# IMPROVING ADVERSARIAL TRANSFERABILITY VIA CLS–PATCH RELATIONSHIP-GUIDED TOKEN GRADIENT SCALING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

The Transferability of adversarial examples pose a practical security threat to Vision Transformers (ViTs), as examples crafted on a surrogate model may also fool unseen black-box models. Existing methods improve transferability by regularizing token gradients or modifying attention and intermediate features. However, intervention targets are often selected using token gradients or attention information without examining how CLS–patch interactions persist or vary across layers. Since these interactions contribute to the class representation, their strength and cross-layer variation can inform a more selective intervention. We therefore propose a method that characterizes CLS–patch relationships using token representations before and after self-attention and assesses their variation across layers. Among tokens with strong relationships, we identify those exhibiting greater variation and scale their gradients at QKV and MLP outputs during backpropagation, without altering forward activations. We randomly select 1,000 images from the ImageNet validation set for evaluation. Preliminary results indicate improved black-box transferability over the MI-FGSM baseline.

## 1 INTRODUCTION

Adversarial examples can induce erroneous predictions by adding small perturbations to clean images (Szegedy et al., 2014; Goodfellow et al., 2015). Compared with white-box attacks against known models, transferable attacks pose a practical security concern: adversarial examples crafted on an accessible surrogate may also fool black-box target models whose parameters are unavailable. Vision Transformers (ViTs) have achieved strong performance in visual recognition (Dosovitskiy et al., 2021), yet they remain vulnerable to adversarial examples. Improving adversarial transferability across ViTs is therefore important for evaluating their security.

Existing transferable attacks seek to reduce overfitting to the surrogate model through gradient stabilization and model refinement (Dong et al., 2018; Xie et al., 2019). Methods designed for ViTs further exploit patch representations and self-attention, for example by varying the patches involved in an attack (Wei et al., 2022), regularizing intermediate token gradients (Zhang et al., 2023; Zhu et al., 2024), adapting attention-related gradient propagation across layers (Ming et al., 2024), or modifying attention maps during forward propagation (Ren et al., 2025a). These methods use token or attention information, but the choice of where to intervene often relies on gradient properties, attention statistics, or random patch selection. Less attention has been given to whether relationships between the CLS token and individual patches remain consistent across layers. This distinction matters when selecting gradients to modify: strong CLS–patch relationships may differ substantially in their cross-layer stability.

To account for this distinction, we study transferable attacks from the perspective of CLS–patch interactions. Using a clean input, we compare CLS and patch-token representations before and after self-attention to characterize their relationship strength, and examine how these relationships vary across selected Transformer layers. We then focus on patches with relatively strong CLS relationships and identify those exhibiting greater cross-layer variation. This procedure uses both relationship strength and stability to determine intervention targets, rather than selecting tokens solely from a single gradient or attention measurement.



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.

For the selected tokens, we scale gradients at the outputs of the QKV and MLP branches during backpropagation. The forward computation remains unchanged, while tokens outside the selected set receive no direct gradient scaling at these locations. The resulting input gradient is used by an MI-FGSM attack (Dong et al., 2018) to update the adversarial image. In preliminary experiments using a ViT-B/16 surrogate and 100 images, the combined QKV–MLP intervention improves mean black-box attack success over the MI-FGSM baseline. These results provide an initial indication of the approach's potential; its advantage over matched random selection and its behavior with other surrogate architectures require further evaluation.

Our main contributions are as follows:

- We introduce a token-selection strategy that considers both the strength and cross-layer variation of CLS–patch relationships when determining where to intervene during a transferable attack.
- We develop a selective gradient intervention that scales the QKV and MLP output gradients of the chosen tokens while leaving forward activations unchanged.

## 2 Related Work

### 2.1 Transfer-Based Attacks on CNNs

Transfer-based attacks initially developed for CNNs have employed gradient optimization, input transformations, and refinements to model or feature representations. Gradient-based methods use information from a surrogate model to generate adversarial perturbations. FGSM performs a single update along the sign of the input gradient (Goodfellow et al., 2015). BIM extends this update to multiple iterations while constraining the perturbation after each step (Kurakin et al., 2017). PGD further uses random initialization and iterative projected updates (Madry et al., 2018). MI-FGSM accumulates gradients across iterations to stabilize the update direction and improve transferability (Dong et al., 2018). SGM attenuates gradients propagated through residual branches, increasing the relative contribution of skip connections during adversarial optimization (Wu et al., 2020).

