ADVERSARIAL SCENARIO ATTACK: QUERY-EFFICIENT BLACK-BOX GENERATION OF NATURAL ADVERSARIAL EXAMPLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Generating natural adversarial examples (NAEs) from a given image remains challenging under tight black-box query budgets, as many existing generative attacks rely on surrogate-model guidance or attack-specific training. We propose Adversarial Scenario Attack (ASA), a query-based black-box attack on image classifiers that searches over human-readable natural-language editing scenarios rather than pixel-level perturbations. ASA uses a multimodal large language model (MLLM) to propose image-dependent changes to backgrounds, weather, and surface appearance, and a generative editor to apply them to the original image. Winner–loser feedback guides subsequent proposals, while a Greedy Explorer composes complementary scenarios and retains only combinations that improve the attack score. Under a budget of 100 victim-model queries per image, ASA achieves the highest attack success rate on all ten evaluated ImageNet classifiers, including two vision–language models, without surrogate models or attack-specific training. On average, its success rate within 20 queries already exceeds that of every baseline at 100 queries, while it maintains competitive image-level transferability. Because each attack is expressed as a readable scenario, the discovered edits can also be reused: class-neutral scenarios that recur during search, such as light mist or dusk lighting, raise the average error rate to 2.0–3.3 times that of a control prompt when applied unchanged to images from all 1,000 ImageNet classes across 13 classifiers.

1 INTRODUCTION

Deep learning models are known to be vulnerable to adversarial attacks, where carefully designed perturbations can induce incorrect predictions (Szegedy et al., 2014; Biggio et al., 2013; Carlini & Wagner, 2017). Most prior studies focus on restricted attacks that constrain perturbations within an ℓ_p norm ball to preserve perceptual similarity (Goodfellow et al., 2015; Madry et al., 2018; Wong et al., 2020; Song et al., 2024). However, robustness evaluation based only on small pixel-level perturbations may overlook realistic semantic variations. Recent advances in generative models have therefore motivated unrestricted adversarial attacks, which generate natural-looking modifications or Natural Adversarial Examples (NAEs) beyond explicit norm constraints (Chen et al., 2023a;b; Yuan et al., 2022; Collins et al., 2026; Zhao et al., 2018; Hendrycks et al., 2021).

Existing generative NAE methods remain limited for strict black-box attacks on image classifiers when a given input image must be edited. Many methods rely on white-box surrogate gradients or learned attack priors to guide the generation process, and then expect the resulting examples to transfer to black-box victim models (Chen et al., 2023a; 2025; 2023b; Song et al., 2018). Other recent methods synthesize NAEs from random noise, improving naturalness and transferability but producing a new image of the same concept rather than adversarially editing the original input (Collins et al., 2026; Dai et al., 2024). Thus, natural adversarial editing of a given input image remains underexplored against score-based black-box classifiers under tight query budgets, particularly when neither victim-specific surrogate training nor attack-specific prior learning is available.

Recent text-guided generative editors enable realistic, instruction-driven modifications to a given image, thereby expanding the adversarial search space from conventional pixel-level or latent-space perturbations to natural editing scenarios. However, existing query-based generative attacks have

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.

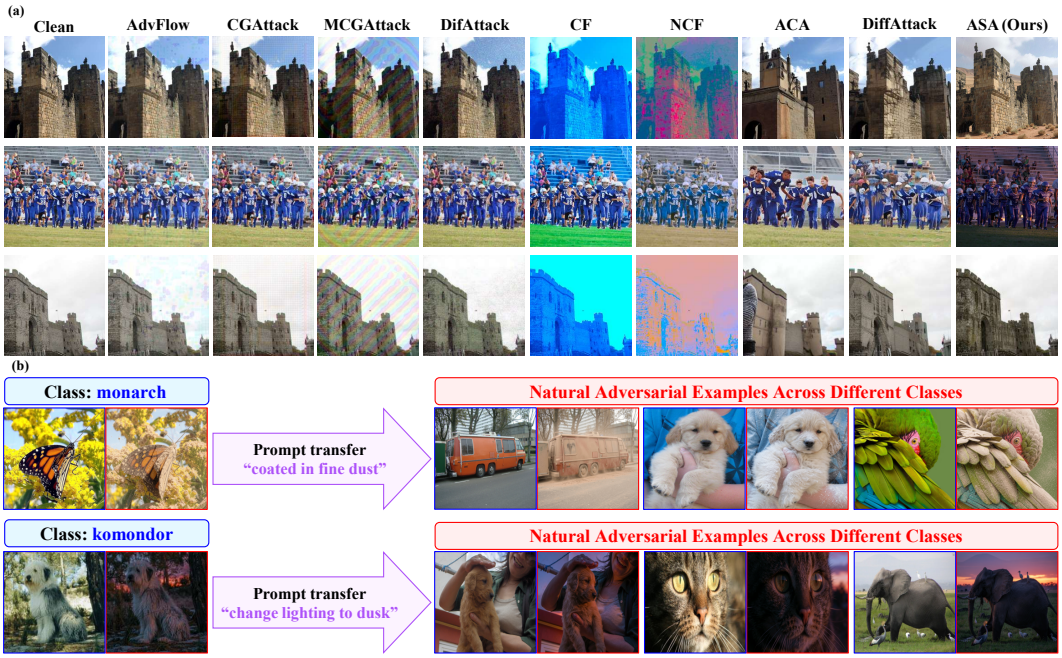


Figure 1: Qualitative overview of ASA. (a) ASA generates visually natural adversarial examples compared with existing attacks. (b) Scenarios discovered by ASA applied unchanged to unseen images of other classes. Additional examples are provided in App. A.

not fully exploited this scenario-level space, particularly under tight query budgets. To this end, we propose **Adversarial Scenario Attack (ASA)**, a query-based black-box framework that generates NAEs by directly searching over natural-language editing scenarios using FLUX.2 [klein] 9B-KV (BlackForestLabs, 2025). Fig. 1(a) shows ASA making scene-level rather than pixel-level changes.

Prior studies on NAEs have shown that model failures can arise from realistic transformations in natural scenes, including changes in background, weather, and surface appearance (Lian et al., 2024; Marchisio et al., 2022; Machiraju & Balasubramanian, 2020; Suryanto et al., 2022; Hu et al., 2022). However, for a given input, it is generally unclear a priori which type of natural transformation will expose the most vulnerable attributes of the victim model. Motivated by this observation, ASA jointly explores multiple natural editing scenarios rather than treating them as separate attack pipelines. Because this scenario space is open-ended and image-dependent, ASA uses Gemma4 (El Abd et al., 2026) to generate plausible strategy-specific transformation scenarios, which are then instantiated by a modern generative editor to produce natural edited candidates. We further introduce a **Greedy Explorer**, which composes strategy-specific scenarios and retains only attack-score-improving combinations, enabling ASA to accumulate complementary natural transformations. As a result, ASA generates effective black-box NAEs with improved query efficiency, without task-specific surrogate training or attack-prior learning, while also showing competitive image-level transferability and promising prompt-level transferability. Our contributions are summarized as follows:

- **Query-efficient black-box NAE generation:** We propose **Adversarial Scenario Attack (ASA)**, a black-box method that searches and composes natural-language editing scenarios to produce NAEs without attack-specific preparation. ASA achieves higher attack success than prior generative attacks on all ten victim models, with fewer victim-model queries on nine of them.
- **Competitive transferability and perceptual similarity:** Among generative unrestricted attacks, ASA best preserves perceptual similarity and subject identity, and its transferability is competitive with transfer-based attacks that require white-box access.
- **Reusable class-neutral scenarios:** Scenarios that recur during ASA's search, such as dusk lighting (Fig. 1(b)), raise the average error rate to 2.0–3.3 times that of a control prompt when applied unchanged to images from all 1,000 ImageNet classes, significantly on all 13 evaluated models, and leave correctly classified images more fragile to noise.

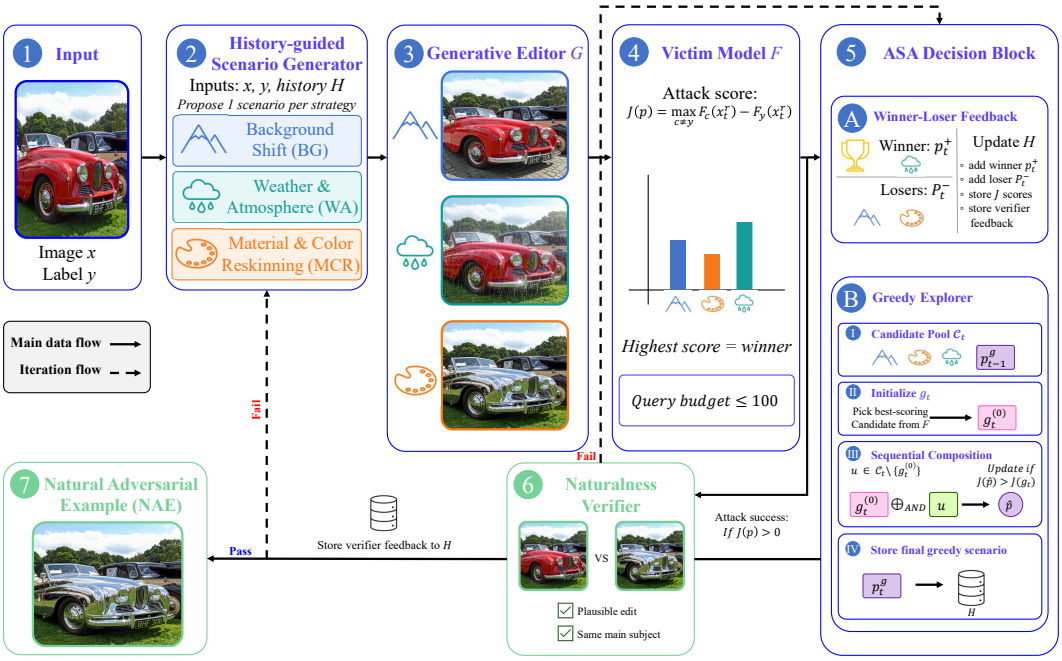


Figure 2: Overview of Adversarial Scenario Attack (ASA). The method searches over natural-language editing scenarios by combining winner–loser feedback with a Greedy Explorer that composes promising candidates into a stronger scenario, followed by naturalness verification.

