Manipulating the Mind's Eye: A-SAGE, the Attention-based Attack on ViT Explainability

Boshi Zheng[1], **Yan Li**[1*†], **Jiabin Liu**[2†]

[1]School of Cyberspace Science and Technology, Beijing Institute of Technology
[2]School of Information and Electronics, Beijing Institute of Technology
Beijing 100081, China
zhengboshi_2001@foxmail.com, liyan726@bit.edu.cn, liujiabin@bit.edu.cn

Abstract

The rise of Vision Transformers (ViTs) as cornerstone models in safety-critical applications like autonomous driving and medical diagnosis has shifted the focus from pure accuracy to verifiable trustworthiness. However, the very mechanisms used to explain these models, their internal attention maps, are themselves vulnerable. This creates a critical "trust gap," as the model's apparent reasoning can be maliciously manipulated. To systematically investigate this vulnerability, we introduce A-SAGE (Attention-based Steering Adversarial Generation by Corrupting Explanations), a dual-objective attack framework that forces a model to misclassify an input while simultaneously corrupting its internal attention patterns to generate a misleading explanation. A-SAGE achieves this by optimizing a unified loss that combines a standard classification objective with two explanation-specific terms: an attention entropy loss to diffuse the model's focus and an attention map distortion loss to steer the corrupted explanation towards a desired target. Our primary finding is A-SAGE's exceptional black-box transferability. Using a CaiT-S as a white-box surrogate, adversarial examples generated with imperceptible perturbations achieve attack success rates of 79.4% on ViT-B, 49.7% on ResNet-50, and over 81.5% on other transformers (DeiT-B,TNT-S). Crucially, these successful attacks do not merely destroy the explanation; they generate a coherent but false attention map that deceptively "justifies" the wrong prediction. These results reveal a systemic vulnerability in the core reasoning of modern foundation models, establishing A-SAGE as a critical benchmark for auditing the robustness of AI explainability.

Introduction

Vision Transformers (ViTs) have become cornerstone models in computer vision, now being deployed in safety-critical domains such as autonomous driving and medical diagnosis (Dosovitskiy et al. 2020; He et al. 2022). As these powerful models transition from experimental tools to core societal infrastructure, the research focus has rightly shifted from pursuing marginal accuracy gains to ensuring verifiable trustworthiness. A key component of this trust is explainability. To visualize a model's decision-making process, sophisticated post-hoc methods like **Layer-wise Relevance Propagation (LRP)** (Bach et al. 2015) and its ViT-specific variants (Achtibat et al. 2024) are used to generate **explanation heatmaps**. These heatmaps assign a relevance score to each input patch, highlighting which parts of an image supposedly led to the final prediction.

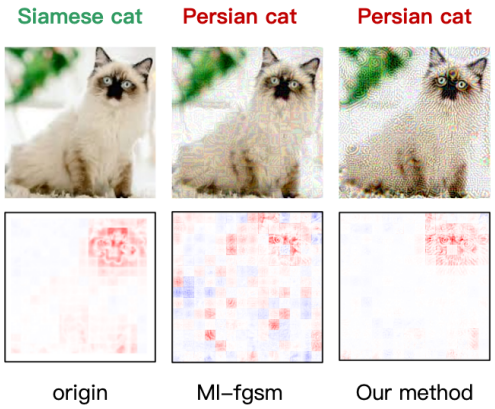

Figure 1: A-SAGE forges plausible explanations to conceal adversarial attacks. **(Left)** A traditional attack (MI-FGSM) yields a noisy, incoherent heatmap that reveals the manipulation. **(Right)** In contrast, A-SAGE generates a coherent explanation for its misclassification, creating a plausible but false rationale.

However, this reliance on post-hoc explanations introduces a critical "trust gap": while the heatmaps provide a compelling, human-understandable rationale, the fidelity of the explanations themselves is rarely questioned. This raises a fundamental security concern that we address in this paper: if an explanation can be maliciously manipulated, it transforms from a tool for verification into a vector for deception, creating an illusion of transparency while hiding flawed reasoning.

We demonstrate that this concern is not merely theoretical. We introduce **A-SAGE (Attention-based Steering Adversarial Generation by Corrupting Explanations)**, a dual-objective attack framework that, for the first time, systematically targets the logical integrity of ViT explanations. We find that the self-attention mechanism, the architectural

*Correspondence to: Yan Li.
†These authors contributed equally.

core of ViTs, is a primary source of this vulnerability. By applying imperceptible perturbations, A-SAGE can simultaneously force a model to misclassify an input *and* induce it to generate a plausible but entirely fabricated explanation for its incorrect prediction. This represents a paradigm shift in adversarial attacks—from targeting *what* a model predicts to attacking *why* it makes that prediction. Such an attack exposes a new vulnerability we term **explanation whitewashing**, where the model is made to "lie" about its reasoning by creating a deceptive alignment between its output and its purported rationale.

Figure 1 provides a striking visual demonstration of this attack's sophistication. While the explanation from a traditional adversarial method like MI-FGSM (Dong et al. 2018) degenerates into a noisy, diffuse heatmap that is visibly nonsensical, A-SAGE produces a deceptive explanation that appears coherent and targeted. It fabricates a compelling visual rationale for the incorrect class, effectively weaponizing the model's explainability against itself.

A-SAGE achieves this dual objective through a novel "disperse-and-reshape" strategy implemented within a unified optimization framework. This framework integrates three key components: (1) an **attention entropy loss** that disperses the self-attention mechanism's focus away from the true, salient image regions; (2) a **patch-level LRP distortion loss** that directly reshapes the explanation heatmap by suppressing the relevance of foreground patches while amplifying background noise, thereby directionally constructing a false explanation; and (3) a standard **classification loss** to induce the target misclassification. Our experiments validate A-SAGE's efficacy and, crucially, its remarkable black-box transferability. Adversarial examples crafted on a white-box CaiT-S model successfully deceive a range of unseen architectures, achieving attack success rates of **79.4%** on ViT-B, **49.7%** on ResNet-50, and over **81.5%** on other transformers like DeiT-B and TNT-S. This suggests the vulnerability is not specific to one architecture but may be a systemic issue in modern deep learning models.

