TPA: EARLY TEXT-EMBEDDING PERTURBATIONS FOR NATURAL ADVERSARIAL DIFFUSION

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural adversarial examples show that classifiers can fail at test time even on images that are clearly recognizable to humans, not just on images modified with small input perturbations. A key challenge is to make adversarial examples work across different models without reducing image quality. Text conditioning influences which visual variations of a class the diffusion model explores. Keeping the text condition fixed may restrict adversarial search to only a subset of these variations. We propose Text-embedding Perturbation Attack (TPA), which temporarily perturbs text embeddings during early denoising and progressively restores the original conditions while guiding sampling with classifier gradients. This exploration changes both the denoising trajectory and the intermediate images on which adversarial guidance acts. On ImageNet-1K, TPA achieves higher black-box attack success rates than existing natural adversarial generation methods across all evaluated surrogate–victim pairs, while outperforming NatADiff on most learned image-quality metrics. It also achieves lower FID to ImageNet-A than both baselines. Matched ablations show consistent transfer gains from early conditioning perturbations across the evaluated black-box classifiers, including adversarially trained models. These results establish early text-conditioning exploration as an effective strategy for generating transferable natural adversarial examples and probing test-time vulnerabilities across classifiers.

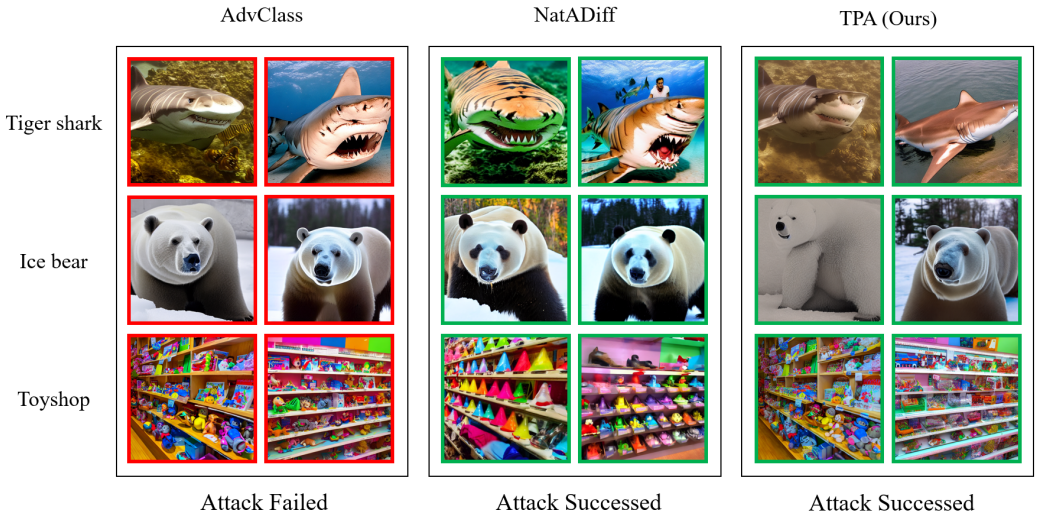


Figure 1: **Qualitative comparison with an RN50 surrogate and an RN152 victim.** AdvClass, NatADiff, and TPA generate examples of the same requested class in each row. The selected TPA outputs retain recognizable class features while fooling the victim. Green/red borders indicate victim attack success/failure.

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.

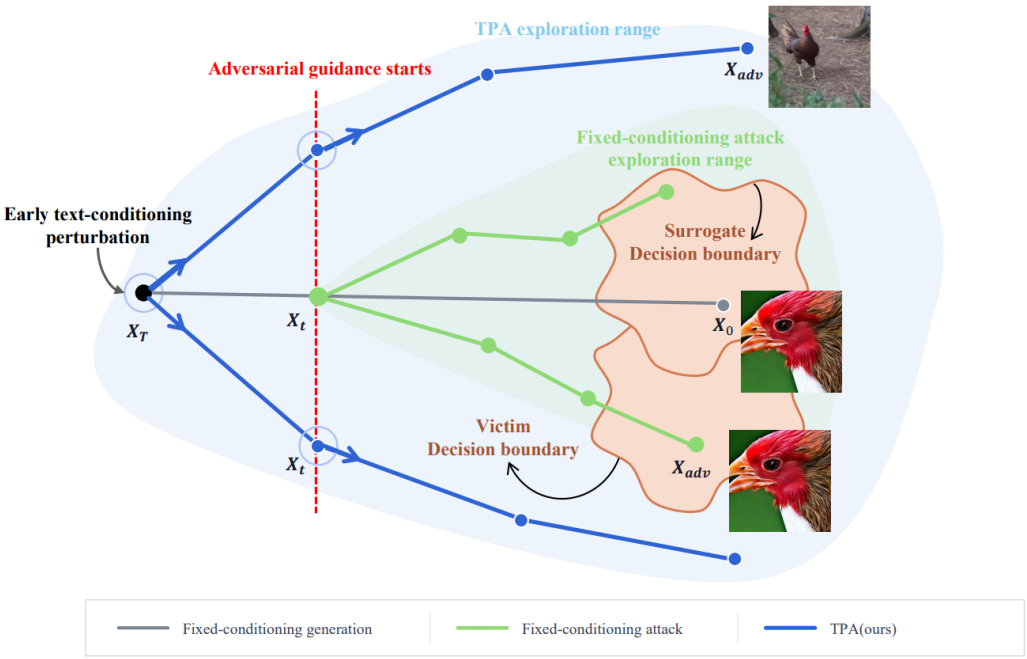


Figure 2: **Early text-conditioning exploration in TPA.** Gray denotes fixed-conditioning generation without adversarial guidance, green a fixed-conditioning attack, and blue TPA. Early text-conditioning perturbations change the sampling state at which adversarial guidance begins (red dashed line). The original conditioning is progressively restored as guided sampling continues. Shading indicates hypothesized exploration ranges; orange contours depict schematic surrogate and victim decision boundaries for the requested class. Only the surrogate supplies attack gradients. X_T and X_t denote initial and intermediate sampling states, while X_0 and X_{adv} denote unguided and adversarial outputs. The depicted ranges, boundaries, and trajectories are illustrative.

1 Introduction

Deep neural networks are vulnerable to adversarial examples, where small, carefully chosen perturbations can cause confident misclassification while remaining nearly imperceptible to humans (Szegedy et al., 2014; Goodfellow et al., 2015). Much subsequent work has formalized this vulnerability through norm-bounded threat models and developed defenses such as adversarial training (Madry et al., 2018). Robustness under such threat models is important, but it characterizes model behavior only within a prescribed perturbation set around an existing input. Unrestricted adversarial attacks broaden this setting by removing the requirement that adversarial examples remain within a small pixel-space norm ball, instead allowing semantic variations or synthesis through generative models (Song et al., 2018; Zhao et al., 2018). This broader setting asks whether a classifier remains reliable not only under small perturbations, but also across semantically valid variations of a class.

Natural adversarial examples make this broader failure mode particularly relevant to model evaluation. Rather than relying on constructed pixel perturbations, ImageNet-A consists of real-world, unmodified examples that reliably induce errors across unseen classifiers, exposing weaknesses shared by different models (Hendrycks et al., 2021). Generative natural adversarial attacks seek to systematically synthesize recognizable class instances that induce analogous classifier failures, providing a complementary way to probe robustness beyond local perturbation threat models. Here, we consider generation from a class prompt and random noise, without a reference image. Diffusion models are well suited to this setting because they provide a strong image prior while allowing classifier feedback to guide sampling toward misclassification (Chen et al., 2023a; Dai et al., 2024). Recent approaches refine adversarial guidance to improve attack effectiveness and visual quality (Kuurila-Zhang et al., 2025). NatADiff further guides sampling toward intersections of source and

adversarial classes to improve transfer and resemblance to natural test-time errors (Collins et al., 2026). Other work optimizes class-token embeddings to induce classifier errors (Lin et al., 2024). Yet generating high-quality class-recognizable examples whose failures transfer beyond the surrogate remains challenging. We investigate a complementary direction: can transiently perturbing text conditioning during early denoising and restoring it later improve the transferability of generated attacks?