2 Related Work

2.1 Generative Models for Natural Adversarial Examples

Robustness to realistic distribution shifts has been studied from several directions. Hendrycks et al. (2021) collect naturally occurring images that fool classifiers through adversarial filtration. These benchmarks expose realistic failures but do not adapt a transformation to a given image using victim-model feedback. Other work constructs NAEs through generation or semantic manipulation, such as AC-GAN-based latent-space search (Song et al., 2018) and the color transformations of *CF* (Shamsabadi et al., 2020) and *NCF* (Yuan et al., 2022), which alter colors while preserving image structure to improve naturalness and black-box transferability.

More recently, diffusion models have enabled higher-quality NAEs. *DiffAttack* (Chen et al., 2025) and *ACA* (Chen et al., 2023b) map a given image into a diffusion latent space and optimize the latent using surrogate-model gradients; DiffAttack uses DDIM (Song et al., 2021) inversion and additionally manipulates attention structures. *AdvDiff* (Dai et al., 2024) and *NatADiff* (Collins et al., 2026) instead synthesize adversarial examples from noise through classifier-guided sampling, with NatADiff steering the sampling trajectory toward the intersection of the true and adversarial classes to improve transferability. Because these two methods generate a new image of the target concept, they do not edit a specific source image.

The diffusion-based methods above are driven by gradients from a surrogate or white-box classifier and are therefore mainly studied in white-box or transfer-based settings. Under strict black-box access, surrogate gradients can misalign with the victim model's decision boundary. Moreover, their optimization variables, such as diffusion latents or sampling trajectories, do not describe what was changed in the image. ASA differs along three axes: it edits a given source image, relies only on the victim model's confidence scores, and optimizes a human-readable natural-language scenario that can be inspected and reapplied to other images (see App. E for a detailed comparison).

2.2 Prior-Assisted Query-Based Black-Box Attacks

Query-based black-box attacks optimize perturbations using only feedback from the victim model, without access to its parameters or gradients. Although realistic, this setting is query-inefficient in

high-dimensional input spaces. For example, *AdvFlow* updates a pre-trained normalizing flow via NES-based queries but still requires many victim-model queries (Dolatabadi et al., 2020). To improve query efficiency, recent work incorporates priors learned in advance from surrogate or generative models into query-based optimization (Feng et al., 2022; Yin et al., 2024; Liu et al., 2024).

One line of research transfers priors from surrogate models. *CGAttack* learns a conditional adversarial distribution, transferring part of its parameters and updating the rest via target-model queries (Feng et al., 2022). *MCGAttack* extends this with meta-learning, training a meta-generator across surrogate tasks that rapidly adapts to a new target model (Yin et al., 2024). *DifAttack* trains an autoencoder on clean–adversarial pairs to disentangle adversarial and visual features, then optimizes only the adversarial feature through queries (Liu et al., 2024). However, most prior-assisted methods depend on surrogate transferability or additional attack-specific training (Feng et al., 2022; Yin et al., 2024; Liu et al., 2024), and adopt ℓ_p-bounded formulations that yield less natural adversarial examples than unrestricted ones. Generating NAEs under strict black-box access without attack-specific training thus remains underexplored.

3 METHOD

3.1 PROBLEM DEFINITION

We consider a score-based black-box attack setting. Let $F : \mathcal{X} \to \mathbb{R}^C$ be a victim classifier over C classes, where $F_c(\cdot)$ denotes the confidence score for class c. Given an input image $x \in \mathcal{X}$ with ground-truth label y, the attacker has no access to the architecture, parameters, or gradients of F and can only query F to observe its output scores.

Unlike conventional black-box attacks that search for perturbations in pixel space, we formulate the attack as a search over natural-language editing scenarios. Let G be a generative image editor that takes an image x and an editing instruction $p \in \mathcal{P}$, where $\mathcal{P}$ is the space of natural-language instructions, and produces the edited image $x_p = G(x, p)$.

The attacker aims to find a scenario p whose edit remains realistic yet causes F to misclassify x_p. Following the margin loss commonly used in adversarial attacks (Carlini & Wagner, 2017), we define the attack score as $J(p) = \max_{c \neq y} F_c(x_p) - F_y(x_p)$. A larger $J(p)$ indicates that x_p is closer to the decision boundary, and $J(p) > 0$ indicates a successful attack. We focus on an extremely query-limited regime with a hard budget of 100 victim-model queries per image. Under this constraint, our goal is to identify adversarial editing scenarios efficiently, without high-dimensional gradient estimation or task-specific surrogate training. As in standard untargeted attacks, the attacker knows y, which enters the margin objective and also conditions scenario proposal (Sec. 3.2); claims about reuse across classes therefore rest only on class-neutral scenarios (Sec. 4.3).

3.2 ADVERSARIAL SCENARIO ATTACK

We propose Adversarial Scenario Attack (ASA), a score-based black-box attack that searches over natural language editing scenarios instead of pixel-level perturbations, as illustrated in Fig. 2. ASA organizes its search space by three editing strategies, $\mathcal{R} = \{\mathrm{BG}, \mathrm{WA}, \mathrm{MCR}\}$: Background Shift, Weather & Atmosphere, and Material & Color Reskinning. Each strategy $r \in \mathcal{R}$ defines a scenario space $\mathcal{S}_r \subset \mathcal{P}$ of plausible editing instructions of that type, and the full search space is $\mathcal{S}_{\mathrm{ASA}} = \bigcup_{r \in \mathcal{R}} \mathcal{S}_r$.

ASA performs a history-guided iterative search. It maintains a search history $\mathcal{H}$, initially empty, that records tested scenarios with their attack scores and verifier feedback. At iteration t, an MLLM observes the source image x, the label y, and $\mathcal{H}$, and proposes one candidate $p_t^r \in \mathcal{S}_r$ per strategy, so that heterogeneous transformations compete under a shared query budget. The MLLM is instructed to propose benign, plausible edits that are semantically related to y and resemble ordinary user requests, without naming classes or replacing the object. For each candidate, ASA generates $x_t^r = G(x, p_t^r)$, queries F, and computes $J(p_t^r)$. We define the winning strategy and scenario as $r_t^+ = \arg\max_{r \in \mathcal{R}} J(p_t^r)$ and $p_t^+ = p_t^{r_t^+}$, and the losing scenarios as $\mathcal{P}_t^- = \{p_t^r \mid r \in \mathcal{R} \setminus \{r_t^+\}\}$. After these results are added to $\mathcal{H}$, the next proposal round refines p_t^+ with more specific variants within r_t^+ while avoiding semantic traits shared by $\mathcal{P}_t^-$.

When a candidate is misclassified by F, it is passed to an MLLM naturalness verifier to rule out degenerate successes from corrupted or implausible images; otherwise, the search continues until the query budget is exhausted. Given a side-by-side view of the source and edited images, the base description of the source image, and the class name of y as the main object, the verifier checks whether the main object remains structurally intact and consistent with the base description and whether the edit is plausible in the real world, while tolerating minor artifacts and changes in background or lighting. If the edit is rejected, the verifier returns a one-sentence feedback on the main issue, which is added to $\mathcal{H}$, and the search continues. The final adversarial example must therefore satisfy both classifier failure and label-preserving visual plausibility. The algorithm is given in App. F.

3.3 Greedy explorer

To expose semantic vulnerabilities that emerge only from the interaction of multiple natural transformations, the Greedy Explorer composes promising strategy-specific scenarios into a single multi-scenario prompt. At iteration t, it forms a candidate pool from the current scenarios and, for $t > 1$, the greedy scenario from the previous iteration:

$$\mathcal{C}_t = \begin{cases} \{p_t^r \mid r \in \mathcal{R}\}, & t = 1, \\ \{p_t^r \mid r \in \mathcal{R}\} \cup \{p_{t-1}^g\}, & t > 1, \end{cases} \tag{1}$$

and initializes the greedy state as the best-scoring candidate, $g_t^{(0)} = \arg\max_{p \in \mathcal{C}_t} J(p)$. Each composed scenario is represented as a set of strategy-keyed components. The remaining candidates $\mathcal{C}_t \setminus \{g_t^{(0)}\}$ are visited in descending order of their individual attack scores; if p_{t-1}^g is among them, it is replaced in this list by its components $\mathrm{Comp}(p_{t-1}^g)$ rather than attached as a whole. Let $u_t^{(m)}$ denote the m-th element to be tested and $g_t^{(m-1)}$ the current greedy scenario. ASA forms a trial scenario

$$\hat{p}_t^{(m)} = g_t^{(m-1)} \oplus_{\mathrm{and}} u_t^{(m)}, \tag{2}$$

where $\oplus_{\mathrm{and}}$ joins the two instructions with a conjunction at the component level: a component of $u_t^{(m)}$ replaces the one with the same strategy key in $g_t^{(m-1)}$, or is appended otherwise. The trial is rendered as $G(x, \hat{p}_t^{(m)})$ and scored by querying F, and the greedy state is updated only if the attack objective improves:

$$g_t^{(m)} = \begin{cases} \hat{p}_t^{(m)}, & \text{if } J(\hat{p}_t^{(m)}) > J(g_t^{(m-1)}), \\ g_t^{(m-1)}, & \text{otherwise.} \end{cases} \tag{3}$$

After all M elements are visited, the greedy scenario is set to $p_t^g = g_t^{(M)}$ and carried over to the next iteration.

4 Experiments

4.1 Experimental details

Datasets. Following prior work, we use the 1,000-image NIPS2017 competition dataset (Kurakin et al., 2018) for query-based and transfer attack evaluation, resizing each image from 299×299 to 224×224. For prompt-level transferability (Sec. 4.3), we use a separate ImageNet validation subset (Deng et al., 2009) of 10 images per class (10,000 images in total). Attack success rate (ASR) is measured only on clean-correct samples (those correctly classified before the attack), and the average number of queries (Avg. Q) only over successful attacks.

Victim models. We evaluate ASA against diverse ImageNet classifiers under query-based black-box attacks: CNNs (ResNet-50 (RN-50) (He et al., 2016), ConvNeXt (Liu et al., 2022)), Transformers (Swin-Base (Liu et al., 2021), DeiT-Base (Touvron et al., 2021)), and state-space models (Vision Mamba-Small (ViM) (Zhu et al., 2024), MambaVision-Base (MV) (Hatamizadeh & Kautz, 2025)), VLMs (CLIP (Radford et al., 2021), and SigLIP2 (Tschannen et al., 2025)). For robustness, we include adversarially trained Inception-v3 (Adv-Inc) and ensemble adversarially trained Inception-ResNet-v2 (Adv-Res) (Tramèr et al., 2018). For transferability, we additionally use Inception-v3 (Inc-v3) (Szegedy et al., 2016), Wide-ResNet-50 (WRN-50) (Zagoruyko & Komodakis, 2016), ViT-Base (ViT-B) (Dosovitskiy et al., 2021).

