Decision-Only Adversarial Editing with Diffusion Models

Yaowen Wang Daniel Cullina
Department of Electrical Engineering, The Pennsylvania State University
{yxw5684, cullina}@psu.edu

Abstract

Subtle visual cues, though often difficult for humans to localize, can critically affect the decisions of modern vision systems. Recent diffusion-based unrestricted adversarial examples have shown strong stealthiness and transferability, but most existing approaches assume white-box access or at least score-based feedback. In many real-world scenarios, however, only a hard-label decision (accept/reject) is available. We address this decision-only *setting and propose* ***AdvDiffEdit****, a diffusion-guided method for adversarial* generation and editing *from a single binary oracle. AdvDiffEdit performs repeated SDEdit-style partial noising and denoising passes. In each pass, we (i) seed a worst-case forward noise at a chosen diffusion strength using a decision-only search, and (ii) progressively weaken guidance and strength across passes to "walk the data manifold" while preserving visual realism. Crucially, our method requires no gradients, logits, or scores—only one-bit feedback—yet reliably produces natural-looking decision-flipping images. Empirically, AdvDiffEdit achieves high attack success with strong visual fidelity and competitive cross-model transfer on ImageNet-class classifiers. Our results position diffusion-guided editing as a practical tool for auditing the robustness of decision-only systems to subtle, naturalistic visual perturbations under unrestricted threat models.*

1. Introduction

Deep learning now powers safety–critical applications such as autonomous driving[9], face recognition[30], and medical image analysis[40]. Despite remarkable accuracy, modern neural networks remain vulnerable to adversarial examples[11]: inputs that induce incorrect decisions through changes that may be small, structured, or difficult for humans to precisely localize. Early work largely studied restricted attacks that minimize an ℓ_p norm in RGB space[2, 5, 22, 26, 36], enabling controlled comparisons between attack and defense methods. However, imperceptibility measured by an ℓ_p norm does not necessarily align with human perception; many ℓ_p-bounded perturbations remain visible. Moreover, such pixel-level perturbations are often not naturalistic[18], and robustness to them does not necessarily imply robustness to the subtle visual cues and shortcuts that can drive model decisions in real-world systems[10, 16, 41].

Unrestricted attacks take a different route: rather than only nudging pixels, they seek to produce natural yet adversarial images by editing content while preserving semantics and realism[31]. With the rise of diffusion models as high-fidelity image generators and editors, a growing line of work perturbs the generative trajectory to produce realistic adversarial outputs[4, 6, 7, 25]. This perspective is especially appealing for *subtle visual robustness auditing*: instead of asking whether a classifier is sensitive to synthetic pixel noise, we can probe whether its decisions can be flipped by visually plausible, naturalistic edits that expose hidden decision-critical cues. Many diffusion-based attacks, however, still rely on model scores or gradients, which are unavailable in realistic hard-label deployments. Real systems often impose tight query limits or expose only binary decisions, limiting the practicality of score- or gradient-based attacks.

In numerous real-world systems (content filters, deployed APIs, privacy-preserving classifiers), the attacker or auditor receives only a one-bit response (accept/reject)[37]. Classical decision-based pixel attacks—Boundary Attack [1] and HSJA [3]—search directly in input space and can succeed, but they typically require large query budgets and often exhibit high-frequency artifacts at moderate budgets. More importantly, their perturbations are optimized for ℓ_p distortion and need not correspond to subtle, naturalistic visual changes.

We introduce *AdvDiffEdit*, a decision-only diffusion editing method for auditing model robustness with binary feedback alone—no logits, no gradients, no scores. The key idea is to (i) seed a worst-case forward noise at a chosen edit strength t via randomized search under the hard-label oracle, and (ii) run repeated SDEdit-style partial noising/denoising passes while progressively reducing edit strength and guidance. This outer-loop annealing ef-

fectively "walks the manifold," yielding natural-looking decision-flipping images that preserve visual realism while exposing subtle vulnerabilities in the target model. A lightweight refinement stage can further reduce artifacts at the cost of modest extra queries.

Empirically, on an ImageNet-compatible 1,000-image subset, AdvDiffEdit achieves reliable attack success under fixed query budgets, produces visually realistic edits with fewer pixel-level artifacts than decision-based pixel attacks at comparable budgets, transfers well across CNN and ViT targets, and remains substantially effective under common preprocessing defenses and adversarially trained models. These results suggest that diffusion-guided editing is not only a practical unrestricted attack mechanism, but also a useful tool for auditing the robustness of decision-only systems to subtle, naturalistic visual perturbations.

Contributions:

- We present a diffusion-guided visual editing pipeline that requires only a hard-label oracle—no score or gradient access—yet reliably produces natural-looking decision-flipping images.
- We propose a binary-feedback seed search at diffusion strength t, coupled with an outer loop that progressively reduces edit strength and guidance; a refinement stage further improves realism with predictable query overhead.
- Against decision-only pixel-space baselines (Boundary Attack, HSJA), AdvDiffEdit produces successful attacks with fewer visible artifacts at comparable query budgets; compared to diffusion-based unrestricted baselines that rely on scores or gradients [4, 6], it attains competitive transferability using only binary feedback.

2. Related Work

Decision-only adversarial attacks. Decision-based attacks study the hard-label setting, where the attacker only observes the model's top-1 decision rather than scores or gradients. A representative early method is Boundary Attack (BA) [1], which starts from an adversarial point and follows the decision boundary while gradually reducing perturbation size in pixel space. HopSkipJumpAttack (HSJA) [3] improves this line by combining boundary search with gradient-free refinement, leading to substantially lower pixel distortion under sufficient query budgets. These methods are strong baselines for the strict decision-only setting, but they optimize perturbations directly in image space and are not designed to preserve visual realism, subtle natural-image structure, or decision-critical cues that remain visually plausible to humans.