We study this question by treating text conditioning as an exploration variable. Text conditioning shapes the denoising trajectory and the intermediate images on which adversarial guidance acts. A requested class admits diverse appearances, poses, and contexts, yet a fixed condition may favor a narrower subset of these configurations. Condition-annealed sampling has shown that temporarily perturbing the conditioning signal can increase generation diversity while largely retaining condition alignment (Sadat et al., 2024). We hypothesize that relaxing text conditioning early in denoising can alter the configurations encountered by adversarial guidance and improve access to recognizable class instances that induce failures across classifiers. The goal is not diversity alone, but a greater chance of reaching class-valid failures shared by the surrogate and unseen victims (Figure 2). Restoring the original condition later reinforces the requested class signal while sampling continues from a trajectory already shaped by the earlier exploration.

We propose TPA, which injects Gaussian noise into text embeddings during early denoising and progressively restores the original conditions while guiding the sampling trajectory with classifier gradients. TPA uses the condition-annealing rule of Sadat et al. (2024) and requires no additional training of the generator or surrogate. On ImageNet-1K, TPA achieves higher black-box attack success rates than NatADiff and AdvClass on every evaluated surrogate–victim pair, while outperforming NatADiff on most learned image-quality metrics and achieving lower FID to the ImageNet-A reference than both baselines. Our contributions are threefold: (i) we introduce transient early text-conditioning exploration as a strategy for transferable natural adversarial generation; (ii) we demonstrate stronger black-box transfer with competitive image quality relative to existing generative attack baselines; and (iii) we isolate the contribution of early conditioning perturbations through matched ablations, showing consistent transfer gains across convolutional and transformer-based black-box classifiers, including adversarially trained models.

2 Related Work

Generative adversarial attacks. Unrestricted adversarial attacks extend adversarial search beyond small pixel-space perturbations through semantic transformations and generative priors (Song et al., 2018; Zhao et al., 2018; Bhattad et al., 2020; Chen et al., 2023b). Diffusion-based approaches include modifying a reference image (Chen et al., 2025; Xue et al., 2023) and synthesizing adversarial images from noise (Chen et al., 2023a; Dai et al., 2024). TPA follows the latter setting, generating images from a requested class prompt without a reference image. Recent methods refine guidance to balance attack effectiveness and image quality. VENOM adaptively activates adversarial guidance and accumulates gradient momentum (Kuurila-Zhang et al., 2025), while NatADiff combines source–adversarial class conditioning, augmented gradients, and time-travel sampling to guide generation toward class intersections (Collins et al., 2026). TPA also uses classifier guidance, but explores conditioning through transient embedding perturbations and restoration, without explicitly conditioning on a source–adversarial class intersection.

Semantic constraints and transferability. Probabilistic attacks express the image variations available to an attack and the classifier errors it seeks as distributions. ProbAttack combines an image-centered similarity distribution with a classifier-induced attack preference (Zhang et al., 2024). ConceptAdv extends this formulation to concept distributions that can represent an individual identity or a broader category (Zhang et al., 2026). DiffAdvMAP uses diffusion priors and scenario-specific constraints to approximate adversarial posteriors, supporting generation with or without a reference image (Pan et al., 2025). Transfer has also been studied through shared non-robust features (Ilyas et al., 2019) and displacement from the source data distribution (Zhu et al., 2022); momentum, input diversity, and translation-invariant gradients improve transfer in pixel-space attacks (Dong et al., 2018; Xie et al., 2019; Dong et al., 2019). TPA studies how temporary changes to conditioning alter the trajectories explored under surrogate guidance.

Text-conditioned adversarial attacks. Text conditioning provides a way to control the content explored by generative attacks. Existing approaches optimize continuous embeddings or search discrete prompts. SD-NAE optimizes class-token embeddings using classifier gradients (Lin et al., 2024), Prompt2Perturb learns prompt embeddings to attack reference ultrasound images (Medghalchi et al., 2025), and Zhu et al. (2024) search for adversarial prompts using classifier queries and semantic consistency. In TPA, embedding noise is sampled independently of classifier feedback; classifier gradients guide the latent trajectory. The original embeddings are progressively restored during sampling.

Conditioning and sampling diversity. In conditional generation, the strength and timing of classifier-free guidance affect sample quality and diversity (Wang et al., 2024; Kynkäänniemi et al., 2024). CADS perturbs the conditioning signal and anneals the noise during sampling, increasing diversity while largely retaining condition alignment (Sadat et al., 2024). TPA adopts the CADS perturbation and rescaling rule to implement early conditioning exploration within adversarial sampling. This changes both the denoising predictions and the intermediate images used to compute attack gradients. Our study evaluates the resulting contribution to cross-model transfer and its image-quality trade-offs.

3 PRELIMINARIES

Natural adversarial generation and threat model. Given a requested (source) class y, natural adversarial generation seeks a natural image that humans recognize as y but a victim classifier misclassifies. Images are generated from a class prompt and a random latent, without a reference image or a pixel-distance constraint. The attacker has gradient access to the frozen text-to-image generator and surrogate classifier f_s. Black-box victims are not queried during generation or output selection. Attack success is evaluated by untargeted misclassification, even when a non-source target is used internally for guidance; class validity is assessed separately.

Conditional diffusion sampling. We use latent diffusion (Rombach et al., 2022) with deterministic DDIM sampling (Song et al., 2021). At sampling iteration k, the latent $z^{(k)}$ corresponds to diffusion timestep t_k, which decreases as k increases. The denoiser $\epsilon_\theta(z, t, c)$ predicts noise given a latent, timestep, and text embedding. We denote the cached embeddings of the source-class and null prompts by c^+ and c^-. At each step, ϵ_k^+ and ϵ_k^- denote the denoiser predictions under the conditions supplied to these two branches. Classifier-free guidance combines them as $\epsilon_k = \epsilon_k^- + w(\epsilon_k^+ - \epsilon_k^-)$, where w is the guidance scale (Ho & Salimans, 2021).

Let $\bar{\alpha}_t = \prod_{i=0}^{t}(1-\beta_i)$ for the forward variance schedule β_i, using zero-based indices, and write $a_k = \sqrt{\bar{\alpha}_{t_k}}$ and $b_k = \sqrt{1-\bar{\alpha}_{t_k}}$. The DDIM transition is

$$z^{(k+1)} = \mathrm{DDIM}_k(z^{(k)}, \epsilon_k) = a_{k+1}\frac{z^{(k)} - b_k\epsilon_k}{a_k} + b_{k+1}\epsilon_k. \tag{1}$$

The fraction gives the predicted clean latent, which we decode to compute adversarial guidance (Section 4.2). The next-state coefficients, including the terminal convention, follow the sampler configuration in Appendix B.1.

4 METHOD

TPA perturbs text embeddings early in sampling and progressively restores them. At guided steps, a surrogate gradient computed from the resulting denoised image estimate modifies the latent update.

