AMEND: ATTENTION MISALIGNMENT FOR NATURAL ADVERSARIAL ATTACKS AGAINST OBJECT DETECTORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Object detectors are increasingly deployed in safety-critical applications such as autonomous driving, surveillance, and robotics. However, their vulnerability to adversarial examples raises serious security concerns. Existing attacks face a fundamental trade-off between attack effectiveness and visual naturalness. Most methods optimize perturbations in image or pixel space, where semantic manipulation is inherently coupled with structural changes. Consequently, stronger attacks often introduce visible artifacts or unwanted distortions. In this paper, we argue that optimizing semantic representations rather than pixels provides a more controllable attack space. We formulate natural adversarial attacks as semantic alignment manipulation between generative models and discriminative detectors. Our key insight is that diffusion cross-attention enables semantic manipulation while preserving image structure. Based on this insight, we propose AMEND, a diffusion-based adversarial attack framework that integrates detector-derived conditioning, semantic attention dispersion, and structural consistency constraints. Experiments on seven object detectors show that AMEND achieves the lowest mAP on six of them, reducing relative mAP to as low as 6.73%, while consistently achieving superior visual naturalness across the evaluated settings.

1 INTRODUCTION

Deep neural networks remain vulnerable to adversarial examples, where carefully crafted inputs can cause incorrect predictions while remaining imperceptible to humans Szegedy et al. (2014); Madry et al. (2018). This vulnerability raises serious security concerns when object detectors are deployed in safety-critical applications, including autonomous driving Eykholt et al. (2018), surveillance, and robotics. Unlike image classifiers that output a single label, object detectors must recognize and localize multiple objects simultaneously Xie et al. (2017). Attacks must therefore disrupt both semantic recognition and spatial localization. Beyond attack effectiveness, practical adversarial examples should also preserve visual naturalness. They should remain perceptually and structurally close to the source image while maintaining plausible visual quality. Visible artifacts, such as artificial patches, unnatural textures, or severe color distortions, can reveal the attack and limit its real-world applicability Li et al. (2023); Zhou et al. (2025).

Existing attacks struggle to achieve both strong attack effectiveness and high visual naturalness. Most methods optimize perturbations in image space or closely related pixel representations Thys et al. (2019); Hu et al. (2021); Miao et al. (2024); Xie et al. (2017); Wu et al. (2020); Huang et al. (2023). In these spaces, semantic manipulation and structural preservation are inherently coupled. Stronger attacks therefore often require changes to image textures or structures, making the perturbations more perceptible. We argue that this limitation stems primarily from the optimization space rather than from the attack algorithms themselves. Existing methods lack a representation that explicitly separates semantic manipulation from structural preservation. This coupling makes it difficult to improve attack effectiveness without sacrificing visual naturalness, resulting in a persistent trade-off between the two.

We propose a different perspective: **Natural adversarial attacks should manipulate semantic correspondence rather than image pixels.** Unlike pixel-space representations, latent diffusion models provide partially decoupled semantic and structural representations through two complementary attention mechanisms. Cross-attention aligns semantic concepts with image regions, while

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

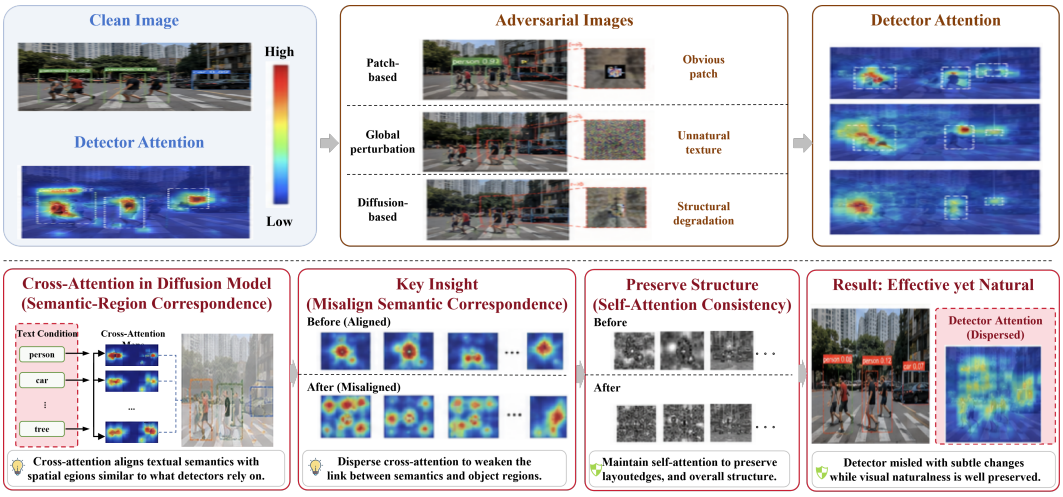


Figure 1: Motivation of AMEND. Existing attacks manipulate image pixels and tend to modify image structures. AMEND instead disperses semantic cross-attention to weaken detector-relevant semantics.

self-attention primarily captures spatial layout and structure. Prior work shows that diffusion features encode the semantic layout of generated images Hertz et al. (2023); Tumanyan et al. (2023). Object detectors also rely on semantic cues to recognize and localize objects. We therefore hypothesize that manipulating semantic correspondence within the diffusion model can affect detector predictions while preserving image structure. DiffAttack Chen et al. (2024) first used attention-based objectives to construct transferable attacks on image classifiers. AMEND extends this idea to object detection, where an attack must suppress multiple localized detections rather than a single classification label. Figure 1 illustrates this intuition. Instead of directly perturbing pixels, AMEND manipulates semantic cross-attention while preserving structural information through self-attention. This design partially decouples semantic manipulation from structural preservation.

Motivated by this insight, we propose AMEND, a diffusion-based framework for natural adversarial attacks that induces semantic attention misalignment. AMEND explicitly separates semantic manipulation from structural preservation during the attack. It follows three principles. First, AMEND disperses diffusion cross-attention to weaken detector-relevant semantic correspondence rather than directly modifying image pixels. Second, it optimizes the unconditional embedding to reconstruct a faithful diffusion trajectory, providing a stable starting point for semantic manipulation. Third, it preserves image structure through self-attention consistency and perceptual constraints. Together, these components assign semantic manipulation and structural preservation to separate objectives, enabling more effective and natural adversarial examples.

Our contributions are summarized as follows:

- We formulate natural adversarial attacks on object detectors as semantic attention manipulation in diffusion models, using diffusion attention to separate semantic manipulation from structural preservation.
- We propose AMEND, which integrates detector-derived conditioning, IoU-matched pre-NMS confidence suppression, and a fixed attention-derived blending mask with attention dispersion and structure-preservation objectives.
- Experiments on seven object detectors show that AMEND achieves the lowest mAP on six detectors while achieving stronger naturalness across the evaluated metrics than the baselines. Ablation studies further characterize how each objective affects the effectiveness–naturalness trade-off.

2 Related Works

Adversarial attacks on object detectors have progressed from image-wide perturbations Xie et al. (2017); Chow et al. (2020) to localized and natural-looking modifications, including physical patches Thys et al. (2019); Xu et al. (2020); Wu et al. (2020), transferable patterns Huang et al. (2023), and generative patch attacks Hu et al. (2021); Miao et al. (2024). More recent methods generate natural-looking adversarial images through semantic manipulation or generative models Chen et al. (2023b;a); Dai et al. (2024); Xue et al. (2023); Wang et al. (2025); Lin et al. (2023); Kuurila-Zhang et al. (2025), but most focus on image classification. Diffusion-based editing further enables structure-preserving manipulation through attention control Mokady et al. (2023); Hertz et al. (2023); Tumanyan et al. (2023). DiffAttack Chen et al. (2024) applies this idea to transferable classification attacks by manipulating cross-attention while preserving self-attention. However, extending attention-based diffusion attacks to object detection requires handling multiple candidate instances rather than a single class-level objective. Our work addresses this gap by deriving conditioning tokens from detector predictions, suppressing IoU-matched pre-NMS candidates, and restricting modifications with an attention-derived mask.