Unrestricted and naturalistic black-box attacks. Beyond small ℓ_p-bounded perturbations, unrestricted attacks aim to generate adversarial examples that remain visually plausible while allowing larger, structured changes. ColorFool [29] demonstrates that modifying global color attributes can already induce misclassification while preserving much of the scene content. Natural Color Fool (NCF) [38] further explores natural color manipulations as a black-box adversarial mechanism, showing that visually mild hue and saturation changes can transfer across models. These methods move toward perceptually natural attacks, but their search space is largely limited to color transformations, which constrains the diversity, locality, and semantic richness of the resulting edits. As a result, they provide only a limited testbed for auditing robustness to broader classes of subtle visual changes.

Diffusion-based adversarial generation. Recent work has shown that diffusion models provide a powerful prior for generating realistic adversarial examples. ACA [4] uses diffusion-based inversion and reconstruction to craft adversarial content edits with explicit optimization signals beyond hard-label feedback. Compared with earlier color-only attacks, diffusion-based methods can introduce richer and more naturalistic modifications while remaining closer to the natural image manifold. This makes diffusion particularly appealing for studying robustness to subtle, visually plausible perturbations rather than only synthetic pixel noise. However, existing approaches typically rely on score access, gradients, or other stronger supervision during optimization.

Our position. Our work lies at the intersection of these directions. Like BA and HSJA, we operate under a strict decision-only oracle with only binary feedback. Like unrestricted and diffusion-based attacks, we seek adversarial examples that remain visually natural rather than merely small in pixel norm. Our key difference is to perform hard-label search through diffusion-guided editing, enabling manifold-aligned adversarial generation without access to scores or gradients. From the perspective of subtle visual robustness auditing, this provides a practical way to test whether black-box vision systems rely on subtle, naturalistic visual cues that are sufficient to flip their decisions.

2.1. Decision-only oracle

We consider an input space $\mathcal{X} \subset [0,1]^{H\times W\times 3}$ and an unknown classifier $f : \mathcal{X} \to \Delta^{K-1}$ that is accessible only through a binary oracle $\mathcal{O} : \mathcal{X} \to \{0,1\}$. We use $\mathcal{O}(x) = 1$ to denote *attack success*. Concretely, in the untargeted setting, $\mathcal{O}(x) = 1$ if the top-1 prediction $h(x) = \arg\max_k f_k(x)$ differs from the clean label; in the targeted setting, $\mathcal{O}(x) = 1$ if $h(x)$ matches a prescribed target class. Otherwise, $\mathcal{O}(x) = 0$. No scores or gradients are available. Any randomized preprocessing (e.g., random crops or compression) is treated as part of $\mathcal{O}$.

Given a clean image x and a query budget B, the attacker adaptively submits $x^{(1)}, x^{(2)}, \ldots$ and observes $\mathcal{O}(x^{(i)}) \in$

$\{0, 1\}$, aiming to find a realistic adversarial example x_{adv} such that $\mathcal{O}(x_{\text{adv}}) = 1$. Let $\mathcal{A} = \{x \in \mathcal{X} : \mathcal{O}(x) = 1\}$ denote the adversarial acceptance region and $\partial\mathcal{A}$ its decision boundary. Classical decision-only pixel attacks such as Boundary Attack and HSJA perform random walks near $\partial\mathcal{A}$ in pixel space. In contrast, AdvDiffEdit searches over *latent* noising directions at a chosen diffusion strength and projects candidates back to image space with an SDEdit-style sampler, biasing updates to stay on the natural-image manifold while crossing into $\mathcal{A}$.

2.2. Diffusion-based editing prior

We leverage SDEdit [23] as a generative prior to keep adversarial examples on the natural-image manifold. Given an input image $x \in \mathbb{R}^{H\times W\times 3}$, we operate in the latent space of a pre-trained autoencoder (VAE), writing $z_0 = \mathcal{E}(x)$. SDEdit applies partial noising followed by guided denoising: noise is added to z_0 up to a diffusion timestep $t \in [0, 1]$ to obtain z_t, and a reverse diffusion sampler maps z_t back to a clean latent z_0^1 conditioned on a text prompt c with guidance scale g (e.g., classifier-free guidance [14]). We treat this as a black-box operator

$$x_{\text{edit}} = \mathcal{D}\big(\mathcal{S}(\mathcal{E}(x), t, c, g)\big), \tag{1}$$

where S is the sampler and $\mathcal{D}$ is the VAE decoder. The strength parameter t acts as a realism–edit knob: larger t allows larger semantic changes (useful for crossing decision boundaries), whereas smaller t yields higher fidelity to the original input.

3. Method: Repeated SDEdit with Decision-Only Seed Search

Figure 1 shows the overall procedure. Given a clean image x, we encode it into a latent $z_0 = \mathcal{E}(x)$, inject forward noise to strength t, search for a noise direction that makes a proxy decode result adversarial under the hard-label oracle, and then denoise with a diffusion sampler to obtain a realistic image. We repeat this add–denoise–refine loop while gradually reducing t in a coarse-to-fine schedule.

3.1. Adversarial seed search

At large edit strengths t, many noise directions can flip the oracle label but faithfulness is weaker; at smaller t the flip is harder to find but the resulting image better preserves fine details. Intuitively, we aim to find the smallest t at which a label flip is still achievable, to maximize fidelity to the original latent z_0:

$$\min t \quad \text{s.t.} \quad \mathcal{O}\big(S_t(z_t)\big) \neq y, \tag{2}$$

where $y = \mathcal{O}(x)$ is the original oracle response, z_t is a noisy latent at strength t, and S_t denotes the reverse-time sampler from t to 0.

