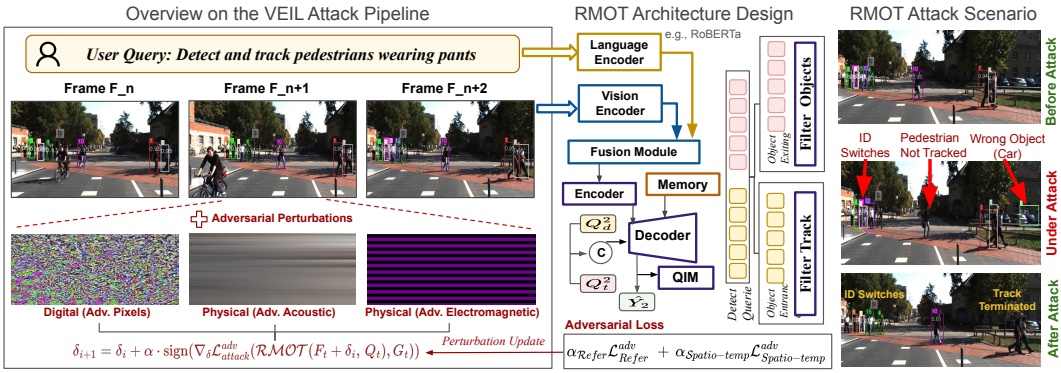
Chasing Shadows: Adversarial Attacks on Referring Multi-Object Tracking Systems

Anonymous



Abstract. Language–vision understanding has driven the development of Referring Multi-Object Tracking (RMOT). However, their security remains underexplored. We examine adversarial vulnerabilities in Transformer-based RMOT, showing that crafted perturbations disrupt both linguistic-visual referring and object-matching components. We introduce VEIL, an adversarial framework that exposes persistent errors in FIFO-based temporal memory and compromises tracking reliability.

Approach. Our method injects carefully crafted digital and physical perturbations in the visual input that propagate into the spatial–temporal reasoning pipeline. VEIL exploits weaknesses in Transformer backbones and FIFO-based memory buffers, leading to cascading tracking errors. The attack design leverages adversarial optimization tailored to RMOT's language–vision alignment, resulting in persistent corruption of both object selection and temporal continuity.

Results. Experiments on the Refer-KITTI dataset show that VEIL significantly degrades RMOT performance. We observe frequent track ID switches, premature terminations, and long-lasting errors persisting across frames. Our results emphasize that robustness against adversarial attacks must be considered a first-class design objective for future multimodal tracking systems.

Table 1: Attack Performance Results across two RMOT models the Refer-KITTI dataset.

Tracker	Attack Strategy	Attack Vector	IDSW ↑	$IDSW_{im}$ ↑	HOTA ↓	AssA ↓	DetA ↓	IDF1 ↓	IDP ↓	IDR ↓
TransRMOT	Clean	–	6.13	0.00	69.66	71.90	65.30	69.54	0.83	0.93
	Adv. Referring	Pixels	9.30 (+3.17)	60.82 –	56.26 (-13.41)	51.86 (-20.03)	59.50 (-5.80)	54.26 (-15.28)	0.64 (-0.19)	0.68 (-0.24)
	Adv. Referring	Physical AAI	8.63 (+2.50)	54.07 –	59.79 (-9.87)	56.69 (-15.21)	61.88 (-3.41)	58.38 (-11.17)	0.67 (-0.15)	0.71 (-0.22)
	Adv. Referring	Physical EAI	8.94 (+2.81)	60.56 –	56.67 (-12.99)	53.32 (-18.58)	59.22 (-6.08)	55.65 (-13.89)	0.67 (-0.16)	0.67 (-0.26)
TempRMOT	Clean	–	0.24	0.00	68.70	67.65	66.60	69.20	0.98	0.98
	Adv. Referring (Spatio-temporal)	Pixels	4.32 (+4.08)	41.07 –	49.89 (-18.81)	46.55 (-21.11)	47.80 (-18.80)	49.99 (-19.21)	0.72 (-0.26)	0.64 (-0.34)
	Adv. Referring (Spatio-temporal)	Physical AAI	2.53 (+2.28)	12.60 –	56.87 (-11.83)	55.69 (-11.97)	55.08 (-11.51)	59.50 (-9.70)	0.89 (-0.10)	0.73 (-0.25)
	Adv. Referring (Spatio-temporal)	Physical EAI	3.20 (+2.96)	14.73 –	51.52 (-17.18)	51.30 (-16.36)	49.72 (-16.88)	55.78 (-13.42)	0.89 (-0.09)	0.70 (-0.28)

Conclusion. We summarize our discovery of design-level vulnerabilities in RMOT systems and presents VEIL as a practical adversarial framework. Our findings call for the integration of security principles into RMOT architectures to ensure safe deployment in real-world scenarios.

Preprint.