Table 1. Average time cost (second) for each instruction. All experiments are conducted with I-FSJ jailbreak. The test samples are mixed with 520 normal samples and 520 jailbreak samples.

Method	LLaMA2-7B		Self Defense	TIM		
	Vanilla	w/ Our Detector	Total	Training	Total	Training/Total
Time	7.18	7.21 (+0.3%)	36.13	0.67	5.49	12.2%

Table 2. The ASR (%) evaluated by LLM of MM-SafetyBench with LLaVA-v1.6-Vicuna-7B. We adopt LLaMA3-8B-Instruct as the evaluator. The results of TIM are reported as ASR / ASR-50.

Method	Vanilla	Adashield	VLGuard	TIM
ASR	36.3	2.4	86.1	0.2/0.0

Table 3. Experimental Results under GCG jailbreak attacks.

	ASR	ODR
LLaMA2-7B	21.5	0.2
+TIM	7.7 (-13.8%)	2.7 (+2.5%)

Table 4. Additional Results with Larger Backbone. The results of TIM are reported as ASR / ASR-50.

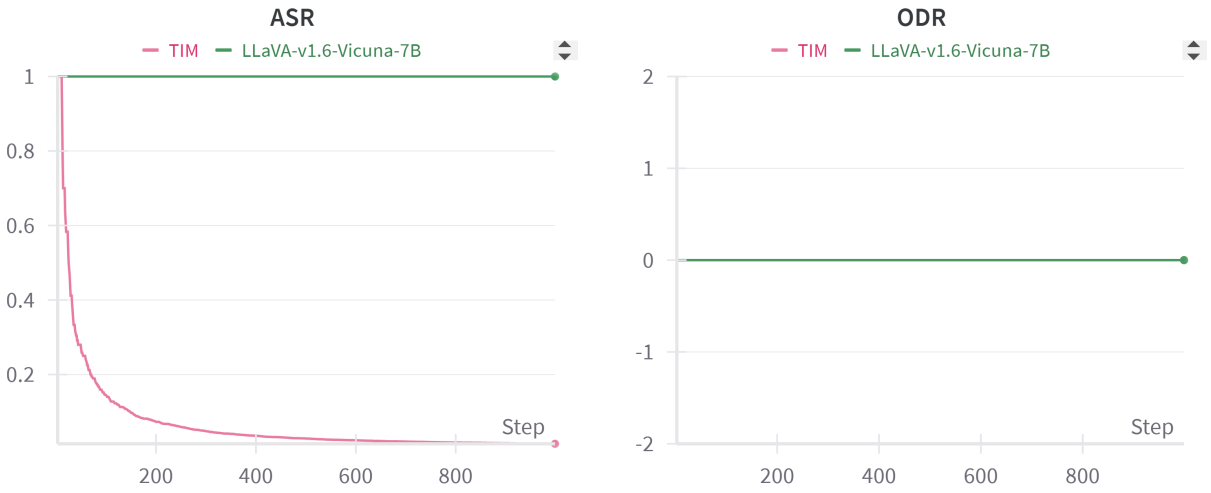
Model	Attack	ASR		ODR	
		Vanilla	TIM	Vanilla	TIM
LLaVA-v1.6-Vicuna-13B	MM-SafetyBench	100	4.8/0.0	0.4	0.4
	Figstep	100	1.8/0.0	0.0	0.4
LLaMA3-8B-Instruct	I-FSJ	94.3	1.0/0.0	0.2	0.2

Table 5. The transferability results. We first adopt TIM on the source jailbreak attack. Then, we freeze the fine-tuned model and evaluate it on the target attack. We report the ASR while adopting the LLaVA-v1.6-Vicuna-7B as the backbone.

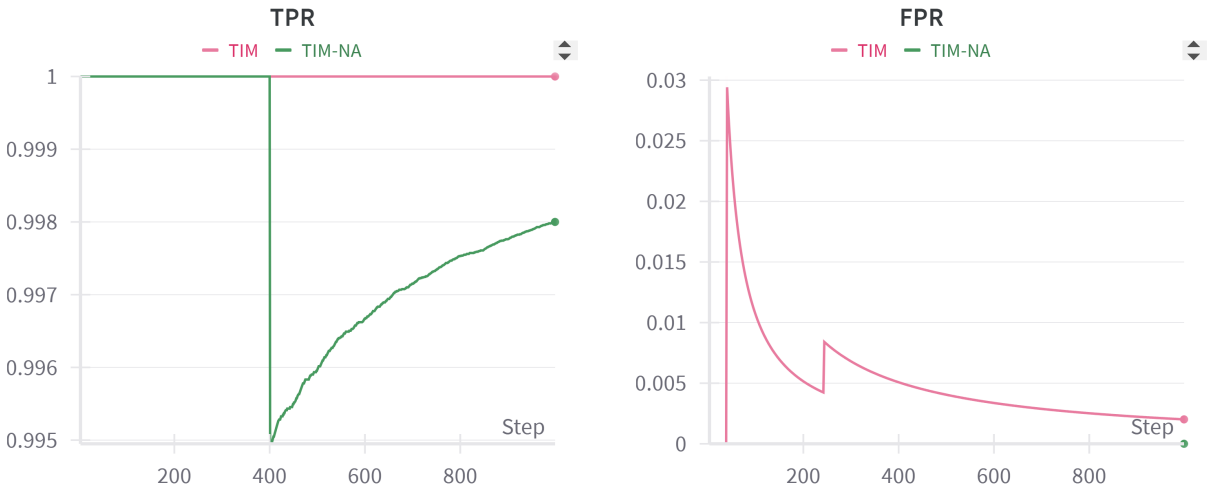
Figstep $\longrightarrow$ MM-SafetyBench	MM-SafetyBench $\longrightarrow$ Figstep
84.3 (-15.5)	0.0 (-100.0)

Table 6. The validation accuracy of the held-out samples from detector training.

