Supplementary Material for
Improving Transferable Targeted Attacks with Feature Tuning Mixup

In the supplementary material, we provide additional experimental results on parameter p in Chapter A.1. Then, in Chapters A.2, A.3, and A.4, we present additional results demonstrating the effectiveness and efficiency of FTM across different surrogate and targeted models. Finally, Chapter A.5 includes visualizations of adversarial examples generated by our FTM.

A. Additional Experiments

A.1. Supplementary results on the analysis of p

We have shown in Section 4.3 (Ablation Study) that the attack success rate when using clean feature mixup drops to near 0 in Figure 6 due to the low clean accuracy of the perturbed surrogate model. Table 4 shows the clean accuracy of the perturbed surrogate model with varying values of p. The results demonstrate that increasing p leads to extremely low clean accuracy when using clean feature mixup in FTM, which explains the rapid decline of the blue line in Figure 6. In contrast, without clean feature mixup, the clean accuracy remains stable, maintaining 65.3% even at $p = 1.0$. Consequently, the corresponding orange line in Figure 6 exhibits relatively stable behavior. Notably, as shown in the left plot of Figure 6 our FTM performs stably around the optimal hyperparameter p and α_{max}.

Ablation	Parameter p									
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
w/ clean feature mixup	51.9	28.3	14.0	10.8	8.2	6.0	5.4	4.9	4.3	3.8
w/o clean feature mixup	79.8	73.3	69.7	68.6	67.0	66.4	66.0	65.8	65.4	65.3

Table 4. Accuracy (%) of clean images on the perturbed surrogate model with varying parameter p.

A.2. Attacking black-box multimodal LLMs

FTM can be used to attack multimodal LLMs. We use four commercial LLMs for evaluation, including Qwen2-VL, Llama-3.2, Claude-3.5, and GPT-4o. We randomly select 100 images from the ImageNet-compatible dataset and generate targeted adversarial examples on ViT using RDI-FTM-E. For each generated targeted adversarial example, we use the prompt "Is this image a photo of {target label}? Yes or No?" to obtain the predictions of the LLMs. Table 5 shows that our method achieves an average attack success rate of 40.5%.

Response	Qwen2-VL	Llama-3.2	Claude-3.5	GPT-4o	Avg
Total	100	100	100	100	100 %
Refuse to Answer	0	10	0	0	2.50 %
Uncertain	1	5	1	0	1.75 %
Attack Failed	52	42	54	73	55.25 %
Attack Succeeded	47	43	45	27	40.50 %

Table 5. Evaluation on multimodal LLMs. The targeted adversarial examples are generated on ViT using RDI-FTM-E.

A.3. Evaluation with different surrogate models

We report the targeted attack success rates when using DN-121 or LeViT as the surrogate model in Table 6. The results demonstrate that, across all black-box attack scenarios, our methods consistently outperform existing approaches, regardless of whether DN-121 or LeViT is used as the surrogate model.

We report the targeted attack success rates using ensemble-based surrogate models in Table 7. We use two settings for the ensemble-based surrogate models: RN-50 + Inc-v3 and RN-50 + LeViT. The results show that our methods can be combined with current ensemble-based methods to further improve the attack success rates.

A.4. Efficiency of FTM on attacking ViT

We show that FTM remains effective and efficient when using ViT as the surrogate model. Unlike other models in Section 4.1, ViT's feature map size does not decrease with depth. As analyzed in Section 3.3, FTM is theoretically efficient and adaptable to various architectures, including ViT. Tables 8 and 9 confirm that FTM maintains low computational complexity when using ViT as the surrogate model. Our RDI-FTM-E-SI$_{m_1=2}$ generates adversarial examples in just 3.82 seconds on average while significantly outperforming existing attacks. In RDI-FTM-E-SI$_{m_1=2}$, the two copies of the surrogate model in FTM-E use scaled inputs (1 and 0.5, respectively), preserving efficiency. The parameter m_1 in RDI-FTM-E-SI$_{m_1=2}$ is set to 2 to correspond with the two copies of the surrogate model used in FTM-E. Overall, our results validate FTM's effectiveness and efficiency across different architectures.

A.5. Visualization of targetd adversarial examples

Figure 7 and Figure 8 present visualizations of adversarial examples crafted by different targeted transfer-based attack methods.

