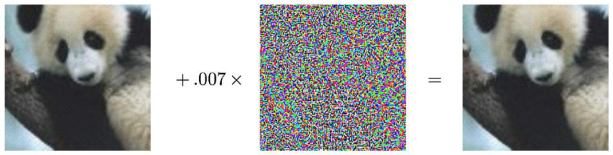
Harnessing the Computation Redundancy in ViTs to Boost Adversarial Transferability

Jiani Liu[1], Zhiyuan Wang[2], Zeliang Zhang[1], Chao Huang[1], Susan Liang[1], Yunlong Tang[1], Chenliang Xu[1]
[1]University of Rochester [2]University of California, Santa Barbara

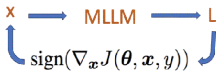
Background & Motivation:

- **Adversarial examples**: imperceptible perturbations that fool deep neural networks (DNNs), threatening safety-critical tasks.

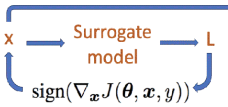


- **White Box vs. Black Box**

—— Why does transferability matters?



White Box: Expensive



Black Box: Cheap

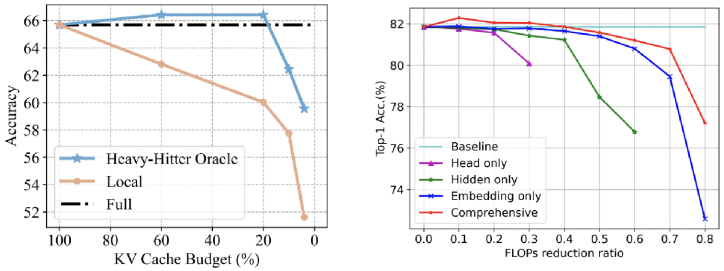
- **How to choose surrogate Model?**

——CNN vs. ViT:

- CNNs: use local convolution filters → strong spatial bias but limited global context.

- ViTs: tokenize images into patches, process them with *multi-head self-attention* and stacked transformer blocks → capture long-range dependencies.

- **How to improve the transferability?** *——The Key: ViT Redundancy*



- *Data-level*: overlapping tokens.
- *Model-level*: redundant attention heads, over-parameterized neurons.

Key idea: exploit this redundancy to enhance adversarial transferability

Our Approach:

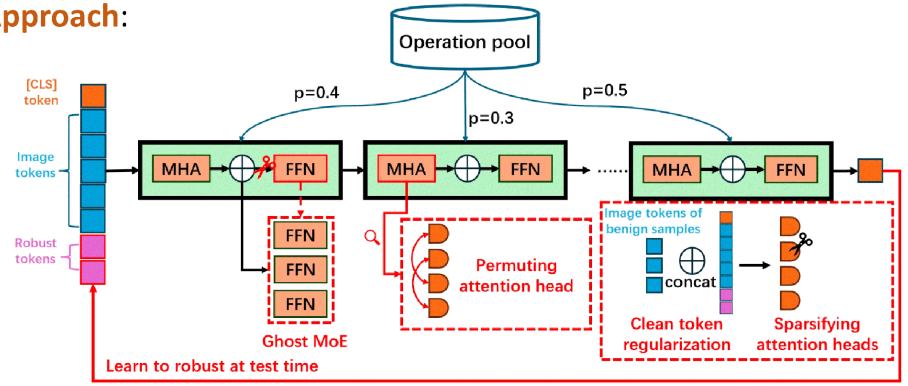


Figure 1: Overview of the proposed attack strategy integrated into Vision Transformers (ViTs). Our method adopts a policy gradient-based framework to selectively apply different operations from an operation pool to each transformer block. These operations include permuting attention heads, sparsifying them, clean token regularization, and activating auxiliary Ghost MoE branches to exploit the computational redundancy within ViTs. Robust tokens are learned at test time to further enhance adversarial transferability.

Result:

Table 2: Our method achieves state-of-the-art performance in attacking diverse models using ViT variants (ViT-B/16, PiT-B, Swin-T) as surrogates. ViT-specific attacks such as TGR, GNS, and FPR are excluded for Swin-T due to unavailable implementations.

