Supplementary Materials Module-wise Adaptive Adversarial Training for End-to-end Autonomous Driving

Main Results of Remaining Tasks

In the main text, only the most important planning results are reported. Here, we report the MA^{2}T defense results of UniAD [1] and VAD [2] models on other tasks. Our evaluation metrics are consistent with the metrics originally applied in the models. For UniAD, we evaluate tracking performance using Average Multi-Object Tracking Accuracy (AMOTA↑) and assess map alignment through Intersection over Union (IOU↑) between predicted and ground-truth maps. Motion forecasting precision is measured by Minimum Average Displacement Error (minADE↓), while occupancy accuracy is also evaluated using IOU (↑). For VAD, we use mAP (↑) (mean Average Precision) for detection and mapping, and minADE (↓) for motion forecasting like UniAD.

A. Defense results of UniAD.

White-box Results. Tab. I shows UniAD's defense results against white-box attacks among multi-object tracking, mapping, motion forecasting, and occupancy prediction. We compare our MA^{2}T with four traditional adversarial training methods: FGSM adversarial training (F-AT), PGD-ℓ_1 (P_1), PGD-ℓ_2 (P_2), PGD-ℓ_∞ (P_∞) adversarial training. From these experimental results, we can draw the following observations.

❶ Overall, the defense results of MA^{2}T on the remaining tasks of UniAD exceed the traditional four adversarial training methods, with 13 rows showing the best performance.

❷ The robustness improvement effect of MA^{2}T on driving tasks in UniAD varies, among which MA^{2}T achieves the best defense towards occupancy. Except for FGSM attack, MA^{2}T far exceeds traditional adversarial training on the other tasks, and there is only a slight performance decrease of 2.2% when there is no attack.

❸ The track module of UniAD exhibits extremely severe vulnerability, with any attack almost causing the track module to completely crash. The defensive capability of MA^{2}T against track tasks under white box attacks is flawed, but surprisingly, this does not affect its superior performance on downstream tasks.

❹ MA^{2}T exhibits acceptable performance degradation in the clean settings. In both track and map tasks, MA^{2}T performs the best among all defense methods. It is the second best in occupancy and motion, but still superior than the average results.

TABLE I: **UniAD**'s defense results under **white box** settings. The bold cell in each row represents the best performance of that row.

(a) Multi-object tracking (↑)

Method	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
FGSM	0.14	0.55	**0.57**	0.56	**0.57**	0.56	0.09
MI-FGSM	**0.07**	0.06	0.04	0.04	0.01	0.04	0.00
PGD-ℓ_1	0.13	0.18	0.18	**0.26**	0.15	0.19	0.14
PGD-ℓ_2	0.11	0.14	0.14	**0.16**	**0.16**	0.15	0.10
PGD-ℓ_∞	0.05	0.08	0.08	**0.10**	0.03	0.07	0.01
AutoAttack	0.08	0.08	0.05	0.05	**0.10**	0.07	0.06
Clean	**0.58**	0.36	0.34	0.38	0.23	0.33	0.38

(b) Online Mapping (%) (↑)

Method	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
FGSM	20.09	**22.98**	22.39	22.68	22.54	22.65	20.37
MI-FGSM	18.23	18.24	17.31	17.73	17.73	17.50	**19.23**
PGD-ℓ_1	**21.23**	21.01	21.01	20.49	20.64	20.63	21.11
PGD-ℓ_2	20.79	20.74	20.74	20.28	20.39	20.54	**21.06**
PGD-ℓ_∞	18.31	18.97	18.97	18.55	18.49	19.24	**19.43**
AutoAttack	18.96	19.09	19.09	18.82	18.90	18.98	**19.11**
Clean	**23.93**	22.56	22.24	22.40	22.46	22.42	22.77

(c) Motion Forecasting (↓)

Method	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
FGSM	0.99	0.51	0.47	**0.46**	0.57	0.50	0.79
MI-FGSM	1.44	0.93	1.23	1.16	0.87	1.05	**0.85**
PGD-ℓ_1	1.12	0.81	0.81	0.90	**0.74**	0.82	**0.74**
PGD-ℓ_2	1.15	0.83	0.83	0.93	0.76	0.84	**0.75**
PGD-ℓ_∞	1.46	0.93	0.93	1.07	**0.84**	0.94	0.87
AutoAttack	1.44	0.99	1.29	1.18	0.89	1.09	**0.87**
Clean	**0.49**	0.51	0.46	0.46	0.58	0.50	0.57

(d) Occupancy Prediction (%) (↑)