Input-transformation methods construct diverse views of the input to reduce overfitting to the surrogate model. DIM randomly resizes and pads the input during iterative optimization (Xie et al., 2019). TIM incorporates translated inputs when computing the update direction, improving robustness to spatial shifts (Dong et al., 2019). SIM aggregates gradients from multiple scaled copies of an image (Lin et al., 2020). SSA applies stochastic transformations in the frequency domain to diversify the gradients used for attack generation (Long et al., 2022).

Other methods refine surrogate representations or the gradients obtained from them. TAP enlarges the distance between clean and adversarial intermediate features while regularizing the perturbation (Zhou et al., 2018). Ghost Networks introduce feature-level variations to construct multiple virtual surrogate models (Li et al., 2020). MIG combines integrated gradients with momentum to improve transfer across CNN and Transformer architectures (Ma et al., 2023). MuMoDIG further refines the integration paths used to estimate transferable gradient directions (Ren et al., 2025b).

### 2.2 Transfer-Based Attacks on ViTs

The token-based architecture of Vision Transformers (ViTs) has motivated attacks designed around patch representations and self-attention (Naseer et al., 2022). PNA blocks gradients through the branch that generates attention maps, whereas PatchOut varies the subset of image patches involved in adversarial optimization (Wei et al., 2022). TGR regulates gradients in intermediate ViT modules and suppresses tokens with extreme backpropagated gradients to reduce gradient variance (Zhang et al., 2023). GNS adjusts gradients in modules such as the query, key, and value projections (Zhu et al., 2024). ATT combines token-gradient adjustment, patch selection, and depth-dependent attention-gradient regulation, including cross-layer gradient variation (Ming et al., 2024). In contrast to these backward-oriented methods, FPR refines attention maps and token representations during forward propagation (Ren et al., 2025a).

Feature-level approaches provide another means of improving transferability. FDAP selects intermediate layers according to feature diversity and suppresses high-frequency components in their

representations (Gao et al., 2024). SAA uses an additional model to encourage alignment between clean and adversarial features at global and local spatial scales (Chen et al., 2025). AIM introduces guide images and corresponding objectives in logit and intermediate-feature spaces to provide additional guidance during adversarial example generation (Li et al., 2025).

These methods demonstrate the value of modifying gradients, attention, or intermediate features in ViT-based attacks. Although cross-layer gradient variation has informed adaptive token tuning (Ming et al., 2024), the strength of CLS–patch relationships and their variation across layers have received less attention when selecting token gradients for intervention. Our method uses these two properties to identify patch tokens for gradient scaling at QKV and MLP outputs, while leaving the forward computation unchanged.

# 3 METHODOLOGY

Our method follows three steps: characterizing token interactions, selecting tokens for gradient intervention, and generating adversarial examples. We first identify patch tokens that have strong but cross-layer variable relationships with the CLS token. We then scale their gradients at selected QKV and MLP outputs during adversarial optimization, and use the resulting input gradient to update the image.

## 3.1 PRELIMINARIES

**Token interactions in ViTs.** Given an image $x$, a Vision Transformer (ViT) divides it into $N$ patches and maps each patch to a token embedding (Dosovitskiy et al., 2021). A classification token (CLS token) is prepended to the patch sequence, followed by positional embeddings:

$$Z^{(0)} = [z_{\text{cls}}^{(0)}; z_1^{(0)}; \ldots; z_N^{(0)}] + E_{\text{pos}}. \tag{1}$$

Within a Transformer block, token representations are projected into query, key, and value vectors. For one attention head, the attention weights and aggregated values are computed as follows (Vaswani et al., 2017):

$$A^{(\ell)} = \text{softmax}\left(\frac{Q^{(\ell)}(K^{(\ell)})^\top}{\sqrt{d_h}}\right), \qquad O_i^{(\ell)} = \sum_{j=0}^{N} A_{ij}^{(\ell)} V_j^{(\ell)}, \tag{2}$$

where index 0 denotes the CLS token and $d_h$ is the dimension of one attention head. Through successive blocks, the CLS token aggregates information from patch tokens to form the representation used for classification (Dosovitskiy et al., 2021; Marouani et al., 2026). Its attention weights describe how token values are weighted during aggregation. However, CLS–patch relationships also depend on value vectors, residual connections, and subsequent transformations; a single attention weight does not fully characterize such a relationship. This motivates the representation-based relationship modeling in Section 3.2.