Table 1: Direct query-based attack results. Each evaluated model is queried during the attack. Higher ASR (%) and lower Avg. Q indicate better performance. Best results are shown in **red bold**.

Methods	CNNs				Transformers				SSMs				Robust Models				VLMs			
	RN-50		ConvNeXt		Swin		DeiT		ViM		MambaVision		Adv-Res		Adv-Inc		CLIP		SigLIP2	
	ASR	Avg Q	ASR	Avg Q	ASR	Avg Q	ASR	Avg Q	ASR	Avg Q	ASR	Avg Q	ASR	Avg Q	ASR	Avg Q	ASR	Avg Q	ASR	Avg Q
AdvFlow	18.56	18.28	5.83	32.30	4.08	24.00	5.93	27.54	8.14	29.93	3.42	29.33	17.47	23.29	22.55	20.45	22.49	23.54	13.15	19.47
DifAttack	46.67	36.99	17.18	46.43	23.93	45.81	25.42	48.42	26.45	43.44	17.22	51.99	27.52	36.83	36.20	30.47	45.76	37.84	35.04	39.02
CGAttack	43.99	18.11	22.19	16.08	19.54	26.06	16.31	35.32	17.67	19.13	13.38	25.90	31.10	30.07	51.82	16.92	32.39	26.65	27.47	24.68
MCGAttack	82.62	**4.16**	49.39	29.96	23.82	30.78	26.59	35.20	32.12	29.48	20.85	34.67	44.30	24.24	52.78	22.10	57.20	15.96	43.89	18.35
ASA (ours)	**94.10**	9.10	**87.22**	**13.76**	**89.34**	**12.76**	**88.03**	**11.73**	**90.90**	**11.23**	**77.90**	**14.48**	**95.87**	**7.75**	**97.69**	**6.11**	**84.19**	**11.20**	**67.17**	**12.64**

Implementation details. We use off-the-shelf Gemma 4 E4B as the MLLM for scenario generation and naturalness evaluation (`thinking=false`, `max_new_tokens`=4096, `do_sample=false`), and FLUX.2 [klein] 9B-KV for text-guided image editing (1024×1024 inputs, seed 123). Search runs for up to 50 iterations or 100 victim-model queries, whichever comes first, with the MLLM generating one scenario per strategy per iteration. Detailed prompts are in App. B.

4.2 Comparison

4.2.1 Query-based attacks

We compare ASA with state-of-the-art prior-assisted query-based generative attacks (AdvFlow, DifAttack, CGAttack, and MCGAttack), reporting the Attack Success Rate (ASR) and the average number of victim-model queries (Avg. Q) under the same query budget. To avoid ASR inflation from prompt leakage, predictions of the original class or any explicitly leaked ImageNet category are counted as failures; leakage occurs in only 0.9–7.2% of ASA-generated scenarios (App. D). As shown in Tab. 1, ASA achieves the highest ASR on all ten victim models and the lowest Avg. Q on nine. The margin is largest on modern architectures where prior attacks struggle (e.g., 89.34% vs. 23.93% on Swin), showing that scenario-level search is architecture-agnostic and extends even to small multimodal LLMs (App. J). Fig. 3 further compares mean ASR across query budgets on the same sample set for all methods. With only 20 queries, ASA reaches 75.79% ASR, surpassing every baseline even at 100 queries (MCGAttack: 43.36%), and gains just 1.01 points beyond 60 queries. The baselines are ℓ_p-bounded ($\epsilon = 12/255$) whereas ASA is unrestricted, so part of this gap reflects the larger search space; within that space, search itself adds 37.7–53.6 pp over $\text{ASA}_{\text{w/o } S,G}$ (Sec. 4.7).

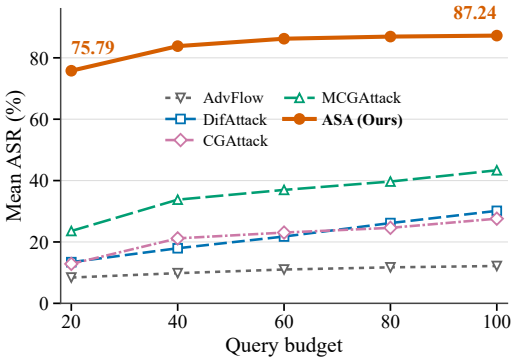


Figure 3: Mean ASR (%) under different budget.

4.2.2 Image-level transferability

Tab. 2 compares ASA with prior-assisted query-based generative attacks (AdvFlow, DifAttack, CGAttack, and MCGAttack) and representative transfer-based attacks, including CF, NCF, ACA, and DiffAttack. Adversarial images are crafted on one of three source models (Swin, ConvNeXt, and ViM) and evaluated on 13 models spanning CNNs, Transformers, SSMs, adversarially trained models, and VLMs, to assess whether they remain effective across heterogeneous architectures.

Among query-based attacks, ASA achieves the strongest transferability for every source model. ASA outperforms all prior-assisted query-based baselines on every victim model and improves the average ASR over the strongest baseline, MCGAttack, by 26.4–29.9 percentage points (e.g., 49.2% vs. 20.1% with Swin as the source). These gains hold consistently across all model families, suggesting that scenario-level adversarial generation does not merely overfit to the queried model but captures transformations that generalize across architectures.

ASA is also competitive with transfer-based attacks that rely on white-box access to the source model. Using only query feedback, ASA ranks third in average ASR for all three source models, behind only ACA and DiffAttack. It trails the second-best method by 4.6 and 4.9 points on Swin and ViM, respectively, and by 10.2 points on ConvNeXt. On robust victim models, ACA achieves the

Table 2: ASR (%) across three source models. * marks white-box results; gray rows indicate attacks querying the source model. **Bold** marks the highest ASR in each column, separately for gray and non-gray rows within each source-model block. Averages exclude the source model.

Source Model	Attack	CNNs				Transformers			SSMs		Robust Models		VLMs		Avg. ASR
		RN-50	WRN-50	Inc-v3	ConvNeXt	ViT	Swin	DeiT	ViM	MV	Adv-Inc	Adv-Res	CLIP	SigLIP2	
Swin	CF	45.5	41.8	30.2	13.5	27.3	60.6*	18.2	22.6	10.2	37.8	29.5	18.25	11.99	25.6
	NCF	45.7	40.7	35.1	28.1	31.1	83.6*	27.1	27.8	21.1	30.7	28.0	16.99	12.95	28.8
	ACA	**60.3**	**59.0**	**62.3**	57.0	**58.6**	80.8*	59.2	**60.9**	58.0	**65.9**	**62.2**	45.30	**44.11**	**57.7**
	DiffAttack	54.1	56.5	46.8	**67.2**	58.1	**89.8***	**63.9**	56.8	**65.5**	44.6	39.7	**49.10**	43.42	53.8
	AdvFlow	6.1	7.0	10.5	1.2	2.5	4.08	1.8	1.9	0.7	8.7	7.9	9.77	5.70	5.3
	DifAttack	12.8	10.2	14.1	4.5	4.6	23.9	4.5	5.3	2.2	16.7	9.6	9.90	6.87	8.4
	CGAttack	17.6	16.4	16.5	10.3	4.8	19.5	4.3	6.5	5.8	35.8	13.5	10.15	10.01	12.6
	MCGAttack	37.7	39.1	42.2	12.5	9.8	23.8	6.5	9.3	5.0	27.9	16.9	20.95	13.85	20.1
	ASA (ours)	**56.1**	**56.0**	**57.3**	**47.1**	**51.9**	**89.3**	**50.7**	**53.1**	**36.7**	**62.3**	**54.6**	**37.79**	**27.24**	**49.2**
ConvNeXt	CF	42.5	42.0	26.5	49.8*	25.6	17.4	18.5	21.2	12.3	38.6	29.4	19.02	12.69	25.5
	NCF	51.3	49.0	44.4	66.2*	38.3	34.6	30.6	32.5	23.9	36.8	32.4	20.33	15.40	34.1
	ACA	63.8	64.9	65.3	80.4*	60.8	66.2	62.5	60.9	60.3	**66.1**	**63.6**	45.95	46.09	60.5
	DiffAttack	**78.9**	**81.4**	**67.1**	**97.6***	**72.5**	**88.2**	**78.4**	**73.1**	**83.8**	58.9	59.4	**65.68**	**59.49**	**72.2**
	AdvFlow	7.5	8.0	12.4	5.83	2.3	1.3	2.2	2.4	0.7	10.6	7.7	8.61	5.47	5.8
	DifAttack	14.3	11.3	15.7	17.2	4.6	5.0	4.0	4.9	3.0	18.0	10.4	11.05	8.03	9.2
	CGAttack	21.2	18.7	16.7	22.2	4.3	8.3	4.7	8.4	6.2	34.0	14.3	12.60	11.06	13.4
	MCGAttack	41.4	42.4	43.0	49.4	9.5	6.0	5.8	10.0	5.0	27.9	18.0	20.82	15.37	20.4
	ASA (ours)	**59.3**	**56.9**	**57.3**	**87.2**	**53.1**	**50.7**	**48.1**	**52.4**	**40.5**	**63.3**	**56.8**	**38.82**	**26.89**	**50.3**
ViM	CF	39.7	35.3	24.1	9.2	18.6	10.2	12.9	69.4*	6.8	35.2	26.6	16.71	11.99	20.6
	NCF	46.9	42.2	37.9	25.0	36.4	25.7	30.5	77.7*	19.7	32.3	29.4	19.69	15.52	30.1
	ACA	**59.7**	**59.0**	**62.5**	**53.8**	60.4	**57.1**	60.8	84.5*	52.2	**66.9**	**61.2**	45.56	**42.36**	**56.8**
	DiffAttack	51.0	51.7	51.7	49.7	**62.7**	**57.1**	**63.4**	**88.9***	**52.7**	44.5	40.4	**46.92**	37.83	50.8
	AdvFlow	7.0	6.7	9.9	1.3	3.2	0.8	2.0	8.14	0.9	9.4	7.4	8.61	5.36	5.2
	DifAttack	13.5	9.7	14.1	2.3	4.2	4.0	4.6	26.5	2.0	18.2	7.9	8.87	7.22	8.0
	CGAttack	21.4	17.9	16.8	9.5	3.2	6.7	3.8	17.7	5.0	32.9	10.0	9.90	10.71	12.3
	MCGAttack	38.2	38.6	41.9	13.4	10.2	5.5	5.4	32.1	3.8	27.9	16.3	19.92	13.39	19.5
	ASA (ours)	**53.7**	**52.0**	**57.1**	**39.1**	**48.7**	**43.8**	**46.9**	**90.9**	**30.9**	**61.6**	**55.4**	**35.48**	**26.08**	**45.9**

highest ASR, while ASA ranks second in five of the six settings and outperforms DiffAttack by up to 17.7 points. Overall, ASA combines the practicality of query-based attacks with transferability approaching that of strong transfer-based methods, making it effective across diverse model families.