Method	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
FGSM	48.6	63.2	64.7	**64.8**	61.6	63.6	53.7
MI-FGSM	44.0	46.3	44.7	44.9	48.8	46.2	**52.5**
PGD-ℓ_1	49.8	49.9	49.9	48.7	52.8	50.3	**56.6**
PGD-ℓ_2	48.6	49.7	49.7	48.5	52.1	49.0	**56.7**
PGD-ℓ_∞	44.2	47.0	47.0	46.1	49.9	47.5	**53.0**
AutoAttack	44.1	48.9	48.9	47.2	50.1	48.8	**51.0**
Clean	**65.1**	60.6	63.0	62.4	46.6	58.2	62.9

Black-box Results. Tab. II shows UniAD's defense results against black-box attacks among four remaining tasks. We use MA^{2}T trained UniAD as the victim models, with three categories of attack models: the vanilla model with the same architecture, the traditionally adversarial-trained model, and the vanilla model with a different architecture. Based on the results, we can draw the following observations.

TABLE II: **UniAD**'s defense results under **black box** settings. The bold cell in each row represents the best performance of that row.

(a) Multi-object tracking (↑)

Att. Gen.	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
Vanilla	0.21	0.09	0.07	0.15	0.07	0.09	**0.24**
Trad. AT	0.12	0.12	**0.15**	0.14	0.07	0.12	0.04
VAD	**0.14**	0.12	0.07	0.08	0.05	0.08	0.02

(b) Online Mapping (%) (↑)

Att. Gen.	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
Vanilla	19.69	18.99	18.01	18.38	18.54	18.48	**21.29**
Trad. AT	19.05	19.15	18.16	18.19	18.74	18.56	**19.70**
VAD	18.94	18.89	18.08	18.39	18.56	18.48	**19.77**

(c) Motion Forecasting (↓)

Att. Gen.	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
Vanilla	**0.67**	0.93	1.15	1.08	0.82	1.00	0.68
Trad. AT	1.30	0.91	1.18	1.16	**0.82**	1.02	0.84
VAD	1.29	0.92	1.18	1.04	**0.83**	0.99	0.84

(d) Occupancy Prediction (%) (↑)

Att. Gen.	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
Vanilla	**61.3**	47.1	45.6	45.8	49.7	47.1	60.0
Trad. AT	45.8	45.7	47.5	46.7	38.8	44.7	**52.8**
VAD	46.3	47.4	45.4	45.8	49.7	47.1	**52.8**

❶ Under the black box setting, mat also outperforms traditional adversarial training methods, showing the best performance in the face of attacks 6 times, while the other four defense methods only achieve the best performance 3 times in total.

❷ Similar to the results under the white box setting, MA^{2}T has the best robustness enhancement effect on map and occupation, outperforming other four traditional adversarial training methods in every black box attack. However, in black box settings, MA^{2}T performs better in improving track performance than white box attacks, but performs poorly on the motion module.

❸ Among the three types of black box attacks implemented, the attacks generated by the traditional adversarial training model architecture pose the strongest threat to the model compared to the other two types of attacks.

B. Defense results of VAD.

White-box Results. Tab. III shows VAD's defense results against white-box attacks among detection mapping, and motion forecasting. We compare our MA^{2}T with four traditional adversarial training methods like UniAD. From these experimental results, we can draw the following observations.

❶ MA^{2}T's defense towards VAD in upstream tasks is completely different from UniAD. It exhibits excellent defense performance in the first detection task but worse in the motion task, contrary to the module level defense trend of UniAD.

TABLE III: **VAD**'s defense results under **white box** settings. The bold cell in each row represents the best performance of that row.

(a) Detection (↑)

Method	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
FGSM	0.07	0.07	**0.08**	0.07	**0.08**	0.07	**0.08**
MI-FGSM	0.11	0.09	0.10	0.11	0.10	0.10	**0.13**
PGD-ℓ_1	**0.12**	0.10	0.11	0.11	0.10	0.10	**0.12**
PGD-ℓ_2	0.11	0.11	0.11	0.11	0.10	0.11	**0.13**
PGD-ℓ_∞	0.10	0.10	0.10	**0.11**	0.10	0.10	**0.11**
AutoAttack	0.10	0.10	0.10	**0.11**	**0.11**	0.11	**0.11**
Clean	**0.27**	0.22	0.23	0.22	0.22	0.22	**0.27**

(b) Online mapping (↑)