Surrogate	Method	RN-50	VGG-16	MN-V2	Inc-v3	ViT-B/16	PiT-B	Vis-S	Swin-T	Avg.
ViT-B/16	MI-	39.4	58.4	57.9	42.2	97.4	40.4	42.0	55.0	54.1
	NI-	40.3	59.2	58.3	44.2	96.8	41.1	44.3	57.4	55.2
	EMI-	57.7	69.7	69.2	60.8	99.3	60.8	65.5	75.4	69.8
	VMI-	50.3	63.6	63.2	52.7	98.3	55.7	57.4	68.1	63.7
	PGN	68.9	75.7	76.3	72.4	97.6	75.6	75.5	80.0	77.8
	DTA	43.5	65.5	64.1	48.0	**99.9**	46.3	49.4	62.1	59.8
	TGR	53.4	72.5	72.4	55.5	97.7	59.2	61.8	74.5	68.4
	GNS	47.5	68.2	68.2	49.6	91.5	50.1	54.8	65.4	61.9
	FPR	52.3	66.6	68.4	52.4	97.5	56.2	60.7	71.0	65.6
	Ours	**77.7**	**90.6**	**91.1**	**79.9**	99.7	**78.9**	**83.5**	**93.5**	**86.9**
PiT-B	MI-	39.4	58.9	56.0	38.7	26.6	95.4	44.6	48.0	44.6
	NI-	39.7	60.4	58.4	37.3	26.0	94.2	45.8	49.4	45.3
	EMI-	58.2	71.6	72.2	57.4	46.0	98.7	66.1	69.6	63.0
	VMI-	54.2	66.7	66.9	55.1	47.2	95.6	61.5	63.2	59.2
	PGN	71.4	77.5	78.4	73.0	69.4	93.9	77.1	79.0	75.1
	DTA	48.6	67.8	67.5	46.4	34.7	**99.9**	54.9	58.4	54.0
	TGR	59.6	78.2	78.8	57.6	49.5	98.2	68.7	71.6	70.3
	GNS	58.9	78.8	77.8	58.8	46.1	98.6	68.9	71.3	69.9
	FPR	58.3	77.5	75.1	67.8	46.1	96.4	64.4	68.6	69.3
	Ours	**86.0**	**91.7**	**93.6**	**81.0**	**74.1**	99.7	**91.7**	**93.4**	**87.4**
Swin-T	MI-	28.8	48.1	52.8	28.8	21.3	27.0	34.1	95.7	42.1
	NI-	30.5	49.5	53.9	28.6	19.8	28.0	34.8	96.4	42.7
	EMI-	42.2	62.4	67.8	42.3	32.4	42.9	52.2	**99.7**	55.2
	VMI-	49.9	61.3	68.1	48.8	46.3	54.1	60.2	97.7	60.8
	PGN	78.5	86.8	87.8	81.8	77.7	83.4	86.9	99.3	85.3
	DTA	31.7	53.0	57.8	29.7	20.6	27.4	35.1	99.5	44.3
	Ours	**85.2**	**90.1**	**91.5**	**89.6**	**85.4**	**88.3**	**92.4**	98.2	**88.9**

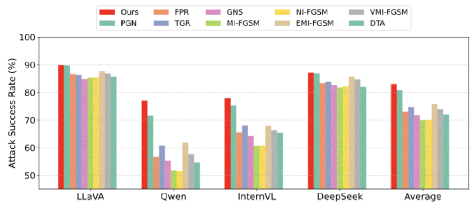


Figure 6: Attack Success Rate on LLaVA-v1.5-7B, Qwen 2.5-VL-7B-Instruct, InternVL 2.5-8B and DeepSeek-VL-7B-Chat. The adversarial examples are generated on ViT-B/16.

State-of-the-art

[1] Zhenyu Zhang, et al. "H_2O: Heavy-Hitter Oracle for Efficient Generative Inference of Large Language Models." NeurIPS 2023.
[2] Chuanyang Zheng, et al. "SAViT: Structure-Aware Vision Transformer Pruning via Collaborative Optimization."NeurIPS 2022.
[3] Jiani Liu, et al. "Harnessing the Computation Redundancy in ViTs to Boost Adversarial Transferability." NeurIPS 2025.