Appendix for "Improving Adversarial Transferability on Vision Transformers via Forward Propagation Refinement"

A. Attack on Online Models

Table 1 presents the Attack Success Rates (ASRs) on two commercial APIs—Baidu Cloud API and Aliyun API—using MIM combined with three advanced surrogate refinement methods: TGR, GNS, and our FPR, evaluated on 50 originally correctly classified images. Among these, FPR attack demonstrates a clear advantage. Specifically, it achieves the highest ASR of 86% on the Baidu Cloud API and 80% on the Aliyun API, outperforming both TGR and GNS. This highlights the superior effectiveness of FPR in cheating commercial models.

Attack (ViT-B)	Baidu Cloud API	Aliyun API
TGR	0.78	0.78
GNS	0.76	0.74
FPR	**0.86**	**0.80**

Table 1. ASRs on two commercial APIs are calculated on 50 originally correctly classified images.

B. More Ablation Studies about Index Set

B.1. Index Set among Different Methods

For fairness, we conduct an ablation study on the index set I and report the performance of various methods using the same index set. Table 2 shows that both TGR and GNS experience a performance drop when using the index set from our method. Additionally, while the performance of AMD improves with the selected index set, it performs worse when applied to all blocks, as AMD makes more significant modifications to the forward propagation. On the other hand, MTE is relatively mild in its modifications, which is why it performs better when applied to all blocks during forward propagation.

Attack (ViT-B)	CNNs	ViTs
TGR (All, Default)	76.2	77.6
TGR (0,1,4,9,11)	60.7	65.6
GNS (All, Default)	74.6	80.0
GNS (0,1,4,9,11)	60.5	64.9
AMD (0,1,4,9,11, Default)	78.4	82.1
AMD (ALL)	47.0	34.1
MTE (All, Default)	79.6	78.4
MTE (0,1,4,9,11)	59.8	63.6
FPR (AMD+MTE)	**83.8**	**84.7**

Table 2. ASRs on two commercial APIs are calculated on 50 originally correctly classified images.

B.2. Sensitivity of Index Set

We analyze the sensitivity of index set used in AMD in Table 3, we can see that block selection is important since our optimal setting largely outperforms the random selection. In addition, the number of selected indexes also matters, since removing any block from our optimal setting slightly decreases the attack performance.

B.3. How to Find the Optimal Index Set

Based on our empirical findings, initially, we suggest starting from block 0 as the baseline. Then, AMD should be applied every 3-5 blocks, as this range tends to provide a balance between the overfitting (increrasing the diversity of attention map) and underfitting (introducing much unfamiliar information to the attention map). After this initial selection, we recommend fine-tuning the index set by replacing each current block with adjacent blocks, as this allows for more targeted adjustments while maintaining the diversity of the perturbations. This iterative refinement process helps in identifying the optimal index set for adversarial transferability.

C. Visualization of Adversarial Examples

Fig 1 provides the illustration of some crafted adversarial examples. It can be observed that, at first, the clean image can be classified with high confidence. However, their corresponding adversarial examples can effectively fool the target models.

Model	Index Set	CNNs	ViTs
ViT-B	Random	70.7	67.7
	1,4,9,11	81.8	83.4
	0,4,9,11	79.7	81.9
	0,1,9,11	83.7	84.4
	0,1,4,9	82.0	83.0
	0,1,4,11	83.5	84.6
	0,1,4,9,11 (Default)	**83.8**	**84.7**
CaiT-S	Random	80.2	84.8
	14,25	84.8	89.8
	2,25	84.3	89.2
	2,14	83.6	89.0
	2,14,25 (Default)	**85.8**	**90.3**
PiT-T	Random	76.0	55.0
	6,11	81.1	63.1
	1,11	81.3	64.6
	1,6	81.1	64.4
	1,6,11 (Default)	**82.4**	**64.7**
DeiT-B	Random	85.9	89.3
	1,5,10,11	85.9	91.0
	0,5,10,11	87.0	91.0
	0,1,10,11	87.0	91.4
	0,1,5,11	87.7	92.2
	0,1,5,10	87.5	92.0
	0,1,5,10,11 (Default)	**87.9**	**92.2**

Table 3. Sensitivity analysis about the index set used in AMD. "Random" refers to randomly selecting the same number of blocks as in AMD.

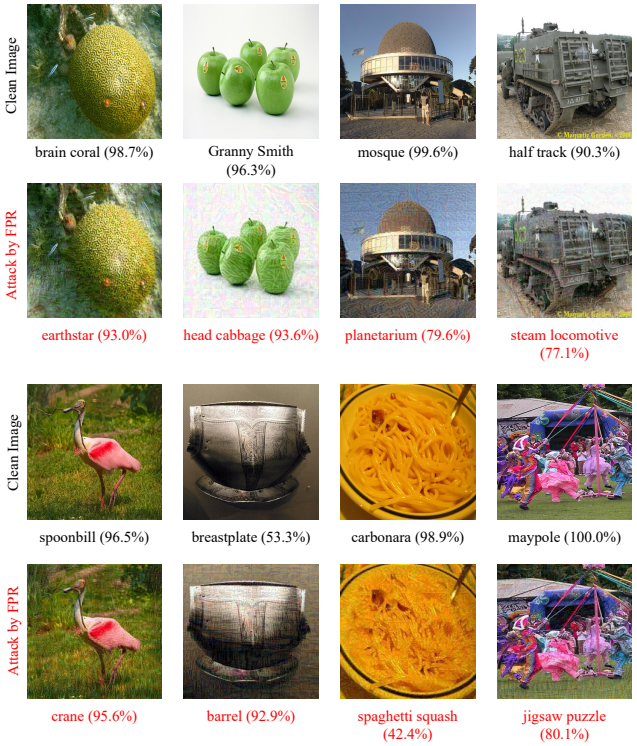

Figure 1. Visualizations for the crafted adversarial examples. Here the substitute model is ViT-B and the target model is RN-18.