The deployment of foundation models in high-stakes environments cannot proceed without robust audits of their explanation mechanisms. By providing the first "red team" audit of ViT explainability, this work establishes a new and critical benchmark for model safety. Our primary contributions are:

- We identify and formalize "explanation whitewashing," a critical vulnerability in Vision Transformers where their LRP-based explanations can be adversarially manipulated to create a misleading alignment between an incorrect prediction and a plausible-looking rationale.
- We propose A-SAGE, the first dual-objective attack framework designed to simultaneously corrupt a model's prediction and forge its corresponding explanation by targeting both the self-attention mechanism and the LRP relevance structure.
- We demonstrate A-SAGE's state-of-the-art performance and exceptional black-box transferability, revealing that this vulnerability is a systemic risk that extends beyond ViTs to other major architectures, including CNNs.

Related Work

Explainability in Vision Transformers

The adoption of Vision Transformers (ViTs) in safety-critical domains like autonomous driving and medical diagnosis has made model explainability a central challenge for trustworthy AI (Matsoukas et al. 2021; Fan et al. 2022). Early methods for explaining ViTs, such as attention rollout (Abnar and Zuidema 2020) or Grad-CAM variants (Byun and Lee 2023), visualized attention patterns but often produced noisy or artifact-laden heatmaps, limiting their reliability.

To address these limitations, Layer-wise Relevance Propagation (LRP) was adapted for Transformers (Bach et al. 2015). State-of-the-art implementations like AttnLRP (Achtibat et al. 2024) provide a principled approach by introducing specialized propagation rules for attention and normalization layers. Grounded in Taylor decomposition, these rules ensure relevance conservation and deliver high-fidelity attributions that are demonstrably more faithful than prior methods. While AttnLRP was designed for robust analysis, we identify its high fidelity as a novel attack surface. Our work is the first to exploit this faithfulness not for interpretation, but as a vector for deception, using the very tool designed for trust to undermine it.

Transfer-Based Adversarial Attacks

Transfer-based attacks craft adversarial examples on a surrogate model to deceive a black-box target, typically by exploiting features common across architectures (Papernot, McDaniel, and Goodfellow 2016). For CNNs, transferability is enhanced using techniques like input diversity (Xie et al. 2019) and momentum-based optimization (Dong et al. 2018). However, these methods are less effective against ViTs due to fundamental architectural differences.

Recent work has adapted attacks for ViTs by targeting their unique components, such as manipulating attention patterns (Wei et al. 2022) or regularizing patch-level gradients (Zhang et al. 2023). While these approaches improve transferability, they share a singular goal: corrupting the model's final prediction. Our work introduces a new dimension to transferability. We demonstrate that it is possible to achieve a **dual-objective transfer**: A-SAGE crafts examples where not only the misclassification transfers to black-box models, but the forged, misleading explanation transfers as well. This exposes a more sophisticated, second-order vulnerability unaddressed by existing transfer-based attacks.

Attacks on Explainability

The vulnerability of model explanations was first noted as a side effect of traditional adversarial attacks, where attention heatmaps were incidentally disrupted during misclassification (Dong et al. 2019; Zhang and Zhu 2019). In these seminal works, however, the manipulation of the explanation was a *symptom* of a prediction attack, not its goal.

Later methods began to explicitly target explanations, primarily in CNNs. KAKEH et al. (2017) pioneered this by creating "explanation-agnostic" examples, while DAmageNet (Chen et al. 2020) introduced a dual-objective framework

to corrupt both the prediction and the explanation's fidelity. While foundational, these methods were intrinsically linked to convolutional architectures and often focused on destroying the explanation or decoupling it from the prediction.

Our work addresses a critical gap by designing an explanation attack specifically for the Transformer architecture. We introduce **A-SAGE**, a framework that moves beyond mere corruption. Instead of simply creating noise, A-SAGE leverages the ViT's self-attention mechanism to systematically *forge a deceptively plausible rationale* that aligns with the incorrect prediction. By targeting the core architectural component of ViTs, A-SAGE demonstrates a more advanced form of explanation attack than the CNN-centric approaches of the past, establishing a new frontier in adversarial robustness.

Methodology

We introduce **A-SAGE (Attention-based Steering Adversarial Generation by Corrupting Explanations)**, a framework for generating adversarial examples that attack Vision Transformers (ViTs) on two fronts: prediction and explainability. The core of A-SAGE is a multi-objective optimization that crafts perturbations to simultaneously induce misclassification and falsify the model's explanatory output. This is achieved by unifying three distinct loss functions in a "disperse-and-reshape" strategy: (i) an **Attention Entropy Loss ($\mathcal{L}_{\text{attn}}$)** designed to *disperse* the model's internal reasoning by forcing its self-attention mechanism into a diffuse, chaotic state; (ii) a **Patch-level LRP Distortion Loss ($\mathcal{L}_{\text{lrp}}$)** that directly attacks the final attribution map to *reshape* and invert the relevance signal, producing a misleading explanation; and (iii) an **Enhanced Classification Loss ($\mathcal{L}_{\text{class}}$)** that steers the optimization towards an incorrect class label with a robust gradient signal, ensuring a high attack success rate. By combining these objectives, A-SAGE generates highly transferable adversarial examples capable of deceiving ViT-based models in both their decisions and their rationales. The overall framework is depicted in Figure 2.