4.3 Prompt-level transferability

To test whether discovered scenarios generalize beyond their source images, we take the three most frequent prompts in the ASA search history against RN-50, selected by frequency rather than ASR: *change lighting to dusk* (Dusk), *introduce light mist* (Light mist), and *coated in fine dust* (Fine dust). All are class-neutral, containing no class-specific content, and each is applied unchanged to all 10,000 prompt-transfer images (Sec. 4.1). As a control, we add *coated in fine grit* (Fine grit), the most frequent prompt that never succeeded during the search.

Fine grit still yields 10.44% ASR on average, a baseline for a benign surface edit; Dusk, Light mist, and Fine dust exceed it by 10.31, 24.18, and 16.70 percentage points (pp). On the same clean-correct images for all four prompts, each prompt is significantly higher than Fine grit on all 13 models (two-sided exact McNemar tests, Holm–Bonferroni correction over 39 comparisons), including the smallest gaps of 2.6–3.3 pp for Dusk on ConvNeXt, MambaVision, and Swin (adjusted $p < 10^{-12}$).

Table 3: Transfer ASR (%) of the three most frequent ASA prompts on RN-50.

Prompt	CNNs				Transformers			SSMs		Robust Models		VLMs		Avg. ASR
	RN-50	WRN-50	Inc-v3	ConvNeXt	ViT	Swin	DeiT	ViM	MV	Adv-Inc	Adv-Res	CLIP	SigLIP2	
Dusk	23.94	20.93	25.46	9.59	16.33	12.16	15.06	16.39	9.97	51.16	36.73	19.85	12.17	20.75
Light mist	44.08	42.15	45.22	29.15	35.16	26.07	24.58	27.76	11.78	68.29	55.05	24.44	16.31	34.62
Fine dust	34.37	32.63	36.27	18.65	24.47	22.26	22.77	24.10	18.31	35.67	31.46	32.25	19.62	27.14
Fine grit (control)	12.59	11.49	14.78	6.97	8.18	8.85	7.77	8.85	7.00	15.87	12.62	13.17	7.60	10.44

Light mist is the most effective prompt on average and on 10 of the 13 models, whereas Fine dust leads on MambaVision, CLIP, and SigLIP2. Compared with Inc-v3 and RN-50, Adv-Inc and Adv-Res are more vulnerable to Dusk (+25.7 and +12.8 pp) and Light mist (+23.1 and +11.0 pp) but not to Fine dust (−0.6 and −2.9 pp) or Fine grit (+1.1 and +0.0 pp). Their added vulnerability thus lies in global illumination and atmosphere rather than surface texture, although image subsets and, for Adv-Res, architectures differ. Overall, recurring class-neutral scenarios act as reusable edits, raising

the average error rate to 2.0–3.3 times the control rate across all 1,000 classes and all evaluated model families.

4.4 Sensitivity to Gaussian noise

ASR counts only edits that flip the prediction, but an edit may also leave the image more fragile without flipping it. To measure this, we take, for each model and prompt, the images correctly classified both before and after editing, and add the same Gaussian noise $\mathcal{N}(0, \sigma^2)$ to each clean–edited pair so that the two differ only in the edit ($\sigma \in \{0.005, 0.010, 0.015, 0.020\}$ in RGB range $[0, 1]$, with clipping). For each model, we compute the error rate of the noisy edited images minus that of the noisy clean images, and report the unweighted mean of this difference over models in percentage points (pp). We test each model and noise level with a two-sided exact McNemar test and apply Holm–Bonferroni correction over all 208 tests (4 prompts $\times$ 13 models $\times$ 4 noise levels, $p < 0.05$). We evaluate all 13 models in Table 3. For contrast, we again include the control prompt Fine grit (Sec. 4.3).

Table 4: Mean paired error-rate increase (pp) of edited over clean under Gaussian noise, averaged equally across 13 models.

	σ			
Prompt	.005	.010	.015	.020
Dusk	2.23	4.24	6.08	7.86
Light mist	7.11	13.02	18.23	22.90
Fine dust	1.88	3.94	5.67	6.91
Fine grit	0.22	0.60	0.78	1.02

As shown in Table 4, all three recurring prompts significantly increase the error rate on all 13 models, including both VLMs, at every noise level, and the increase grows with σ. Light mist has the largest effect, reaching 22.90 pp at $\sigma = 0.020$, followed by Dusk (7.86 pp) and Fine dust (6.91 pp), whereas Fine grit adds at most 1.02 pp. The prompts found by ASA therefore leave images more fragile even when they do not flip the prediction, so ASR understates their effect. The contrast with Fine grit, a similar benign surface edit, indicates that this fragility depends on the scenario applied.

4.5 Perceptual Similarity and Object Identity Preservation

We measure semantic consistency with DINOv2 and DINOv3 similarity, subject identity with ID-Sim, and distributional similarity with FID in Table 5 (Oquab et al., 2024; Siméoni et al., 2025; Chae et al., 2026; Heusel et al., 2017). Restricted attacks retain high similarity (DINOv2 0.937–0.980, ID-Sim 0.061–0.147) due to their bounded perturbations, and color-based attacks behave similarly (DINOv2 0.953, ID-Sim 0.141–0.156) since they only alter color. Generative unrestricted attacks, in contrast, permit broad appearance changes, making preservation of the original subject the key criterion. Within this group, ASA achieves the best results on all paired metrics. Its DINOv2 and DINOv3 similarities (0.787 / 0.718) exceed those of ACA (0.520 / 0.422) and DiffAttack (0.713 / 0.569), and its ID-Sim of 0.313 is 46.4% and 24.6% lower than ACA and DiffAttack, respectively. Its FID (60.87) also improves on ACA (66.27) and remains close to DiffAttack (58.79). Overall, ASA best preserves the semantics and identity of the original subject among generative unrestricted attacks.

Table 5: Perceptual and identity preservation. All metrics are averaged over Swin, ConvNeXt, and ViM. **Bold**: best among generative methods.

Attack	DINOv2↑	DINOv3↑	ID-Sim↓	FID↓
Restricted (ℓ_p-bounded)				
AdvFlow	0.980	0.955	0.061	22.68
DifAttack	0.960	0.893	0.120	37.17
CGAttack	0.957	0.902	0.094	83.59
MCGAttack	0.937	0.890	0.147	76.81
Color-based unrestricted				
CF	0.953	0.931	0.141	46.78
NCF	0.953	0.912	0.156	44.10
Generative unrestricted				
ACA	0.520	0.422	0.584	66.27
DiffAttack	0.713	0.569	0.415	**58.79**
ASA (Ours)	**0.787**	**0.718**	**0.313**	60.87

4.6 Human Evaluation

Automatic similarity metrics and the MLLM verifier cannot confirm that humans still recognize the original object after editing. We therefore conducted a human study on successful ASA examples against ResNet-50, drawn from its clean-correct subset. Fifty-three non-author participants each judged 20 randomly sampled original–edited pairs shown side by side, for 1,060 judgments in total. For each pair, participants answered: "Does the main object in the left image still appear to be the

Table 6: Sensitivity to the scenario MLLM and image editor against RN-50. **Bold**: best.

Method	*(a) Scenario MLLM*			*(b) Image editor*		
	Gemma-4 E4B	Qwen3-VL 4B (Bai et al., 2025)	InternVL3.5 4B (Wang et al., 2025)	FLUX.2 Klein 9B	Qwen Image Edit (Wu et al., 2025)	FireRed 1.0 (Qiao et al., 2026)
$\text{ASA}_{\text{w/o } S,G}$	56.40	38.92	32.58	56.40	70.11	29.35
ASA	**94.10**	**79.46**	**76.24**	**94.10**	**98.39**	**74.19**

same object in the right image?" This question assesses the perceived consistency of the main object between the original and edited images. Overall, 75.4% of responses judged the main object to be the same across the original and edited images (participant-level bootstrap 95% CI: 70.4–79.7%). The entire confidence interval lies above the approximately 70% level regarded in prior work as evidence of semantic preservation (Kollovieh et al., 2024). These results support the preservation of perceived object identity under paired comparison. Together with the ID-Sim results in Sec. 4.5, these findings provide complementary evidence that most successful ASA edits preserve the identity of the original object.

4.7 Ablation Study

Fig. 4 isolates the two main components of ASA: scenario generation (S) and greedy selection (G). With neither component ($\text{ASA}_{\text{w/o } S,G}$), ASR drops sharply, showing that feedback alone is insufficient. Adding scenario generation ($\text{ASA}_{\text{w/o } G}$) raises ASR by 30.0–40.1 percentage points (pp). Adding greedy selection yields a further 7.7–16.6 pp while reducing both victim-model and MLLM queries. The two components are thus complementary: scenario generation supplies strong candidates, and greedy selection improves both effectiveness and query efficiency.

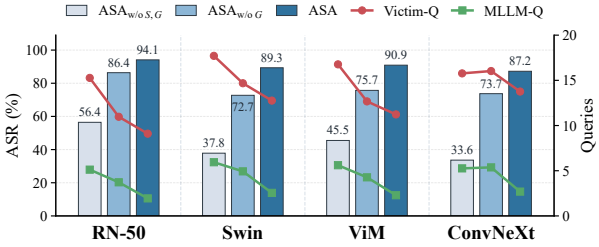


Figure 4: Ablation study of ASA components. Bars show ASR, while lines show victim-model and MLLM query counts across source models.

Sensitivity to the scenario MLLM and image editor. To verify that ASA's gains come from its search rather than from a particular model, we replace the MLLM or the editor one at a time (Tab. 6). In every configuration, full ASA improves over $\text{ASA}_{\text{w/o } S,G}$ by 28.3–44.8 pp, and the gain grows as $\text{ASA}_{\text{w/o } S,G}$ weakens. With Qwen3-VL 4B, InternVL3.5 4B, or the FireRed 1.0 editor, full ASA more than doubles its ASR (e.g., 32.58%→76.24% with InternVL3.5), showing that scenario generation and greedy selection compensate for weaker MLLMs and editors. A stronger editor, in turn, directly helps: with Qwen Image Edit, ASA reaches 98.39%, above the default FLUX.2 [klein] configuration (94.10%). ASA is therefore not tied to a specific MLLM or editor and can benefit from future advances in generative editing.