Attack	DN-121 ⇒											
	VGG-16	RN-50	Inc-v3	DN-121*	IR-v2	Inc-v4	Xcep	ViT	LeViT	ConViT	Twins	PiT
DI	37.1	44.4	7.1	**98.7**	4.3	8.3	5.2	0.2	3.0	0.4	1.0	1.1
RDI	42.1	55.7	20.8	98.5	12.8	18.8	10.1	0.8	8.5	1.3	3.7	4.5
RDI-Admix	53.2	67.6	31.1	98.3	20.1	26.5	17.8	1.0	14.7	1.7	6.8	7.4
RDI-Admix$_5$	49.6	65.3	41.0	98.6	28.9	34.3	21.6	2.4	21.8	3.4	10.5	14.2
RDI-SI	45.4	60.1	34.3	98.6	22.0	25.8	16.1	2.0	16.1	2.4	8.2	11.7
RDI-VT	47.7	62.1	31.5	98.6	25.4	27.2	20.3	2.2	19.2	3.5	8.3	11.7
RDI-ODI	64.2	71.7	52.8	98.0	39.8	45.9	31.4	3.3	26.9	7.4	14.7	21.9
RDI-CFM	76.2	83.9	56.1	97.8	43.6	53.8	41.1	3.6	32.8	6.4	17.3	21.1
RDI-FTM	77.3	85.6	62.8	98.2	46.9	58.4	46.0	3.7	40.0	8.2	21.6	26.5
RDI-FTM-E	**79.5**	**87.6**	**64.6**	98.0	**48.7**	**60.1**	**47.6**	**4.2**	**43.4**	**8.9**	**24.4**	**26.7**

Attack	LeViT ⇒											
	VGG-16	RN-50	Inc-v3	DN-121	IR-v2	Inc-v4	Xcep	ViT	LeViT*	ConViT	Twins	PiT
DI	1.6	2.5	4.2	2.7	1.8	2.6	2.3	0.6	**100**	4.2	9.3	10.3
RDI	3.0	3.5	5.4	4.1	3.6	4.6	2.3	1.1	**100**	7.9	13.8	22.1
RDI-Admix	6.5	8.3	9.9	8.8	5.8	7.2	5.5	3.3	**100**	10.7	23.3	30.9
RDI-Admix$_5$	5.3	8.1	13.9	12.4	9.9	8.4	7.2	8.0	99.9	20.6	30.0	47.2
RDI-SI	3.4	6.3	10.6	10.0	6.4	5.4	5.1	4.3	**100**	18.1	24.1	38.4
RDI-VT	5.3	7.2	12.0	10.1	8.3	8.8	8.9	6.3	99.9	18.7	27.2	40.5
RDI-ODI	21.0	25.0	40.9	38.3	25.7	31.2	26.3	15.3	98.7	34.1	43.8	66.1
RDI-CFM	27.3	30.3	39.8	39.0	23.6	30.1	27.2	18.4	**100**	45.5	63.8	75.7
RDI-FTM	41.3	41.9	56.9	54.2	39.8	46.5	42.2	31.1	99.9	63.2	77.5	86.7
RDI-FTM-E	**50.1**	**52.4**	**62.8**	**63.1**	**48.4**	**53.2**	**49.2**	**40.3**	99.9	**72.2**	**86.2**	**93.0**

Table 6. Targeted attack success rates (%) using DN-121 or LeViT as surrogate model. All methods are combined with MI-TI. The best results are shown in bold, and the second best results are underlined. The surrogate models are marked with * in the first column.

Attack	RN-50 + Inc-v3 ⇒											
	VGG-16	RN-50*	Inc-v3*	DN-121	IR-v2	Inc-v4	Xcep	ViT	LeViT	ConViT	Twins	PiT
DI	63.5	**99.0**	**99.2**	77.0	16.6	24.0	12.5	0.2	7.3	0.7	3.1	2.6
RDI	66.9	98.4	98.7	83.1	35.6	42.5	25.4	0.8	22.4	3.2	9.8	10.8
RDI-SI	70.1	98.0	98.9	89.3	53.3	57.3	40.0	4.5	40.7	8.8	21.4	27.4
RDI-VT	66.4	98.4	98.8	83.1	51.6	56.6	40.7	5.0	39.6	8.4	21.6	22.7
RDI-Admix	72.9	98.4	99.1	88.0	46.3	58.3	36.5	1.8	33.4	4.5	15.4	17.1
RDI-ODI	70.9	94.0	96.1	83.4	65.4	69.9	57.6	9.1	52.9	18.6	28.6	42.8
RDI-CFM	86.1	98.3	97.0	92.1	71.7	79.9	72.0	10.3	66.5	20.3	41.9	45.8
RDI-FTM	87.5	97.8	96.7	92.4	75.4	81.7	74.0	14.5	71.0	25.2	49.3	54.4
RDI-FTM-E	**88.2**	98.1	97.1	**93.4**	**76.7**	**83.4**	**77.1**	**18.4**	**74.9**	**30.1**	**53.4**	**58.0**