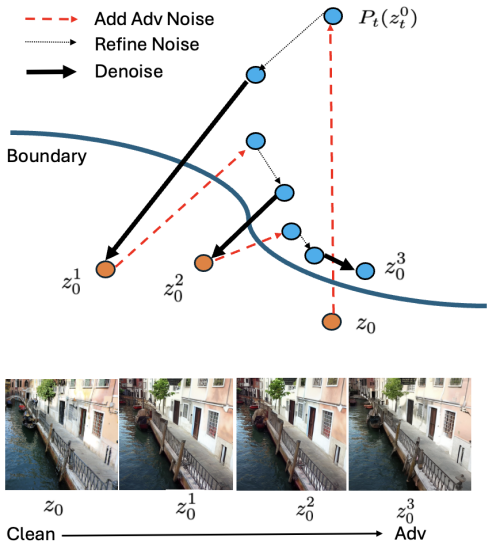

Figure 1. AdvDiffEdit add–refine–denoise loop in latent and image space. Starting from clean z_0, we add noise, refine the direction using the proxy P_t, and denoise back to t=0, iterating until we cross the decision boundary while staying near the image manifold (bottom: clean → adversarial).

Directly evaluating $\mathcal{O}(S_t(z_t))$ requires a full denoising rollout for every candidate and is computationally expensive. Instead, we rely on the standard clean-image proxy $\hat{z}_0(z_t)$ obtained from the UNet's noise prediction. Let $\alpha(t)$ and $\delta(t)$ denote scheduler parameters and let $\hat{\varepsilon}_\theta(z_t, t, c; g)$ be the guided noise estimate. We form

$$P_t(z_t) = \hat{z}_0(z_t) \;=\; \frac{z_t - \delta(t)\,\hat{\varepsilon}_\theta(z_t, t, c)}{\alpha(t)}, \tag{3}$$

which removes instantaneous noise at level t while preserving object-level content that governs the hard-label decision. Each proxy evaluation requires only a single UNet call, avoiding repeated sampling. The geometry of this proxy-based search is illustrated in Fig. 2.

We use the standard SDEdit forward map to reach time t:

$$z_t(z_0, \varepsilon) = \alpha(t)\, z_0 + \sigma(t)\, \varepsilon, \tag{4}$$

where $\varepsilon \sim \mathcal{N}(0, I)$ and $\sigma(t)$ is the noise scale and $\sigma(t) = \sqrt{1 - \alpha^2(t)}$. Given a strength t and original label $y = \mathcal{O}(x)$, we draw a small pool $\{\varepsilon^{(k)}\}_{k=1}^K$, form noisy latents $z_t(z_0, \varepsilon^{(k)})$, compute proxy decodes $P_t(z_t(z_0, \varepsilon^{(k)}))$, decode them to images, and query the oracle. We accept the first seed whose proxy decode flips the oracle:

$$\mathcal{O}\big(D(P_t(z_t(z_0, \varepsilon^{(k)})))\big) \neq y.$$

If no candidate flips at this t, we increase t and repeat with a fresh pool. The accepted seed $\varepsilon_{\text{seed}}$ is then passed to the refinement stage.

Algorithm 1 ADVDIFFEDIT: repeated SDEdit with decision-only seeding

Inputs: image x, encoder $\mathcal{E}$, Sampler S, oracle $\mathcal{O}$, initial strength t_{start}, minimum strength $t_{\min}$, decay $\gamma \in (0,1)$, pool size K, max passes M, guidance g.
$z_0 \leftarrow \mathcal{E}(x_{\text{in}}), \quad t \leftarrow t_{\text{start}}$
for $p = 1$ **to** M **do**
  Seed search at t: *seed_found* $\leftarrow$ **false**
  for $k = 1$ **to** K **do**
    sample $\varepsilon^{(k)} \sim \mathcal{N}(0, I)$
    $z_t^p(\varepsilon^{(k)}) = \alpha(t)\, z_0^p + \sigma(t)\, \varepsilon^{(k)}$
    if $\mathcal{O}\big(P_t(z_t^p(\varepsilon^{(k)}, z_0^p))\big) = 1$ **then**
      seed_found $\leftarrow$ **true**; $\varepsilon^\star \leftarrow \text{Refine}\big(\varepsilon^{(k)}; z_0^p\big)$
      break
    end if
  end for
  if *seed_found* **is true then**
    Roll-down: $z_0^p \leftarrow S_{t\to 0}\big(z_t^p(\varepsilon^\star); g\big)$
    if $\mathcal{O}(z_0^p) = 1$ **then**
      return z_0^p (success)
    else
      $t \leftarrow \max(t_{\min},\ \gamma t)$ (coarse $\to$ fine)
    end if
  else
    $g \leftarrow \gamma g$
    $t \leftarrow \min(t_{max}, t/\gamma)$ (retry bigger strength)
  end if
end for
return last x_{out}

3.2. Seed noise refinement

The initial adversarial seed is drawn from a Gaussian and may introduce unnecessary visual drift. To bias the solution toward faithfulness while preserving the flip, we perform a short feasible pattern search around the seed in noise space.

Let $I(\varepsilon) = D\big(P_t(z_t(z_0, \varepsilon))\big)$ denote the proxy decode result used during search, Given the benign reference image x, we refine a flipping seed $\varepsilon_{\text{seed}}$ by solving

$$\min_{\delta}\ d\big(I(\varepsilon_{\text{seed}} + \delta),\, x\big) \quad \text{s.t.} \quad \text{O}(I(\varepsilon_{\text{seed}} + \delta)) = 1, \quad (5)$$

where $d(\cdot,\cdot)$ is an image-space distance (we use ℓ_2 in practice).