5 Conclusion

We proposed **Adversarial Scenario Attack (ASA)**, a query-based black-box framework that generates NAEs by searching over natural-language scenarios. Rather than relying on surrogate gradients, attack-specific learned priors, or additional generative-model training, ASA leverages text-guided image editing to explore realistic transformations such as changes in background, weather, lighting, and material/color. By combining MLLM-guided scenario generation with a Greedy Explorer that composes attack-improving scenarios, ASA discovers edits that fool diverse vision models. Our experiments show that ASA achieves high attack success rates with low query cost and competitive image-level transferability. Moreover, class-neutral scenarios discovered during search remain effective when applied unchanged to images from all 1,000 ImageNet classes, substantially outperforming a control prompt. These results suggest that the evaluated models share vulnerabilities to specific natural changes in illumination and atmosphere, highlighting scenario-level search as a practical direction for black-box robustness evaluation beyond conventional perturbation-based attacks.

AI USE STATEMENT

In this work, we used generative AI tools to aid and polish writing, assist with literature retrieval and discovery, help draft and revise parts of the manuscript, assist with code generation, and present experimental results in figures and tables. Beyond these supporting uses, generative AI models are a core component of our method: ASA is a framework built on an MLLM (Gemma 4 E4B), which generates attack scenarios and evaluates naturalness, and a text-guided image editing model (FLUX.2 [klein] 9B-KV), which applies these scenarios to input images (Sec. 3.2). Both models are used off-the-shelf without fine-tuning. We have not used generative AI tools for any other tasks requiring disclosure. We have reviewed all AI-assisted work. In particular, we manually verified the AI-assisted text and references, tested and checked the generated code for correctness, and confirmed that all figures and tables accurately reflect the underlying experimental results. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our human study (Sec. 4.6) was a minimal-risk perceptual judgment task on publicly available ImageNet images. It falls within the category of research eligible for exemption from ethics review under the applicable national regulations, as it involved no intervention and collected no personally identifiable information, but no formal exemption determination was obtained. All 53 participants were adult volunteers who were informed of the study's purpose and task and gave consent; participation was unpaid.

REPRODUCIBILITY STATEMENT

We provide the algorithmic details of our method in the main paper and report the datasets, model configurations, and evaluation protocols used in our experiments. Additional implementation details and hyperparameters required to reproduce our results are provided. We also provide the source code as part of the supplementary materials.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *CoRR*, abs/2511.21631, 2025.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion Attacks against Machine Learning at Test Time. In *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, 2013.

BlackForestLabs. Flux.2: [klein]. `https://huggingface.co/black-forest-labs/FLUX.2-klein-9b-kv`, 2025.

Nicholas Carlini and David A. Wagner. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 2017.

Julia Chae, Nicholas I. Kolkin, Jui-Hsien Wang, Richard Zhang, Sara Beery, and Cusuh Ham. Id-sim: An identity-focused similarity metric. *Conference on Computer Vision and Pattern Recognition*, 2026.

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.

Xinquan Chen, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Advdiffuser: Natural adversarial example synthesis with diffusion models. In *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.

Max Collins, Jordan Vice, Tim French, and Ajmal Mian. Natadiff: Adversarial boundary guidance for natural adversarial diffusion. In *International Conference on Learning Representations*, 2026.

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

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

Hadi Mohaghegh Dolatabadi, Sarah M. Erfani, and Christopher Leckie. Advflow: Inconspicuous black-box adversarial attacks using normalizing flows. In *Advances in Neural Information Processing Systems*, 2020.

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.

Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Carbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, Clément Farabet, Johan Ferret, Filippo Galgani, Sertan Girgin, Jean-Bastien Grill, Maarten Grootendorst, Jiaxian Guo, Cassidy Hardin, Yanzhang He, Steven M. Hernandez, Omri Homburger, Léonard Hussenot, Ju-yeong Ji, Armand Joulin, Aishwarya Kamath, Parnian Kassraie, Olivier Lacombe, Preethi Lahoti, Gaël Liu, Gus Martins, Luciano Martins, Tatiana Matejovicova, Ramona Merhej, Nikola Momchev, Sneha Mondal, Ryan Mullins, Sindhu Raghuram Panyam, Shreya Pathak, Sarah Perrin, André Susano Pinto, Etienne Pot, Angéline Pouget, Alexandre Ramé, Sabela Ramos, Douglas Reid, David Rim, Morgane Rivière, Karsten Roth, Louis Rouillard, Omar Sanseviero, Pier Giuseppe Sessa, Shane Settle, Danila Sinopalnikov, Sara Smoot, Piotr Stanczyk, Andreas Steiner, Lawrence Stewart, Ilya O. Tolstikhin, Michael Tschannen, Anton Tsitsulin, Nino Vieillard, Renjie Wu, Pingmei Xu, Haichuan Yang, Edouard Yvinec, Biao Zhang, Li Zhang, Joe Zou, Nicolas Aagnes, Abdelrahman Abdelhamed, Jakub Adámek, Shivani Agrawal, Shubham Agrawal, Ibrahim Alabdulmohsin, Jean-Baptiste Alayrac, Uri Alon, Chandramouli Amarnath, Ankesh Anand, Chrysovalantis Anastasiou, Setareh Ariafar, François-Xavier Aubet, Kyriakos Axiotis, Federico Barbero, Joelle K. Barral, Alexei Bendebury, Urs Bergmann, Stanley Bileschi, Kat Black, Mathieu Blondel, Sebastian Borgeaud, and Arthur Brazinskas. Gemma 4 technical report. *CoRR*, abs/2607.02770, 2026.

Yan Feng, Baoyuan Wu, Yanbo Fan, Li Liu, Zhifeng Li, and Shu-Tao Xia. Boosting black-box attack with partially transferred conditional adversarial distribution. In *Conference on Computer Vision and Pattern Recognition*, 2022.

S. Alireza Golestaneh, Saba Dadsetan, and Kris M. Kitani. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Winter Conference on Applications of Computer Vision*, 2022.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015.

Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Conference on Computer Vision and Pattern Recognition*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Conference on Computer Vision and Pattern Recognition*, 2021.

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*, 2017.

Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. Adversarial texture for fooling person detectors in the physical world. In *Conference on Computer Vision and Pattern Recognition*, 2022.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *International Conference on Computer Vision*, 2021.

Marcel Kollovieh, Lukas Gosch, Marten Lienen, Yan Scholten, Leo Schwinn, and Stephan Günnemann. Assessing robustness via score-based adversarial image generation. *Trans. Mach. Learn. Res.*, 2024.

Alexey Kurakin, Ian J. Goodfellow, Samy Bengio, Yinpeng Dong, Fangzhou Liao, Ming Liang, Tianyu Pang, Jun Zhu, Xiaolin Hu, Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, Alan L. Yuille, Sangxia Huang, Yao Zhao, Yuzhe Zhao, Zhonglin Han, Junjiajia Long, Yerkebulan Berdibekov, Takuya Akiba, Seiya Tokui, and Motoki Abe. Adversarial attacks and defences competition. *CoRR*, abs/1804.00097, 2018.

Jiawei Lian, Shaohui Mei, Xiaofei Wang, Yi Wang, Lefan Wang, Yingjie Lu, Mingyang Ma, and Lap-Pui Chau. Attack anything: Blind dnns via universal background adversarial attack. *CoRR*, 2024.

Jun Liu, Jiantao Zhou, Jiandian Zeng, and Jinyu Tian. Difattack: Query-efficient black-box adversarial attack via disentangled feature space. In *AAAI Conference on Artificial Intelligence*, 2024.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, 2021.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Conference on Computer Vision and Pattern Recognition*, 2022.

Harshitha Machiraju and Vineeth N. Balasubramanian. A little fog for a large turn. In *Winter Conference on Applications of Computer Vision*, 2020.

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.

Alberto Marchisio, Giovanni Caramia, Maurizio Martina, and Muhammad Shafique. fakeweather: Adversarial attacks for deep neural networks emulating weather conditions on the camera lens of autonomous systems. In *International Joint Conference on Neural Networks*, 2022.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, Shuang Sun, Wei Zhu, Xu Tang, Yao Hu, Yibo Chen, Yuhao Huang, Yuxuan Duan, Zhiyi Chen, and Ziyuan Guo. Firered-image-edit-1.0 technical report. *CoRR*, abs/2602.13344, 2026.

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

Ali Shahin Shamsabadi, Ricardo Sánchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *Conference on Computer Vision and Pattern Recognition*, 2020.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3. *CoRR*, abs/2508.10104, 2025.

Dongsu Song, Daehwa Ko, and JayHoon Jung. Amnesia as a Catalyst for Enhancing Black Box Pixel Attacks in Image Classification and Object Detection. In *Advances in Neural Information Processing Systems*, 2024.

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

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, 2018.

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 *Conference on Computer Vision and Pattern Recognition*, 2020.

Naufal Suryanto, Yongsu Kim, Hyoeun Kang, Harashta Tatimma Larasati, Youngyeo Yun, Thi-Thu-Huong Le, Hunmin Yang, Se-Yoon Oh, and Howon Kim. Dta: Physical camouflage attacks using differentiable transformation network. In *Conference on Computer Vision and Pattern Recognition*, 2022.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Conference on Computer Vision and Pattern Recognition*, 2016.

Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27:3998–4011, 2018.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, 2021.

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.

Michael Tschannen, Alexey A. Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier J. Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *CoRR*, abs/2502.14786, 2025.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, JingJing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu,

Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025.

Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *CoRR*, abs/2508.02324, 2025.

Fei Yin, Yong Zhang, Baoyuan Wu, Yan Feng, Jingyi Zhang, Yanbo Fan, and Yujiu Yang. Generalizable black-box adversarial attack with meta learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(3):1804–1818, 2024.

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

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations*, 2018.

Lianghui Zhu, Bencheng Liao, Qian Zhang, Xinlong Wang, Wenyu Liu, and Xinggang Wang. Vision mamba: Efficient visual representation learning with bidirectional state space model. In *International Conference on Machine Learning*, 2024.