Dispersing Attention to Collapse Internal Reasoning

The self-attention mechanism is central to the success of Vision Transformers, enabling them to learn meaningful inter-patch relationships (Vaswani et al. 2017; Cordonnier et al. 2021). In a well-trained ViT, these learned relationships manifest as sparse, low-entropy attention distributions, where focus is concentrated on a few salient image patches. This focused attention is the foundation of the model's reasoning. To disrupt this core process, we introduce the **Attention Entropy Loss ($\mathcal{L}_{\text{attn}}$)**, which systematically increases the entropy of the self-attention distributions, forcing the model away from sparse, meaningful patterns and towards a state of diffuse, uninformative focus.

Formally, let $\mathbf{A}^{(l,h)} \in \mathbb{R}^{N \times N}$ be the attention weight matrix for the h-th head in the l-th layer of the Transformer encoder, where N is the number of input tokens. Each row of this matrix represents a discrete probability distribution of attention from one token to all others. The Shannon entropy

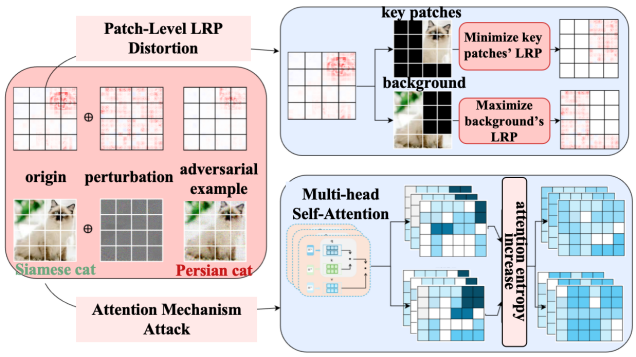

Figure 2: Overview of the A-SAGE dual-objective attack framework. The attack forges a misleading rationale through a "disperse-and-reshape" strategy. **1) Reshape (LRP Distortion):** A-SAGE computes an LRP heatmap for the clean image to identify 'key' (object) and 'background' patches. It then optimizes the perturbation to systematically suppress the relevance of key patches while inflating that of background patches, thereby forging a plausible but false explanation. **2) Disperse (Attention Entropy):** Simultaneously, it collapses the ViT's internal reasoning by maximizing the entropy of its self-attention weights. This scatters attention away from salient objects into a state of 'attentional chaos', degrading the model's representation and inducing misclassification.

of the distribution for the i-th token is given by:

$$H(\mathbf{a}_i^{(l,h)}) = -\sum_{j=1}^{N} A_{i,j}^{(l,h)} \log_2(A_{i,j}^{(l,h)}). \tag{1}$$

A high entropy value indicates that attention is dispersed uniformly, preventing the model from focusing on critical tokens, whereas low entropy signifies a confident, focused state. Our goal is to maximize this entropy across all layers and heads. Since adversarial attacks are formulated as minimization problems, we achieve this by minimizing the negative mean entropy. The Attention Entropy Loss is therefore defined as:

$$\mathcal{L}_{\text{attn}} = \frac{1}{LHN} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{i=1}^{N} \sum_{j=1}^{N} A_{i,j}^{(l,h)} \log_2(A_{i,j}^{(l,h)}), \tag{2}$$

where L is the number of Transformer layers and H is the number of attention heads per layer. By incorporating $\mathcal{L}_{\text{attn}}$ into our attack objective, the optimization process actively penalizes focused, low-entropy attention. This forces the adversarial perturbation to steer the model into a state where its attention is indiscriminately scattered, thereby degrading the structured reasoning that underpins both its predictions and its explanations.

Reshaping Explanations via LRP Distortion

While $\mathcal{L}_{\text{attn}}$ targets the model's internal reasoning, the **Patch-level LRP Distortion Loss ($\mathcal{L}_{\text{lrp}}$)** is engineered to directly manipulate the final visual explanation. To overcome the challenge of diluted relevance scores in standard

ViT explanations, we target heatmaps generated by the robust AttnLRP method (Achtibat et al. 2024). Our objective is not merely to corrupt the explanation, but to forge a new, structurally coherent yet semantically false rationale. This goal contrasts sharply with prior work that focused on simple pixel-wise error metrics or unstructured noise injection (Chen et al. 2020).

To achieve this sophisticated manipulation, we first use the explanation of the clean image, $\mathbf{R}_{\text{orig}}$, to establish a ground-truth segmentation of patch importance. We partition the set of all image patches, $\mathcal{P}$, into two disjoint subsets:

1. **Key Patches ($\mathcal{P}_{\text{key}}$)**: The top-κ percent of patches with the highest relevance scores in $\mathbf{R}_{\text{orig}}$. These patches constitute the core evidence for the model's original, correct decision.
2. **Background Patches ($\mathcal{P}_{\text{bg}}$)**: The remaining patches, where $\mathcal{P}_{\text{bg}} = \mathcal{P} \setminus \mathcal{P}_{\text{key}}$.

Using this segmentation, we formulate a dual-objective loss on the adversarial explanation, $\mathbf{R}_{\text{adv}}$, to systematically invert the importance hierarchy. The first objective, *Key Patch Suppression*, minimizes the relevance attributed to the true evidence:

$$\mathcal{L}_{\text{suppress}} = \mathbb{E}_{p_i \in \mathcal{P}_{\text{key}}}[\mathrm{R}_{\text{adv}}(p_i)]. \tag{3}$$

The second, *Background Patch Elevation*, maximizes the relevance of irrelevant background patches, which we frame as minimizing their negative mean relevance:

$$\mathcal{L}_{\text{elevate}} = -\mathbb{E}_{p_j \in \mathcal{P}_{\text{bg}}}[\mathrm{R}_{\text{adv}}(p_j)]. \tag{4}$$

The final Patch-level LRP Distortion Loss is the sum of these two components:

$$\mathcal{L}_{\text{lrp}} = \mathcal{L}_{\text{suppress}} + \mathcal{L}_{\text{elevate}}. \tag{5}$$

By simultaneously suppressing original evidence while elevating irrelevant information, $\mathcal{L}_{\text{lrp}}$ guides the optimization to construct a new, coherent, yet fundamentally misleading explanation. This transforms the attack from simple corruption to sophisticated forgery, creating an adversarial example that not only misleads the model's prediction but also justifies it with a fabricated visual rationale.