Acceptance-only pattern search. Because the constraint in (5) is binary, we use a derivative-free accept–reject scheme. Starting from $\varepsilon_{\text{best}} = \varepsilon_{\text{seed}}$ and $D_{\text{best}} = d(I(\varepsilon_{\text{seed}}), x)$ with step size α, we iterate a few rounds of local proposals. At each iteration we sample a random direction $u \sim \mathcal{N}(0, I)$ and consider symmetric candidates $\varepsilon_{\text{cand}}^{(\pm)} = \varepsilon_{\text{best}} \pm \alpha\, u$. For each candidate we compute $\text{O}(I(\varepsilon_{\text{cand}}^{(\pm)}))$ and $D_{\text{cand}}^{(\pm)} = d(I(\varepsilon_{\text{cand}}^{(\pm)}), x)$. If any candidate satisfies $\text{O}(I(\varepsilon_{\text{cand}}^{(\pm)})) = 1$ and $D_{\text{cand}}^{(\pm)} < D_{\text{best}}$, we accept it, set $\varepsilon_{\text{best}} \leftarrow \varepsilon_{\text{cand}}^{(\pm)}$ and $D_{\text{best}} \leftarrow D_{\text{cand}}^{(\pm)}$, and mildly increase the step size $\alpha \leftarrow \gamma_{\text{grow}}\alpha$; otherwise we shrink it, $\alpha \leftarrow \gamma_{\text{shrink}}\alpha$. The loop stops after a small fixed budget or when α falls below a threshold. This procedure is zero-order and decision-only: we never differentiate through the oracle, and the constraint is enforced purely by acceptance. In practice, 5–10 iterations with a few proposals per iteration suffice to noticeably reduce drift without sacrificing attack success.

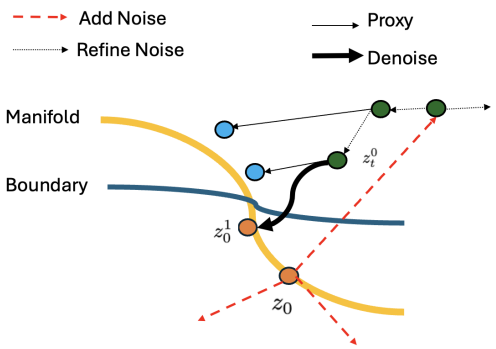

Figure 2. Role of the proxy P_t in seed search. Noise moves z_0 off the manifold; $P_t(z_t)$ provides a cheap clean proxy used to query the oracle and refine the noise direction before full denoising, so the trajectory returns to the manifold near the decision boundary.

3.3. Full denoising and outer coarse-to-fine loop

Given a flipping seed at strength $t \in (0,1)$, i.e., a noise vector ε such that the proxy decode result at time t predicts the target label, we run the reverse denoising dynamics from t down to 0 to obtain $\mathcal{S}_{t\to 0}\big(z_t(z_0, \varepsilon);\ g\big)$, where $\mathcal{S}_{t\to 0}$ denotes the conditional reverse-time sampler, and g is the classifier-free guidance (CFG) scale. If a text prompt is provided we apply CFG during the trajectory; otherwise we set g=0.

Guidance scheduling during roll-down. Because the oracle is decision-only, we monitor a lightweight proxy prediction at intermediate times and detect when the flip is being lost. Upon detection, we reduce g to allow more drift and preserve the adversarial trajectory, trading prompt adherence for attack robustness.

Outer coarse-to-fine loop. A single pass may fail: seeds found at a given t can be "purified" by subsequent denoising. We therefore iterate a coarse-to-fine schedule over the strength parameter

$$t \ \leftarrow\ \max\big(t_{\min},\, \gamma\, t\big), \qquad \gamma \in (0,1),$$

starting from $t_0 \in [0.3, 0.7]$ and decaying toward $t_{\min} \in [0.2, 0.3]$. Large t edits coarse structure to move across decision regions; after a successful move, smaller t concen-

trates edits on fine details, improving fidelity while maintaining the flip. Each pass proceeds as:

1. **Seed at t:** sample and refine ε until the proxy at time t flips.
2. **Roll-down:** integrate the reverse dynamics $t \to 0$, decreasing g if the flip is lost mid-trajectory.
3. **Check:** if the oracle flips on x_{out}, terminate; otherwise update t via the decay rule and repeat.

The loop stops on success or when a query budget is exhausted. In practice, $3-8$ passes suffice: early passes at larger t establish adversarial coarse geometry, while later passes at smaller t preserve naturalness and refine details along the diffusion manifold. For completeness, Algorithm 1 summarizes the full AdvDiffEdit procedure in pseudocode.

4. Informal Intuition

We briefly explain why a *coarse→fine* schedule (large t to small t) helps both *discover* and *preserve* adversarial flips.

Setup. Let $x_t = \alpha(t)x_0 + \sigma(t)\varepsilon$ be the VP marginal, and let $S_t(x_t)$ denote the final output after full reverse dynamics. Let $P_t(x_t)$ denote the proxy used at time t for margin evaluation, as defined in Sec. 3.1. Recall that P_t is the standard one-step clean estimate induced by the diffusion denoiser, used there as a cheap surrogate for the fully denoised output; this view is closely related to denoising and Tweedie-style interpretations of diffusion models [15, 27, 32, 34].