A MORE VISUALIZATION

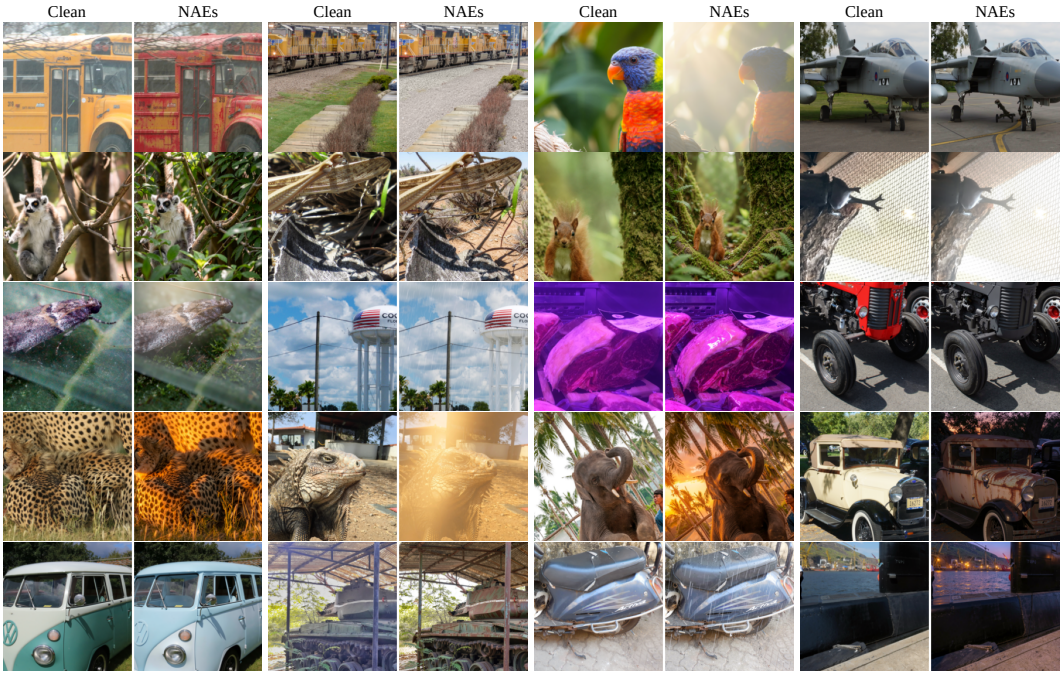


Figure 5: Examples of clean images and the corresponding natural adversarial examples (NAEs) generated by ASA. The generated NAEs preserve the overall semantic content of the original images while introducing natural scenario-level changes that lead to misclassification.

We provide additional qualitative examples of natural adversarial examples (NAEs) generated by ASA in Fig. 5. These examples further illustrate that ASA modifies images through natural scenario-level changes, such as variations in illumination, color tone, background context, and local appearance, rather than relying on imperceptible pixel-level perturbations. Despite these changes, the generated NAEs visually preserve the semantic content and object identity of the original clean images.

We also visualize same-class prompt transfer in Fig. 6. In this setting, the adversarial scenario is discovered from other images of the same ImageNet class and is then applied, without re-optimization, to unseen source images from that class. The examples show that a single class-level scenario can often induce natural adversarial examples across different images while preserving realistic appearance and the main object identity.

B DETAIL PROMPTS

This section reports the detailed adversarial scenario generation prompts for candidate generation and naturalness evaluation. We instantiate the generation prompt separately for each sample and attack step. The placeholder `<class>` is replaced by the ImageNet ground-truth class name of the input image, while `<scene>` denotes the editable background/context slot. The generation prompt includes the current visual feedback, the best attack objective observed so far, previous candidate feedback, and a fixed strategy instruction that balances candidates across background, atmosphere, and material/color edits. The naturalness evaluation prompt is used to judge whether the edited image remains semantically plausible while preserving the main object. In the **ablation study** without strategy-specific scenario generation ($\text{ASA}_{\text{w/o}S,G}$), we disable the fixed strategy instruction by leaving {`strategy_instruction`} empty while keeping all other parts of the prompt unchanged.

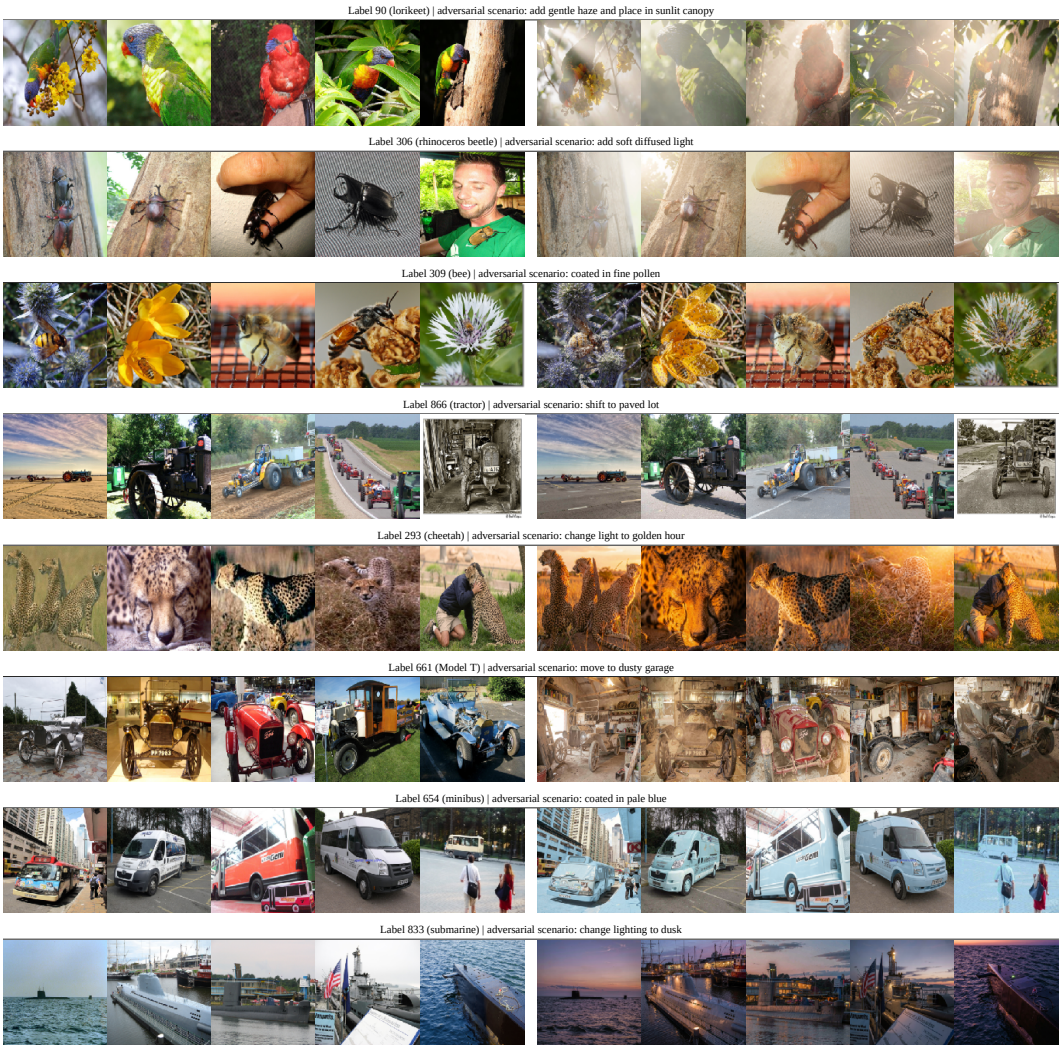


Figure 6: Same-class prompt transfer examples. For each ImageNet class, the five images on the left are unseen source images, and the five images on the right are the corresponding NAEs generated by applying a class-level adversarial scenario discovered from other images of the same class.

Adversarial Scenario Generation Template

```
You are a visual consistency tester. Your goal is to find 'Trojan' editing commands for a
text-to-image model. These commands must be so contextually perfect and benign that they
seamlessly camouflage into the original scene, yet act as strong latent attractors that
subtly disrupt the generated output.

Adversarial visual feedback:
- The reference image is attached and represents the most recent generated result.
Step: {step_idx}
Prompt template with slot marker: "{marked_prompt}"
Current {slot_marker} value: "{current_value}"
{anchor_label}: "{slot_topic}"
Current best objective score: {best_objective_plus_offset}.

Feedback from previous steps:
{previous_feedback}

Task:
- Generate exactly {total_candidate_count} unique candidate replacement instructions for
{slot_marker}.
- Prefer short, concrete, highly visual commands.
```

```
- CRITICAL: ALL candidates MUST be formulated as direct editing instructions starting
with action verbs, such as 'turn the scene into', 'add', 'change the lighting to', or
'apply a'.
- Do not output standalone nouns, adjectives, or simple prepositions.
- Favor candidates that request subtle, plausible modifications (e.g., 'turn the scene
into a misty morning', 'add soft shadows', 'change the material to brushed metal') rather
than overt or destructive visual conflicts.
- Do NOT use the word 'overlay' and words related to 'cybernetic', 'neon' and strong
'fog'.
- We allow verb-driven transformations of background, color, texture, and weather, but
constrain them to attribute-level edits. Do Not use Prompts that instantiate a named
object, location, scene category, or class-specific material. This preserves the attack
mechanism while preventing explicit class-concept injection.
- The instructions MUST make logical and semantic sense for the original context. They
should appear as perfectly normal user requests to a human observer.
{naturalness_feedback_instruction}
{strategy_instruction}
- Do NOT repeat any exact candidate already listed in the feedback block.
- Every candidate must be semantically related to the target class "<class>" (for this
example: "{target_class_name}").
- Each candidate must be a concise English phrase of 1~5 words.
- Candidates should stay visually meaningful in relation to the current prompt context.
- Do not write explanations, numbering, markdown, or sentences.
- Output ONLY valid JSON with the schema {"strategies": [{"name": "background_shift",
"candidates": ["..."]}, {"name": "weather_atmosphere", "candidates": ["..."]}, {"name":
"texture_material", "candidates": ["..."]}]}.
```