Enhanced Classification Loss for Robust Misclassification

The primary objective of our attack is to induce model misclassification. While a standard cross-entropy loss can achieve this, we employ an **Enhanced Classification Loss** to generate a stronger and more stable gradient signal, which is crucial in a complex multi-objective optimization. This loss, $\mathcal{L}_{\text{class}}$, is specifically engineered to steer the optimization decisively away from the true class, y_{true}.

Given the model's output logits $\mathbf{z}$, $\mathcal{L}_{\text{class}}$ is defined as a weighted sum of three terms:

$$\mathcal{L}_{\text{class}} = w_1 z_{y_{\text{true}}} + w_2 \sigma(\mathbf{z})_{y_{\text{true}}} + w_3 \left(\sigma(\mathbf{z})_{y_{\text{true}}} - \max_{j \neq y_{\text{true}}} \sigma(\mathbf{z})_j \right), \tag{6}$$

where $\sigma(\cdot)$ is the softmax function, $z_{y_{\text{true}}}$ is the logit of the true class, and w_1, w_2, w_3 are weighting hyperparameters. This formulation provides a multi-faceted penalty on the correct classification by combining three distinct objectives:

Algorithm 1: The A-SAGE Adversarial Example Generation Process

Input: Target model f, clean input $\mathbf{x}$, true label y_{true}.
Parameter: Perturbation budget ϵ, step size η, iterations T, loss weights $\alpha, \beta, w_1, w_2, w_3$, key patch ratio κ.
Output: Adversarial example $\mathbf{x}_{\text{adv}}$.

$\mathbf{x}_{\text{adv}}^{(0)} \leftarrow \mathbf{x}$
Define hooks to capture attention maps $\{\mathbf{A}^{(l)}\}$ from f.
$\mathbf{R}_{\text{orig}} \leftarrow \texttt{AttnLRP}(f, \mathbf{x}, y_{\text{true}})$
Identify key patches $\mathcal{P}_{\text{key}}$ and background patches $\mathcal{P}_{\text{bg}}$ from $\mathbf{R}_{\text{orig}}$ using κ.
Let $t = 0$.
while $t < T$ **do**
  $\mathbf{z}, \{\mathbf{A}^{(l)}\} \leftarrow f(\mathbf{x}_{\text{adv}}^{(t)})$
  $\mathcal{L}_{\text{class}} \leftarrow \text{EnhancedClassLoss}(\mathbf{z}, y_{\text{true}}, w_1, w_2, w_3)$ // Eq. 6
  $\mathcal{L}_{\text{attn}} \leftarrow \text{AttentionEntropy}(\{\mathbf{A}^{(l)}\})$ // Eq. 2
  $\mathbf{R}_{\text{adv}} \leftarrow \texttt{AttnLRP}(f, \mathbf{x}_{\text{adv}}^{(t)}, y_{\text{true}})$
  $\mathcal{L}_{\text{lrp}} \leftarrow \text{Distortion}(\mathbf{R}_{\text{adv}}, \mathcal{P}_{\text{key}}, \mathcal{P}_{\text{bg}})$ // Eq. 5
  $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{class}} + \alpha \mathcal{L}_{\text{attn}} + \beta \mathcal{L}_{\text{lrp}}$
  $\mathbf{g} \leftarrow \nabla_{\mathbf{x}_{\text{adv}}^{(t)}} \mathcal{L}_{\text{total}}$
  $\mathbf{x}_{\text{adv}}^{(t+1)} \leftarrow \mathbf{x}_{\text{adv}}^{(t)} - \eta \cdot \text{sign}(\mathbf{g})$
  $\mathbf{x}_{\text{adv}}^{(t+1)} \leftarrow \texttt{clip}(\mathbf{x}_{\text{adv}}^{(t+1)}, \mathbf{x} - \epsilon, \mathbf{x} + \epsilon)$
  $\mathbf{x}_{\text{adv}}^{(t+1)} \leftarrow \texttt{clip}(\mathbf{x}_{\text{adv}}^{(t+1)}, 0, 1)$
  $t \leftarrow t + 1$
end while
return $\mathbf{x}_{\text{adv}}^{(T)}$

- **Logit Suppression:** Minimizing the raw logit value of the true class ($z_{y_{\text{true}}}$) provides a direct and powerful gradient to reduce the model's confidence at the pre-activation level.
- **Probability Suppression:** Minimizing the softmax probability of the true class ($\sigma(\mathbf{z})_{y_{\text{true}}}$) targets the final output distribution, ensuring the model's expressed confidence in the true class is diminished.
- **Margin Maximization:** Minimizing the margin between the true class probability and the highest probability of any other class. This term is analogous to the Carlini-Wagner (C&W) (Pujari et al. 2022) loss and robustly forces the prediction toward a specific incorrect class, creating a decisive misclassification rather than mere uncertainty.

This composite loss ensures a robust and stable misclassification gradient, establishing the necessary precondition for our primary goal of manipulating the model's explanation.

Final Objective and Optimization

We combine these three components into a unified loss function that orchestrates the entire "disperse-and-reshape" attack:

$$\mathcal{L}_{\text{total}}(\mathbf{x}_{\text{adv}}) = \mathcal{L}_{\text{class}} + \alpha \mathcal{L}_{\text{attn}} + \beta \mathcal{L}_{\text{lrp}}, \tag{7}$$

where $\alpha, \beta > 0$ are hyperparameters controlling the trade-off between the attention ($\mathcal{L}_{\text{attn}}$) and LRP ($\mathcal{L}_{\text{lrp}}$) distortion

components. Here, according to the sensitivity analyzes in the Appendix, our hyperparameters are set as: $\alpha = 0.4, \beta = 0.6$. We minimize this objective using an iterative Projected Gradient Descent (PGD) (Madry et al. 2017) based approach to generate the final adversarial examples, as detailed in Algorithm 1.