4.1 CONDITIONING PERTURBATIONS ALONG ADVERSARIAL TRAJECTORIES

Text conditioning shapes both the denoising trajectory and the intermediate image estimates used to compute adversarial guidance. Conditional sampling under a fixed prompt can favor a limited range of image configurations (Sadat et al., 2024). We therefore vary conditioning during early image formation and progressively return to the original class condition as generation proceeds.

Figure 2 schematically illustrates this design by contrasting TPA with fixed-conditioning generation and attacks.

TPA starts from a source-class prompt and a random latent. During early denoising, it injects Gaussian noise into the source and null text embeddings and progressively reduces the perturbations. At each guided step, the current embeddings determine the denoising prediction and the image estimate evaluated by the surrogate. The surrogate gradient is then backpropagated through this estimate to guide the latent update. Conditioning perturbations therefore alter both the diffusion update and the image-dependent attack direction. The embedding noise is sampled independently of classifier feedback; attack gradients act on the latent trajectory.

As the perturbations vanish, sampling continues under the original class conditioning from the current latent state. In our main setting, adversarial guidance begins before the conditioning perturbations vanish and continues afterward. Section 5.3 evaluates the contribution of this intervention to cross-model transfer under matched guidance settings.

4.2 Early text-embedding perturbations and restoration

We implement conditioning exploration using the CADS perturbation and rescaling rule (Sadat et al., 2024). At step k, each cached source or null embedding $c \in \{c^+, c^-\}$ is perturbed as

$$\widetilde{c}_k = \sqrt{\gamma_k} c + s_c \sqrt{1-\gamma_k}\,\xi_k, \qquad \xi_k \sim \mathcal{N}(0, I). \tag{2}$$

Here s_c is the noise amplitude, and γ_k increases from zero to one so that the perturbations vanish as sampling proceeds. Noise is drawn independently for each step and embedding. Each perturbation is applied to the cached original embedding, rather than accumulated across steps.

Following CADS, we rescale the perturbed embedding using the original mean and standard deviation:

$$\bar{c}_k = \mu(c) + s(c)\,\frac{\widetilde{c}_k - \mu(\widetilde{c}_k)}{\max\{s(\widetilde{c}_k), \epsilon_c\}}. \tag{3}$$

Here μ and s are the scalar mean and population standard deviation across embedding tokens and channels, and $\epsilon_c > 0$ is a numerical floor. The resulting embeddings $\bar{c}_k^{\pm}$ determine the classifier-free noise prediction ϵ_k defined in Section 3. Appendix B.1 specifies the annealing schedule and numerical settings.

Following training-free diffusion guidance (Bansal et al., 2024; Shen et al., 2024), we evaluate the attack on the denoised image estimate

$$\widehat{x}_0^{(k)} = D\left(\frac{z^{(k)} - b_k \epsilon_k}{a_k}\right), \tag{4}$$

where D includes VAE decoding and RGB conversion. This estimate links the conditioning rule to adversarial guidance: changing the embeddings affects both $\widehat{x}_0^{(k)}$ and its derivative with respect to $z^{(k)}$, through which the attack reward is backpropagated. The corresponding latent update is defined next.

4.3 Adversarial guidance along the trajectory

For a sampler with N steps, let S denote the guidance onset index. At each step $k \geq S$, we evaluate the surrogate on M deterministic views of the image estimate in Eq. (4), obtaining logits $\ell_{k,u} = f_s(T_u(\widehat{x}_0^{(k)}))$ and their mean $\bar{\ell}_k = M^{-1}\sum_u \ell_{k,u}$. Multiple views encourage consistent misclassification under transformations (Athalye et al., 2018; Shen et al., 2024; Collins et al., 2026).

A non-source target can provide a consistent direction for adversarial generation, motivating targeted guidance in prior generative attacks (Dai et al., 2024; Kuurila-Zhang et al., 2025; Collins et al., 2026). We use a two-stage objective: an initial untargeted phase leaves the competing class open, and a later fixed-target phase focuses the remaining updates on one competitor. At the switch index K, we select the strongest non-source class in the current mean logits, $q = \arg\max_{j \neq y} \bar{\ell}_{K,j}$, and hold it fixed thereafter. The selected target therefore depends on the trajectory reached during the first

phase. Throughout sampling, the source text condition refers to the requested class y; q enters only the classifier reward. With σ denoting the sigmoid function, we use

$$R_k = \begin{cases} \dfrac{1}{M}\displaystyle\sum_u \log \sigma\left(\max_{j \neq y} \ell_{k,u,j} - \ell_{k,u,y}\right), & S \leq k < K, \\ \bar{\ell}_{k,q} - \bar{\ell}_{k,y}, & k \geq K. \end{cases} \tag{5}$$

This schedule is a design choice for the untargeted generation task. Appendix E.2 compares it with targeted-only and untargeted-only variants, examining the resulting ASR–quality trade-offs.

Let $d_k = \nabla_{z^{(k)}} R_k$. We obtain g_k by scaling d_k in the first phase and its L_2-normalized direction in the second; $g_k = 0$ before S. The onset, switch, and guidance strengths are hyperparameters; Appendix B.3 gives the exact scaling, and Appendix B.2 lists the settings. Both phases apply reward ascent through the noise prediction (Dhariwal & Nichol, 2021), using the DDIM transition defined in Eq. (1):

$$\epsilon_k^{\mathrm{adv}} = \epsilon_k - b_k g_k, \qquad z^{(k+1)} = \mathrm{DDIM}_k(z^{(k)}, \epsilon_k^{\mathrm{adv}}). \tag{6}$$

At each guided step, we differentiate through the current denoiser, decoder, views, and surrogate with respect to the incoming latent, keeping model parameters frozen. Algorithm 1 produces one output per latent/condition seed pair, without retries or output filtering.

Algorithm 1 Generating one image with TPA

Require: Source label y; frozen denoiser ϵ_θ, decoder D, and surrogate f_s;
latent/condition seeds; N, S, K, w and remaining settings (Appendix B).

Cache source/null embeddings c^+, c^-; draw $z^{(0)} \sim \mathcal{N}(0, I)$.
for $k = 0, \ldots, N-1$ **do**
  Obtain $\bar{c}_k^+, \bar{c}_k^-$ from cached c^+, c^- using independent noise (Eqs. (2)–(3)).
  $\epsilon_k^{\pm} \leftarrow \epsilon_\theta(z^{(k)}, t_k, \bar{c}_k^{\pm})$; $\epsilon_k \leftarrow \epsilon_k^- + w(\epsilon_k^+ - \epsilon_k^-)$.
  $g_k \leftarrow 0$.
  if $k \geq S$ **then**
    $\widehat{x}_0^{(k)} \leftarrow D\big((z^{(k)} - b_k \epsilon_k)/a_k\big)$ (Eq. (4)).
    $\ell_{k,u} \leftarrow f_s(T_u(\widehat{x}_0^{(k)}))$ for $u = 1, \ldots, M$; $\bar{\ell}_k \leftarrow M^{-1} \sum_{u=1}^M \ell_{k,u}$.
    if $k = K$ **then**
      $q \leftarrow \arg\max_{j \neq y} \bar{\ell}_{k,j}$; hold q fixed for subsequent steps.
    end if
    Compute R_k using the untargeted ($k < K$) or fixed-target ($k \geq K$) reward in Eq. (5).
    $d_k \leftarrow \nabla_{z^{(k)}} R_k$; scale d_k to obtain g_k using Eq. (8) (Appendix B.3).
  end if
  $\epsilon_k^{\mathrm{adv}} \leftarrow \epsilon_k - b_k g_k$ (Eq. (6)).
  $z^{(k+1)} \leftarrow \mathrm{DDIM}_k(z^{(k)}, \epsilon_k^{\mathrm{adv}})$ (Eqs. (1), (6)).
end for
return $D(z^{(N)})$.