Model	LLaVA-v1.6-Vicuna-7B	LLaVA-v1.6-Mistral-7B	LLaVA-v1.6-Vicuna-13B	LLaMA2-7B
Accuracy	100.0	100.0	99.6	99.9



(a) Accumulated ASR

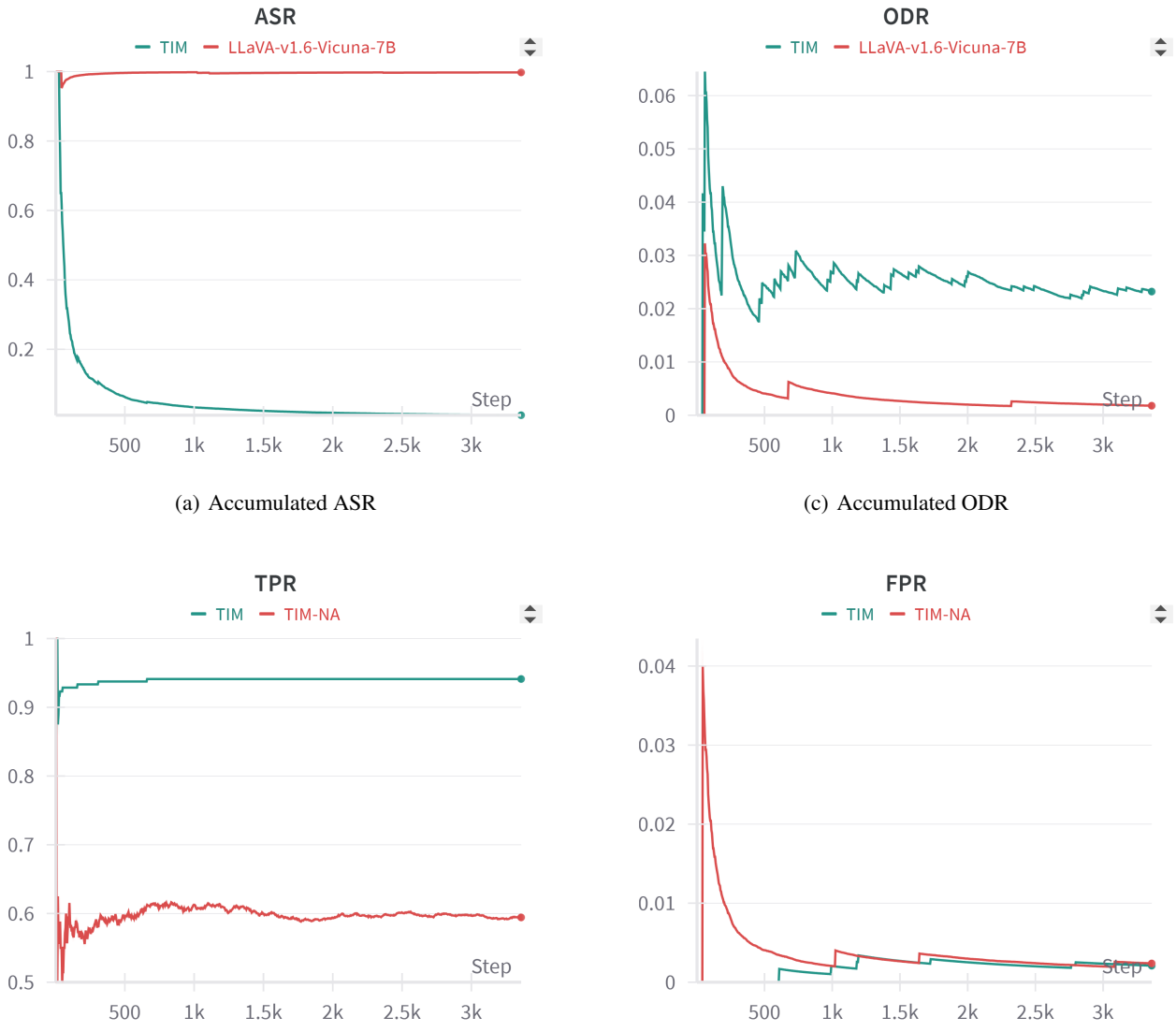


(b) Accumulated TPR

(c) Accumulated ODR

(d) Accumulated FPR

Figure 1. Changes in metrics during the test process against Figstep.



(a) Accumulated ASR

(c) Accumulated ODR

(b) Accumulated TPR

(d) Accumulated FPR

Figure 2. Changes in metrics during the test process against MM-SafetyBench.