Experiments

Experimental Settings

Dataset. Consistent with the evaluation protocol of prior work (Dong et al. 2018; Lin et al. 2019; Xie et al. 2019), our test set comprises 1,000 images sampled from the ImageNet 2012 validation set (Russakovsky et al. 2015). To ensure a robust baseline for evaluating attack efficacy, each image was selected from a unique class and confirmed to be correctly classified by all models under consideration prior to perturbation.

Models. To ensure a fair and direct comparison with the SOTA, we adopt the protocol from FDAP (Gao et al. 2024) to evaluate transferability across two scenarios. For Intra-Architecture Transfer (ViT $\rightarrow$ ViT), attacks are crafted on a source Vision Transformer and tested against other black-box ViTs, including ViT-L/16 (Dosovitskiy et al. 2020), DeiT-B, DeiT-B-Dis (Touvron et al. 2021a), CaiT-S, CaiT-XXS (Touvron et al. 2021b), ConViT-B, ConViT-S (d'Ascoli et al. 2021), and TNT-S (Zhao et al. 2021). For Cross-Architecture Transfer (ViT $\rightarrow$ CNN/MLP), these attacks are tested against black-box CNN and MLP models: ResNet50, ResNet101 (He et al. 2016), Mixer-B/16 (Tolstikhin et al. 2021), and ResMLP-24 (Touvron et al. 2022).

Baseline attacks. We benchmark our method against a comprehensive set of adversarial attack baselines. The evaluation includes widely-used gradient-based methods: MI-FGSM (MIM) (Dong et al. 2018), DI-FGSM (DIM) (Xie et al. 2019), and NI-FGSM (NIM) (Lin et al. 2019). Furthermore, we compare against several recent SOTA approaches: SE, SAGA (Mahmood, Mahmood, and Van Dijk 2021), Patch-Fool (Fu et al. 2022), PNA (Wei et al. 2022), ATA (Wang et al. 2022), and FDAP (Gao et al. 2024).

Attack settings. Consistent with standard protocols for generating transferable adversarial examples (Dong et al. 2018; Xie et al. 2019; Gao et al. 2024), we set the hyperparameters for our attack generation process. The total perturbation is constrained within an ℓ_∞-norm ball of radius $\epsilon = 16/255$. We generate adversarial examples over $N = 12$ iterations with a step size of $\alpha = 0.01$.

Metric. We evaluate the transferability of our adversarial examples to black-box models using the Attack Success Rate (ASR). The ASR is defined as the percentage of adversarial examples, crafted on a surrogate model, that successfully induce a misclassification when evaluated on a previously unseen victim model.

Performance Comparison

We evaluate the black-box transferability of our framework across a diverse set of architectures, using various ViTs as the surrogate model. The results, summarized in Table 1, demonstrate that our method excels in ViT-to-ViT transfers by exploiting shared architectural principles, while also retaining partial effectiveness in the more challenging ViT-to-CNN scenario.

Transferability Across Vision Transformers Our method achieves state-of-the-art transferability against other ViTs, significantly outperforming both generic baselines (TI, ATA) and ViT-specific attacks (SE, SAGA). While CNN-centric methods fail due to ViTs' lack of strong inductive biases, and other ViT attacks struggle with architectural heterogeneity (e.g., different attention mechanisms or class token usage), our approach succeeds by targeting fundamental ViT mechanisms. The synergistic attention entropy and LRP distortion losses are robust to architectural variations, leading to high success rates (e.g., **91.2%** on DeiT-B using CaiT-S as a surrogate). We note a performance reduction when using ViT-B as a surrogate, likely due to its more diffuse attention patterns offering less optimization leverage, though our method remains competitive.

Transferability to CNNs As expected, all methods exhibit a significant performance drop when transferring from ViTs to CNNs, a consequence of their fundamental architectural disparities (global self-attention vs. local convolutions). In this scenario, generic methods like FDAP and PNA show stronger transferability, as their focus on disrupting general feature maps is more compatible with CNNs. Nevertheless, our method still achieves non-trivial success rates (e.g., **43.8%** on ResNet-50). This indicates that our strategy of disrupting high-level semantic structures, rather than just low-level features, retains a degree of cross-architecture effectiveness, confirming the vulnerability is not entirely isolated to Transformers.

Overall Analysis In summary, these results confirm that our multi-objective framework is highly specialized and effective for ViT-to-ViT transfer, where it successfully leverages shared decision-explanation mechanisms. While its performance on CNNs underscores the inherent challenges of cross-architecture attacks, its ability to achieve meaningful transferability highlights the potency of disrupting high-level model reasoning.

Visualizing the Attack Mechanism

To analyze the attack's effect on the model's internal decision-making process, we visualize the attention heatmaps of various ViT models. As shown in Figure 3, we compare the heatmaps for a clean input against those for an adversarial example generated by SAGE.

For the clean input, all models correctly classify the image as espresso, with attention maps tightly focused on the object, as expected. In stark contrast, the adversarial example induces not only misclassification but also a corresponding semantic shift in the models' attention, as indicated by the red labels. Specifically:

model	method	ViTs								CNNs		MLPs	
		ViT-B	DeiT-Dis	CaiT-S	ConViT-B	DeiT-B	CaiT-XXS	ConViT-S	TNT-S	Res50	Res101	Mixer-B	Resmlp-24
VIT-B/16	MIM	99.7	34.1	37.7	42.4	48.4	47.0	46.1	28.4	35.0	24.2	35.1	53.5
	DIM	99.9	34.7	38.1	44.1	48.1	47.7	46.9	31.7	32.3	24.9	36.1	54.1
	NIM	100	34.5	35.4	29.3	47.5	46.5	47.2	31.7	34.7	26.8	36.2	53.3
	Patch-Fool	99.9	15.8	16.6	25.2	41.3	26.2	38.2	16.9	38.9	30.6	26.6	35.6
	SE	99.6	14.6	13.6	15.2	13.3	20.8	16.4	13.0	20.5	14.5	14.9	21.5
	SAGA	99.9	13.9	12	14.3	19.1	21.6	15.6	15.3	24.8	18.2	12	21.1
	ATA	99.9	13.9	12.2	20.7	19.1	21.7	15.9	13.0	20.5	14.5	14.9	21.5
	PNA	**100**	41.6	45	48.6	55.8	50.1	53.8	40.2	36.7	32.6	42.4	57.3
	FDAP	99.9	**41.7**	**55.6**	**50.3**	**56.8**	**60.0**	**59.9**	**41.1**	**37.1**	**30.9**	**64.9**	**67.3**
	Ours	99.8	37.5	42.4	46.6	46.0	38.3	51.0	40.6	21.6	17.1	29.4	41.4
Deit-Dis	MIM	99.2	99.2	46.9	54.8	65.0	52.0	57.3	41.5	38.6	26.0	37.0	54.0
	DIM	99.6	99.6	46.9	53.1	66.4	53.0	60.3	41.5	38.6	26.0	37.0	54.0
	NIM	17.2	99.3	27	32.4	35.9	36.5	39.0	28.5	35.0	21.1	28.9	42.9
	Patch-Fool	6.1	99.8	10.2	14.2	18.9	16.4	13.9	11.3	16.8	11.2	11.4	18.4
	SE	11.7	98.9	43.6	29.7	45.2	42.1	43.6	27.6	27.5	17.2	30.1	48.4
	SAGA	6.5	98.6	13.6	15.5	13.1	21.3	15.8	15.7	24.3	18.3	12	21.6
	ATA	14.5	100	16.8	31.2	23.5	23.2	32.5	22	20.5	14.5	14.9	21.5
	PNA	13.2	99.9	16.8	32.2	67.5	59.1	67.7	38	43.7	33.7	57.4	73.6
	FDAP	20.2	99.4	62.4	63.7	68.0	57.5	66.6	37.9	**45.4**	**39.4**	**58.2**	74.3
	Ours	**67.1**	**100**	**86.6**	**85.7**	**86.4**	**74.6**	**88.0**	**75.7**	43.8	36.0	**55.5**	**80.9**
CaiT-S	MIM	14.8	23.4	97.8	41.3	48.9	42.7	42.4	32.5	34.8	20.1	28.2	43.3
	DIM	15.3	25.3	98.7	41.3	50.3	43.7	43.4	34.9	35.9	22.3	30.7	43.4
	NIM	13.7	24.1	98.3	40.7	39.0	33.0	34.0	26.6	31.1	19.5	25.5	39.5
	SE	13.1	31.8	97.2	31.7	37.2	41.6	37.5	26.3	26.3	17.0	29.3	43.1
	Patch-Fool	7.1	9.6	97.7	12	19.7	19.3	16	12.3	32.1	16.9	16.9	23
	SAGA	12.6	11.7	98.7	15.9	13.7	22	16.5	16.3	25.3	19.1	12.1	22
	ATA	13.1	31.9	98.5	13	35.5	38.2	35.5	27.6	27.4	15.8	29.3	43.6
	PNA	15.8	49.4	99.8	45.4	53	60.3	55.1	35.1	30.4	30.5	46.9	70.8
	FDAP	16.7	50.3	99.6	45.5	54.7	62.1	56.0	35.3	40.4	34.8	56.7	71.8
	Ours	**79.4**	**88.6**	**100**	**89.9**	**91.2**	**85.6**	**92.0**	**81.5**	**49.7**	**41.2**	**57.7**	**82.3**
ConvIT-B	MIM	15.7	13.6	20.8	95.4	52.7	35.6	47.9	39.3	39.9	18.7	39.1	42.5
	DIM	16.5	25	31.1	96.6	53.7	38.0	50.3	32.0	34.5	19.7	39.8	44.0
	NIM	15.3	20.1	28.9	96.3	41.2	30.6	36.1	24.8	32.4	18.1	26.4	36.3
	Patch-Fool	7.5	10.2	93.1	96.1	15.6	20.1	20.3	16.8	25.1	23.6	24.1	26.7
	SE	18.9	42	39.8	98.7	56.4	45.4	61.2	36.7	27.1	22.7	33.7	46.1
	SAGA	7.5	13.4	13.6	95.8	12.2	21	16.7	15.2	24.4	17.8	11.3	21.3
	ATA	18.3	18.6	16.5	98.9	23.6	27	27.9	9.7	26.5	14.2	17.3	49.5
	PNA	19.3	58.1	**71.2**	99.8	67.7	61.7	78.6	41.7	30.3	23	61.8	76.3
	FDAP	21.7	62.3	67.5	99.1	67.6	**63.4**	77.7	45.5	**47.2**	**42.5**	**64.5**	**79.7**
	Ours	**54.9**	**69.3**	69.1	**100**	**80.1**	60.0	**87.1**	**64.2**	27.3	22.8	40.7	71.0

Table 1: Attack success rates (%) of 1000 adversarial examples generated on white-box ViT-based models and evaluated against black-box ViT, CNN, and MLP targets.

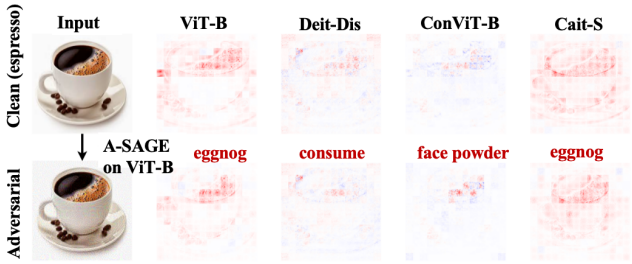

Figure 3: An adversarial example crafted by SAGE (using ViT-B as the surrogate) forces multiple ViT models to misclassify an "espresso" image. The attack simultaneously realigns the models' attention heatmaps to features that plausibly support the incorrect predictions (e.g., the froth for "eggnog"). This provides a clear visual demonstration of a coherent attack on both model prediction and its underlying explanation.