5 Experiments

5.1 Experimental setup

Data and sample generation. We evaluate on all 1,000 ImageNet classes (Russakovsky et al., 2015), generating images from class prompts and random latents without reference images. All methods use the same source prompts and two fixed latent seeds per class, yielding 2,000 final outputs for each method–surrogate combination. Every final output is evaluated without filtering by attack success or visual appearance.

Surrogate and victim models. We use ResNet-50 (RN50), Inception-v3, and ViT-H/14 as surrogates in separate experiments. Each generated set is evaluated on the same nine-classifier panel, covering convolutional and transformer architectures and including two adversarially trained models. The generating surrogate provides the white-box result; the other eight classifiers are black-box

victims and are not used for generation or output selection. Checkpoints and preprocessing are fixed across methods and specified in Appendix C.

Diffusion model and TPA settings. All methods use Stable Diffusion 1.5 (SD1.5) at 512×512 with classifier-free guidance scale 7.5. TPA uses 50 DDIM steps, with 35 guided steps at diffusion timesteps $0 < t < 700$, and one sampling attempt per output. The conditioning schedule, surrogate-specific guidance hyperparameters, and implementation details are given in Appendix B.

Comparison methods. We compare with NatADiff and its similarity-target AdvClass baseline (Collins et al., 2026), which address the same reference-free generation setting. AdvClass provides a classifier-guided baseline, while NatADiff additionally uses mixed-class conditioning and time-travel sampling. Both select a fixed non-source target by CLIP text similarity and use a 200-step DDIM grid with up to five attempts per output. Only the final image from each attempt sequence is retained, including failures at the retry limit. Comparisons therefore use the same number of final outputs with method-specific sampling budgets. Baseline hyperparameters, retry procedures, differences from the public NatADiff runner, and computational costs are detailed in Appendix B.

Evaluation metrics. Following NatADiff (Collins et al., 2026), attack success rate (ASR) is the percentage of final outputs whose top-1 prediction differs from the requested class. This untargeted criterion applies to all methods, including those using targeted guidance. *Average ASR* is the arithmetic mean over all nine classifiers, including the surrogate; the eight non-surrogate entries measure black-box transfer.

We assess the same outputs with seven learned no-reference image-quality (NR-IQA) metrics and BRISQUE, reporting per-image means in Table 2. Fréchet Inception Distance (Heusel et al., 2017) measures set-level similarity to ImageNet validation images (FID-VAL) and ImageNet-A natural adversarial examples (FID-A) (Hendrycks et al., 2021; Collins et al., 2026). FID-A compares our 1,000-class outputs with a 200-class reference set. Appendix C specifies the reference sets and metric implementations. Human recognition of the requested class is evaluated separately in Appendix F.

5.2 Main Results

Tables 1 and 2 summarize attack success, cross-model transfer, and image quality on ImageNet-1K.

Attack success and transferability. TPA achieves the highest Average ASR in all three surrogate settings, exceeding NatADiff by 2.20–6.06 percentage points (Table 1). It outperforms both NatADiff and AdvClass in all 24 evaluated black-box surrogate–victim pairs, including adversarially trained victims. TPA's white-box ASR is lower than NatADiff's in every setting, so its higher Average ASR comes from improved transfer to other classifiers. These comparisons use the method-specific generation and retry procedures in Section 5.1; Section 5.3 isolates the contribution of text perturbations under matched guidance settings.

Transfer across architectures and training regimes. With every surrogate, TPA improves over NatADiff in each of three victim groups: standard convolutional, adversarially trained, and transformer-based models. With RN-50 or Inc-v3 as the surrogate, transformer-based victims have the lowest group-mean ASR for all three methods, and TPA's margin over NatADiff is smallest for this group. With ViT-H as the surrogate, TPA achieves similar mean ASRs across the three groups: 61.2%, 61.9%, and 61.1%, respectively. Group means exclude the surrogate; Appendix C specifies group membership. The transfer gains thus extend beyond victims from the surrogate's architecture family. Under non-adaptive DiffPure purification, TPA also retains higher ASR than AdvClass in all three settings and than NatADiff in two of the three settings (Appendix D).

Image quality. The transfer gains over NatADiff are accompanied by higher mean scores on six of the seven learned NR-IQA metrics with either convolutional surrogate and on all seven with ViT-H (Table 2). NIMA-AVA is the only learned metric favoring NatADiff in the convolutional-surrogate settings. Relative to AdvClass, TPA offers stronger transfer with metric-dependent quality trade-offs: it has higher HyperIQA, ARNIQA, and MUSIQ-KonIQ means in all three settings, whereas AdvClass has higher NIMA-AVA and lower BRISQUE. Figures 1 and 3 provide selected visual comparisons and TPA examples, respectively; Appendix C.3 describes their selection.

Table 1: **Attack success rates (%) on ImageNet-1K.** † marks white-box entries. Average ASR includes all nine evaluation models. Bold/underline mark the best/second-best distinct displayed values within each surrogate and column, including ties. AdvRes/AdvInc denote adversarially trained Inception-ResNet-v2/Inception-v3.

Surrogate model	Method	Victim model ASR (%)									Average ASR
		RN-50	Inc-v3	RN-152	AdvRes	AdvInc	ViT-H	Max-ViT	Swin-B	ViT-B	
RN-50	AdvClass	**100.00**†	38.15	38.50	35.10	37.15	27.00	33.35	33.75	35.55	42.06
	NatADiff	99.65†	66.05	68.10	62.95	64.65	43.80	60.50	62.05	62.75	65.61
	TPA	97.05†	**72.70**	**74.40**	**67.30**	**70.70**	**46.10**	**64.25**	**65.60**	**64.70**	**69.20**
Inc-v3	AdvClass	34.10	**100.00**†	33.55	35.35	38.15	25.90	31.40	32.10	33.70	40.47
	NatADiff	58.40	99.35†	56.00	61.60	65.65	39.00	51.30	51.50	53.60	59.60
	TPA	**65.30**	98.70†	**63.50**	**69.20**	**76.15**	**42.85**	**56.95**	**58.55**	**59.75**	**65.66**
ViT-H	AdvClass	33.25	35.50	32.25	32.35	34.30	98.80†	30.50	31.80	33.35	40.23
	NatADiff	57.85	60.25	56.50	58.10	59.00	**99.15**†	54.80	55.45	60.10	62.36
	TPA	**60.75**	**62.40**	**60.40**	**60.70**	**63.00**	90.45†	**59.60**	**60.15**	**63.55**	**64.56**

Distributional similarity. TPA achieves the lowest FID-A in every surrogate setting, indicating closer set-level similarity to the ImageNet-A reference under the shared feature representation. The ordering differs for FID-VAL: AdvClass obtains the lowest FID-VAL in all three settings, and TPA improves over NatADiff only with the ViT-H surrogate.

Table 2: **Image quality and distributional similarity on ImageNet-1K.** Evaluated on the same outputs as Table 1. The seven learned NR-IQA metrics and BRISQUE report per-image means; FID-VAL and FID-A are set-level statistics. Arrows indicate favorable directions. Bold/underline mark the best/second-best distinct displayed values within each surrogate and metric, including ties. FID-A uses a different reference class mixture (Section 5.1).