3 Method

3.1 Overview

Threat model. A detector $\mathcal{F}(\cdot)$ maps an image x to $\mathcal{F}(x) = \{(\hat{b}_i, \hat{y}_i, \hat{s}_i)\}_{i=1}^{\hat{N}}$, where $\hat{b}_i$, $\hat{y}_i$, and $\hat{s}_i$ denote the predicted box, category, and confidence, respectively. We adopt a white-box, digital setting in which the attacker has access to the detector architecture, parameters, and gradients. Let y denote the target category to be suppressed, selected from the categories predicted by $\mathcal{F}(x)$. We consider a category-specific vanishing attack on a target category y: given the clean detections $\mathcal{B} = \{b_1, \ldots, b_N\}$ of y (i.e., predictions of $\mathcal{F}$ on x, not ground-truth annotations), the goal is to suppress the detections of y while retaining the natural appearance of the input. On INRIA Person, $y =$ person. By *naturalness*, we mean that x^{adv} remains perceptually and structurally close to x and free of visible artifacts.

Overview. AMEND performs the attack in the latent space of a pretrained diffusion model $\mathcal{D}$ and consists of two stages (Figure 2). Stage 1 constructs a faithful, image-specific reference trajectory: the clean image x is encoded and inverted with DDIM, and the unconditional embedding is optimized so that reverse diffusion reproduces the inverted trajectory. Stage 2 injects a learnable perturbation δ into the reference latent at the attack step and rolls out the remaining denoising steps, conditioned on the target category y. δ is optimized to suppress detections and disperse category-token cross-attention, while self-attention and perceptual constraints preserve structure. The decoded image is finally blended with the clean image through an attention-derived mask M:

$$x^{\text{adv}} = M \odot \tilde{x}^{\text{adv}} + (1 - M) \odot x, \qquad M \in [0,1]^{H \times W}, \tag{1}$$

where $\tilde{x}^{\text{adv}}$ is the decoded adversarial image before blending and $\odot$ denotes element-wise multiplication. Thus, AMEND combines detector-aware semantic disruption with localized, structure-preserving image manipulation.

3.2 Faithful Diffusion Reconstruction

Before semantic manipulation, AMEND establishes a faithful diffusion trajectory that preserves the input-specific semantics and structure. We derive the condition directly from detector predictions without an external captioning model: we run $\mathcal{F}$ on the clean image, select the target category y from its predictions, and use the category name of y as the textual conditioning, whose embedding we denote by c.

We index diffusion time so that generation proceeds from $t{=}T$ (noise) to $t{=}0$ (clean). DDIM inversion Song et al. (2021) yields a deterministic reference trajectory $\{z'_t\}_{t=0}^{T} \leftarrow \text{DDIM_Inversion}(z_0)$, which provides a stable latent initialization for subsequent optimization. Direct DDIM reconstruction with a generic unconditional embedding introduces errors because it lacks image-specific information. Following null-text inversion Mokady et al. (2023), we align the two trajectories by optimizing the

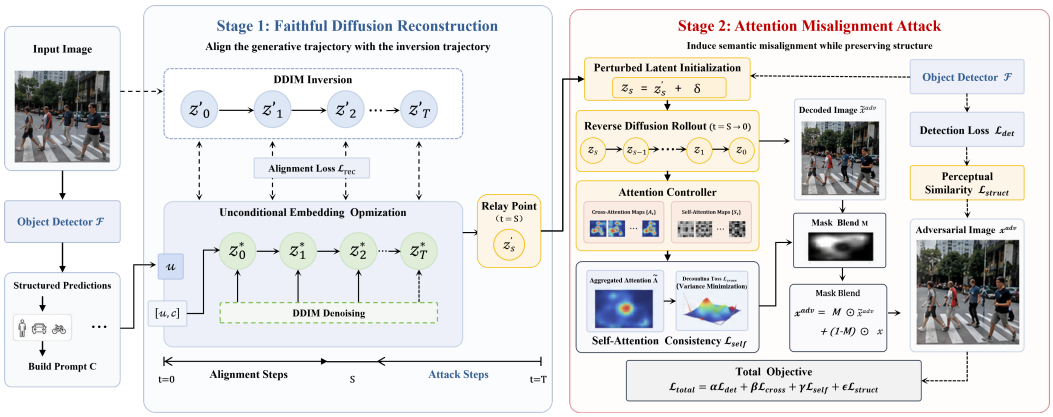


Figure 2: Overview of AMEND. Stage 1 builds a faithful diffusion trajectory via DDIM inversion and unconditional-embedding optimization; Stage 2 perturbs the latent to disperse cross-attention for the attack, while self-attention and perceptual constraints preserve image structure.

unconditional embedding u while keeping c fixed. Let $z^*_{t-1} \leftarrow \text{DDIM_Step}(z^*_t, [u, c], t)$ denote the latent recovered under the joint condition $[u, c]$, with $z^*_T = z'_T$.

Prior work observes that perturbations around $t = 0.8T$ are effective for adversarial optimization Chen et al. (2023a). We accordingly inject the attack perturbation at the high-noise latent z_S with $S = 0.8T$, and align only the pre-attack denoising steps (those reducing the noise level from T to S):

$$\mathcal{L}_{\text{align}} = \sum_{t=S+1}^{T} \left\| z^*_{t-1} - z'_{t-1} \right\|_2^2, \quad S = 0.8T. \tag{2}$$

Restricting the alignment to these steps also reduces computation and memory cost. We freeze u once $\mathcal{L}_{\text{align}}$ converges. The effects of embedding optimization and of the attack step S are analyzed in Appendix D.1.

3.3 Attention Misalignment Attack

Cross-attention associates semantic conditions with spatial image regions during diffusion generation. Such semantic-region correspondence provides cues that are also exploited by discriminative detectors. AMEND targets this correspondence and introduces controlled misalignment to weaken detector-relevant semantics: rather than altering image content directly, it redistributes the semantic focus encoded by cross-attention.

Let $A_t \in \mathbb{R}^{H \times W}$ denote the aggregated cross-attention map at denoising step t for the conditioned category token(s). Following common practice Hertz et al. (2023); Tumanyan et al. (2023), we collect the cross-attention tensors from all U-Net blocks at the 16×16 resolution, average over attention heads, and keep only the column(s) corresponding to the category token(s) of y; maps from multiple blocks and tokens are averaged into a single map. We use the 16×16 resolution because it carries the most reliable object-level semantics, and each A_t is min–max normalized to $[0, 1]$. The attack rollout evaluates the U-Net at steps $t = S, \ldots, 1$, producing $z_{S-1}, \ldots, z_0$; all sums over $t = 1, \ldots, S$ below range over these steps. We aggregate over the rollout steps into a global distribution

$$\bar{A} = \frac{1}{S} \sum_{t=1}^{S} A_t, \tag{3}$$

and define the cross-attention dispersion loss as its spatial variance,

$$\mathcal{L}_{\text{cross}} = \text{Var}(\bar{A}) = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(\bar{A}_{i,j} - \mu \right)^2, \quad \mu = \frac{1}{HW} \sum_{i,j} \bar{A}_{i,j}. \tag{4}$$

Cross-attention exhibits high responses on regions matching the conditioned object concepts. Following DiffAttack Chen et al. (2024), we minimize the spatial variance of the normalized map $\bar{A}$

to weaken this semantic contrast, rather than enforce a particular spatial distribution. Appendix C.1 confirms that the reduced variance corresponds to a flatter, less object-concentrated response rather than a collapse onto a few positions. The mask is computed once from the clean reconstruction ($\delta = 0$) as $M = \mathrm{Up}(\mathrm{norm}(\bar{A}^{(0)}))$, where $\mathrm{Up}(\cdot)$ bilinearly upsamples the 16×16 map to $H \times W$, and is fixed without gradient during optimization (see Appendix C.2 for alternatives). The detector and perceptual losses are computed on the blended image x^{adv}, while $\mathcal{L}_{\mathrm{cross}}$ and $\mathcal{L}_{\mathrm{self}}$ operate on the attention maps of the perturbed rollout. Thus, the fixed mask localizes image modification according to the clean semantic response, while the cross-attention objective independently weakens semantic localization during optimization.