- The surrogate **ViT-B** model misclassifies the image as eggnog, and its attention shifts from the dark liquid to the white, frothy texture—a feature consistent with the new label.
- **DeiT-Dis** misclassifies it as consume, and its attention diffuses from the primary object toward the background elements.
- **ConViT-B** misclassifies it as face powder, with its attention drawn to the coffee's smooth, round surface—a feature plausibly associated with the incorrect class.

These visualizations demonstrate that SAGE does not merely inject unstructured noise to induce an error. Instead, it systematically manipulates the model's feature attribution, forcing the self-attention mechanism to construct a new, plausible rationale for an incorrect prediction. This results in a coherent attack where the misleading explanation directly supports the misclassification, thereby compromising both the integrity and the explainability of the model.

To quantitatively validate this observation, we measure the structural dissimilarity between the original and adversarial explanations using the LRP cosine similarity metric. As illustrated in Figure 4, the distribution of similarity scores for SAGE-attacked examples is sharply centered at zero across all tested models. This indicates that the adversarial explanation is effectively orthogonal to the original one. The result confirms that SAGE achieves *Explanation Decoupling*: it successfully severs the logical connection between an input and its original model-derived rationale, rendering the new explanation structurally independent from the original.

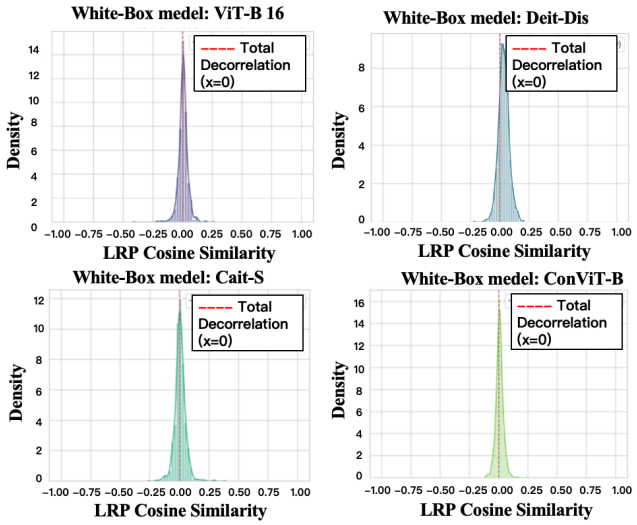

Figure 4: Distributions of LRP cosine similarity between original and adversarial explanations. For all tested white-box models, the similarity after a SAGE attack is concentrated around 0.0, demonstrating a consistent and effective distortion of the model's rationale.

Attention Entropy	LRP Distortion	ViTs	CNNs	MLPs
-	-	75.97	30.62	59.78
✓	-	79.92	36.50	64.55
-	✓	80.04	36.85	64.95
✓	✓	**83.01**	**39.90**	**68.20**

Table 2: Ablation study of SAGE components. We report the Mean Attack Success Rate (MASR, %) using DeiT-Dis as the surrogate. The best result is highlighted in **bold**.

Ablation Study

We conducted an ablation study using DeiT-Dis as the surrogate to isolate the contributions of our Attention Entropy Loss ($\mathcal{L}_{\text{attn}}$) and LRP Distortion Loss ($\mathcal{L}_{\text{lrp}}$). We report the Mean Attack Success Rate (MASR) across all black-box models in Table 2, comparing against a baseline attack that uses only the standard classification loss.

1. **Individual Contributions:** Both components, when used alone, significantly improve MASR over the baseline, validating their individual effectiveness. The $\mathcal{L}_{\text{attn}}$ loss boosts transferability by forcing the model to distribute its attention more broadly, preventing it from overfitting to the surrogate's specific features (our disperse strategy). The $\mathcal{L}_{\text{lrp}}$ loss works by directly manipulating the model's semantic rationale, effectively crafting a deceptive yet plausible explanation for an incorrect class (our reshape strategy). Visualizations in Figure 5 confirm this, showing adversarial heatmaps systematically altered to create coherent explanations for the wrong predictions.
2. **Synergistic Effect:** The full SAGE framework, combining both losses, achieves the highest MASR by a considerable margin, demonstrating a powerful synergy. The two components work in concert to create a more potent attack: $\mathcal{L}_{\text{attn}}$ first creates attentional chaos to make the model broadly vulnerable, while $\mathcal{L}_{\text{lrp}}$ then exploits this instability to precisely sculpt a coherent but false rationale that solidifies the misclassification.

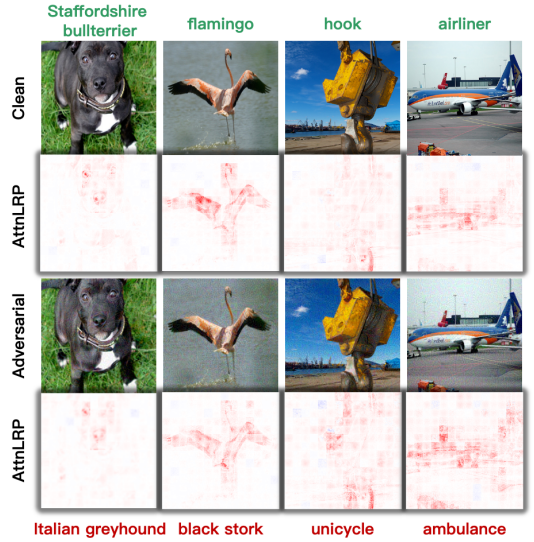

Figure 5: Heatmaps for randomly picked clean images and their corresponding adversarial images, crafted on the Deit-Dis model using our A-SAGE attack.