**Transfer-based adversarial attacks.** Let $f_s$ be an accessible surrogate model, $f_t$ an unseen target model, and $y$ the true label of a clean image $x$. Under an untargeted $\ell_\infty$ constraint, the attacker constructs an adversarial image by increasing the surrogate classification loss:

$$\max_{x' \in [0,1]^{C\times H\times W}} \mathcal{L}(f_s(x'), y) \quad \text{subject to} \quad \|x' - x\|_\infty \le \epsilon. \tag{3}$$

The target model's parameters and gradients are unavailable during attack generation. Iterative attacks such as MI-FGSM (Dong et al., 2018) use the surrogate input gradient to determine each perturbation update. Modifying the propagation of internal gradients can therefore change the input gradient used for optimization. To make this modification selective, we first determine which patch tokens should receive the intervention.

## 3.2 CLS–PATCH RELATIONSHIP MODELING AND TOKEN SELECTION

Self-attention enables the CLS token to aggregate patch information (Dosovitskiy et al., 2021; Marouani et al., 2026), suggesting that CLS–patch relationships can inform token selection. Relationship strength at a single layer, however, does not reveal whether a relationship remains consistent across layers. We therefore use both strength and cross-layer variation: patches with strong

CLS relationships form a candidate set, from which we select those exhibiting greater variation for subsequent gradient intervention.

For a clean image, consider a predefined set of scoring layers $\mathcal{L}_s$. Let $c_{\mathrm{pre}}^{(\ell)}$ and $p_{i,\mathrm{pre}}^{(\ell)}$ denote the CLS and $i$-th patch-token representations immediately before self-attention in layer $\ell$. Their representations after the attention output is added to the residual stream are $c_{\mathrm{post}}^{(\ell)}$ and $p_{i,\mathrm{post}}^{(\ell)}$. The selected layers retain a consistent spatial correspondence between patch indices, so that index $i$ refers to the same image region across these layers.

We characterize the CLS–patch relationship using both post-attention representational alignment and the increase in alignment across the attention operation:

$$\begin{aligned} a_i^{(\ell)} &= \frac{1+\cos\left(c_{\mathrm{post}}^{(\ell)}, p_{i,\mathrm{post}}^{(\ell)}\right)}{2}, \\ \Delta_i^{(\ell)} &= \cos\left(c_{\mathrm{post}}^{(\ell)}, p_{i,\mathrm{post}}^{(\ell)}\right) - \cos\left(c_{\mathrm{pre}}^{(\ell)}, p_{i,\mathrm{pre}}^{(\ell)}\right), \\ b_i^{(\ell)} &= \frac{[\Delta_i^{(\ell)}]_+}{[\Delta_i^{(\ell)}]_+ + \tau}, \qquad r_i^{(\ell)} = \frac{2a_i^{(\ell)} b_i^{(\ell)}}{a_i^{(\ell)} + b_i^{(\ell)} + \xi}, \end{aligned} \tag{4}$$

where $[z]_+ = \max(z, 0)$ and $\tau, \xi > 0$ are numerical constants. A high $r_i^{(\ell)}$ requires both strong post-attention alignment and a positive increase in alignment. This score describes a relationship between token representations, rather than a direct pairwise attention weight.

Because relationship scores may have different scales across layers, we convert $r_i^{(\ell)}$ to its percentile rank $u_i^{(\ell)}$ among the patch tokens in layer $\ell$. Tied scores receive their average rank. We then measure the mean relative relationship strength and its cross-layer variation:

$$\bar{u}_i = \frac{1}{|\mathcal{L}_s|} \sum_{\ell \in \mathcal{L}_s} u_i^{(\ell)}, \qquad v_i = \frac{1}{|\mathcal{L}_s|} \sum_{\ell \in \mathcal{L}_s} \left(u_i^{(\ell)} - \bar{u}_i\right)^2. \tag{5}$$

A larger $\bar{u}_i$ indicates a consistently higher relative CLS–patch relationship strength, whereas a larger $v_i$ indicates greater variation in that relative relationship across layers.

We first retain the $2K$ patches with the highest mean relationship strength and then select the $K$ most variable patches from this candidate set:

$$\mathcal{C}(x) = \mathrm{Top}_{2K}\{\bar{u}_i\}_{i=1}^{N}, \qquad \mathcal{U}(x) = \mathrm{Top}_K\{v_i \mid i \in \mathcal{C}(x)\}, \quad 2K \le N. \tag{6}$$

The resulting set $\mathcal{U}(x)$ specifies the targets of the gradient intervention in Section 3.3. It is computed once from the clean image and remains fixed throughout attack optimization. Cross-layer variation serves as a selection criterion; it is not, by itself, evidence that a relationship is specific to the surrogate model.

### 3.3 Relationship-Guided Gradient Intervention

Methods such as TGR (Zhang et al., 2023) and GNS (Zhu et al., 2024) show that regulating gradients within ViT modules can change the gradient used to generate adversarial examples: TGR suppresses extreme token gradients, while GNS applies finer-grained normalization and scaling to internal gradients. Motivated by this approach to backward-propagation control, we use the CLS–patch relationships established above to specify the tokens whose gradients are scaled.