Attack	RN-50 + LeViT ⇒											
	VGG-16	RN-50*	Inc-v3	DN-121	IR-v2	Inc-v4	Xcep	ViT	LeViT*	ConViT	Twins	PiT
DI	71.3	**99.2**	29.5	80.3	16.2	23.7	15.4	1.2	**100**	6.7	18.2	20.8
RDI	75.3	98.7	55.7	87.8	35.1	41.9	29.6	4.2	99.4	14.6	34.3	41.1
RDI-SI	78.1	98.9	74.0	92.8	53.1	57.4	45.5	10.6	99.9	30.5	52.2	67.1
RDI-VT	77.4	98.9	66.4	88.1	50.7	56.0	48.4	14.6	99.1	30.3	53.9	60.7
RDI-Admix	82.3	98.8	64.8	91.2	43.6	53.9	41.0	7.6	99.9	17.4	43.9	48.8
RDI-ODI	81.9	96.6	81.6	89.6	69.9	73.4	67.9	27.8	97.4	52.7	66.5	81.0
RDI-CFM	88.5	98.8	84.1	92.1	69.5	78.7	72.1	26.4	99.8	51.6	77.5	81.3
RDI-FTM	89.2	98.4	86.4	93.0	74.2	81.0	77.1	41.0	99.5	66.1	85.5	89.2
RDI-FTM-E	**90.0**	98.1	**86.7**	**93.9**	**76.7**	**82.4**	**80.1**	**51.5**	99.7	**72.6**	**87.9**	**90.8**

Table 7. Targeted attack success rates (%) using ensemble-based surrogate models. All methods are combined with MI-TI. The best results are shown in bold, and the second best results are underlined. The surrogate models are marked with * in the first column.

Source	Attack	Time (s)	VGG-16	RN-18	RN-50	DN-121	Xcep	MB-v2	EF-B0	IR-v2	Inc-v3	Inc-v4	**Avg.**
ViT	RDI	1.76	2.4	2.4	3.1	5.2	3.5	3.6	10.6	4.4	5.3	4.3	4.5
	RDI-VT	10.20	2.5	3.6	4.1	7.1	5.9	4.1	13.4	7.4	5.5	5.4	5.9
	RDI-Admix	5.11	5.4	5.4	6.9	11.5	6.7	6.9	18.7	8.1	10.1	8.0	8.8
	RDI-SI	8.38	4.2	8.2	9.5	17.6	10.2	9.1	24.7	13.3	18.0	12.1	12.7
	RDI-ODI	4.53	10.2	12.7	14.3	22.1	18.6	12.1	32.2	22.5	22.3	20.7	18.8
	RDI-CFM	1.80	10.8	14.2	14.9	21.0	14.9	15.0	31.9	15.0	18.9	17.1	17.4
	RDI-FTM	1.92	11.6	15.1	14.6	22.1	16.7	18.6	31.8	15.8	20.5	18.0	18.5
	RDI-FTM-E	3.81	13.4	18.1	18.3	26.5	19.6	19.5	37.2	19.8	23.0	21.3	21.7
	RDI-FTM-E-SI$_{m_1=2}$	3.82	**24.0**	**29.1**	**30.0**	**41.5**	**28.9**	**31.3**	**54.7**	**30.9**	**39.3**	**31.8**	**34.2**

Table 8. Targeted attack success rates (%) against ten target models, with ViT as the surrogate model. All methods are combined with MI-TI. The best results are shown in bold, and the second best results are underlined. Time (s) denotes the average computation time required to attack a single image. RDI-FTM-E-SI$_{m_1=2}$ denotes using scaled input for RDI-FTM-E .

Source	Attack	Time (s)	ViT	LeViT	ConViT	Twins	PiT	**Avg.**
ViT	RDI	1.76	**99.5**	24.3	23.3	11.7	26.0	37.0
	RDI-VT	10.20	98.3	32.0	29.7	15.3	35.7	42.2
	RDI-Admix	5.11	99.1	39.0	34.6	18.5	40.5	46.3
	RDI-SI	8.38	98.4	58.4	69.2	38.8	63.8	65.7
	RDI-ODI	4.53	87.6	45.9	42.1	29.5	49.3	50.9
	RDI-CFM	1.80	96.5	47.2	50.1	27.4	48.8	54.0
	RDI-FTM	1.92	94.3	50.5	53.3	31.0	52.6	56.3
	RDI-FTM-E	3.81	96.0	54.9	55.6	36.3	55.8	59.7
	RDI-FTM-E-SI$_{m_1=2}$	3.82	96.3	**71.2**	**74.8**	**54.7**	**74.6**	**74.3**

Table 9. Targeted attack success rates (%) against five transformer-based DNNs, with ViT as the surrogate model. All methods are combined with MI-TI. The best results are shown in bold, and the second best are underlined. Time (s) denotes the average time required to attack a single image. RDI-FTM-E-SI$_{m_1=2}$ denotes using scaled input for RDI-FTM-E .