Let the oracle margin be $m(z) = \inf_{z':\mathcal{O}(z')=1} d(z, z')$ By construction, m is 1-Lipschitz and that the proxy and final margins are aligned up to a proxy–final gap $\Delta(t)$:

$$\sup_{z_t} \left| m(S_t(z_t)) - m(P_t(z_t)) \right| \;\le\; \Delta(t).$$

We further model $\Delta(t)$ as nonincreasing in t, with $\Delta(0) = 0$. This is motivated by the fact that P_t is a local one-step surrogate for the same reverse denoising process defining S_t, while the remaining reverse horizon becomes shorter as $t \to 0$; from the diffusion ODE/SDE and numerical-solver viewpoint, this leaves less room for accumulated trajectory drift and integration error, making a smaller proxy–final discrepancy at finer strengths natural to expect [17, 21, 32].

Define the seed set $\mathcal{E}_t = \{\varepsilon : \; m(P_t(z_t(\varepsilon))) \ge 0\}$, and let $p_t = \mathbb{P}(\mathcal{E}_t)$ denote the probability that a random seed flips the proxy margin at time t.

Coarse exploration. At coarser strengths, the injected noise has larger scale, so perturbations explore a broader neighborhood around the current point. Under a local linear view of the proxy margin in the noise direction, this increases the chance that a random seed crosses the proxy decision boundary. Thus, coarse strengths are more effective for discovering adversarial seeds, while finer strengths are better suited for preserving them once found.

Proposition 1 (Fine persistence) *For any t, if the proxy margin has a buffer, $m(P_t(z_t)) \ge \Delta(t)$, then the final margin flips:*

$$m(S_t(z_t)) \;\ge\; m(P_t(z_t)) - \Delta(t) \;\ge\; 0.$$

Moreover, since $\Delta(t') \le \Delta(t)$ for $t' < t$, the buffer required for a proxy flip to survive full denoising is no larger at finer strengths.

Lemma 1 (Carryover to finer strengths) *Assume for fixed ε that $t \mapsto m(P_t(z_t(\varepsilon)))$ is A-Lipschitz and that $\Delta(t)$ is nonincreasing. If at some t the proxy margin has a buffer $\rho > 0$:*

$$m(P_t(z_t(\varepsilon))) \;\ge\; \Delta(t) + \rho,$$

then for any $t' \in [t-\delta,\, t]$ with $\delta \le \rho/(2A)$,

$$m(P_{t'}(z_{t'}(\varepsilon))) \;\ge\; \Delta(t').$$

Hence once a seed is found at a coarse t, it remains a seed for a nontrivial range of finer strengths $t' \le t$.

Discussion. The picture is therefore simple: coarse noise improves exploration by enlarging the chance of crossing the proxy decision boundary, while finer strengths shorten the remaining reverse horizon and reduce the proxy–final gap, making discovered flips easier to preserve. Our schedule exploits exactly this asymmetry: start coarse to find seeds, then move fine to stabilize them.

5. Experimental Results

5.1. Setup

We evaluate on an ImageNet-compatible subset of 1,000 images sampled from the validation set, a standard benchmark for adversarial attacks [18] and widely used in recent unrestricted-attack work [4, 6, 29].

We adopt a strict decision-only oracle

$$\mathcal{O}(x) = \mathbf{1}\left[\arg\max_k f_k(x) \neq y\right],$$

which returns 1 if the classifier's top-1 prediction differs from the ground-truth label y, and 0 otherwise. Unless noted, we use the untargeted objective and count an attack as successful when $\mathcal{O}(x_{\text{adv}}) = 1$. One query corresponds to one evaluation of $\mathcal{O}$.

We report attack success rate (ASR, %), average queries and average time cost. To assess visual fidelity, we report FID [13] and LPIPS [39] with respect to the original images. For decision-only pixel-space comparisons, we also report the standard ℓ_2 distortion.

All experiments use PyTorch on a single NVIDIA A100 GPU. We sample with DDIM using $T = 50$ denoising

Figure 3. Qualitative comparison under a hard-label oracle (4 ImageNet samples). Rows: **Original**; **BA1k** = Boundary Attack with 1000 queries shows visible high-frequency noise; **BA20k** = Boundary Attack with 20,000 queries shows artifacts largely suppressed; **AdvDiffEdit** = our decision-only diffusion editing at a comparable budget, producing more natural edits while flipping the decision.

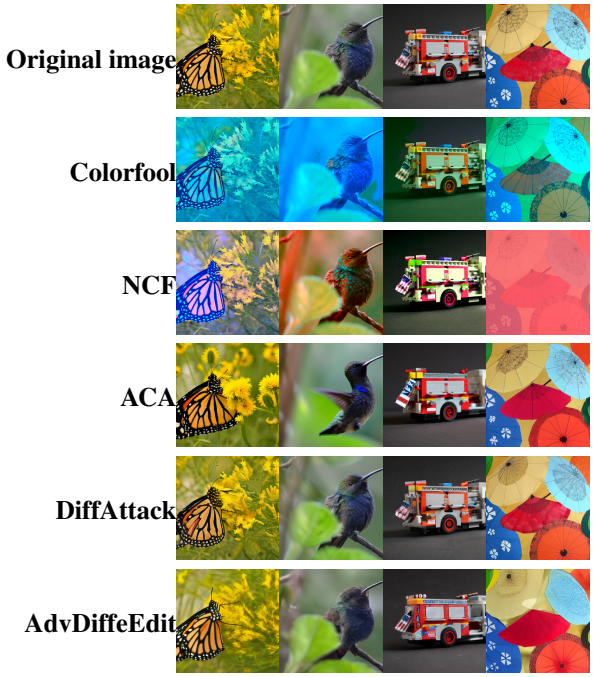

Figure 4. Qualitative comparison on 4 ImageNet samples. Rows: Original; ColorFool; Natural Color Fool(NCF); Adversarial Content Attack(ACA) ; AdvDiff; AdvDiffEdit (ours). Color-based attacks alter global hues; diffusion-based edits preserve structure while flipping the decision.