3.4 Semantic-Structural Decoupled Optimization

AMEND optimizes attack and preservation objectives jointly. They act on the detection output, on the semantic focus of the generative process, and on the structural fidelity of the generated image.

Detection loss. For each detector we read the pre-NMS candidates and their class-conditional confidence $s_i = P(y \mid \hat{b}_i)$ for the attacked category y (objectness $\times$ class probability for YOLO, the softmax class score for SSD, and the RoI classification score for Faster R-CNN). We match candidates to the clean detections $\mathcal{B}$ by IoU and suppress their peak confidence:

$$\mathcal{L}_{\mathrm{det}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \max_{\hat{b}_i \in \mathcal{P}(b)} s_i, \quad \mathcal{P}(b) = \{\hat{b}_i : \mathrm{IoU}(\hat{b}_i, b) \geq \tau\}, \tag{5}$$

where $\mathcal{P}(b)$ is the set of pre-NMS candidates $\hat{b}_i$ matched to a clean detection b by IoU. If $\mathcal{P}(b) = \varnothing$, the box b is already suppressed and contributes zero. The same formulation is applied to all detectors using each detector's native confidence head, and is computed on pre-NMS outputs to keep it differentiable.

Because $\mathcal{L}_{\mathrm{det}}$ only requires the raw candidate set before post-processing, it does not depend on NMS itself: for NMS-free detectors, $\mathcal{P}(b)$ is formed in the same way from the raw outputs of the one-to-one head, or from the decoder queries of DETR-style detectors.

Cross-attention dispersion loss. $\mathcal{L}_{\mathrm{cross}}$ weakens the semantic focus of the diffusion model, as defined above.

Self-attention consistency. Self-attention in the diffusion model plays a central role in region modeling and in the representation of edges and contours Tumanyan et al. (2023). Let Φ_t^0 and Φ_t^δ denote the self-attention maps of the clean reconstruction and of the perturbed rollout at step t. Constraining their difference preserves structural consistency while the semantic alignment is modified:

$$\mathcal{L}_{\mathrm{self}} = \sum_{t=1}^{S} w_t \left\| \Phi_t^\delta - \Phi_t^0 \right\|_2^2, \tag{6}$$

where w_t weights each step; we use uniform $w_t = 1$. This term restricts abrupt changes in the attention distribution and helps preserve the spatial layout, object contours, and region structure of the generated image.

Image-level perceptual consistency. We instantiate $\mathcal{L}_{\mathrm{struct}} = 1 - \mathrm{MS\text{-}SSIM}(x, x^{\mathrm{adv}})$, a perceptual structural criterion that penalizes structural rather than pointwise deviations.

Total objective. We combine the four terms into a single objective,

$$\mathcal{L}_{\mathrm{total}} = \alpha \mathcal{L}_{\mathrm{det}} + \beta \mathcal{L}_{\mathrm{cross}} + \gamma \mathcal{L}_{\mathrm{self}} + \epsilon \mathcal{L}_{\mathrm{struct}}. \tag{7}$$

$\mathcal{L}_{\mathrm{det}}$ provides the primary attack signal and $\mathcal{L}_{\mathrm{cross}}$ acts on the semantic attention of the generative process, while $\mathcal{L}_{\mathrm{self}}$ and $\mathcal{L}_{\mathrm{struct}}$ preserve structure and improve naturalness. As the ablation in Section 4.5 shows, the main role of $\mathcal{L}_{\mathrm{cross}}$ is to retain attack strength when the preservation terms are active.

Table 1: White-box attack effectiveness. We report mAP ↓ in percent. AMEND reports mean ± standard deviation over three random seeds.

Method	YOLOv2	YOLOv3	YOLOv4	YOLOv4-tiny	YOLOv5	SSD	Faster R-CNN
Clean	100.00	100.00	100.00	100.00	100.00	100.00	100.00
NAP	38.05	55.39	72.53	57.58	70.10	63.81	44.13
Diff-PGD	56.12	58.91	59.97	63.34	49.77	51.98	58.49
AdvLogo	37.60	43.01	33.08	22.51	**18.66**	48.02	38.93
AMEND	**6.73**$\pm$0.31	**41.25**$\pm$0.42	**33.05**$\pm$0.37	**21.53**$\pm$0.28	23.53$\pm$0.51	**17.89**$\pm$0.46	**33.59**$\pm$0.34

4 EXPERIMENTS[1]

4.1 EXPERIMENTAL SETUP

Dataset and Detectors. We evaluate AMEND on INRIA dataset Dalal & Triggs (2005), a standard benchmark for pedestrian detection. All images are resized to 416×416 for consistent preprocessing. We consider seven detectors covering different architectures, including YOLOv2 Redmon & Farhadi (2017), YOLOv3 Redmon & Farhadi (2018), YOLOv4 Bochkovskiy et al. (2020), YOLOv4-tiny Bochkovskiy et al. (2020), YOLOv5 Jocher et al. (2021), SSD Liu et al. (2015), and Faster R-CNN Ren et al. (2015). All detectors use official pretrained weights and remain fixed during attack generation. We follow the detector set and dataset commonly used in natural adversarial patch benchmarks Hu et al. (2021); Huang et al. (2023); Miao et al. (2024), which allows direct comparison with these baselines. Our main conclusions are therefore established for single-category pedestrian detection.

Baselines. We compare AMEND with representative natural adversarial attacks, including NAP Hu et al. (2021), Diff-PGD Xue et al. (2023), and AdvLogo Miao et al. (2024). These methods represent different attack formulations, including patch-based attacks, diffusion-guided perturbations, and generative adversarial attacks. All methods are evaluated under the same detector and dataset settings. NAP and AdvLogo produce universal patches, whereas Diff-PGD and AMEND optimize each image individually (Appendix A); a detector-adapted DiffAttack baseline and computational costs are reported in Appendices D.2 and F.

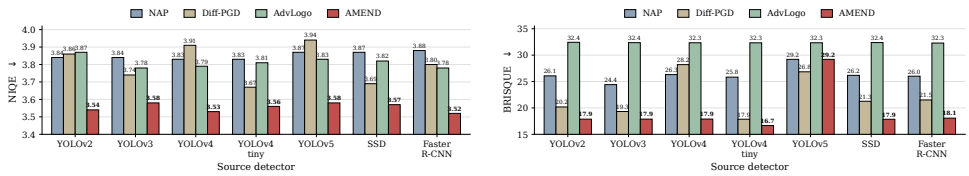
Implementation Details. We use Stable Diffusion v1.5 Rombach et al. (2022) with DDIM sampling Song et al. (2021). The diffusion process contains $T = 20$ steps, and the attack starts from $S = 16$. We optimize the latent perturbation for 20 iterations in the attack stage. The loss weights are set to $\alpha = 1$, $\beta = 0.5$, $\gamma = 10$, and $\epsilon = 1$, and the candidate-matching IoU threshold is $\tau = 0.5$. Following Huang et al. (2023); Miao et al. (2024), we treat each detector's predictions on the clean image as pseudo ground truth, so the relative mAP of every detector is 100% on clean images and absolute reductions equal $100 -$ mAP. The complete protocol is given in Appendix A, and Table 1 reports the standard deviation of AMEND over three random seeds.