Surrogate	ResNet-50			Inception-v3			ViT-H		
Metric	AdvClass	NatADiff	TPA	AdvClass	NatADiff	TPA	AdvClass	NatADiff	TPA
HyperIQA ↑	0.6245	0.5814	**0.6330**	0.6196	0.5820	**0.6321**	0.6124	0.5580	**0.6125**
DBCNN ↑	0.6216	0.5797	**0.6295**	0.6169	0.5797	**0.6292**	**0.6081**	0.5527	0.6079
ARNIQA ↑	0.7077	0.6847	**0.7087**	0.7062	0.6877	**0.7088**	0.7045	0.6725	**0.7077**
MUSIQ-AVA ↑	**5.298**	5.240	5.278	5.256	5.217	**5.265**	5.191	5.111	**5.232**
NIMA-AVA ↑	**5.407**	5.310	5.299	**5.375**	5.301	5.287	**5.320**	5.249	5.316
MUSIQ-KonIQ ↑	70.15	66.64	**70.41**	70.01	67.29	**70.39**	69.76	67.39	**70.40**
TReS ↑	**80.04**	73.95	79.94	79.52	74.26	**79.91**	**78.59**	71.21	78.06
FID-VAL ↓	**28.65**	30.35	32.18	**28.36**	29.14	31.75	**28.24**	35.48	33.73
FID-A ↓	76.83	73.51	**67.53**	77.64	77.87	**67.96**	82.63	74.11	**65.62**
BRISQUE ↓	**16.29**	16.81	17.33	**16.75**	17.44	17.26	**17.49**	24.04	19.25

5.3 Ablation study

To isolate the contribution of early text perturbations, we compare TPA with a variant that uses the original text conditions throughout sampling. For each surrogate, both variants use the same guidance onset, objective and gradient-normalization schedules, and guidance strengths, with 35 guided steps and 2,000 outputs matched by class and latent seed. Target selection follows the same rule on each variant's own trajectory.

Text perturbations increase Average ASR by 8.49–10.88 percentage points across the three surrogates (Table 3). ASR increases on every non-surrogate classifier in all three settings (Appendix E.1). Using the victim groups defined in Section 5.2, mean transfer gains range from 9.1 to 11.4 percentage points across the nine surrogate–group combinations. The matched comparisons thus show consistent transfer gains across architecture families, including adversarially trained victims.



Figure 3: **Adversarial examples generated by TPA.** Selected ImageNet examples from the main 35-step guidance setting, grouped by surrogate. Class denotes the requested class; Pred. is the black-box ResNet-152 victim's prediction. All displayed examples are misclassified by both their surrogate and ResNet-152.

These gains come with a quality cost: mean scores decrease on all seven learned NR-IQA metrics, and FID-VAL and BRISQUE increase, while FID-A decreases. This trade-off within TPA is distinct from the comparison of complete methods in Table 2, where TPA outperforms NatADiff on most learned quality metrics.

Table 3: **Effect of early text-embedding perturbations in TPA.** Guidance settings are matched with 35 guided steps. Removing text perturbation omits noise and moment restoration. Model-wise ASR and quality differences appear in Appendix E.1.

Surrogate	Method	Avg. ASR ↑	TReS ↑	BRISQUE ↓	FID-VAL ↓	FID-A ↓
RN50	TPA w/o text perturbation	60.02	82.14	15.53	30.99	78.89
	TPA	69.20	79.94	17.33	32.18	67.53
Inception-v3	TPA w/o text perturbation	57.17	81.97	15.39	30.58	79.99
	TPA	65.66	79.91	17.26	31.75	67.96
ViT-H/14	TPA w/o text perturbation	53.67	80.17	17.67	30.47	77.43
	TPA	64.56	78.06	19.25	33.73	65.62

Appendices E.2 and E.3 examine the objective schedule and guidance onset, respectively.

6 Conclusion

We presented TPA, which combines early text-embedding perturbations with progressive condition restoration to shape adversarial diffusion trajectories. This exploration changes both the denoising predictions and the intermediate images from which classifier guidance is computed. On ImageNet-1K, TPA achieves higher black-box ASR than NatADiff and AdvClass across every evaluated surrogate–victim pair, while outperforming NatADiff on most learned image-quality metrics. It also obtains lower FID to ImageNet-A than both baselines. Matched ablations show that early conditioning perturbations consistently improve cross-model transfer, with an accompanying cost in image-quality metrics. These findings identify early text conditioning as an effective means of shaping adversarial search and support its use in generating transferable natural adversarial examples to probe test-time vulnerabilities across classifiers.

AI USE STATEMENT

Generative AI assistants were used during research planning, experiment design, code development and testing, analysis of recorded results, and manuscript preparation. They were also used for internal review and editing. The authors reviewed and verified all AI-assisted code, results, claims, citations, and text against the underlying experiments and source materials. Human ratings were collected independently and were not generated or inferred by AI. The authors are responsible for the final content of the submission.

REPRODUCIBILITY STATEMENT

Implementation and evaluation details are provided in the paper and appendices. Section 4 describes the TPA sampling procedure, Appendix B lists the implementation settings and hyperparameters, and Appendix C specifies the evaluation setup. Additional ablations are reported in Appendix E. All experiments use fixed checkpoints, prompts, and random seeds, and all generated images are kept for evaluation. We will release the code, experiment configurations, and evaluation scripts publicly.

ETHICS STATEMENT

This study measures classifier failures on generated images to support robustness evaluation. It does not test attacks against deployed services or collect personal data. Outputs can reflect biases of the pretrained generator and classifiers. The human-study protocol and observed response proportions are reported in Appendix F. Failure rates alone do not certify that an image is a valid member of the requested class.

REFERENCES

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, and Alberto Del Bimbo. ARNIQA: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 189–198, January 2024. URL https://openaccess.thecvf.com/content/WACV2024/html/Agnolucci_ARNIQA_Learning_Distortion_Manifold_for_Image_Quality_Assessment_WACV_2024_paper.html.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 284–293. PMLR, 2018. URL https://proceedings.mlr.press/v80/athalye18b.html.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pzpWBbnwiJ.

Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A. Forsyth. Unrestricted adversarial examples via semantic manipulation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=Sye_OgHFwH.

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. doi: 10.1109/TPAMI.2024.3480519. URL https://doi.org/10.1109/TPAMI.2024.3480519.

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)*, pp. 4562–4572, October 2023a. URL https://openaccess.thecvf.com/content/ICCV2023/html/Chen_AdvDiffuser_Natural_Adversarial_Example_Synthesis_with_Diffusion_Models_ICCV_2023_paper.html.

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*, volume 36, pp. 51719–51733, 2023b. doi: 10.52202/075280-2253. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a24cd16bc361afa78e57d31d34f3d936-Abstract-Conference.html.

Max Collins, Jordan Vice, Tim French, and Ajmal Mian. NatADiff: Adversarial boundary guidance for natural adversarial diffusion. In *International Conference on Learning Representations*, pp. 12099–12145, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/147117db5bbde42ce9bdf87740bafb9e-Abstract-Conference.html.

Xuelong Dai, Kaisheng Liang, and Bin Xiao. AdvDiff: Generating unrestricted adversarial examples using diffusion models. In *Computer Vision – ECCV 2024*, volume 15104 of *Lecture Notes in Computer Science*, pp. 93–109. Springer, 2024. doi: 10.1007/978-3-031-72952-2_6. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6243_ECCV_2024_paper.php.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/hash/49ad23d1ec9fa4bd8d77d02681df5cfa-Abstract.html.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9185–9193, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Dong_Evading_Defenses_to_Transferable_Adversarial_Examples_by_Translation-Invariant_Attacks_CVPR_2019_paper.html.