steps. To align with continuous-time notation, we index diffusion strength by a fraction $t \in [0, 1]$ and map it to the scheduler's discrete index via $t^\star = \lfloor t\,(N_{\text{train}} - 1) \rfloor$, with $N_{\text{train}} = 1000$. Unless otherwise noted, all reported results use the same default search schedule: $t_0 = 0.3, t_{\min} = 0.2, M = 5, K = 50$, together with a lightweight refinement budget of 50 queries per stage. We choose this configuration as a quality-oriented operating point: larger initial noise levels can often find adversarial seeds more quickly, but they also tend to introduce stronger semantic drift and visibly worse image quality. Our default setting therefore favors visually natural edits over aggressive high-noise search.

5.2. Source-Model Attacks under a Fixed Search Schedule

We compare against decision-only pixel-space methods: Boundary Attack (BA) [1] and the accelerated HopSkipJumpAttack (HSJA) [3]. BA typically requires many iterations and often introduces visible high-frequency artifacts, whereas HSJA achieves lower pixel distortion through local boundary refinement in image space.

Table 1. ℓ_2 distortion on ImageNet under a matched query budget.

Method	1k Queries	20k Queries
Boundary Attack	79.6	40.9
HSJA	71.4	10.9
AdvDiffEdit (Ours)	**62.1**	–[†]

[†] AdvDiffEdit terminates once a successful edit is found.

AdvDiffEdit is not designed as an anytime pixel-space optimizer. Instead, it follows a fixed coarse-to-fine search schedule that starts from $t_0 = 0.3$, progressively shrinks the edit strength, and applies a small refinement budget at each stage, leading to roughly 1k oracle queries per image in our default configuration. Under this operating point, AdvDiffEdit achieves an average ℓ_2 distortion of 62.1, compared with 79.6 for Boundary Attack and 71.4 for HSJA (Table 1). We report "–" for AdvDiffEdit at 20k queries because additional queries do not play the same role as they do in BA or HSJA. Pixel-space attacks admit continued local boundary refinement and can therefore keep reducing ℓ_2 as the budget grows. In contrast, our method searches for the first successful edit along the diffusion manifold and then stops. The goal is not to asymptotically minimize pixel distortion, but to obtain a successful edit that remains visually natural.

This distinction is evident qualitatively in Fig. 3. At its default operating point, AdvDiffEdit typically produces mild color, texture, or semantic adjustments, whereas BA and HSJA often yield grainy high-frequency perturbations. This perceptual difference is also reflected in substantially lower FID than BA/HSJA and color-based baselines (Table 2), while maintaining LPIPS comparable to diffusion-based attacks.

5.3. Cross-Architecture Transferability

While the previous section demonstrated that AdvDiffEdit produces more realistic images than pixel-space decision-only attacks, we now investigate whether this "realism" pro-

Table 2. Transferability comparison on ImageNet (Source: ResNet-50). Entries report ASR (%) evaluated on the final saved adversarial images; the diagonal corresponds to same-model re-evaluation, while off-diagonal entries measure cross-model transferability. **Top Block:** Hard-label methods (ours vs. baselines). **Bottom Block:** Score/gradient-based methods. Classical hard-label attacks (BA/HSJA) exhibit very limited transferability, whereas **AdvDiffEdit** uniquely combines hard-label access with high transferability.

		ASR (%) on Target Models				Quality & Cost			
Access	**Method**	**Res-50**	**MN-v2**	**ViT-B**	**Swin-B**	**Avg†**	**Avg Time**	**FID↓**	**LPIPS↓**
Hard Label	Boundary Attack (20k)	93.0	10.0	7.5	7.0	8.2	124.3	126.0	0.052
	HSJA (5k)	92.0	10.5	5.5	5.3	7.1	42.6	115.0	0.043
	AdvDiffEdit (Ours)	96.8	71.8	56.4	50.3	59.5	24.1	68.2	0.144
Score/Gradient	ColorFool	90.4	41.6	20.3	15.0	25.6	12.1	72.3	0.231
	NCF	89.6	71.4	37.8	21.6	43.6	10.5	70.9	0.383
	ACA	88.8	65.2	52.7	48.3	55.4	65.3	63.4	0.137
	DiffAttack	96.3	75.9	52.9	56.4	61.7	32.2	62.2	0.127

† Avg excludes the surrogate (Res-50).

vides functional benefits.

Traditional decision-only attacks (like HSJA) typically fail to transfer because they overfit to the specific decision boundary of the target model via high-frequency noise. In contrast, by editing strictly along the diffusion manifold, we hypothesize that AdvDiffEdit captures robust semantic features that should generalize across models. To test this, we compare against recent unrestricted attacks: two *color–manipulation* methods—ColorFool [29] and Natural Color Fool (NCF) [38]—which utilize score-based black-box access; and two *diffusion-based content* attacks—Content-based unrestricted ad- versarial attack (denoted as ACA) [6] and DiffAttack [4]. Targets cover two CNNs (ResNet-50 [12], MobileNetV2 [28]) and two Vision Transformers (ViT-B/16 [8], Swin-B [20]).