QKV projections produce the query, key, and value representations used in self-attention, while the MLP branch further transforms token features. Gradients propagated through both branches contribute to the eventual input gradient. At each attack iteration, we retain the surrogate's ordinary forward computation and intervene only during backpropagation at selected QKV and MLP outputs. For a predefined set of intervention layers $\mathcal{L}_g$, let $G_i^{(\ell,b)}$ be the gradient arriving at the output for token $i$ in branch $b \in \{\mathrm{QKV}, \mathrm{MLP}\}$. The gradient passed to earlier operations is

$$\widetilde{G}_i^{(\ell,b)} = \left(1 - (1-\eta_b)\mathbf{1}[i \in \mathcal{U}(x)]\right) G_i^{(\ell,b)}, \qquad \ell \in \mathcal{L}_g, \tag{7}$$

**Algorithm 1** Relationship-Guided Token Gradient Intervention

**Require:** Surrogate model $f_s$ and loss $\mathcal{L}$; clean image $x$ and label $y$; perturbation budget $\epsilon$, step size $\alpha$, iterations $T$, and momentum factor $\mu$; scoring layers $\mathcal{L}_s$, intervention layers $\mathcal{L}_g$, token number $K$ with $2K \le N$; constants $\tau, \xi > 0$; scaling factors $\eta_{\mathrm{QKV}} = \eta_{\mathrm{MLP}} = 0.5$

**Ensure:** Adversarial image $x_{\mathrm{adv}}$

Run an ordinary forward pass on $x$ and collect the CLS and patch representations before and after attention in each $\ell \in \mathcal{L}_s$.
**for** $\ell \in \mathcal{L}_s$ **do**
    Compute $r_i^{(\ell)}$ for all patch tokens using Eq. equation 4.
    Convert $\{r_i^{(\ell)}\}_{i=1}^N$ to percentile ranks $\{u_i^{(\ell)}\}_{i=1}^N$.
**end for**
Compute $\bar{u}_i$ and $v_i$ for each patch token using Eq. equation 5.
$\mathcal{C}(x) \leftarrow \mathrm{Top}_{2K}\{\bar{u}_i\}_{i=1}^N$.
$\mathcal{U}(x) \leftarrow \mathrm{Top}_K\{v_i \mid i \in \mathcal{C}(x)\}$.
Initialize $x_0 \leftarrow x$ and $m_0 \leftarrow 0$.
**for** $t = 0, \dots, T-1$ **do**
    Compute the ordinary forward loss $\mathcal{L}_t \leftarrow \mathcal{L}(f_s(x_t), y)$.
    During backpropagation, for each $\ell \in \mathcal{L}_g$ and $b \in \{\mathrm{QKV}, \mathrm{MLP}\}$, replace the gradient at branch output $b$ by
    $\widetilde{G}_i^{(\ell,b)} \leftarrow \big(1 - (1-\eta_b)\mathbf{1}[i \in \mathcal{U}(x)]\big) G_i^{(\ell,b)}, \quad i = 0, \dots, N.$
    Continue backpropagation to obtain the modified input gradient $\widetilde{q}_t$.
    $m_{t+1} \leftarrow \mu m_t + \dfrac{\widetilde{q}_t}{\|\widetilde{q}_t\|_1 + \xi}$.
    $x_{t+1} \leftarrow \Pi_{[0,1]\cap B_\infty(x,\epsilon)}\big(x_t + \alpha\ \mathrm{sign}(m_{t+1})\big)$.
**end for**
**return** $x_{\mathrm{adv}} \leftarrow x_T$.

where $0 < \eta_b \le 1$ is the scaling factor for branch $b$. In the current joint QKV–MLP configuration, both factors are set to $0.5$. The CLS token and patches outside $\mathcal{U}(x)$ receive no direct scaling at these locations.

The intervention reduces the backward signal through the selected branch outputs while leaving their forward activations unchanged. Since attention mixing and residual connections also affect backpropagation, Eq. equation 7 specifies a local operation; it does not imply that the final input gradient is scaled uniformly by $0.5$. Backpropagation then continues to the input image, providing the gradient used in Section 3.4.