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*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1220–1230, January 2022. URL https://openaccess.thecvf.com/content/WACV2022/html/Golestaneh_No-Reference_Image_Quality_Assessment_via_Transformers_Relative_Ranking_and_Self-Consistency_WACV_2022_paper.html.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL https://research.google/pubs/explaining-and-harnessing-adversarial-examples/.

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)*, pp. 770–778, June 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15262–15271, June 2021. URL

https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

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*, volume 30, pp. 6626–6637, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=qw8AKxfYbI.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/e2c420d928d4bf8ce0ff2ec19b371514-Abstract.html.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5148–5157, October 2021. URL https://openaccess.thecvf.com/content/ICCV2021/html/Ke_MUSIQ_Multi-Scale_Image_Quality_Transformer_ICCV_2021_paper.html.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=BJm4T4Kgx.

Hui Kuurila-Zhang, Haoyu Chen, and Guoying Zhao. VENOM: Text-driven unrestricted adversarial example generation with diffusion models. *arXiv preprint arXiv:2501.07922*, 2025. URL https://arxiv.org/abs/2501.07922.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 122458–122483, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/dd540e1c8d26687d56d296e64d35949f-Abstract-Conference.html.

Yueqian Lin, Jingyang Zhang, Yiran Chen, and Hai Li. SD-NAE: Generating natural adversarial examples with stable diffusion. In *The Second Tiny Papers Track at ICLR 2024*, 2024. URL https://openreview.net/forum?id=D87rimdkGd.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_Swin_Transformer_V2_Scaling_Up_Capacity_and_Resolution_CVPR_2022_paper.html.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Yasamin Medghalchi, Moein Heidari, Clayton Allard, Leonid Sigal, and Ilker Hacihaliloglu. Prompt2Perturb (P2P): Text-guided diffusion-based adversarial attack on breast ultrasound images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28564–28574, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Medghalchi_Prompt2Perturb_P2P_Text-Guided_Diffusion-Based_Adversarial_Attack_on_Breast_Ultrasound_Images_CVPR_2025_paper.html.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, December 2012. doi: 10.1109/TIP.2012.2214050.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pp. 16805–16827. PMLR, 2022. URL https://proceedings.mlr.press/v162/nie22a.html.

Zhengzhao Pan, Hua Chen, and Xiaogang Zhang. DiffAdvMAP: Flexible diffusion-based framework for generating natural unrestricted adversarial examples. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 47590–47614. PMLR, 2025. URL https://proceedings.mlr.press/v267/pan25b.html.

Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in GAN evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11410–11420, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Parmar_On_Aliased_Resizing_and_Surprising_Subtleties_in_GAN_Evaluation_CVPR_2022_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Rombach_High-Resolution_Image_Synthesis_With_Latent_Diffusion_Models_CVPR_2022_paper.html.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015. doi: 10.1007/s11263-015-0816-y.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations*, pp. 23723–23755, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6769bfdeb7313ded45714aabf759adfe-Abstract-Conference.html.

Yifei Shen, Xinyang Jiang, Yifan Yang, Yezhen Wang, Dongqi Han, and Dongsheng Li. Understanding and improving training-free loss-based diffusion guidance. In *Advances in Neural Information Processing Systems*, volume 37, pp. 108974–109002, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c4edc5113b4ffd4632718558fb66b9ef-Abstract-Conference.html.

Mannat Singh, Laura Gustafson, Aaron Adcock, Vinicius de Freitas Reis, Bugra Gedik, Raj Prateek Kosaraju, Dhruv Mahajan, Ross Girshick, Piotr Dollár, and Laurens van der Maaten. Revisiting weakly supervised pre-training of visual perception models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 804–814, June 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Singh_Revisiting_Weakly_Supervised_Pre-Training_of_Visual_Perception_Models_CVPR_2022_paper.html.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8312–8323, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/8cea559c47e4fbdb73b23e0223d04e79-Abstract.html.

Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqiu Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3667–3676, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Su_Blindly_Assess_Image_Quality_in_the_Wild_Guided_by_a_CVPR_2020_paper.html.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. URL https://research.google/pubs/intriguing-properties-of-neural-networks/.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2826, June 2016. URL https://openaccess.thecvf.com/content_cvpr_2016/html/Szegedy_Rethinking_the_Inception_CVPR_2016_paper.html.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-ResNet and the impact of residual connections on learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1):4278–4284, February 2017. doi: 10.1609/aaai.v31i1.11231. URL https://doi.org/10.1609/aaai.v31i1.11231.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, August 2018. doi: 10.1109/TIP.2018.2831899.

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*, 2018. URL https://openreview.net/forum?id=rkZvSe-RZ.

Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. MaxViT: Multi-axis vision transformer. In *Computer Vision – ECCV 2022*, volume 13684 of *Lecture Notes in Computer Science*, pp. 459–479. Springer, 2022. doi: 10.1007/978-3-031-20053-3_27. URL https://doi.org/10.1007/978-3-031-20053-3_27.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernández Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=SUMtDJqicd.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2730–2739, June 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Xie_Improving_Transferability_of_Adversarial_Examples_With_Input_Diversity_CVPR_2019_paper.html.

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*, volume 36, pp. 2894–2921, 2023. doi: 10.52202/075280-0129. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/088463cd3126aef2002ffc69da42ec59-Abstract-Conference.html.

Andi Zhang, Mingtian Zhang, and Damon Wischik. Constructing semantics-aware adversarial examples with a probabilistic perspective. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136259–136285, 2024. URL

https://proceedings.neurips.cc/paper_files/paper/2024/hash/f620c653a8f196076f9a2fbc3c9d7efb-Abstract-Conference.html.

Andi Zhang, Xuan Ding, Steven McDonagh, and Samuel Kaski. Concept-based adversarial attack: a probabilistic perspective. In *International Conference on Learning Representations*, pp. 73613–73646, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/778cc31583604ebf246726055e9f737f-Abstract-Conference.html.

Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, January 2020. doi: 10.1109/TCSVT.2018.2886771.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1BLjgZCb.

Xiaopei Zhu, Peiyang Xu, Guanning Zeng, Yinpeng Dong, and Xiaolin Hu. Natural language induced adversarial images. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10872–10881. ACM, 2024. doi: 10.1145/3664647.3680902. URL https://doi.org/10.1145/3664647.3680902.

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. doi: 10.1109/TIP.2022.3211736. URL https://doi.org/10.1109/TIP.2022.3211736.

A LIMITATIONS

TPA's fixed target can be semantically close to the requested class, so some successful attacks involve fine-grained confusions, a limitation also noted by NatADiff (Collins et al., 2026). Selecting targets with explicit semantic separation is a useful direction for future work. Conditioning perturbations improve transfer at a cost in image-quality metrics, and the human study assesses class recognition on a limited image subset. The observed transfer gains do not by themselves establish that conditioning perturbations broaden the class-valid region explored during generation. Our evaluation uses SD1.5 and ImageNet classifiers that informed development; extending the method to other generators, domains, and independent classifier panels remains to be tested.

B IMPLEMENTATION AND HYPERPARAMETERS

B.1 SAMPLING AND CONDITIONING

We use frozen Stable Diffusion 1.5 in FP16 with deterministic DDIM ($\eta = 0$) at 512×512. The 50-step sampling grid is $t_k = 981 - 20k$, $k = 0, \ldots, 49$, with 1,000 training timesteps and scaled-linear betas $[0.00085, 0.012]$. The terminal update uses $a_{50} = \sqrt{\bar{\alpha}_0}$ and $b_{50} = \sqrt{1 - \bar{\alpha}_0}$ in Eq. (1). Source prompts are "a photo of a {label}".

