010	SurFree	19.0	[11.0, 28.0]	1.89×10^{-4}
Laplace	0.010	CGBA	21.0	[13.0, 29.0]	1.09×10^{-4}

ConvNeXt-Tiny: separate 18-test family. These comparisons use the ConvNeXt-Tiny image list, not the ResNet50 list.

Table 18: ImageNet–ConvNeXt-Tiny BAPS minus adapted SurFree/CGBA at $\epsilon = 5$, $Q = 20$K (100 image pairs per row). p-values are Holm-adjusted within the separate 18-test family.

Noise	σ	Baseline	Δ	95% CI	Holm p
Gaussian	0.001	SurFree	34.0	[25.0, 43.0]	1.86×10^{-9}
Gaussian	0.001	CGBA	28.0	[19.0, 37.0]	9.69×10^{-8}
Gaussian	0.005	SurFree	23.0	[15.0, 32.0]	1.55×10^{-5}
Gaussian	0.005	CGBA	20.0	[12.0, 28.0]	7.68×10^{-5}
Gaussian	0.010	SurFree	16.0	[9.0, 24.0]	4.88×10^{-4}
Gaussian	0.010	CGBA	14.0	[8.0, 21.0]	4.88×10^{-4}
Uniform	0.001	SurFree	35.0	[26.0, 44.0]	9.90×10^{-10}
Uniform	0.001	CGBA	29.0	[20.0, 38.0]	5.22×10^{-8}
Uniform	0.005	SurFree	23.0	[15.0, 32.0]	2.62×10^{-6}
Uniform	0.005	CGBA	19.0	[12.0, 27.0]	3.05×10^{-5}
Uniform	0.010	SurFree	15.0	[8.0, 22.0]	3.66×10^{-4}
Uniform	0.010	CGBA	13.0	[6.0, 20.0]	9.77×10^{-4}
Laplace	0.001	SurFree	37.0	[28.0, 47.0]	2.62×10^{-10}
Laplace	0.001	CGBA	31.0	[22.0, 40.0]	1.40×10^{-8}
Laplace	0.005	SurFree	24.0	[16.0, 32.0]	1.43×10^{-6}
Laplace	0.005	CGBA	20.0	[13.0, 28.0]	1.72×10^{-5}
Laplace	0.010	SurFree	14.0	[8.0, 21.0]	4.88×10^{-4}
Laplace	0.010	CGBA	15.0	[8.0, 22.0]	3.66×10^{-4}

Fixed-$\epsilon = 5$ ResNet50 PSJ adaptation: separate nine-test family. This family evaluates our adapted PSJ implementation. PSJ's high-noise CIFAR advantages are reported separately below.

Table 19: ImageNet–ResNet50 BAPS minus the fixed radius-constrained PSJ adaptation at $\epsilon = 5$, $Q = 20$K. Each row has 100 paired images; p-values are Holm-adjusted within the separate nine-test family.

Noise	σ	Δ	95% CI	Holm p
Gaussian	0.001	53.0	[43.0, 63.0]	1.78×10^{-15}
Gaussian	0.005	37.0	[28.0, 47.0]	5.82×10^{-11}
Gaussian	0.010	21.0	[13.0, 29.0]	2.86×10^{-6}
Uniform	0.001	54.0	[44.0, 64.0]	9.99×10^{-16}
Uniform	0.005	39.0	[29.0, 49.0]	2.18×10^{-11}
Uniform	0.010	20.0	[13.0, 28.0]	3.81×10^{-6}
Laplace	0.001	53.0	[43.0, 63.0]	1.78×10^{-15}
Laplace	0.005	39.0	[30.0, 49.0]	2.18×10^{-11}
Laplace	0.010	19.0	[12.0, 27.0]	3.81×10^{-6}

CIFAR-10 PSJ adaptation: clustered nine-test family. Each row uses 100 images, with five attack seeds retained inside each image cluster during resampling and sign flips. Negative differences favor adapted PSJ.

Table 20: CIFAR-10 BAPS minus the fixed radius-constrained PSJ adaptation at $\epsilon = 1$, $Q = 10$K. Intervals and differences are percentage points. Holm correction covers all nine conditions, including those favoring PSJ.