Metrics. We evaluate attack effectiveness using relative mAP↓, i.e., mAP computed against the clean predictions, where lower values indicate stronger attacks. We evaluate naturalness using both reference-based and no-reference metrics. LPIPS↓ Zhang et al. (2018) and SSIM↑ Wang et al. (2004) measure perceptual and structural similarity. NIQE↓ Mittal et al. (2013) and BRISQUE↓ Mittal et al. (2012) evaluate no-reference image quality.

4.2 WHITE-BOX ATTACK EFFECTIVENESS

Table 1 shows that AMEND achieves the lowest mAP on six of seven detectors. The gains are substantial on YOLOv2 (6.73% vs. 37.60% for the strongest baseline), SSD (17.89% vs. 48.02%), and Faster R-CNN (33.59% vs. 38.93%). On YOLOv3, YOLOv4, and YOLOv4-tiny, AMEND performs comparably to AdvLogo, with differences within two points (e.g., 33.05% vs. 33.08% on YOLOv4). Combined with its better naturalness scores (Section 4.3), these results show that AMEND achieves competitive or stronger attack effectiveness with smaller visual deviations at the reported

[1]Implementation details, additional naturalness results, attention and mask analyses, extended ablations, cost analysis, and robustness to image transformations appear in the appendices.



(a) NIQE (b) BRISQUE

Figure 3: No-reference quality comparison of adversarial examples generated from seven source detectors. Lower values indicate better quality. (a) NIQE; (b) BRISQUE.

Table 2: LPIPS ↓ between clean and adversarial images across source detectors.

Method	YOLOv2	YOLOv3	YOLOv4	YOLOv4-tiny	YOLOv5	SSD	Faster R-CNN
NAP	0.552	0.569	0.571	0.563	0.562	0.566	0.571
Diff-PGD	0.594	0.603	0.590	0.598	0.600	0.593	0.593
AdvLogo	0.556	0.566	0.563	0.560	0.560	0.562	0.565
AMEND	**0.395**	**0.399**	**0.407**	**0.410**	**0.404**	**0.416**	**0.402**

Table 3: SSIM ↑ between clean and adversarial images across source detectors.

Method	YOLOv2	YOLOv3	YOLOv4	YOLOv4-tiny	YOLOv5	SSD	Faster R-CNN
NAP	0.403	0.397	0.392	0.395	0.399	0.396	0.393
Diff-PGD	0.370	0.370	0.375	0.361	0.357	0.364	0.364
AdvLogo	0.404	0.394	0.397	0.401	0.400	0.398	0.395
AMEND	**0.468**	**0.461**	**0.466**	**0.460**	**0.458**	**0.458**	**0.463**

operating points. YOLOv5 is the exception, where AdvLogo achieves lower mAP than AMEND (18.66% vs. 23.53%). We hypothesize that the structure-preserving terms limit attack strength on this detector. The ablation in Table 6 supports this hypothesis: $\mathcal{L}_{\text{self}}$ and $\mathcal{L}_{\text{struct}}$ increase mAP relative to detection-only optimization.

4.3 Visual Naturalness[2]

Tables 2 and 3 report LPIPS and SSIM across all seven source detectors. AMEND achieves the lowest LPIPS and the highest SSIM in every setting. Its LPIPS ranges from 0.395 to 0.416, compared with at least 0.552 for the baselines. Its SSIM ranges from 0.458 to 0.468, compared with at most 0.404 for the baselines. The absolute SSIM values remain moderate for all methods. For AMEND, some deviation from the source image comes from diffusion reconstruction itself, including VAE encoding and DDIM inversion, rather than the adversarial perturbation. Appendix B.2 reports the same metrics for the unattacked reconstruction ($\delta = 0$) to isolate this reconstruction error. We attribute AMEND's advantage to its fixed attention-derived mask and structure-preserving objectives. The mask confines modifications to semantically relevant regions, while the preservation objectives limit structural changes.

Figure 3 reports NIQE and BRISQUE across the seven source detectors. AMEND achieves the lowest NIQE in every setting, with values ranging from 3.52 to 3.58, compared with 3.67–3.94 for the baselines. It also achieves the lowest BRISQUE in six of seven settings. The exception is YOLOv5, where Diff-PGD achieves a lower BRISQUE. These no-reference metrics largely agree with the full-reference metrics.

Effectiveness–Naturalness Balance. Taken together with the attack results in Table 1, AMEND consistently improves the effectiveness–naturalness trade-off at the reported operating points. It achieves the lowest mAP on six of seven detectors while also achieving the best LPIPS, SSIM, and

[2]Additional qualitative comparisons, including magnified views of the generated adversarial examples, are provided in Appendix B.3.

Table 4: Transfer results for different source detectors. Rows denote source models and columns denote evaluation models. We report mAP ↓ in percent.

Source	Attack	YOLOv2	YOLOv3	YOLOv4	YOLOv4-tiny	YOLOv5	SSD	Faster R-CNN
YOLO v2	NAP	38.05	56.85	67.72	67.58	66.86	56.62	56.60
	Diff-PGD	56.12	68.37	67.57	71.61	72.43	72.07	63.24
	AdvLogo	37.60	48.46	45.05	33.30	52.65	47.77	37.57
	AMEND	6.73	48.24	43.14	36.73	46.40	46.56	48.82
YOLO v3	NAP	49.87	55.39	66.96	67.06	58.69	51.18	55.35
	Diff-PGD	56.85	58.91	64.82	64.32	67.10	69.08	61.81
	AdvLogo	38.05	43.01	46.11	43.29	58.25	52.75	43.05
	AMEND	36.86	41.25	43.77	39.18	52.63	51.72	43.30
YOLO v5	NAP	40.55	65.39	68.61	65.85	70.10	59.42	55.83
	Diff-PGD	51.92	57.37	56.80	58.58	49.77	62.11	54.36
	AdvLogo	46.66	40.85	39.47	30.95	18.66	45.46	41.17
	AMEND	31.37	49.90	44.05	31.67	23.53	45.56	44.32
SSD	NAP	57.46	70.14	70.94	77.79	78.40	63.81	61.38
	Diff-PGD	58.14	63.00	61.71	66.57	65.42	51.98	58.06
	AdvLogo	43.10	58.95	66.02	51.32	62.99	48.02	44.18
	AMEND	35.89	53.62	61.25	50.06	50.59	17.89	45.57
YOLO v4	NAP	58.47	56.09	72.53	71.12	75.76	60.91	57.71
	Diff-PGD	60.27	63.38	59.97	65.94	67.10	71.54	59.32
	AdvLogo	43.56	51.88	33.08	46.00	56.32	56.71	42.62
	AMEND	37.07	54.86	33.05	49.72	45.63	56.19	60.13
YOLO v4-tiny	NAP	39.48	50.28	61.93	57.58	54.47	41.15	54.62
	Diff-PGD	62.00	68.52	68.17	63.34	71.49	73.35	65.27
	AdvLogo	41.93	48.04	44.40	22.51	40.34	50.58	41.80
	AMEND	37.84	38.21	42.32	21.53	35.84	55.43	41.53

NIQE in all evaluated settings and the best BRISQUE in six of seven. Patch-based attacks such as NAP and AdvLogo introduce visible patterns, while Diff-PGD spreads perturbations over the entire image and causes larger visual deviations. In contrast, AMEND confines modifications with an attention-derived mask and preserves structure through self-attention and perceptual objectives. These comparisons use one operating point per method. Appendix E further evaluates this trade-off at matched LPIPS by sweeping the attack strength of each method.