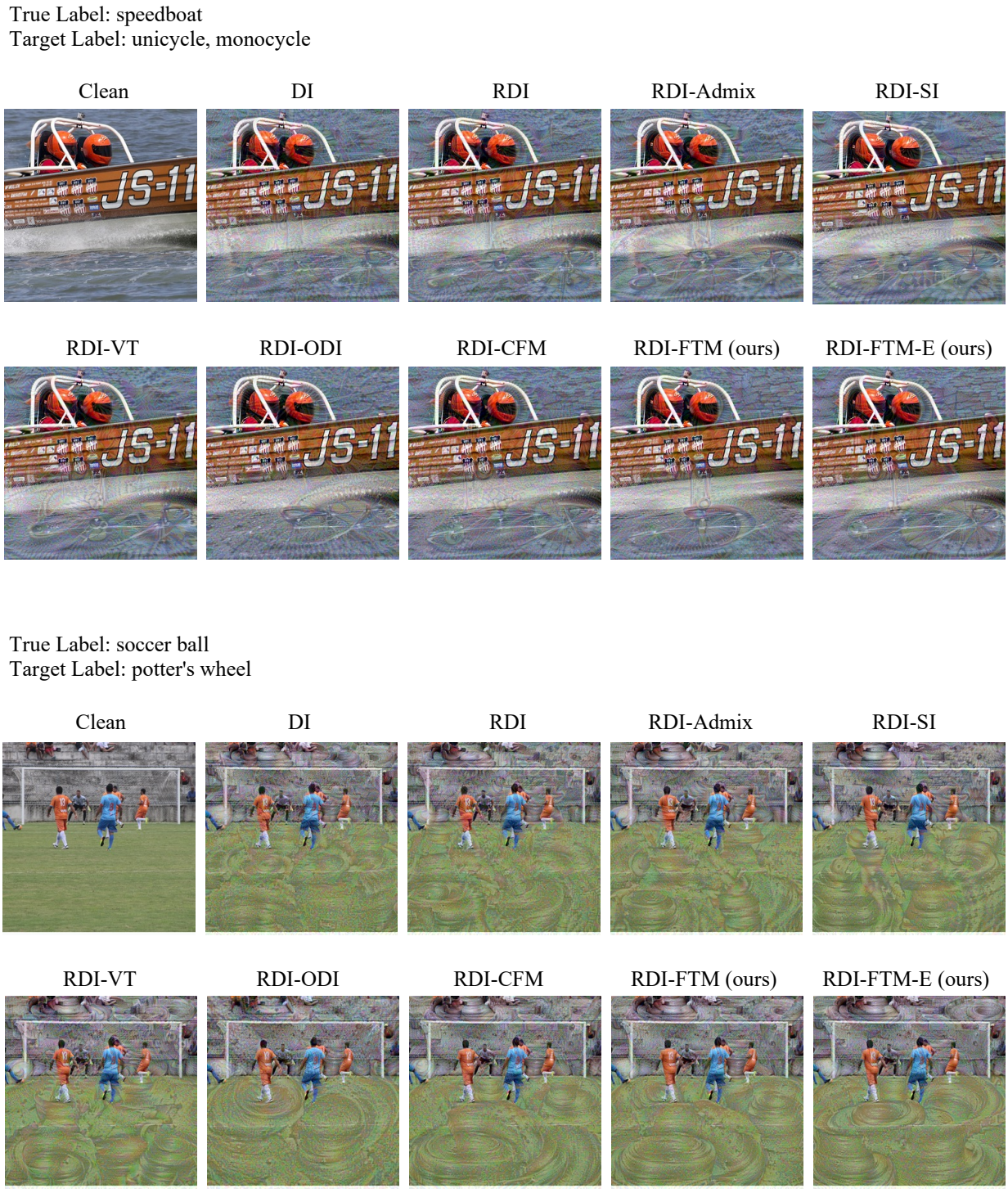

Figure 7. Visualization of targeted adversarial examples generated by different attack methods. The surrogate model used for attack generation is an ensemble of RN-50 and Inc-v3.

Figure 8. Visualization of targeted adversarial examples generated by different attack methods. The surrogate model used for attack generation is an ensemble of RN-50 and LeViT.