Following [4, 6], we measure *transferability*. For each source model f_s, we generate an unrestricted adversarial set and then evaluate the attack success rate (ASR, %) on every target $f_t \neq f_s$ under a hard-label oracle. For context, we also list the white-box diagonal from prior methods. Crucially, our method uses *only* binary feedback—no scores or gradients. We parameterize SDEdit strength by a fraction $t \in [0, 1]$ of the forward-noising schedule and cap at $t_{\max} = 0.6$; if success requires $t > t_{\max}$, we stop and count a failure to avoid excessive semantic drift. A lightweight refinement stage suppresses artifacts with modest extra queries.

Color-based attacks mainly shift global hues and saturation. ACA uses null-text inversion and reconstruction, while DiffAttack augments its loss with attention consistency. Consequently, both diffusion baselines explicitly optimize fidelity using scores and gradients.

The top block reports strict hard-label methods, while the bottom block includes score- or gradient-based methods with stronger access. These comparisons are intended to contextualize the transferability/realism trade-off rather than to claim equal-oracle fairness.

(a) Original (b) No refine adv (c) Refine 1k adv (d) Refine 2k adv

Figure 5. Refinement ablation under a hard-label oracle. (a) Original. (b) No refinement: adversarial but with visible artifacts. (c) +1k refinement queries and (d) +2k refinement queries progressively reduce artifacts and improve realism while retaining the adversarial effect.

As shown in Table 2, AdvDiffEdit achieves strong transferability under a strict hard-label setting, remaining competitive with score- and gradient-based baselines. Across transfer targets, it substantially exceeds classical black-box attacks such as ColorFool, outperforms ACA on all targets, and is comparable to DiffAttack overall, including a higher ASR on ViT-B. While gradient-based diffusion methods achieve slightly lower FID and LPIPS, our method maintains a good balance between transferability and realism, with a practical runtime of 24.1 s/image under only 1,000 hard-label queries.

5.4. Transferability to Defended Targets

To further verify the robustness of each attack method, we evaluate the performance of the crafted adversarial examples on defense approaches. The surrogate model is ViT-B.

Preprocessing defenses. We evaluate three preprocessing defenses—HGD [19], R&P [35], and DiffPure [24]. For DiffPure, we match the published noise magnitude settings. Our adversarial images retain strong ASR against these defenses (see Table 3), indicating that manifold-aligned edits can survive common input transformations.

Adversarial training. We further test adversarially trained models (Inc-v3ens3, Inc-v3ens4, Adv-Inc-v3, Inc-v2ens) from [33], using ViT-B as the surrogate for transfer. Despite lower absolute rates relative to undefended models,

Table 3. Robustness against preprocessing defenses (HGD, R&P, DiffPure) and adversarially trained models (Adv-Inc-v3, Inc-v3$_{ens3/ens4}$, IncRes-v2$_{ens}$) under a hard-label oracle. Entries are ASR (%, higher = less robust). Surrogate: ViT-B. Classical hard-label attacks (Boundary Attack / HSJA) are included for comparison.

		HGD	R&P	DiffPure	Adv-Inc-v3	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Hard Label	Boundary Attack(20k)	1.60	7.28	20.85	8.08	10.69	13.32	5.34
	HSJA(5k)	2.03	6.58	17.74	7.68	10.24	12.44	5.22
	AdvDiffEdit(Ours)	46.3	51.5	**67.5**	**61.9**	58.4	59.3	46.3
Score/ Gradient	ColorFool	15.1	13.4	33.6	40.0	17.6	22.5	10.8
	NCF	29.6	33.6	30.8	51.2	32.8	31.0	21.5
	ACA	**52.0**	**57.3**	64.7	56.9	**59.0**	**62.7**	**53.1**
	DiffAttack	38.5	35.9	54.2	54.0	56.2	56.9	41.7

Table 4. Ablation on maximum edit strength $t_{\max}$. Entries are ASR (%, higher is better). Increasing $t_{\max}$ improves success, while refinement improves visual quality (Fig. 5) with only a marginal cost to ASR.

	ASR (%)	
$t_{\max}$	No Refinement	With Refinement (1k)
0.3	5.6	5.4
0.4	35.5	30.7
0.5	75.4	72.4
0.6	97.3	93.7

our method maintains competitive ASR under these robust targets. Notably, DiffPure—conceptually related to SDEdit—still permits substantial residual attack success (e.g., 67.5% in our setting), suggesting that diffusion-style purification does not necessarily return images to the correct decision region.

5.5. Ablation Studies

Edit Strength. We control edit strength via the diffusion fraction $t \in [0, 1]$. Our search seeks the smallest t necessary to find a valid adversarial seed. We sweep $t_{\max} \in \{0.3, 0.4, 0.5, 0.6\}$ and report ASR, where failure indicates no seed was found below the cap.

As shown in Table 4, ASR increases monotonically with $t_{\max}$. For $t_{\max} < 0.3$, random search rarely succeeds, whereas gradient-based methods often succeed here by exploiting high-frequency, imperceptible noise. Our method requires slightly higher strengths ($t \in [0.5, 0.6]$) precisely because it induces *semantic* flips rather than pixel-noise perturbations. This results in adversarial examples that are structurally faithful and robust, yielding a favorable realism–success trade-off.

Noise Refinement. The refinement stage—consisting of short manifold walks—suppresses artifacts at the cost of additional queries. The total overhead scales linearly as $Q_{\text{refine}} \approx P \times (I \cdot W)$, where P is passes, I is steps, and W is walks (typically adding $\sim$1,000 queries).

In practice, this is a high-value trade-off: refinement improves visual fidelity and stabilizes the output on the manifold (see Fig. 5), with only a marginal drop in ASR (see Table 4). To further bias towards realism, we prioritize successful candidates with lower ℓ_2 mix with LPIPS distance to x_{in} during the selection process.