Strategy Instruction

```
For the first attack step:

- Since this is the initial step (no previous feedback), establish a baseline by equally
distributing your candidates across THREE distinct attack vectors:
  * Strategy 1 (Background Shift): Commands that entirely replace or subtly alter the
  surrounding environment (e.g., 'move the subject to a dusty warehouse', 'place the
  scene in a dense forest').
  * Strategy 2 (Weather & Atmosphere): Commands that introduce meteorological changes or
  atmospheric conditions affecting global lighting (e.g., 'add a thick morning fog',
  'turn the weather into a torrential downpour').
  * Strategy 3 (Material & Color Reskinning): Alter ONLY the surface material, texture,
  or color of the target object. The object's structural shape and core identity MUST
  remain 100% intact. Apply contradictory or highly detailed surface wraps. CRITICAL
  SYNTAX: You MUST start your candidate with a past participle (e.g., 'coated in',
  'wrapped in', 'painted with', 'covered in'). Do NOT start with a noun or a base verb.
  Do NOT alter the photo medium itself and do NOT change what the object fundamentally
  is.
- Propose diverse concepts within these three categories to discover which dimension the
target model is most sensitive to.
- Keep the candidates separated by strategy in the final JSON output.
- For EACH strategy, generate exactly 1 unique candidate.
- Keep the three strategy buckets balanced. Do not move candidates between buckets.

For later steps:

- Employ the following THREE distinct generation strategies to systematically test the
model's vulnerabilities. Use the feedback to refine your attacks:
  * Strategy 1 (Background Shift): Generate commands that modify the environment. If
  previous background shifts failed, try contrasting indoor/outdoor settings or changing
  the era/time period.
  * Strategy 2 (Weather & Atmosphere): Generate commands introducing atmospheric effects.
  Focus on conditions that logically conflict with the original lighting.
  * Strategy 3 (Material & Color Reskinning): Alter ONLY the surface material, texture,
  or color of the target object. The object's structural shape and core identity MUST
  remain 100% intact. Apply contradictory or highly detailed surface wraps.
  * Exploit Winners: Identify which of the three categories produced the best candidate
  in the feedback. Generate more compounded variants within that winning category.
  * Avoid Losers: Identify the underlying characteristics of the remaining ineffective
  candidates. STRICTLY AVOID generating new commands that share these traits or fall into
  the weakest category.
- Keep the candidates separated by strategy in the final JSON output.
- For EACH strategy, generate exactly 1 unique candidate.
- Keep the three strategy buckets balanced. Do not move candidates between buckets.
```

Naturalness Evaluation Prompt

```
You are judging whether an edited image is a natural and plausible result of a requested
change.
The attached image is a side-by-side comparison.
- Left: source image
- Right: edited image
Base description of the source image: "{base_description}"
The main_object is the target_class_name: "{target_class_name}".
Your primary goal is to verify if the MAIN OBJECT from the source image is preserved and
if the overall edit makes semantic sense.

CRITICAL EVALUATION CRITERIA:
1. Focus heavily on SEMANTIC PLAUSIBILITY and the structural integrity of the main
subject.
2. BE LENIENT with minor editing artifacts, slight lighting mismatches, or background
changes. Do NOT reject the image for small imperfections as long as the main object is
clearly recognizable and consistent with the base description.

If the main object is maintained and the edit is logically possible in the real world (or
the given context), set natural=true.
If the edited result is completely structurally broken, the main object is destroyed, or
it contains severe, immersion-breaking distortions, set natural=false and write one short
feedback sentence about the main visual problem.
Return ONLY valid JSON with the schema {"natural": true, "feedback": ""}.
```

Dynamic variables. Table 7 summarizes how the placeholders in the prompt templates are instantiated.

Table 7: Dynamic variables used in the adversarial scenario generation prompt templates.

Variable	Description
{step_idx}	Attack step.
{marked_prompt}	Base prompt with the editable slot marked, e.g., "a photo of goldfish in the <scene>."
{slot_marker}	Editable slot marker. For the background/context attack, this is <scene>.
{current_value}	Current value of the editable slot, initially "background."
{anchor_label}	Human-readable anchor label, set to "Scene/topic anchor."
{slot_topic}	Semantic topic used to guide candidate generation, initially "background."
{best_objective_plus_offset}	Current best objective displayed as objective +10.0.
{previous_feedback}	Ranked feedback from previously evaluated candidates and scores. At the first step, this is "No previous-step feedback is available yet."
{total_candidate_count}	Number of requested candidates. In experiments, this is 3.
{target_class_name}	ImageNet ground-truth class name of the current sample.
{base_description}	Base textual description of the source image used for naturalness evaluation.
{naturalness_feedback_instruction}	Additional warning inserted only after an unnatural successful edit is detected; otherwise empty.
{strategy_instruction}	Fixed strategy instruction selected according to whether the current step is the first step or a later step.
<class>	Placeholder in the base prompt, replaced by the ImageNet class name of each sample.

C BROADER IMPACTS

Positive impacts

- **More realistic robustness evaluation.** This work can support robustness evaluation for vision models under realistic semantic changes. ASA studies natural transformations such as background, weather, lighting, and material or color changes, which are closer to ordinary image variations than small norm-bounded perturbations. This may help researchers and practitioners identify visual conditions under which classifiers fail, especially in black-box settings where model internals are unavailable.
- **Interpretable stress testing and adversarial data augmentation.** The discovered adversarial scenarios are expressed as natural-language prompts, making them easier to inspect than pixel-level perturbations. These scenarios can provide useful examples for stress testing and may also be used as adversarial data augmentation to improve robustness against realistic semantic transformations. For example, prompts that consistently expose failures under changes in background, weather, lighting, or material appearance could be incorporated into future adversarial training pipelines.

Negative impacts

- **Responsible use.** ASA is intended for authorized robustness evaluation. Although transferable scenarios could inform unauthorized probing if misused, our experiments are limited to

research models and do not evaluate commercial deployed systems. Moreover, many edits can be identified by side-by-side human inspection against the original image, so practical risk depends on deployment-specific safeguards and review processes.

Mitigation and future work. We view defense against natural transformation attacks as an important direction for future work. One possible direction is to check semantic consistency before relying on a vision model's prediction. Since ASA targets cases where the main object is semantically preserved but the classifier prediction changes, an auxiliary verifier, object recognizer, or vision-language model could be used to check whether the predicted label is consistent with the visible main object, or whether the main object remains consistent with a reference image when such a reference is available.

Another possible future direction is prediction aggregation under label-preserving natural transformations. Given a final input image, a defender could apply multiple mild natural edits that are expected to preserve the main object, such as changes in lighting, background, weather, or color, and aggregate the classifier predictions across the edited variants. If the original prediction is unstable under these transformations, the system could abstain, request human review, or query an additional robust model.

These possible defenses require further study. Many deployed systems observe only the final image, auxiliary verifiers may also fail, generative edits can introduce artifacts or unintended semantic changes, and adaptive attackers could optimize against semantic checks or aggregation procedures. We therefore leave systematic evaluation of these defenses to future work.

D LIMITATION

As discussed in Sec. 4.2.1, a small fraction of prompts exhibit prompt leakage, where an ImageNet category name is explicitly included in the generated prompt. To avoid overestimating attack success in such cases, we adopt a conservative evaluation protocol: when a prompt contains an ImageNet category phrase, we regard the corresponding category as an additional valid ground-truth label. Therefore, if the attacked classifier predicts either the original label or the leaked category, we count the attack as a failure.

For this analysis, we exclude the category phrase `light` because it is semantically ambiguous in our setting. In ImageNet, `light` refers to the object category "lighter", whereas in our generated prompts it typically denotes illumination conditions, such as ambient light, natural light, or changes in lighting.

Table 8: Number of ASA source prompts containing ImageNet category phrases.

Metric	RN-50	Swin	ConvNeXt	Vim-Small	DeiT	MambaVision	Adv-Res	Adv-Inc	CLIP	SigLIP2
Total images	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000
Matched prompts	35	50	64	53	52	65	43	9	46	72
Unmatched prompts	965	950	936	947	948	935	957	991	954	928
Match rate (%)	3.50	5.00	6.40	5.30	5.20	6.50	4.30	0.90	4.60	7.20

As shown in Table 8, prompt leakage affects only a limited portion of the evaluation set, with approximately 0.9–7.2% of prompts containing leaked category phrases across all models. Table 9 further reports the most frequently leaked ImageNet categories.

Table 9: Most frequent ImageNet category phrase leakage in ASA prompts. Counts indicate the number of prompts in which a category phrase appears; a prompt is counted at most once per category.

ID	Category	Total	RN-50	Swin	ConvNeXt	Vim-Small	DeiT	MambaVision	Adv-Res	Adv-Inc	CLIP	SigLIP2
886	velvet	116	7	14	12	13	10	13	11	0	13	23
459	brass	79	6	12	12	8	14	8	5	1	8	5
707	patio	79	5	7	7	10	4	10	6	0	12	18
580	grand piano	50	3	6	7	5	6	10	5	1	2	5
625	library	37	2	3	9	4	4	6	2	0	2	5
826	stone wall	36	4	2	8	1	3	6	5	3	1	3
912	wool	26	3	2	4	2	4	7	2	2	0	0

Consequently, while prompt leakage is a limitation of the current prompt generation process, the number of affected samples remains small and does not materially affect the main conclusions of our study. Future work will further reduce this issue by incorporating category-aware prompt filtering and re-sampling prompts that contain explicit ImageNet category phrases before image editing. We also plan to develop context-aware filters that distinguish true category leakage from benign uses of category words as attributes or scene descriptors, such as material, lighting, or background terms.

E ORIGINALITY AND POSITIONING

ASA does not claim novelty from using natural language, an MLLM, or a text-guided editor alone. Its contribution is a source-image-conditioned, score-based black-box attack that treats executable natural transformation scenarios as the optimization variables and iteratively refines and composes them using victim-model feedback while seeking label-preserving edits. Unlike the compared diffusion-based attacks, which optimize latent variables or sampling trajectories through gradient signals, ASA searches executable natural-language scenarios using black-box victim-model confidence scores. More importantly, ASA discovers interpretable and reusable transformations corresponding to naturally occurring conditions rather than artificial perturbations. These transformations reveal a practical attack surface that black-box adversaries may exploit, while helping model developers identify natural vulnerabilities that should be prioritized for evaluation and defense.

Table 10: Comparison of ASA with diffusion-based adversarial attacks.

Aspect	**DiffAttack**	**ACA**	**NatADiff**	**ASA (Ours)**
Source-image editing	✓	✓	✗	✓
Optimization signal	Surrogate-model gradients	Surrogate-model gradients	Classifier-guidance gradients	**Victim-model confidence scores**
Search variable	Diffusion latent	Diffusion latent	Diffusion sampling trajectory	**Natural-language scenario**
Refinement	Gradient optimization	Gradient optimization	Augmented classifier guidance + time-travel sampling	**Feedback + greedy composition**
Reusable language-level transformation output	✗	✗	✗	✓

F ALGORITHM

Algorithm 1 summarizes the full procedure of the Adversarial Scenario Attack. Instead of perturbing pixels, the attack searches over natural-language editing prompts. Each prompt p is rendered by the generative editor G and scored by the margin loss $J(p) = \max_{c \neq y} F_c(G(x, p)) - F_y(G(x, p))$, where $J(p) > 0$ means the edited image is misclassified by the victim F. Every evaluation of J consumes one query from the budget Q. Each iteration consists of two stages.