Noise	σ	Δ	95% CI	Holm p
Gaussian	0.001	13.4	[7.2, 20.0]	0.00200
Gaussian	0.005	11.2	[3.4, 18.8]	0.0375
Gaussian	0.010	-7.2	[-16.2, 2.0]	0.403
Uniform	0.001	14.6	[7.8, 21.6]	0.00135
Uniform	0.005	10.0	[2.6, 17.4]	0.0560
Uniform	0.010	-6.2	[-15.2, 2.8]	0.403
Laplace	0.001	13.6	[6.4, 21.0]	0.00245
Laplace	0.005	10.6	[2.4, 18.6]	0.0590
Laplace	0.010	-3.2	[-12.6, 6.2]	0.540

C Frozen Nine-Condition Mechanism Ablation

The mechanism-ablation protocol was frozen on 19 September 2026 before the formal runs (protocol SHA-256 prefix `0168ec0026570639`). All nine noise cells, 100 images, and six variants completed: 5,400/5,400 image–condition–variant records. All records passed the analysis checks. The primary estimand is the mean over 100 images of each image's equal-weight average strict-success difference across the nine cells. All nine outcomes for an image remain together during resampling.

Switches and controls. The *full* arm retains probability-boundary calibration and paired geometric direction information. *Geometry-only* removes boundary calibration but keeps paired information. *Anchor-only* keeps calibration but suppresses the directional contrast passed to the geometry update while still paying for paired observations; *neither* removes calibration and suppresses that contrast. These two arms are placebo-information ablations. *Repeat-only* uses the same warm directions, samples low-frequency random shell proposals around them, spends its remaining search allowance on repeated 31-call archive rankings, and omits the anchor, paired normal, tangent projection, and boundary line search. *Warm-only* is diagnostic. All arms use $\epsilon = 5$, $Q = 20$K, and independent 200-draw final verification. The geometry switch bundles several operations.

Table 21: Strict-success counts out of 100 images per ImageNet–ResNet50 noise cell for the six frozen arms. The last row is equal-weight nine-cell macro strict ASR (%).

Noise	σ	Full	Anchor	Geometry	Neither	Repeat	Warm
Gaussian	0.001	53	35	47	35	36	35
	0.005	37	18	33	16	18	14
	0.010	22	9	14	9	11	5
Uniform	0.001	54	34	44	32	33	32
	0.005	39	18	35	17	19	16
	0.010	20	11	13	10	10	3
Laplace	0.001	53	35	47	35	37	35
	0.005	40	20	36	15	19	14
	0.010	20	9	13	10	10	2
Macro ASR (%)		37.56	21.00	31.33	19.89	21.44	17.33

Table 22: Mean raw defended-model calls over the 900 image–condition records per arm. All means include 200 final calls. Warm-only uses fewer calls.

Arm	Mean raw calls	Mean warm-start calls
Full	19,973.3	2,970.0
Repeat-only	19,972.0	2,970.0
Anchor-only	19,972.7	2,970.0
Geometry-only	19,957.4	2,970.0
Neither	19,956.3	2,970.0
Warm-only	3,232.0	2,970.0

Pre-registered image-level tests. The three primary full-minus-control contrasts use 20,000 paired image-bootstrap resamples for pointwise percentile intervals, 100,000 two-sided image-level sign flips with an add-one Monte Carlo correction for p-values, and Holm correction across exactly those three tests. The smallest reportable raw Monte Carlo p-value is about 10^{-5}; a value at that floor is not an exact tail probability. Neither and warm-only are secondary descriptive contrasts.

Table 23: Frozen nine-cell paired ablation contrasts. Δ is full minus control in strict-ASR percentage points. Positive/negative/tied counts refer to the per-image mean difference across nine cells. Secondary intervals are descriptive.

Control	Δ (pp)	95% CI (pp)	Pos./neg./tie	Raw p	Holm p
Repeat-only	16.11	[12.22, 20.11]	52/1/47	1.00×10^{-5}	3.00×10^{-5}
Anchor-only	16.56	[12.67, 20.78]	51/1/48	1.00×10^{-5}	3.00×10^{-5}
Geometry-only	6.22	[4.00, 8.67]	35/4/61	1.00×10^{-5}	3.00×10^{-5}
Neither	17.67	[13.56, 21.89]	53/0/47	–	–
Warm-only	20.22	[16.11, 24.44]	58/0/42	–	–

The three primary contrasts all pass the pre-registered Holm family at level 0.05. The main text reports the earlier six Gaussian/Uniform cells as an exploratory subset; this appendix reports the subsequently completed, pre-registered nine-cell estimand. Geometry-only retains much of the full arm's strict ASR, while restoring calibration on top of geometry adds 6.22 percentage points. The warm-only contrast does not isolate a query-matched warm-start effect.