TPA applies the CADS rule (Sadat et al., 2024) to cached source and null embeddings, with independent noise draws at every step. The annealing schedule in Eq. (2) is

$$\gamma_k = \mathrm{clip}\left(\frac{\tau_2 - t_k/999}{\tau_2 - \tau_1}, 0, 1\right). \tag{7}$$

Moment restoration (Eq. (3)) uses $\epsilon_c = 10^{-6}$ and mixing factor one. Noise vanishes at $k = 20$; the main sampler continues the moment-restoration arithmetic thereafter. Latent and conditioning noise use separate seeded random streams; matched comparisons reuse the same class and latent seeds.

B.2 HYPERPARAMETER SETTINGS

Tables 4 and 5 collect the shared and surrogate-specific settings. The noise schedule is fixed across surrogates; only the objective switch and late guidance strength differ.

Table 4: Shared TPA hyperparameters. Indices are zero-based.

Parameter	Value
Denoising steps / CFG scale w	50 / 7.5
Condition-noise scale s_c	0.25
Annealing thresholds (τ_1, τ_2)	$(0.6, 0.9)$
Condition perturbation interval	$k = 0$–19
Guidance onset S / active steps	15 / 35
Views per guided step	12
Embedding moment-restoration factor	1
Outputs per class / attempts per output	2 / 1

Surrogate	Switch K	λ_{pre}	λ_{late}
RN-50	40	2.5	100
Inc-v3	40	2.5	100
ViT-H	30	2.5	200

Table 5: Surrogate-specific guidance settings. The untargeted phase uses $S \leq k < K$ and the fixed-target phase uses $K \leq k < 50$, with $S = 15$ for all surrogates.

The ViT-H switch and late strength were selected using surrogate ASR on a 100-class development subset. Guidance onset $S = 15$ follows NatADiff's attack timestep range while reducing the number of gradient evaluations; Appendix E.3 compares this choice with full-window guidance.

B.3 GUIDANCE AND CLASSIFIER INPUTS

For $d_k = \nabla_{z^{(k)}} R_k$, the guidance in Eq. (6) is

$$g_k = \begin{cases} 0, & k < S, \\ \lambda_{\text{pre}} d_k, & S \le k < K, \\ \lambda_{\text{late}} \dfrac{d_k}{\max\{\|d_k\|_2, 10^{-12}\}}, & k \ge K. \end{cases} \tag{8}$$

Each guided step differentiates through the frozen UNet, VAE, and surrogate, without backpropagating through earlier sampling steps. Decoded RGB values are clamped to $[0, 1]$ before classifier preprocessing. The 12 views are five crops (corners and center), their horizontal flips, grayscale, and identity. Crops have side $\lfloor 0.6H \rfloor$ and are resized to input height H. Surrogate resize/crop sizes are 232/224 for RN50, 342/299 for Inception-v3, and 224/224 for ViT-H/14; normalization and checkpoints are given in Table 8.

B.4 BASELINE SETTINGS

AdvClass and NatADiff select the nearest non-source label by CLIP EOS-embedding cosine distance using full-synonym labels (Radford et al., 2021; Collins et al., 2026). Their mixed prompt is "a photo of a {target} and a {source}", with mixture weight 0 for AdvClass and 0.2 for NatADiff. Both source and mixed CFG scales are 7.5. The baseline implementation evaluates three conditioning branches; TPA uses source and null branches.

Table 6: RN50 attack settings. p_s is the softmax of mean-view logits.

Setting	AdvClass	NatADiff	TPA
DDIM steps / active raw t	200 / $(0, 200)$	200 / $(0, 700)$	50 / $(0, 700)$
Attack reward	$\log p_s(q)$	$\log p_s(q)$	Eq. (5)
Target selection	CLIP fixed	CLIP fixed	At $k = 40$
Gradient / views	L_2 / 1	L_2 / 12	Raw then L_2 / 12
Initial strength	500	50	2.5 then 100
Retry increment / cap	250 / 5 attempts	25 / 5 attempts	None / 1 attempt
Retry early stop	Target hit	Target hit 5 repeats in	None
Time travel	None	$[500, 800]$	None

NatADiff's initial strength/retry increment is 50/20 for Inception-v3 and 100/50 for ViT-H/14; AdvClass uses 500/250 for every surrogate. Our controlled baselines use FP16, source phrases, and per-image seeds; the public NatADiff runner uses BF16, bare labels, and a shared seed stream.

B.5 COMPUTATIONAL EFFICIENCY

Table 7 reports average generation runtime over 2,000 final outputs per method with an RN50 surrogate. Timings exclude initialization and post-generation evaluation, and include retries and time travel where applicable. The sampling configurations are specified in Table 6.

Table 7: Average recorded generation runtime with an RN50 surrogate.

Method	Average runtime (s/image)
AdvClass	16.41
NatADiff	118.60
TPA	10.28

C Evaluation Protocol

C.1 Classifiers and generated samples

Sample inclusion and Average ASR follow Section 5.1. Supplementary black-box averages exclude the generating surrogate. Table 8 specifies the checkpoints and transforms shared by all methods. ViT-H, Max-ViT, Swin-B, and ViT-B abbreviate ViT-H/14, MaxViT-T, Swin-V2-B, and ViT-B/16, respectively (Dosovitskiy et al., 2021; Singh et al., 2022; Tu et al., 2022; Liu et al., 2022).

For group-level transfer analysis, standard convolutional victims are RN-50, Inc-v3, and RN-152; adversarially trained victims are AdvRes and AdvInc; and transformer-based victims are ViT-H, Max-ViT, Swin-B, and ViT-B (He et al., 2016; Szegedy et al., 2016; Kurakin et al., 2017; Szegedy et al., 2017; Tramèr et al., 2018). The transformer group includes the convolution–attention hybrid Max-ViT. Group means average per-model ASRs after excluding the generating surrogate.

Table 8: Evaluation checkpoints and transforms. Unmarked models use TorchVision. Inputs are bilinearly resized with antialiasing, center cropped, and normalized. I denotes ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$; H uses mean and standard deviation $(0.5, 0.5, 0.5)$.

Model	Weights / pretrained model identifier	Resize / crop	Norm.
RN50	`IMAGENET1K_V2`	232 / 224	I
Inception-v3	`IMAGENET1K_V1`	342 / 299	I
ResNet152	`IMAGENET1K_V2`	232 / 224	I
Adv. Inc-Res-v2 (timm)	`inception_resnet_v2.tf_ens_adv_in1k`	333 / 299	H
Adv. Inception-v3 (timm)	`adv_inception_v3.tf_adv_in1k`	342 / 299	H
ViT-H/14	`IMAGENET1K_SWAG_LINEAR_V1`	224 / 224	I
MaxViT-T	`IMAGENET1K_V1`	224 / 224	I
Swin-V2-B	`IMAGENET1K_V1`	272 / 256	I
ViT-B/16	`IMAGENET1K_V1`	256 / 224	I

C.2 Image-quality metrics

The seven learned NR-IQA metrics and BRISQUE use `pyiqa` 0.1.13 at the saved image resolution, with each model's internal preprocessing (Su et al., 2020; Zhang et al., 2020; Agnolucci et al., 2024; Ke et al., 2021; Talebi & Milanfar, 2018; Golestaneh et al., 2022; Mittal et al., 2012). TReS uses its KonIQ checkpoint and 50 test crops, without external 256×256 resizing. FID-VAL and FID-A use the same 2,048-dimensional `torch-fidelity` Inception features, comparing each generated set with all 50,000 ImageNet validation images and all 7,500 ImageNet-A images, respectively. All methods use this common evaluation pipeline (Parmar et al., 2022). FID-A compares different class mixtures and measures set-level similarity, rather than class preservation.