Responsible Use. Our goal is to support robustness auditing under realistic hard-label access, not to facilitate misuse. AdvDiffEdit is intended for controlled security evaluation, red teaming, and the study of decision-only vulnerabilities in deployed vision systems. Following standard practice in adversarial-robustness research, we report aggregate results and focus on evaluation protocols rather than deployment-oriented misuse scenarios. We also note that stronger attack capability should be paired with improved defenses and more careful assessment of real-world model exposure.

6. Conclusion

In this work, we presented **AdvDiffEdit**, a diffusion-guided framework for adversarial editing in the challenging decision-only setting. By integrating a randomized adversarial seed search with a coarse-to-fine diffusion refinement loop, our method effectively "walks the natural image manifold" to discover subtle, naturalistic visual perturbations that flip model decisions without requiring gradients or confidence scores.

Empirical evaluations on ImageNet demonstrate that AdvDiffEdit offers a strong trade-off between query budget, attack success, and visual fidelity compared to traditional pixel-space baselines. While methods such as HSJA often rely on high-frequency perturbations that degrade visual quality, our diffusion-guided approach produces more naturalistic edits that better preserve image realism. Crucially, we show that this manifold alignment yields transferability competitive with gradient-based attacks, despite using only hard-label feedback. Overall, our results suggest that diffusion-guided editing is not only an effective unrestricted attack mechanism, but also a practical tool for auditing whether black-box vision systems are vulnerable to subtle, visually plausible cues under strict information constraints.

References

[1] 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 (ICLR)*, 2018.

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, San Jose, CA, USA, 2017. IEEE.

[3] Jianbo Chen, Michael I. Jordan, and Martin J. Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1277–1294, 2020.

[4] Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):961–977, 2025.

[5] Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4562–4572, 2023.

[6] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 51719–51733, 2023.

[7] Xuelong Dai, Kaisheng Liang, and Bin Xiao. Advdiff: Generating unrestricted adversarial examples using diffusion models. In *Computer Vision – ECCV 2024*, pages 93–109. Springer, Cham.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.

[9] Shuo Feng, Haowei Sun, Xintao Yan, Haojie Zhu, Zhengxia Zou, Shengyin Shen, and Henry X. Liu. Dense reinforcement learning for safety validation of autonomous vehicles. *Nature*, 615(7953):620–627, 2023.

[10] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, San Diego, CA, USA, 2015. OpenReview.net.

[12] 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 (CVPR)*, pages 770–778, Las Vegas, NV, USA, 2016. IEEE.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017.

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

[15] Jonathan Ho, Ajay N. Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems 33*, 2020.

[16] 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*, 2019.

[17] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems 35*, 2022.

[18] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations (ICLR), Workshop Track*, Toulon, France, 2017.

[19] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, Salt Lake City, UT, USA, 2018. Computer Vision Foundation / IEEE.

[20] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, Montreal, QC, Canada, 2021. IEEE.

[21] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *Advances in Neural Information Processing Systems 35*, 2022.

[22] 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 (ICLR)*, Vancouver, BC, Canada, 2018. OpenReview.net.

[23] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[24] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pages 16805–16827, Baltimore, Maryland, USA, 2022. PMLR.

[25] Zhengzhao Pan, Hua Chen, and Xiaogang Zhang. Diffadvmap: Flexible diffusion-based framework for generating natural unrestricted adversarial examples. In *International Conference on Machine Learning (ICML)*, 2025. ICML 2025 Poster.

[26] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Prac-

tical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security (ASIACCS)*, pages 506–519, Abu Dhabi, United Arab Emirates, 2017. ACM.

[27] Chicago Y. Park, Michael T. McCann, Cristina Garcia-Cardona, Brendt Wohlberg, and Ulugbek S. Kamilov. Random walks with tweedie: A unified view of score-based diffusion models. *IEEE Signal Processing Magazine*, 42(3): 40–51, 2025.

[28] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, Salt Lake City, UT, USA, 2018. IEEE.

[29] Ali Shahin Shamsabadi, Ricardo Sánchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[30] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1528–1540, Vienna, Austria, 2016. ACM.

[31] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2018.

[32] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[33] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018. OpenReview.net.

[34] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011.

[35] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan L. Yuille. Mitigating adversarial effects with randomization. In *International Conference on Learning Representations (ICLR)*, Vancouver, BC, Canada, 2018. OpenReview.net.

[36] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[37] Fan Yin, Philippe Laban, Xiangyu Peng, Yilun Zhou, Yixin Mao, Vaibhav Vats, Linnea Ross, Divyansh Agarwal, Caiming Xiong, and Chien-Sheng Wu. Bingoguard: Llm content moderation tools with risk levels. In *The Thirteenth International Conference on Learning Representations*, 2025. Published as a conference paper at ICLR 2025.

[38] Shengming Yuan, Qilong Zhang, Lianli Gao, Yaya Cheng, and Jingkuan Song. Natural color fool: Towards boosting black-box unrestricted attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[39] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[40] Yilan Zhang, Fengying Xie, and Jianqi Chen. Tformer: A throughout fusion transformer for multi-modal skin lesion diagnosis. *Computers in Biology and Medicine*, 157:106712, 2023.

[41] Yao Zhu, Yuefeng Chen, Xiaodan Li, Kejiang Chen, Yuan He, Xiang Tian, Bolun Zheng, Yaowu Chen, and Qingming Huang. Toward understanding and boosting adversarial transferability from a distribution perspective. *IEEE Transactions on Image Processing*, 31:6487–6501, 2022.