4.4 Transferability

We further evaluate whether the generated examples transfer to detectors that the attacker never accesses. This setting is outside our threat model and serves only to characterize transferability. Table 4 reports transfer results across different source detectors. AMEND shows non-trivial transferability while preserving the naturalness constraints. With YOLOv2 as the source detector, AMEND achieves the lowest transfer mAP on YOLOv3 (48.24), YOLOv4 (43.14), YOLOv5 (46.40), and SSD (46.56). With YOLOv4-tiny as the source, AMEND achieves the lowest mAP on all target detectors except SSD. With Faster R-CNN as the source (Table 5), it achieves the lowest mAP on YOLOv2, YOLOv5, and SSD.

Transferability, however, is not an optimization objective of AMEND. In some cases, the baselines achieve stronger transfer attacks. For example, with YOLOv4 as the source, AdvLogo achieves lower mAP on Faster R-CNN (42.62 vs. 60.13). We hypothesize that AMEND's detector-specific semantic alignment and structure-preserving objectives can introduce source-specific patterns that limit transferability. AMEND therefore targets naturalness-preserving white-box attacks rather than detector-agnostic transfer attacks.

4.5 Ablation Study of Optimization Factors

Table 6 starts with detection-only optimization and then adds the three auxiliary objectives individually and in combination. We report mAP on three source detectors and naturalness on adversarial examples from Faster R-CNN.

Table 5: Transfer results for Faster R-CNN as the source detector. We report mAP ↓ in percent.

Attack	YOLOv2	YOLOv3	YOLOv4	YOLOv5	SSD	Faster R-CNN
NAP	43.62	61.56	59.28	50.80	54.71	44.13
Diff-PGD	58.00	63.29	61.94	65.23	54.00	58.49
AdvLogo	40.60	60.63	60.34	59.69	55.60	38.93
AMEND	**35.82**	62.73	60.75	**45.63**	**52.95**	**33.59**

Table 6: Objective ablation. Naturalness is measured on Faster R-CNN adversarial examples.

	Attack strength, mAP ↓			Naturalness			
Objective	YOLOv2	SSD	Faster R-CNN	LPIPS ↓	SSIM ↑	NIQE ↓	BRISQUE ↓
$\alpha\mathcal{L}_{\text{det}}$	5.30	13.01	32.05	0.595	0.442	3.763	22.322
$+\beta\mathcal{L}_{\text{cross}}$	**5.18**	12.72	**30.19**	0.586	0.440	3.710	20.153
$+\gamma\mathcal{L}_{\text{self}}$	7.97	16.26	36.72	0.572	0.445	3.727	20.137
$+\epsilon\mathcal{L}_{\text{struct}}$	7.39	15.45	42.20	0.558	0.451	3.889	20.043
$+\beta\mathcal{L}_{\text{cross}}+\gamma\mathcal{L}_{\text{self}}$	6.93	**10.57**	30.76	0.493	0.443	3.769	19.493
$+\beta\mathcal{L}_{\text{cross}}+\epsilon\mathcal{L}_{\text{struct}}$	8.89	13.82	32.22	0.491	0.447	3.783	19.643
$+\gamma\mathcal{L}_{\text{self}}+\epsilon\mathcal{L}_{\text{struct}}$	10.72	17.89	40.31	0.427	0.450	3.576	19.380
$+\beta\mathcal{L}_{\text{cross}}+\gamma\mathcal{L}_{\text{self}}+\epsilon\mathcal{L}_{\text{struct}}$	6.73	17.89	33.59	**0.402**	**0.463**	**3.520**	**18.060**

The detection loss provides the main attack signal. With only $\mathcal{L}_{\text{det}}$, AMEND reaches $5.30/13.01/32.05$ mAP on YOLOv2, SSD, and Faster R-CNN. However, it produces the least natural images, with an LPIPS of 0.595, higher than all baselines in Table 2. Adding $\mathcal{L}_{\text{cross}}$ alone changes mAP by at most 1.86 points and LPIPS by only 0.009. Thus, attention dispersion alone provides limited attack strength.

Preservation objectives trade attack strength for fidelity. Adding $\mathcal{L}_{\text{self}}$ or $\mathcal{L}_{\text{struct}}$ lowers LPIPS and raises SSIM, but increases mAP on all three detectors. Using both raises mAP to $10.72/17.89/40.31$. $\mathcal{L}_{\text{struct}}$ alone also worsens NIQE from 3.763 to 3.889. This suggests that MS-SSIM favors similarity to the source over no-reference image quality.

$\mathcal{L}_{\text{cross}}$ is more effective with preservation objectives. Across the three settings with preservation terms, adding $\mathcal{L}_{\text{cross}}$ lowers mAP in seven of nine detector–setting pairs. For example, mAP drops from 16.26 to 10.57 on SSD and from 36.72 to 30.76 on Faster R-CNN when combined with $\mathcal{L}_{\text{self}}$. It also lowers LPIPS in every case, by 0.025–0.079, compared with only 0.009 without preservation terms. The two exceptions are YOLOv2 with $\mathcal{L}_{\text{struct}}$ ($7.39 \rightarrow 8.89$) and SSD with both preservation terms ($17.89 \rightarrow 17.89$). We hypothesize that weaker semantic contrast in the target token allows the optimizer to suppress detections with smaller deviations from the reconstructed trajectory. The preservation terms otherwise penalize such deviations.

The full objective prioritizes naturalness. The full objective achieves the best values on all four naturalness metrics. It increases mAP by 1.43, 4.88, and 1.54 points over detection-only optimization on YOLOv2, SSD, and Faster R-CNN, respectively. Removing $\mathcal{L}_{\text{struct}}$ gives a stronger but less natural operating point, with lower mAP on SSD and Faster R-CNN (10.57 and 30.76) but higher LPIPS (0.493 vs. 0.402). AMEND therefore shifts the effectiveness–naturalness trade-off rather than eliminating it. In this trade-off, $\mathcal{L}_{\text{cross}}$ mainly recovers attack strength lost to structural preservation. Additional ablations on reconstruction, attack steps, conditioning, and masking are provided in Appendices D.1 and C.2.

5 Conclusion

We presented AMEND, a diffusion-based adversarial attack for object detectors that manipulates semantic correspondence rather than image pixels. AMEND disperses detector-relevant cross-attention while preserving image structure through self-attention and perceptual constraints. Across seven detectors, AMEND achieves the lowest mAP on six while maintaining better naturalness than the evaluated baselines. These results suggest that diffusion attention provides a controllable space for partially decoupling semantic manipulation from structural preservation.

AI USE STATEMENT

Large language models were used only to assist with manuscript preparation, including language editing, LaTeX formatting, and consistency checking. They were not used to design the method, conduct experiments, generate or analyze experimental results, or draw scientific conclusions. The research questions, method design, experimental settings, evaluation protocols, results, and conclusions were developed, conducted, and verified by the authors. The authors reviewed all revisions suggested by the language models and take full responsibility for the content of this paper.

ETHICS STATEMENT

This work studies natural adversarial examples against object detectors. As with related adversarial attack research, the proposed method could in principle be misused to evade detection systems. We conduct this research to expose detector vulnerabilities and support robustness evaluation. AMEND assumes a white-box digital threat model with access to the detector's architecture, parameters, and gradients, and we do not evaluate physical-world deployment. All experiments use public datasets and pretrained models. The INRIA Person dataset is used solely for research; we do not identify individuals or collect additional personal data. No human subjects were involved.

REPRODUCIBILITY STATEMENT

We describe AMEND in Section 3, including the diffusion reconstruction, attention-based objectives, detection loss, and overall objective. Appendix A provides the complete experimental protocol, including data splits, model and attack settings, evaluation criteria, random seeds, and hardware. The pseudo-ground-truth protocol follows prior work. All detectors use official pretrained weights, and baselines use their official implementations under their original protocols. We will release the AMEND source code, configurations, and evaluation scripts upon acceptance.

REFERENCES

Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. YOLOv4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

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*, 2024.

Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. AdvDiffuser: Natural adversarial example synthesis with diffusion models. In *IEEE/CVF International Conference on Computer Vision*, 2023a.