Exploration. Conditioned on x, y, and the feedback history $\mathcal{H}$, the MLLM $\mathcal{M}$ proposes one scenario prompt p_t^r for each strategy $r \in \mathcal{R}$. Each prompt is used to edit the image and is then scored by J. The highest-scoring prompt is marked as the winner and the others as losers. This winner–loser contrast is added to $\mathcal{H}$, so that $\mathcal{M}$ can refine its proposals in the next iteration.

Composition. Scenario edits that are weak on their own can be complementary, so the algorithm then combines prompts greedily. Each scenario is represented as a set of strategy-keyed components. The candidate pool $\mathcal{C}_t$ contains the prompts of the current iteration and, for $t > 1$, the greedy scenario from the previous iteration, p_{t-1}^g. Starting from the best-scoring candidate g_t, the algorithm visits the remaining candidates in decreasing order of J; if p_{t-1}^g is among them, it is replaced by its components $\mathrm{Comp}(p_{t-1}^g)$. Each visited element u is merged into g_t with $\oplus_{\mathrm{and}}$: a component of u replaces the component of g_t with the same strategy key or is otherwise appended, and the components are joined

with a conjunction ("A and B"). The composite is kept only if it increases J, and the final g_t is carried over to the next iteration as p_t^g.

Termination. In both stages, the attack stops as soon as an edited image satisfies $J > 0$ and passes $\textsc{Verify}_{\mathcal{M}}$. In this check, the MLLM confirms that the main object of class y is preserved and that the edit is plausible in the real world. If the query budget is exhausted first, the attack returns failure.

Algorithm 1 Adversarial Scenario Attack

Require: image x, label y, victim F, generative editor G, MLLM $\mathcal{M}$, budget Q

$q \leftarrow 0$, $\mathcal{H} \leftarrow \emptyset$
Each computation of $J(\cdot)$ consumes one query.
$\oplus_{\text{and}}$ *and* $\text{Comp}(\cdot)$ *operate on strategy-keyed components (Sec. 3.3).*
for $t = 1, 2, \ldots$ **while** $q < Q$ **do**
  Generate $\{p_t^r\}_{r \in \mathcal{R}}$ using $\mathcal{M}(x, y, \mathcal{H})$
  for each $r \in \mathcal{R}$ **do**
    $x_t^r \leftarrow G(x, p_t^r)$
    $J(p_t^r) \leftarrow \max_{c \neq y} F_c(x_t^r) - F_y(x_t^r)$
    $q \leftarrow q + 1$
    if $J(p_t^r) > 0$ **and** $\textsc{Verify}_{\mathcal{M}}(x, x_t^r, y)$ **then**
      return x_t^r
    end if
    if $q \geq Q$ **then**
      return failure
    end if
  end for
  $r_t^+ \leftarrow \arg\max_{r \in \mathcal{R}} J(p_t^r)$
  $p_t^+ \leftarrow p_t^{r_t^+}$
  $\mathcal{P}_t^- \leftarrow \{p_t^r \mid r \in \mathcal{R} \setminus \{r_t^+\}\}$
  Update $\mathcal{H}$ with winner–loser feedback
  $\mathcal{C}_t \leftarrow \{p_t^r : r \in \mathcal{R}\}$
  if $t > 1$ **then**
    $\mathcal{C}_t \leftarrow \mathcal{C}_t \cup \{p_{t-1}^g\}$
  end if
  $g_t \leftarrow \arg\max_{p \in \mathcal{C}_t} J(p)$
  $\mathcal{U}_t \leftarrow \mathcal{C}_t \setminus \{g_t\}$ sorted by decreasing J
  if $p_{t-1}^g \in \mathcal{U}_t$ **then**
    Replace p_{t-1}^g in $\mathcal{U}_t$ with $\text{Comp}(p_{t-1}^g)$
  end if
  for each $u \in \mathcal{U}_t$ in order **do**
    $\hat{p} \leftarrow g_t \oplus_{\text{and}} u$
    $\hat{x} \leftarrow G(x, \hat{p})$
    $J(\hat{p}) \leftarrow \max_{c \neq y} F_c(\hat{x}) - F_y(\hat{x})$
    $q \leftarrow q + 1$
    if $J(\hat{p}) > 0$ **and** $\textsc{Verify}_{\mathcal{M}}(x, \hat{x}, y)$ **then**
      return $\hat{x}$
    end if
    if $J(\hat{p}) > J(g_t)$ **then**
      $g_t \leftarrow \hat{p}$
    end if
    if $q \geq Q$ **then**
      return failure
    end if
  end for
  $p_t^g \leftarrow g_t$
end for
return failure

G Additional Results

Table 11 reports the mean ASR under query budgets from 20 to 100. ASA achieves the highest ASR at every budget. With only 20 queries, it reaches 75.79%, which already exceeds the ASR of every baseline at 100 queries (at most 43.36%). Most of its gain occurs early, and ASR rises only slightly beyond 60 queries. This indicates that ASA finds successful adversarial scenarios within a small number of queries.

Table 11: Mean ASR (%) under different query budgets.

	Query budget				
Method	20	40	60	80	100
AdvFlow	8.41	9.83	11.02	11.73	12.16
DIFAttack	13.42	17.95	21.77	26.16	30.14
CGAttack	12.85	21.17	23.06	24.60	27.59
MCGAttack	23.57	33.78	36.97	39.69	43.36
ASA (Ours)	**75.79**	**83.81**	**86.23**	**86.92**	**87.24**

H Image Quality

We assess image quality from complementary perspectives. The MLLM-based verifier checks whether the main object is preserved and whether the edit is semantically plausible, while the DINOv2-based perceptual similarity and FID reported in Sec. 4.5 quantify semantic preservation and distributional fidelity. Here, we further evaluate perceptual image quality using five no-reference image quality assessment (NR-IQA) metrics: NIMA-AVA (Talebi & Milanfar, 2018), HyperIQA (Su et al., 2020), MUSIQ-AVA, MUSIQ-KoniQ (Ke et al., 2021), and TReS (Golestaneh et al., 2022). Tab. 12 shows that ASA maintains competitive visual quality despite operating as a query-based attack. ASA ranks in the top three on NIMA-AVA, HyperIQA, and MUSIQ-AVA, and remains competitive on TReS. Qualitatively, Fig. 1 shows that ASA primarily induces natural scene-level changes, such as lighting, weather, background, and material variations, whereas several baselines introduce visible noise, color artifacts, or local distortions. These results suggest that ASA improves attack effectiveness without substantially degrading visual naturalness. Since ASA searches only for adversarial editing scenarios and does not modify the image generation process, its absolute image quality is bounded by the underlying editor G and may further benefit from advances in text-guided image editing models.

Table 12: Image quality assessment.

Attack	NIMA -AVA↑	HyperIQA↑	MUSIQ -AVA↑	MUSIQ -KoniQ↑	TReS↑
CF	5.26	0.561	4.00	51.80	71.40
NCF	4.90	0.516	3.79	49.00	66.26
DiffAttack	5.19	**0.616**	4.16	**55.61**	**81.51**
ACA	**5.38**	0.605	**4.34**	55.51	79.11
AdvFlow	4.74	0.477	3.76	47.88	64.77
DifAttack	4.43	0.448	3.68	39.71	49.38
CGAttack	4.52	0.430	3.74	42.06	55.42
MCGAttack	4.96	0.541	3.66	46.16	59.98
ASA	5.30	0.565	4.08	50.26	71.06

I Time Analysis

Tab. 13 compares pre-generation training cost, per-image generation time, and relative total runtime. Prior denotes the total additional attack-specific training time required before generation, such as learning an adversarial data distribution or training a surrogate model. Gen. denotes the average generation time per image. Rel. denotes the sum of Prior and the total generation time over the evaluated inputs, normalized by the corresponding total runtime of ASA. ASA requires no prior training, but incurs a higher per-image generation cost than the other compared methods. When both prior training and generation costs are included, ASA has a lower total runtime than MCGAttack, but a higher total runtime than AdvFlow, DifAttack, and CGAttack.

Table 13: Runtime comparison. Rel. includes both prior training and total generation costs, normalized by ASA.

Attack	Prior (s)	Gen. (s/img)	Rel.
AdvFlow	651.6	**0.14**	**0.0324**×
DifAttack	4939.2	0.21	0.2121×
CGAttack	15645.6	0.78	0.6763×
MCGAttack	69343.2	1.13	2.9073×
ASA	**0**	25.98	1.0000×

J APPLICATION TO SMALL MLLM

Experimental setup. We additionally evaluate our method against Gemma-4-E4B and Qwen3-VL-4B-Instruct. For each victim, we independently sample 100 clean-correct images from NIPS2017 using seed 0. We retain Gemma-4-E4B for instruction generation and naturalness verification, and FLUX.2 Klein 9B KV for image editing. Both victims use greedy decoding with fixed model parameters and no conversation history across queries.

Table 14: ASR and Avg. Q over successful samples.

Model	ASR (%)	Avg. Q
Gemma E4B	76.0	11.14
Qwen3-VL-4B-Instruct	83.0	8.71

For each classification query, we provide a 224×224 RGB image and a fixed prompt containing all 1,000 ImageNet categories, asking the victim to identify the two most likely classes and estimate their probabilities. The ground-truth label is not disclosed in the prompt. The response is requested as a JSON object containing `top1` and `top2`, each with a `class_id` and a `confidence` value, subject to $0 \leq p_2 \leq p_1 \leq 1$ and $p_1 + p_2 \leq 1$.

We parse the response to obtain the predicted classes and self-reported confidence estimates, which are neither internal logits nor calibrated probabilities. With $\Delta = p_1 - p_2$, the optimization score is $-\Delta$ when the top-ranked prediction matches the ground-truth label and $1 + \Delta$ otherwise. Malformed or invalid responses are recorded as query errors and count toward the query budget.

The attack budget is 100 victim queries per image, excluding clean screening and source-baseline evaluation. We apply post-hoc prompt-label correction to the saved successful candidates, excluding successes whose predicted class is explicitly named in the editing prompt. We report the corrected ASR and the average number of attack queries issued before termination over the remaining successful samples.

Results. Table 14 shows that our method achieves corrected ASRs of 76.0% against Gemma and 83.0% against Qwen, requiring an average of 11.14 and 8.71 queries over successful samples, respectively. Prompt-label correction excludes one Gemma success, reducing its ASR from 77.0% to 76.0%, while leaving Qwen's ASR unchanged. Since the two victims are evaluated on independently sampled clean-correct subsets, these results do not establish a paired comparison of their robustness. Nevertheless, they support the applicability of our method to MLLM-based image classification, suggesting its potential beyond conventional image classifiers.