### 3.4 Adversarial Example Generation

The intervention leaves the surrogate's forward loss computation unchanged but modifies the gradient obtained when that loss is propagated to the input. Let $\widetilde{q}_t$ denote the resulting input gradient at attack iteration $t$. Starting from $x_0 = x$ and $m_0 = 0$, we update the momentum and adversarial image using MI-FGSM (Dong et al., 2018):

$$\begin{aligned} m_{t+1} &= \mu m_t + \frac{\widetilde{q}_t}{\|\widetilde{q}_t\|_1 + \xi}, \\ x_{t+1} &= \Pi_{[0,1]\cap B_\infty(x,\epsilon)}\left(x_t + \alpha\ \mathrm{sign}(m_{t+1})\right), \end{aligned} \tag{8}$$

where $\mu$ is the momentum decay factor, $\alpha$ is the step size, and $\Pi$ projects the image onto the valid pixel range and perturbation budget. Every iteration uses the same clean-image token set $\mathcal{U}(x)$: relationship modeling determines where gradients are scaled, the intervention changes the input update direction, and MI-FGSM generates the adversarial image under the prescribed constraint. The final image is evaluated on unseen target models.

The complete procedure is summarized in Algorithm 1.

# 4 EXPERIMENTS

## 4.1 EXPERIMENTAL SETUP

**Dataset and evaluation protocol.** Following the experimental setting of our previous study, we select 1,000 images from the ILSVRC2012 validation set (Russakovsky et al., 2015), with one image from each of the 1,000 categories. All experiments consider untargeted attacks. Images are resized to $224 \times 224$ before adversarial example generation. We use attack success rate (ASR) to evaluate transferability and report the ASR on each target model. The average ASR excludes the surrogate model when that model also appears among the targets.

**Models.** For the proposed method, adversarial examples are generated independently using four ImageNet-pretrained surrogate models: ViT-B/16 (Dosovitskiy et al., 2021), CaiT-S/24 (Touvron et al., 2021b), PiT-Ti (Heo et al., 2021), and distilled DeiT-B (Touvron et al., 2021a). We evaluate their transferability on the 13 target models used in our previous study. The seven Transformer-based targets are ViT-B/16, CaiT-S/24, PiT-B, Visformer-S (Chen et al., 2021), Swin-T (Liu et al., 2021), DeiT-T, and CoaT-T (Xu et al., 2021). The six CNN-based targets are ResNet-18 (He et al., 2016), VGG-16 (Simonyan & Zisserman, 2015), DenseNet-121 (Huang et al., 2017), EfficientNet-B0 (Tan & Le, 2019), MobileNet-V3 (Howard et al., 2019), and ResNeXt-50 (Xie et al., 2017). All model parameters remain fixed during attack generation and evaluation. For distilled DeiT-B, the distillation token is excluded from patch-token selection.

**Parameter settings.** We use MI-FGSM (Dong et al., 2018) as the base attack. Under an $\ell_\infty$ constraint, the perturbation budget is $\epsilon = 16/255$, the number of iterations is $T = 10$, the step size is $\alpha = 1.6/255$, and the momentum decay factor is $\mu = 1.0$. For the proposed method, we select $K = 64$ patch tokens on a $14 \times 14$ token grid. For PiT-Ti, token selection is performed separately at each stage, and $K$ is adjusted according to the number of patch tokens at that stage. The numerical offset used to compute the positive relationship gain is set to $0.1$. During backpropagation, the gradients associated with selected patch tokens are scaled by $0.5$ in the QKV and MLP branches. Token selection is performed on the clean image and remains fixed across attack iterations.

## 4.2 TRANSFERABILITY ACROSS TARGET MODELS

Table 1 presents the original comparison results alongside rows reserved for the proposed method. The previously obtained baseline results and their surrogate-model labels are retained. In particular, PiT-B and Visformer-S in the original results must not be interpreted as PiT-Ti and distilled DeiT-B, respectively. The rows marked "Pending" require results from the new 1,000-image experiment and are not used to draw performance conclusions at this stage.

For ViT-B/16 and CaiT-S/24, the corresponding target-model entries measure white-box performance; the other target entries measure black-box transferability. PiT-Ti and distilled DeiT-B are distinct from the PiT-B and DeiT-T targets, so all 13 target entries in their new rows measure black-box transferability. Once the pending results are available, the transfer performance of the proposed method can be analyzed across Transformer-based and CNN-based targets.

In a preliminary experiment using ViT-B/16 as the surrogate model, the evaluated token-gradient scaling strategy increases the average attack success rate across 12 black-box target models from 45.7% to 47.1% compared with MI-FGSM under the same settings. Specifically, it achieves gains of 4.2 and 3.1 percentage points on VGG-16 and EfficientNet-B0, respectively, indicating improved transferability to certain target models. The results on CaiT-S, PiT-B, and MobileNet-V3 remain unchanged, suggesting that the transferability gains depend on the target architecture.