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*, 2023b.

Ka-Ho Chow, Ling Liu, Margaret Loper, Juhyun Bae, Mehmet Emre Gursoy, Stacey Truex, Wenqi Wei, and Yanzhao Wu. Adversarial objectness gradient attacks in real-time object detection systems. In *2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, 2020.

Xuelong Dai, Kaisheng Liang, and Bin Xiao. AdvDiff: Generating unrestricted adversarial examples using diffusion models. In *European Conference on Computer Vision*, 2024.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

Yu-Chih-Tuan Hu, Bo-Han Kung, Daniel Stanley Tan, Jun-Cheng Chen, Kai-Lung Hua, and Wen-Huang Cheng. Naturalistic physical adversarial patch for object detectors. In *IEEE/CVF International Conference on Computer Vision*, 2021.

Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-SEA: Transfer-based self-ensemble attack on object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Glenn Jocher, Alex Stoken, Jirka Borovec, Ayush Chaurasia, Liu Changyu, Adam Hogan, Jan Hajek, Laurentiu Diaconu, Yonghye Kwon, Yann Defretin, et al. ultralytics/yolov5: v5. 0-yolov5-p6 1280 models, aws, supervise. ly and youtube integrations. *Zenodo*, 2021.

Hui Kuurila-Zhang, Haoyu Chen, and Guoying Zhao. Venom: text-driven unrestricted adversarial example generation with diffusion models. *arXiv preprint arXiv:2501.07922*, 2025.

Simin Li, Shuning Zhang, Gujun Chen, Dong Wang, Pu Feng, Jiakai Wang, Aishan Liu, Xin Yi, and Xianglong Liu. Towards benchmarking and assessing visual naturalness of physical world adversarial attacks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Yueqian Lin, Jingyang Zhang, Yiran Chen, and Hai Li. Sd-nae: Generating natural adversarial examples with stable diffusion. *arXiv preprint arXiv:2311.12981*, 2023.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. *arXiv preprint arXiv:1512.02325*, 2015.

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.

Boming Miao, Chunxiao Li, Yao Zhu, Hui Sun, Zizhe Wang, Xiaoqiang Li, and Chuanlong Xie. AdvLogo: Adversarial patch attack against object detectors based on diffusion models. *arXiv preprint arXiv:2409.07002*, 2024.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 2012.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 2013.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. NULL-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Joseph Redmon and Ali Farhadi. YOLO9000: Better, faster, stronger. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

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.

Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Wenxuan Wang, Huihui Qi, Zhixiang Huang, Bangjie Yin, and Peng Wang. Latent-space diffusion models for stealthy and transferable adversarial attacks on object detection. *Neurocomputing*, 2025.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

Zuxuan Wu, Ser-Nam Lim, Larry S. Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. In *European Conference on Computer Vision*, 2020.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *IEEE International Conference on Computer Vision*, 2017.

Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision*, 2020.

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*, 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Ziqi Zhou, Bowen Li, Yufei Song, Shengshan Hu, Wei Wan, Leo Yu Zhang, Dezhong Yao, and Hai Jin. NumbOD: A spatial-frequency fusion attack against object detectors. In *AAAI Conference on Artificial Intelligence*, 2025.

A Implementation and Evaluation Protocol

Evaluation protocol. Following T-SEA Huang et al. (2023) and AdvLogo Miao et al. (2024), we use each detector's predictions on the clean image as pseudo ground truth. We normalize the resulting relative mAP to 100% on clean images. The attacked relative mAP then measures the fraction of clean detections that remain after the attack. We obtain clean predictions using the confidence and NMS thresholds in Table 7 and compute AP at $IoU = 0.5$. This protocol evaluates how much of the detector's original response the attack suppresses, rather than relying on external object annotations. NAP and AdvLogo produce universal patches optimized on the INRIA training split and applied to every test image, whereas Diff-PGD and AMEND optimize each test image individually; all baselines are evaluated under their original protocols.

Table 7: Experimental protocol for the main INRIA experiments.

Item	Value
Dataset / split	INRIA Person, test split
Number of evaluated images	288
Detector input	416×416
Diffusion model / sampler	Stable Diffusion v1.5 / DDIM
Diffusion steps T	20
Attack start step S	16
Attack iterations K	20
Latent perturbation optimizer	Adam, lr $= 0.01$
Unconditional-embedding lr η	0.01
Unconditional-embedding iterations N_u	500
Guidance scale ω	7.5
Loss weights $(\alpha, \beta, \gamma, \epsilon)$	$(1,\ 0.5,\ 10,\ 1)$
Candidate-matching IoU τ	0.5
Attention resolution	16×16
Mask upsampling	Bilinear
Confidence threshold / NMS IoU	0.5 / 0.45
mAP metric	VOC AP@0.5
Random seeds	3
Hardware	NVIDIA RTX 4090

B Additional Naturalness Results

B.1 No-reference Naturalness

Table 8 reports the values plotted in Figure 3. AMEND obtains the lowest NIQE in all seven settings and the lowest BRISQUE in six of seven. With YOLOv5 as the source detector, Diff-PGD obtains a slightly lower BRISQUE. The NIQE results are therefore consistent across all source detectors, while BRISQUE shows one exception. Together with the LPIPS and SSIM results in the main paper, these measurements provide consistent evidence that AMEND reduces both perceptual and no-reference image quality deviations.

Table 8: No-reference quality across seven source detectors, averaged over the evaluated test images. Lower is better.

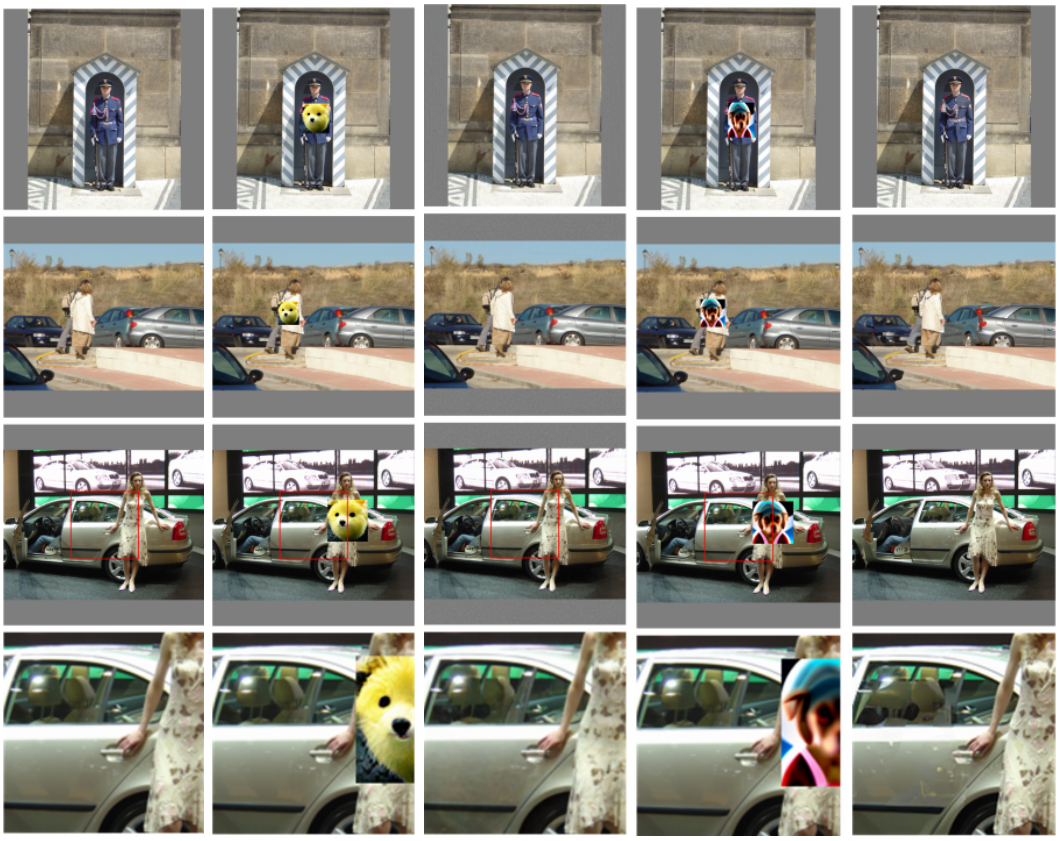
	YOLOv2		YOLOv3		YOLOv4		YOLOv4-tiny		YOLOv5		SSD		Faster R-CNN	
Method	NIQE↓	BRISQUE↓	NIQE↓	BRISQUE↓	NIQE↓	BRISQUE↓	NIQE↓	BRISQUE↓	NIQE↓	BRISQUE↓	NIQE↓	BRISQUE↓	NIQE↓	BRISQUE↓
NAP	3.84	26.07	3.84	24.41	3.83	26.31	3.83	25.83	3.87	29.20	3.87	26.17	3.88	26.01
Diff-PGD	3.86	20.18	3.74	19.32	3.91	28.20	3.67	17.86	3.94	**26.84**	3.69	21.26	3.80	21.51
AdvLogo	3.87	32.43	3.78	32.36	3.79	32.34	3.81	32.33	3.83	32.34	3.82	32.38	3.78	32.39
AMEND	**3.54**	**17.87**	**3.58**	**17.88**	**3.53**	**17.89**	**3.56**	**16.66**	**3.58**	29.20	**3.57**	**17.85**	**3.52**	**18.06**