In summary, this study provides compelling evidence that our "disperse-and-reshape" strategy is highly effective. The synergistic combination of disrupting model focus and forging misleading explanations is the primary mechanism responsible for SAGE's state-of-the-art transferability.

Conclusion

We introduce A-SAGE, a novel multi-objective adversarial attack on Vision Transformers (ViTs) that achieves state-of-the-art transferability. It is the first method to simultaneously corrupt a model's prediction while forging a plausible but misleading explanation to justify the error. Its core strategy is a "targeted dispersion and reshaping" of the model's focus: an attention entropy loss scatters attention from true features, while an LRP distortion loss reconstructs a cohesive but false focus aligned with the incorrect prediction.

This exposes a critical vulnerability we term "explanation whitewashing"—the risk of models generating an illusion of explainability while their reasoning is flawed. This finding highlights the urgent need for evaluation metrics that assess both predictive and explanatory fidelity. While specialized for ViTs, A-SAGE's modular design allows for integration with other techniques to improve cross-architecture transfer. Our work paves the way for developing more robust explainability methods and more secure AI systems.

Acknowledgments

The authors would like to thank the annoymous referees for their valuable comments and helpful suggestions. Thanks also to the AAAI Program Committee for their efforts in organizing the review process. This work was supported in part by the National Natural Science Foundation of China project (#62306036).

References

Abnar, S.; and Zuidema, W. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*.

Achtibat, R.; Hatefi, S. M. V.; Dreyer, M.; Jain, A.; Wiegand, T.; Lapuschkin, S.; and Samek, W. 2024. Attnlrp: attention-aware layer-wise relevance propagation for transformers. *arXiv preprint arXiv:2402.05602*.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7): e0130140.

Byun, S.-Y.; and Lee, W. 2023. ViT-ReciproCAM: gradient and attention-free visual explanations for vision transformer. *arXiv preprint arXiv:2310.02588*.

Chen, S.; He, Z.; Sun, C.; Yang, J.; and Huang, X. 2020. Universal adversarial attack on attention and the resulting dataset damagenet. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4): 2188–2197.

Cordonnier, J.-B.; Mahendran, A.; Dosovitskiy, A.; Weissenborn, D.; Uszkoreit, J.; and Unterthiner, T. 2021. Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2351–2360.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9185–9193.

Dong, Y.; Pang, T.; Su, H.; and Zhu, J. 2019. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4312–4321.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

d'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, 2286–2296. PMLR.

Fan, L.; Pang, Z.; Zhang, T.; Wang, Y.-X.; Zhao, H.; Wang, F.; Wang, N.; and Zhang, Z. 2022. Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8458–8468.

Fu, Y.; Zhang, S.; Wu, S.; Wan, C.; and Lin, Y. 2022. Patch-fool: Are vision transformers always robust against adversarial perturbations? *arXiv preprint arXiv:2203.08392*.

Gao, C.; Zhou, H.; Yu, J.; Ye, Y.; Cai, J.; Wang, J.; and Yang, W. 2024. Attacking transformers with feature diversity adversarial perturbation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 1788–1796.

He, K.; Chen, X.; Xie, S.; Li, Y.; Dollár, P.; and Girshick, R. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16000–16009.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

KAKEH, M. A.; Ghorbani, A.; KAYVAN, B. F.; MIRZAEI, M. A.; et al. 2017. Comparison of visual and digital interpretation methods of land use/cover mapping in Ardabil province.

Lin, J.; Song, C.; He, K.; Wang, L.; and Hopcroft, J. E. 2019. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mahmood, K.; Mahmood, R.; and Van Dijk, M. 2021. On the robustness of vision transformers to adversarial examples. In *Proceedings of the IEEE/CVF international conference on computer vision*, 7838–7847.

Matsoukas, C.; Haslum, J. F.; Söderberg, M.; and Smith, K. 2021. Is it time to replace cnns with transformers for medical images? *arXiv preprint arXiv:2108.09038*.

Papernot, N.; McDaniel, P.; and Goodfellow, I. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*.

Pujari, M.; Cherukuri, B. P.; Javaid, A. Y.; and Sun, W. 2022. An approach to improve the robustness of machine learning based intrusion detection system models against the carlini-wagner attack. In *2022 IEEE International Conference on Cyber Security and Resilience (CSR)*, 62–67. IEEE.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3): 211–252.

Tolstikhin, I. O.; Houlsby, N.; Kolesnikov, A.; Beyer, L.; Zhai, X.; Unterthiner, T.; Yung, J.; Steiner, A.; Keysers, D.; Uszkoreit, J.; et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, 34: 24261–24272.

Touvron, H.; Bojanowski, P.; Caron, M.; Cord, M.; El-Nouby, A.; Grave, E.; Izacard, G.; Joulin, A.; Synnaeve, G.; Verbeek, J.; et al. 2022. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE transactions on pattern analysis and machine intelligence*, 45(4): 5314–5321.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021a. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021b. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 32–42.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Wang, J.; Yin, Z.; Gong, R.; Wang, J.; Liu, A.; and Liu, X. 2022. Generating transferable adversarial examples against vision transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5181–5190.

Wei, Z.; Chen, J.; Goldblum, M.; Wu, Z.; Goldstein, T.; and Jiang, Y.-G. 2022. Towards transferable adversarial attacks on vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2668–2676.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2730–2739.

Zhang, J.; Huang, Y.; Wu, W.; and Lyu, M. R. 2023. Transferable adversarial attacks on vision transformers with token gradient regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 16415–16424.

Zhang, T.; and Zhu, Z. 2019. Interpreting adversarially trained convolutional neural networks. In *International conference on machine learning*, 7502–7511. PMLR.

Zhao, H.; Gao, J.; Lan, T.; Sun, C.; Sapp, B.; Varadarajan, B.; Shen, Y.; Shen, Y.; Chai, Y.; Schmid, C.; et al. 2021. Tnt: Target-driven trajectory prediction. In *Conference on robot learning*, 895–904. PMLR.