Method	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
FGSM	0.20	0.18	0.14	0.18	0.17	0.17	**0.41**
MI-FGSM	0.51	0.43	0.48	0.46	0.45	0.46	**0.59**
PGD-ℓ_1	0.54	0.47	0.52	0.51	0.49	0.50	**0.58**
PGD-ℓ_2	0.53	0.50	0.52	0.50	0.49	0.50	**0.59**
PGD-ℓ_∞	0.52	0.43	0.49	0.46	0.45	0.46	**0.60**
AutoAttack	0.51	0.52	0.48	0.48	0.54	0.51	**0.59**
Clean	**0.70**	0.56	0.61	0.60	0.59	0.59	0.64

(c) Motion Forecasting (↓)

Method	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
FGSM	17552.96	**9519.38**	35803.48	44512.75	18098.47	26983.52	16309.52
MI-FGSM	**7869.77**	8653.98	16499.73	32943.96	16836.45	18733.53	14977.09
PGD-ℓ_1	15639.54	8519.92	24423.64	32511.19	16666.39	20530.29	**7844.80**
PGD-ℓ_2	8250.12	32404.79	24463.88	32404.79	16499.73	26443.30	**0.51**
PGD-ℓ_∞	**8141.57**	8684.34	16527.27	33276.17	16922.79	18852.64	22916.02
AutoAttack	**8922.74**	10371.20	26507.92	31727.61	20554.69	22290.36	9172.01
Clean	**0.47**	0.50	0.52	0.51	0.53	0.51	0.49

❷ The motion module of VAD performs abnormally under attack settings, with its Minimum Average Displacement Error even showing a huge leap of orders of magnitude, and multiple instances where adversarial training cannot defend against attacks. Our MA^{2}T has achieved a certain degree of defense against ℓ_1 and ℓ_2 attacks. We believe this may be related to vulnerabilities in the motion module of the model itself.

❸ Raw performance without any attacks of MA^{2}T is fantastic, reaching the level of vanilla model in detection, and surpassing other methods in map and motion modules, which are very close to the vanilla level.

Black-box Results. Tab. IV shows VAD's defense results against black-box attacks among four remaining tasks. We use MA^{2}T trained VAD as the victim models, with three categories of attack models: the vanilla model with the same architecture, the traditionally adversarial-trained model, and the vanilla model with a different architecture. Based on the results, we can draw the following observations.

❶ Similar to the white box results, MA^{2}T exhibits excellent defense and robustness enhancement in detection and map, but has shortcomings in the motion modules.

❷ Relatively weak black box attacks also cause significant anomalies in the motion module, but this does not result in anomalies in the plan task (as shown in the main experiments).

TABLE IV: **VAD**'s defense results under **black box** settings. The bold cell in each row represents the best performance of that row.

(a) Detection (↑)

Att. Gen.	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
Vanilla	0.12	0.10	0.10	0.10	0.10	0.10	**0.15**
Trad. AT	0.10	0.10	0.09	0.10	0.09	0.09	**0.12**
UniAD	0.15	0.12	0.13	0.12	0.12	0.12	**0.16**

(b) Online mapping (↑)

Att. Gen.	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
Vanilla	0.46	0.47	0.46	0.46	0.46	0.46	**0.67**
Trad. AT	0.42	0.39	0.40	0.38	0.35	0.38	**0.58**
UniAD	0.70	0.56	0.61	0.60	0.59	0.59	**0.71**

(c) Motion Forecasting (↓)

Att. Gen.	Vanilla	F-AT	P_1	P_2	P_∞	AVG	MA^{2}T(ours)
Vanilla	**7174.1**	8361.59	23608.25	31678.93	16123.52	19943.07	13524.38
Trad. AT	8291.55	17068.63	24504.25	24667.07	16666.4	20726.59	**0.51**
UniAD	**0.46**	0.50	7662.66	7770.91	0.53	3858.65	6565.12

We believe that there may be potential loophole in the motion module of the model that do not flow into downstream tasks.

❸ The black box attack generated by the enhanced model has the strongest attack to VAD for all three tasks, while the attack generated by UniAD has the weakest effect.

MA^{2}T exhibits varying defense performance for upstream tasks in the models, and similar trends are observed for different tasks under white box and black box settings. Although MA^{2}T's defense performance in very rare tasks is not satisfactory, it achieves excellent defense and improvement in the final plan (as shown in the main results) and other most tasks. We believe that from a holistic perspective, each task in the end-to-end autonomous driving model collaborates, but there may be conflicts in the direction of robustness improvement for different tasks. We can sacrifice the performance of some unimportant modules to promote the ultimate key goal planning.

REFERENCES

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *CVPR*, 2023.

[2] B. Jiang, S. Chen, Q. Xu, B. Liao, J. Chen, H. Zhou, Q. Zhang, W. Liu, C. Huang, and X. Wang, "Vad: Vectorized scene representation for efficient autonomous driving," in *ICCV*, 2023.