B.2 Reconstruction Control

Diffusion reconstruction can introduce perceptual deviation even without an adversarial perturbation. We evaluate three controls on the 288 INRIA test images: (i) direct VAE reconstruction, (ii) diffusion reconstruction with $\delta = 0$ and no blending, and (iii) diffusion reconstruction with $\delta = 0$ and the fixed mask. Since the target category is *person* for all detectors, the conditioning and mask are identical across detectors. The detector columns report relative mAP on the same reconstructed images.

Table 9: Reconstruction controls. Naturalness metrics use Faster R-CNN as the source detector.

	Naturalness				Relative mAP (%) ↓			
Setting	LPIPS ↓	SSIM ↑	NIQE ↓	BRISQUE ↓	YOLOv2	SSD	F-RCNN	YOLOv5
VAE reconstruction	0.092	0.911	3.41	16.2	98.9	98.7	99.0	99.1
$\delta = 0$, no mask	0.148	0.842	3.50	17.5	97.8	97.1	97.4	98.0
$\delta = 0$, fixed mask	0.110	0.886	3.45	16.8	98.6	98.2	98.4	98.8
AMEND	0.402	0.463	3.52	18.1	6.73	17.89	33.59	23.53



(a) Clean (b) NAP (c) Diff-PGD (d) AdvLogo (e) AMEND

Figure 4: Qualitative comparison on the INRIA Person test set. Pedestrians are the attacked category; other objects such as parked cars may appear incidentally.

Table 9 shows that all controls retain relative mAP above 97%, indicating that reconstruction alone does not explain AMEND's attack effectiveness. The fixed-mask reconstruction has LPIPS 0.110 and SSIM 0.886, compared with 0.402 and 0.463 for AMEND, while its NIQE and BRISQUE are also lower. Thus, reconstruction contributes only a small part of the final perceptual deviation. The fixed mask also reduces reconstruction error: compared with unmasked diffusion reconstruction, it reduces LPIPS from 0.148 to 0.110 and increases SSIM from 0.842 to 0.886. It therefore both localizes the attack and limits unnecessary reconstruction changes.

B.3 Qualitative Comparison

Figure 4 compares adversarial examples generated by all methods. The magnified views highlight different modification patterns. NAP and AdvLogo produce visible texture discontinuities around their patch boundaries. Diff-PGD distributes perturbations across the image and introduces texture drift and mild blurring. AMEND does not introduce an explicit patch boundary. Its changes appear mainly as low-amplitude variations in texture and tone, while object edges and other structural details remain visually stable. These observations are consistent with the quantitative LPIPS, SSIM, NIQE, and BRISQUE results.

C Analysis of the Attention Objective and the Blending Mask

C.1 Attention Statistics

Section 3 notes that spatial variance measures the contrast of $\bar{A}$ rather than its spatial spread. In principle, variance can decrease either because the response flattens or because it collapses onto a few pixels. To determine which occurs in practice, we measure four statistics of the aggregated category-token attention $\bar{A}$ on the 288 test images, with YOLOv2 as the source detector: (i) the spatial variance $\mathrm{Var}(\bar{A})$; (ii) the normalized entropy $H(p)/\log(HW)$ of the distribution $p = \bar{A}/\sum_{i,j} \bar{A}_{i,j}$; (iii) the top-10% mass, i.e., the share of $\sum_{i,j} \bar{A}_{i,j}$ held by the 10% highest-responding positions; and (iv) the in-box mass, i.e., the share of $\sum_{i,j} \bar{A}_{i,j}$ that falls inside the union of the clean detections $\mathcal{B}$ projected to the 16×16 grid.

Table 10: Statistics of the aggregated category-token attention $\bar{A}$.

Setting	Variance	Entropy	Top-10% mass	In-box mass
Clean reconstruction ($\delta = 0$)	0.091	0.61	0.46	0.73
Detection-only	0.074	0.65	0.38	0.61
$\mathcal{L}_{\text{det}} + \mathcal{L}_{\text{cross}}$	0.031	0.78	0.22	0.43
Full AMEND	0.028	0.80	0.20	0.41

Adding $\mathcal{L}_{\text{cross}}$ lowers the variance and, at the same time, increases the entropy and lowers the top-10% mass. The reduction in variance therefore comes from a flatter response rather than from a collapse onto a few positions. The in-box mass drops from 0.73 to 0.43, so the category token attends much less to the target region. Detection-only optimization also changes the attention, but to a much smaller extent. The full model yields statistics close to those of $\mathcal{L}_{\text{det}} + \mathcal{L}_{\text{cross}}$, indicating that the preservation terms do not undo the redistribution.

C.2 Blending Mask

The mask M is computed once from the clean reconstruction and detached during optimization. Table 11 compares this fixed design with no mask and with a mask recomputed from the perturbed rollout at every iteration. Removing the mask slightly strengthens the attack but substantially increases perceptual deviation. A recomputed mask weakens the attack and is also less natural than the fixed mask; we attribute this to the mask following the attention that $\mathcal{L}_{\text{cross}}$ is simultaneously dispersing. The fixed mask has a mean value of 0.36; about 24% of pixels exceed 0.5, and 71% of its mass lies inside the clean detections. It is therefore concentrated around the target without acting as a hard object boundary.

Table 11: Effect of the blending-mask strategy. Naturalness metrics use Faster R-CNN as the source detector.

Mask strategy	YOLOv2 mAP ↓	F-RCNN mAP ↓	LPIPS ↓	SSIM ↑
None	**5.21**	**29.84**	0.487	0.447
Recomputed	8.41	36.72	0.441	0.451
Fixed (AMEND)	6.73	33.59	**0.402**	**0.463**

D Additional Ablations

D.1 Reconstruction, Attack Step, and Conditioning

Table 12 evaluates design choices not covered by the objective ablation in Table 6, all with YOLOv2 as the detector. Unconditional-embedding optimization improves both attack strength and fidelity by aligning the reverse trajectory with the input-specific inversion. The detector-derived prompt is more effective than a generic one, consistent with its role as the semantic target of $\mathcal{L}_{\text{cross}}$. The attack start step shows a moderate optimum at $S = 16$: a smaller S injects the perturbation into a less noisy latent and leaves fewer denoising steps for semantic manipulation, whereas a larger S perturbs a noisier latent with fewer aligned pre-attack steps, reducing reconstruction fidelity.