## 4.3 ABLATIONS AND VALIDATION

The following analyses use the same 100-image pilot set as the ViT-B/16 results in Table 1. For each target model, the attack success rate is computed over images that the target classifies correctly before the attack. The reported averages are macro-averages over six black-box Transformer targets, six CNN targets, or all 12 black-box targets, respectively.

Table 1: Untargeted attack success rates (%) on Transformer-based and CNN-based target models. Numerical baseline entries are retained from the original experiments. "Pending" denotes an unmeasured result of the proposed method; "–" denotes a surrogate setting not used for the proposed method. The Average column excludes the surrogate model when it is also a target.

| Surrogate | Attack | ViT-B | CaiT-S | PiT-B | Vis-S | Swin-T | DeiT-T | CoaT-T | RN-18 | VGG-16 | DN-121 | EN-B0 | MN-V3 | RNX-50 | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ViT-B | MI-FGSM | 100.0 | 73.0 | 35.0 | 49.0 | 52.0 | 54.8 | 35.1 | 42.0 | 44.8 | 41.0 | 50.0 | 38.6 | 33.0 | 45.7 |
| | PatchOut | 100.0 | 68.9 | 44.5 | 48.8 | 58.6 | 68.9 | 46.9 | 56.6 | 60.6 | 54.9 | 53.7 | 54.7 | 40.8 | 54.8 |
| | TGR | 99.5 | 79.8 | 58.5 | 62.9 | 73.9 | 88.7 | 60.0 | 72.4 | 74.4 | 70.9 | 74.2 | 73.5 | 54.5 | 70.3 |
| | GNS | 99.8 | 80.9 | 59.5 | 62.6 | 73.9 | 86.8 | 59.4 | 68.8 | 72.2 | 68.2 | 70.5 | 70.2 | 51.9 | 68.7 |
| | FPR | 98.0 | 77.1 | 48.6 | 56.6 | 73.5 | 93.3 | 58.5 | 75.8 | 79.9 | 70.5 | 78.0 | 77.0 | 52.3 | 70.1 |
| | Ours | 100.0 | 73.0 | 35.0 | 51.0 | 54.0 | 55.9 | 36.1 | 44.0 | 49.0 | 42.0 | 53.1 | 38.6 | 34.0 | 47.1 |
| CaiT-S | MI-FGSM | 65.3 | 98.5 | 49.9 | 56.2 | 68.0 | 75.8 | 54.2 | 61.2 | 67.1 | 60.5 | 60.6 | 56.2 | 47.5 | 60.2 |
| | PatchOut | 76.5 | 99.6 | 59.4 | 65.9 | 78.6 | 86.3 | 64.5 | 70.5 | 74.3 | 69.4 | 72.1 | 64.0 | 54.6 | 69.7 |
| | TGR | 89.0 | 100.0 | 76.3 | 80.7 | 90.5 | 97.5 | 80.1 | 88.6 | 86.8 | 84.0 | 88.2 | 81.9 | 69.4 | 84.4 |
| | GNS | 91.3 | 100.0 | 78.3 | 82.2 | 91.6 | 97.8 | 81.1 | 86.3 | 87.0 | 85.0 | 89.9 | 82.5 | 71.1 | 85.3 |
| | FPR | 92.4 | 99.9 | 81.4 | 85.2 | 94.3 | 98.7 | 86.5 | 88.3 | 87.6 | 87.3 | 90.9 | 82.4 | 77.0 | 87.7 |
| | Ours | | | | | | | | – | | | | | | |
| PiT-B | MI-FGSM | 30.3 | 36.0 | 38.6 | 43.3 | 50.7 | 75.0 | 55.2 | 70.0 | 69.2 | 61.5 | 64.6 | 71.6 | 40.7 | 55.7 |
| | PatchOut | 36.7 | 44.9 | 46.4 | 51.2 | 58.8 | 83.6 | 62.0 | 76.6 | 76.2 | 68.0 | 74.8 | 78.0 | 47.5 | 63.2 |
| | TGR | 39.8 | 48.0 | 46.6 | 53.8 | 62.2 | 87.1 | 64.9 | 82.9 | 82.0 | 75.9 | 82.0 | 88.7 | 54.0 | 68.4 |
| | GNS | 37.4 | 44.4 | 46.3 | 51.7 | 59.3 | 84.4 | 63.1 | 79.8 | 78.1 | 71.8 | 75.8 | 84.4 | 48.8 | 64.9 |
| | FPR | 45.3 | 56.7 | 55.5 | 61.0 | 69.4 | 92.1 | 72.5 | 88.0 | 85.6 | 80.0 | 88.0 | 90.6 | 59.8 | 74.1 |
| | Ours | | | | | | | | – | | | | | | |
| Vis-S | MI-FGSM | 81.5 | 92.5 | 67.3 | 71.7 | 80.3 | 87.3 | 66.2 | 68.3 | 70.5 | 70.3 | 71.2 | 68.1 | 58.1 | 73.5 |
| | PatchOut | 84.9 | 94.8 | 70.9 | 74.9 | 83.7 | 89.6 | 70.0 | 72.2 | 72.9 | 74.5 | 74.1 | 65.8 | 61.9 | 76.3 |
| | TGR | 91.4 | 97.7 | 82.2 | 84.8 | 93.1 | 98.4 | 83.2 | 85.8 | 85.2 | 85.3 | 89.9 | 84.5 | 74.9 | 87.6 |
| | GNS | 90.7 | 98.0 | 80.5 | 83.6 | 91.1 | 98.0 | 79.9 | 83.5 | 80.6 | 83.2 | 86.9 | 80.8 | 71.0 | 85.4 |
| | FPR | 92.8 | 98.3 | 82.8 | 86.8 | 94.6 | 99.6 | 84.6 | 89.1 | 88.6 | 87.4 | 91.8 | 86.4 | 79.2 | 89.6 |
| | Ours | | | | | | | | – | | | | | | |