C.3 Qualitative selection

Figures 1 and 3 are selected visual illustrations rather than random samples. The quantitative evaluation follows Section 5.1.

D EVALUATION UNDER DIFFPURE

Setup. We apply DiffPure (Nie et al., 2022) to the same generated image sets, without adapting the attacks to the defense. We use the official unconditional ImageNet 256×256 ADM checkpoint with reverse-SDE purification at $t = 0.15$ and one draw per image, with purification randomness matched across methods. Images are resized to 256 pixels before purification and 224 pixels afterward, followed by each classifier's fixed preprocessing. The resize-only baseline uses the same resizing without purification.

Table 9: Average ASR (%) under DiffPure. For generated attacks, *No defense* reproduces Table 1; the other columns report resize-only and purification evaluations. Bold marks the highest ASR within each surrogate and column.

Surrogate	Method	No defense	Resize only	DiffPure
RN-50	AdvClass	42.06	41.85	34.41
	NatADiff	65.61	65.98	**60.15**
	TPA	**69.20**	**69.46**	59.73
Inc-v3	AdvClass	40.47	39.96	34.42
	NatADiff	59.60	59.19	54.83
	TPA	**65.66**	**65.68**	**57.58**
ViT-H	AdvClass	40.23	40.16	33.63
	NatADiff	62.36	63.12	60.62
	TPA	**64.56**	**65.56**	**62.75**

Results. DiffPure reduces Average ASR for every attack and surrogate setting (Table 9). TPA retains higher ASR than AdvClass in all three settings and than NatADiff with Inc-v3 and ViT-H surrogates. With RN-50, TPA and NatADiff obtain similar post-purification ASR, with NatADiff higher by 0.42 percentage points. Thus, TPA's advantage over NatADiff persists under this defense in two of the three settings. These results measure attack persistence under the stated non-adaptive purification protocol.

E SUPPLEMENTARY ABLATIONS

E.1 EFFECT OF TEXT PERTURBATIONS

The main ablation fixes class/latent seeds, guidance onset, switch, strengths, and classifier panel for each surrogate. The control omits both embedding noise and moment restoration. Each arm selects its target from its own trajectory. Tables 10 and 11 extend Table 3 with model-wise ASR changes and all quality differences. ASR increases on every non-surrogate classifier, while mean scores decrease on all seven learned image-quality metrics. FID-A improves while FID-VAL and BRISQUE worsen.

Table 10: TPA minus TPA w/o text perturbation ASR in percentage points at the main 35-step onset. Columns specify the surrogate; † marks white-box entries. All other entries measure black-box transfer.

Victim model	RN-50	Inc-v3	ViT-H
RN-50	$+3.70^{\dagger}$	+9.15	+11.00
Inc-v3	+10.90	$+2.45^{\dagger}$	+10.00
RN-152	+9.70	+9.55	+11.50
AdvRes	+9.85	+9.10	+10.60
AdvInc	+10.60	+9.90	+12.25
ViT-H	+8.00	+7.95	$+8.80^{\dagger}$
Max-ViT	+10.65	+9.05	+11.75
Swin-B	+9.00	+8.90	+10.50
ViT-B	+10.20	+10.35	+11.55

Table 11: Quality changes from adding text perturbations at the main 35-step onset. Entries are TPA minus TPA w/o text perturbation. Higher is favorable for learned metrics; lower is favorable for BRISQUE and FIDs.

Metric	RN-50	Inc-v3	ViT-H
HyperIQA	-0.0137	-0.0127	-0.0135
DBCNN	-0.0105	-0.0100	-0.0108
ARNIQA	-0.0047	-0.0047	-0.0029
MUSIQ-AVA	-0.095	-0.099	-0.069
NIMA-AVA	-0.125	-0.122	-0.088
MUSIQ-KonIQ	-0.37	-0.28	-0.27
TReS	-2.20	-2.06	-2.11
FID-VAL	+1.19	+1.17	+3.26
FID-A	-11.36	-12.03	-11.81
BRISQUE	+1.80	+1.87	+1.57

E.2 OBJECTIVE SCHEDULE

With RN-50, each objective uses matched class/latent seeds, condition noise, 35 guided steps, twelve views, strengths, and normalization switch at $K = 40$. Targeted-only selects the strongest non-source class at $S = 15$ and uses its mean target-minus-source logit throughout. Untargeted-only uses the log-sigmoid source-away reward throughout. TPA selects its fixed target at $K = 40$, so targeted-only also changes target-selection time.

Targeted-only increases average ASR by 1.22 points relative to TPA, with lower TReS and higher FID-VAL. Untargeted-only lowers ASR by 3.68 points. Table 12 therefore supports a design trade-off, not uniform superiority of the two-stage schedule.

Table 12: **Effect of the adversarial objective on RN-50.** All variants retain TPA text perturbations and the same attack window, strengths, and normalization schedule. ASR follows Section 5.1.

Objective	Avg. ASR ↑	TReS ↑	BRISQUE ↓	FID-VAL ↓	FID-A ↓
Two-stage (TPA)	69.20	79.94	17.33	32.18	67.53
Targeted-only	70.42	79.42	17.04	39.01	60.16
Untargeted-only	65.52	79.90	17.51	32.49	67.35

E.3 GUIDANCE ONSET

We use $S = 15$, corresponding to $0 < t < 700$, to follow NatADiff's attack timestep range (Collins et al., 2026) and reduce gradient evaluations from 50 to 35 per image. Table 13 compares this computational trade-off with full-window guidance under matched classes, latent seeds, and all other settings. Delayed guidance reduces gradient evaluations by 30%, with a 0.45–1.71 percentage-point reduction in mean black-box ASR; TPA still exceeds NatADiff in each surrogate setting. Quality changes are mixed. The main setting therefore prioritizes a smaller guidance budget while retaining strong transfer.

Table 13: Full-window (50) versus delayed (35) guidance on ImageNet-1K. Black-box ASR (%) averages over the eight non-surrogate models; differences are delayed minus full-window. The 35-step setting is used in the main tables.

	Black-box ASR (%)			TReS ↑	
Surrogate	50 steps	35 steps	Δ (pp)	50 steps	35 steps
RN-50	66.54	65.72	-0.83	80.04	79.94
Inc-v3	61.98	61.53	-0.45	79.93	79.91
ViT-H	63.03	61.32	-1.71	77.63	78.06

F HUMAN STUDY

Following NatADiff (Collins et al., 2026), we generated 256 TPA images with an RN50 surrogate from the same recognizable ImageNet class subset. Ten participants each evaluated a randomly sampled subset of 60 images from this pool, yielding 600 judgments in total. For each image, participants selected the requested source class, the adversarial target class, or neither.

Across the 600 judgments, 91.83% selected the source class, 3.00% the target class, and 5.17% neither. These ratings support source-class recognition on the evaluated image set. The percentages describe individual responses, rather than image-level majority votes, photorealism scores, or human-validated ASR. The study draws from a pool of 256 generated images, with each participant evaluating 60 randomly sampled images; therefore, the 600 judgments do not constitute 600 independent images. Because the images and participants differ from those in NatADiff, the response rates do not establish a direct comparison between methods.

Table 14: Human source-class recognition results aggregated over 600 judgments. Repeated evaluations of an image are counted separately.

Response	Count	Percentage
Source class	551	91.83%
Adversarial target class	18	3.00%
Neither	31	5.17%
Total	600	100.00%