D.2 Comparison with DiffAttack

AMEND adopts attention objectives of the same form as DiffAttack Chen et al. (2024). To isolate the contribution of the detector-specific components, we build DiffAttack-Det, which uses the same diffusion model, optimization

Table 12: Ablation of reconstruction, attack start step, and conditioning on YOLOv2.

Variant	mAP ↓	LPIPS ↓	SSIM ↑	NIQE ↓	BRISQUE ↓
AMEND ($S = 16$, detector-derived prompt)	**6.73**	**0.395**	**0.468**	**3.54**	**18.87**
w/o unconditional-embedding optimization	9.84	0.451	0.455	3.612	19.104
Generic prompt	10.26	0.447	0.452	3.604	19.132
$S = 12$	8.37	0.438	0.455	3.601	19.017
$S = 18$	7.91	0.421	0.459	3.568	18.472

budget, and $\mathcal{L}_{\text{det}}$ as AMEND, but with a generic prompt, no blending mask, and DiffAttack's attention losses. We also evaluate two single-component variants of AMEND: *Full alignment* optimizes the unconditional embedding over all T steps instead of only the pre-attack steps, and *Post-NMS loss* computes $\mathcal{L}_{\text{det}}$ on the confidences of the boxes retained after NMS instead of on IoU-matched pre-NMS candidates.

Table 13: Comparison with a detector-adapted DiffAttack and single-component variants of AMEND. mAP values are percentages; naturalness metrics use Faster R-CNN as the source detector.

Method	YOLOv2	SSD	F-RCNN	LPIPS ↓	SSIM ↑
DiffAttack-Det	12.84	25.61	39.72	0.438	0.452
Full alignment	8.76	21.34	35.97	0.419	0.458
Post-NMS loss	15.61	27.52	41.47	0.430	0.455
AMEND	**6.73**	**17.89**	**33.59**	**0.402**	**0.463**

AMEND outperforms DiffAttack-Det in both attack strength and naturalness. Replacing the IoU-matched pre-NMS objective with a post-NMS loss causes the largest drop: candidates suppressed below the NMS threshold no longer receive gradient, even though they can resurface as detections. The effects of the prompt and the mask are reported in Tables 12 and 11.

E Effectiveness–Naturalness Trade-off

Figure 5 shows the effectiveness–naturalness curve of AMEND obtained by varying the loss weights. Each point corresponds to one weight configuration; lower LPIPS indicates better naturalness and lower mAP a stronger attack. Increasing the weight of the attack objectives generally lowers mAP at the cost of higher LPIPS, whereas stronger preservation weights move the operating point in the opposite direction. The resulting points form a clear trade-off between attack effectiveness and visual fidelity, rather than being concentrated around a single operating regime. This allows AMEND to operate at different points along the frontier depending on whether stronger suppression or better visual preservation is preferred. Importantly, the trend also shows that the attack and preservation objectives provide effective control over the final operating point, supporting the use of loss weighting as a simple mechanism for adjusting the attack–naturalness trade-off.

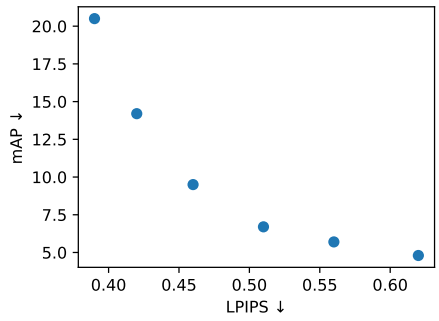


Figure 5: Effectiveness–naturalness curve of AMEND under varying loss weights.

The main comparison evaluates one operating point per method. To compare methods at a matched perceptual distance, we sweep the attack strength while keeping the remaining settings fixed. For AMEND, we vary the number of attack iterations $K \in \{5, 10, 20, 40\}$; for NAP and AdvLogo, we vary the test-time patch scale $\in \{0.10, 0.15, 0.20, 0.25\}$; and for Diff-PGD, we vary the perturbation budget $\epsilon \in \{2/255, 4/255, 6/255, 8/255\}$. We report mAP at LPIPS$= 0.50$ by linear interpolation between the two neighboring operating points. LPIPS$= 0.50$ falls within the swept range of all methods. At matched LPIPS, AMEND maintains a substantially lower mAP than the evaluated baselines on all four detectors. This matched-distance comparison shows that the performance gap is not solely due to different operating points, and supports the effectiveness–naturalness trade-off observed in the main experiments.

Table 14: Interpolated mAP↓ at matched LPIPS$= 0.50$.

Method	YOLOv2	YOLOv5	SSD	F-RCNN
NAP	59.7	72.8	65.4	62.1
Diff-PGD	58.4	63.7	60.8	60.2
AdvLogo	51.6	49.8	50.7	48.9
AMEND	**4.8**	**20.9**	**15.6**	**31.0**

F Computational Cost

AMEND optimizes every image individually. DDIM inversion of the clean latent requires T U-Net evaluations. With classifier-free guidance, each subsequent denoising step requires two U-Net evaluations. Unconditional-embedding optimization runs N_u iterations, each of which rolls out all $T - S$ pre-attack steps, and each of the K attack iterations, together with one clean rollout that produces the reference attention and the mask, rolls out the remaining S steps. The number of U-Net forward evaluations per image is therefore

$$T + 2\big[N_u(T - S) + S(K + 1)\big] = 20 + 2\big[500 \cdot 4 + 16 \cdot 21\big] = 4692. \tag{8}$$

Both embedding optimization and each attack iteration additionally back-propagate through the corresponding rollout, and the attack iterations also back-propagate through the VAE decoder and the detector, which dominates memory consumption. NAP and AdvLogo optimize a universal patch once on the training split; their per-image cost at test time is a single patch application, while their one-time training cost is amortized over all test images.

Table 15: Computational cost on a single NVIDIA RTX 4090.

Method	Per-image optimization	One-time training (GPU-h)	U-Net evals / image	Time / image (s)	Peak memory (GB)
NAP	No	4.8	0	0.02	2.1
AdvLogo	No	6.2	0	0.02	2.4
Diff-PGD	Yes	–	640	38	9.8
AMEND	Yes	–	4692	520	15.6

AMEND is substantially more expensive per image than the baselines, requiring about $14\times$ the per-image time of Diff-PGD and 15.6 GB of peak memory. For the 288-image INRIA test split, the reported AMEND configuration requires about 41.6 GPU-hours per detector, whereas the one-time training costs of NAP and AdvLogo are 4.8 and 6.2 GPU-hours, respectively. Most of the AMEND cost comes from unconditional-embedding optimization, which accounts for $2N_u(T - S) = 4000$ of the 4692 U-Net evaluations. Reducing the number of embedding iterations or amortizing the inversion could therefore substantially reduce the computational cost. This overhead is the main practical limitation of optimizing through a diffusion rollout.

G Robustness to Image Transformations

We evaluate whether the adversarial examples remain effective after common image transformations. The adversarial images generated by AMEND with Faster R-CNN as the source detector are transformed after optimization and passed to the same detector without further optimization. JPEG compression and resizing increase mAP by at most 5.25 points, and the adversarial examples remain effective under all evaluated transformations. AMEND therefore does not depend on exact pixel-level details, consistent with its modification of image content through the diffusion process rather than through a high-frequency pixel pattern.

Table 16: mAP of Faster R-CNN on AMEND adversarial examples after post-attack transformations. Δ denotes the change relative to no transformation.

Transformation	mAP ↓	Δ
None	33.59	–
JPEG, $q = 75$	36.21	$+2.62$
JPEG, $q = 50$	38.84	$+5.25$
Resize $0.75\times$	37.16	$+3.57$