Table 2: Pilot ablation results on 100 images. Values are attack success rates (%). Random QKV+MLP results are reported as mean ± standard deviation across three random seeds.

| Method | Transformer | CNN | All black-box |
|---|---|---|---|
| MI-FGSM | 49.8 | 41.6 | 45.7 |
| Identity hook | 49.8 | 41.6 | 45.7 |
| Nonstable QKV | 50.8 | 41.9 | 46.4 |
| Nonstable MLP | 50.1 | 41.7 | 45.9 |
| Nonstable QKV+MLP | 50.8 | 43.4 | 47.1 |
| Random QKV+MLP (3 seeds) | $49.9 \pm 0.1$ | $42.2 \pm 0.5$ | $46.1 \pm 0.3$ |

**Branch ablation and random-token control.** We first examine the contributions of gradient scaling in the QKV and MLP branches. We also compare the selected nonstable patch tokens with randomly selected patch tokens under the combined intervention. Table 2 summarizes the results.

Scaling gradients in either branch alone yields a smaller improvement than scaling both branches. The combined intervention increases the average black-box attack success rate from 45.7% to 47.1%, with a larger observed gain on CNN targets than on Transformer targets. Its average success rate also exceeds that of the three random-token controls. These observations suggest that both the intervention branches and the choice of patch tokens contribute to the pilot result; the limited sample and number of random seeds do not yet establish that this advantage will persist at a larger scale.

**Intervention and perturbation checks.** The identity hook yields the same attack success rates as MI-FGSM, and its adversarial images are identical to the MI-FGSM outputs for all 100 pilot images. In contrast, each active gradient-intervention setting produces adversarial images that differ from the baseline outputs. The perturbation audit reports no violations of the $\ell_\infty$ budget across the 100 images in any tested setting; the maximum observed perturbation magnitude is approximately $16/255$, up to numerical precision. These checks confirm that the active interventions change the generated examples while maintaining the specified perturbation constraint.

**Uncertainty of the observed gains.** We further assess the pilot results using paired bootstrap resampling over the 100 images. For the combined nonstable-token intervention, the estimated image-level mean gain across black-box targets is 1.42 percentage points, with a 95% bootstrap confidence interval of [0.21, 2.70] percentage points. The corresponding estimated gain is 1.90

percentage points for CNN targets, with an interval of [0.07, 3.83], and 1.00 percentage point for Transformer targets, with an interval of [-0.67, 2.70]. At the individual-target level, VGG-16 shows an observed gain of 4.17 percentage points, with an interval of [1.04, 8.33]. The uncertainty on several other targets remains substantial, so the observed model-specific gains should be interpreted as pilot findings.

## 5 Conclusion

Transferable adversarial attacks against Vision Transformers (ViTs) often rely on gradients specific to the surrogate model, while existing gradient intervention strategies may insufficiently account for information exchange among tokens. To address this issue, we propose a token-interaction-guided gradient intervention method for generating more transferable adversarial examples. The method measures changes in CLS–patch relationships across Transformer blocks and characterizes patch tokens by their interaction strength and cross-layer variation. It then selects tokens with strong interactions and substantial variation, and scales their backpropagated gradients in the QKV and MLP branches during MI-FGSM optimization. This intervention adjusts the gradient signal used to update the perturbation while preserving the model's forward computation. In preliminary experiments using ViT-B/16 as the surrogate model, joint intervention in the two branches improves black-box attack success on several Transformer and CNN targets relative to MI-FGSM. Branch ablations and random-token controls further indicate that both the intervention location and token selection influence the observed transferability. These findings provide initial evidence for the value of token-specific gradient modulation, although the gains vary across target architectures. The available experiments use a limited image set and an earlier nonstable-token selection variant, and therefore do not yet establish the effectiveness of the complete method. Furthermore, a fixed token budget and gradient scaling strength may limit adaptation to different Transformer architectures. Future work will evaluate the full method on 1,000 images using four surrogate models and investigate adaptive token selection and gradient intervention.

### Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62602591 and Grant 62472323; in part by the Natural Science Foundation of Zhejiang province under Grant LQN26F020036; in part by the ShaoXing Key Research Project under Grant 2025B41013; in part by Scientific Research Fund of Zhejiang Provincial Education Department under Grant Y202558717; in part by Science Foundation of Zhejiang Sci-Tech University (ZSTU) under Grant 25222016-Y.

## References

Zhaoyu Chen, Haijing Guo, Kaixun Jiang, Jiyuan Fu, Xinyu Zhou, Dingkang Yang, Hao Tang, Bo Li, and Wenqiang Zhang. Boosting adversarial transferability with spatial adversarial alignment. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zhengsu Chen, Lingxi Xie, Jianwei Niu, Xuefeng Liu, Longhui Wei, and Qi Tian. Visformer: The vision-friendly transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 589–598, 2021.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9185–9193, 2018.

Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4312–4321, 2019.

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 16×16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Chenxing Gao, Hang Zhou, Junqing Yu, Yuteng Ye, Jiale Cai, Junle Wang, and Wei Yang. Attacking transformers with feature diversity adversarial perturbation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3):1788–1796, 2024. doi: 10.1609/aaai.v38i3.27947.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11936–11945, 2021.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1314–1324, 2019.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4700–4708, 2017.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations Workshop*, 2017.

Teng Li, Xingjun Ma, and Yu-Gang Jiang. AIM: Additional image guided generation of transferable adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(5):4941–4949, 2025. doi: 10.1609/aaai.v39i5.32523.

Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(7):11458–11465, 2020. doi: 10.1609/aaai.v34i07.6810.

Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E. Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *Proceedings of the European Conference on Computer Vision*, pp. 549–566, 2022. doi: 10.1007/978-3-031-19772-7_32.

Wenshuo Ma, Yidong Li, Xiaofeng Jia, and Wei Xu. Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4630–4639, 2023.

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.

Alexis Marouani, Oriane Siméoni, Hervé Jégou, Piotr Bojanowski, and Huy V. Vo. Revisiting [CLS] and patch token interaction in vision transformers. In *International Conference on Learning Representations*, 2026.

Di Ming, Peng Ren, Yunlong Wang, and Xin Feng. Boosting the transferability of adversarial attack on vision transformer with adaptive token tuning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0658.

Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Fahad Shahbaz Khan, and Fatih Porikli. On improving adversarial transferability of vision transformers. In *International Conference on Learning Representations*, 2022.

Yuchen Ren, Zhengyu Zhao, Chenhao Lin, Bo Yang, Lu Zhou, Zhe Liu, and Chao Shen. Improving adversarial transferability on vision transformers via forward propagation refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25071–25080, 2025a.

Yuchen Ren, Zhengyu Zhao, Chenhao Lin, Bo Yang, Lu Zhou, Zhe Liu, and Chao Shen. Improving integrated gradient-based transferable adversarial examples by refining the integration path. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):6731–6739, 2025b. doi: 10.1609/aaai.v39i7.32722.

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

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

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

Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357. PMLR, 2021a.

Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 32–42, 2021b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Zhipeng Wei, Jingjing Chen, Micah Goldblum, Zuxuan Wu, Tom Goldstein, and Yu-Gang Jiang. Towards transferable adversarial attacks on vision transformers. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(3):2668–2676, 2022. doi: 10.1609/aaai.v36i3.20169.

Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with ResNets. In *International Conference on Learning Representations*, 2020.

Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2730–2739, 2019.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1492–1500, 2017.

Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9981–9990, 2021.

Jianping Zhang, Yizhan Huang, Weibin Wu, and Michael R. Lyu. Transferable adversarial attacks on vision transformers with token gradient regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16415–16424, 2023.

Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision*, pp. 452–467, 2018.

Zhiyu Zhu, Xinyi Wang, Zhibo Jin, Jiayu Zhang, and Huaming Chen. Enhancing transferable adversarial attacks on vision transformers through gradient normalization scaling and high-frequency adaptation. In *International Conference on Learning Representations*, 2024.