Centre for Frontier AI Research
A*STAR CFAR

Advancing Trustworthy and Responsible AI

Resilient & Safe AI

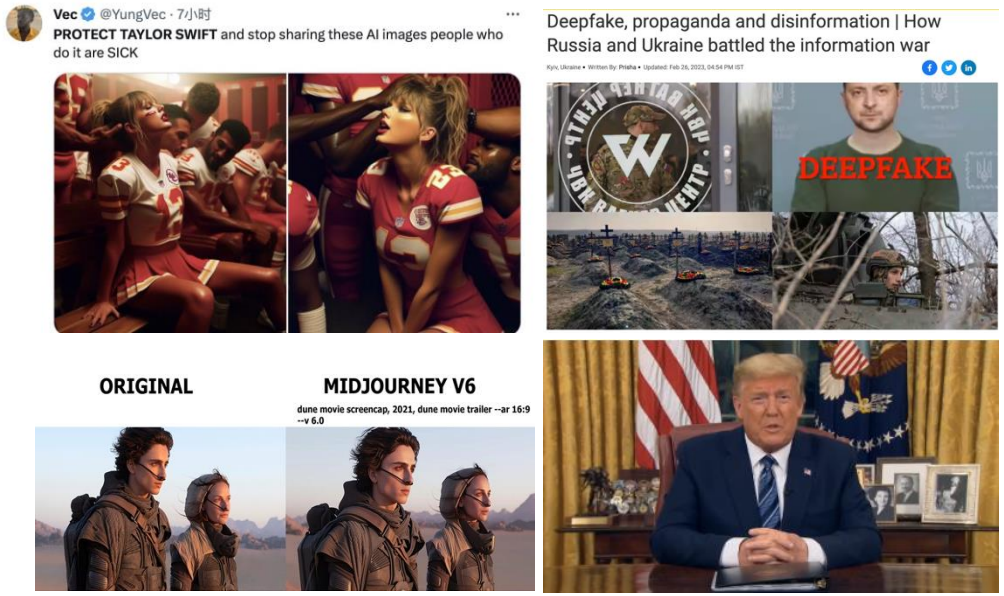
ZHANG Jie
Scientist & Innovation Lead
Aug, 2025

Why Trustworthy AI?

Autonomous Vehicle

Safety Concern

Generative AI



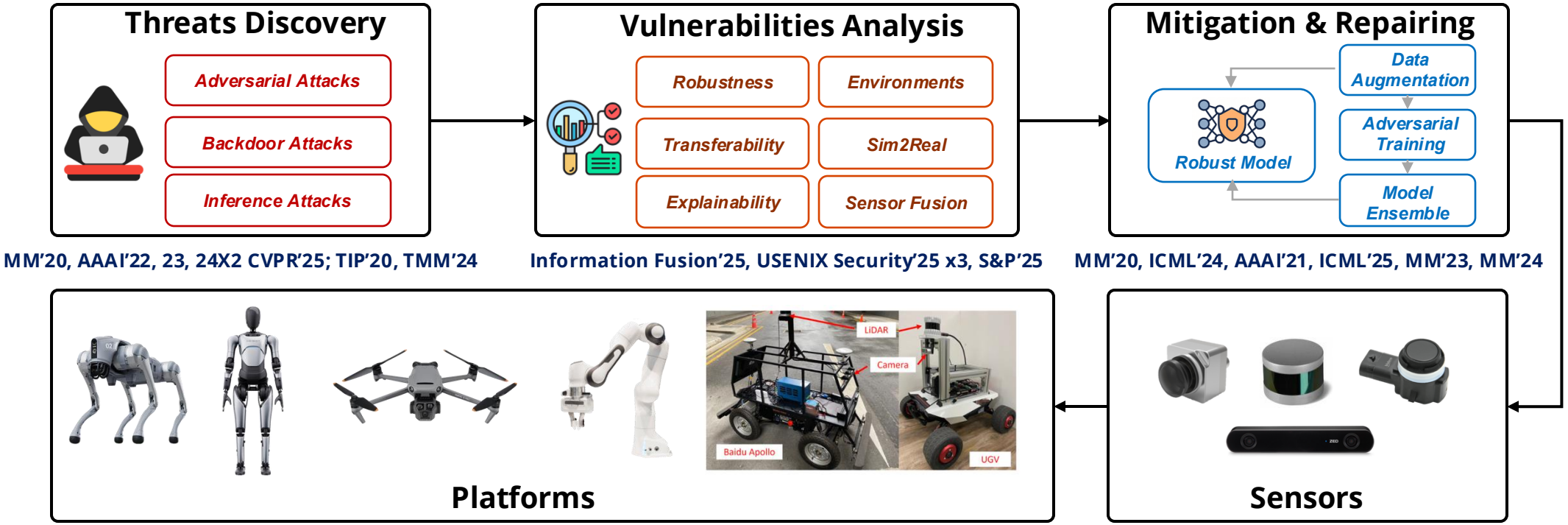
Ethical Challenge

Trustworthy and Responsible AI is no longer optional, it is **essential**.

Build powerful AI that is safe, secure, and truly trustworthy for people and society

Safety & Robustness

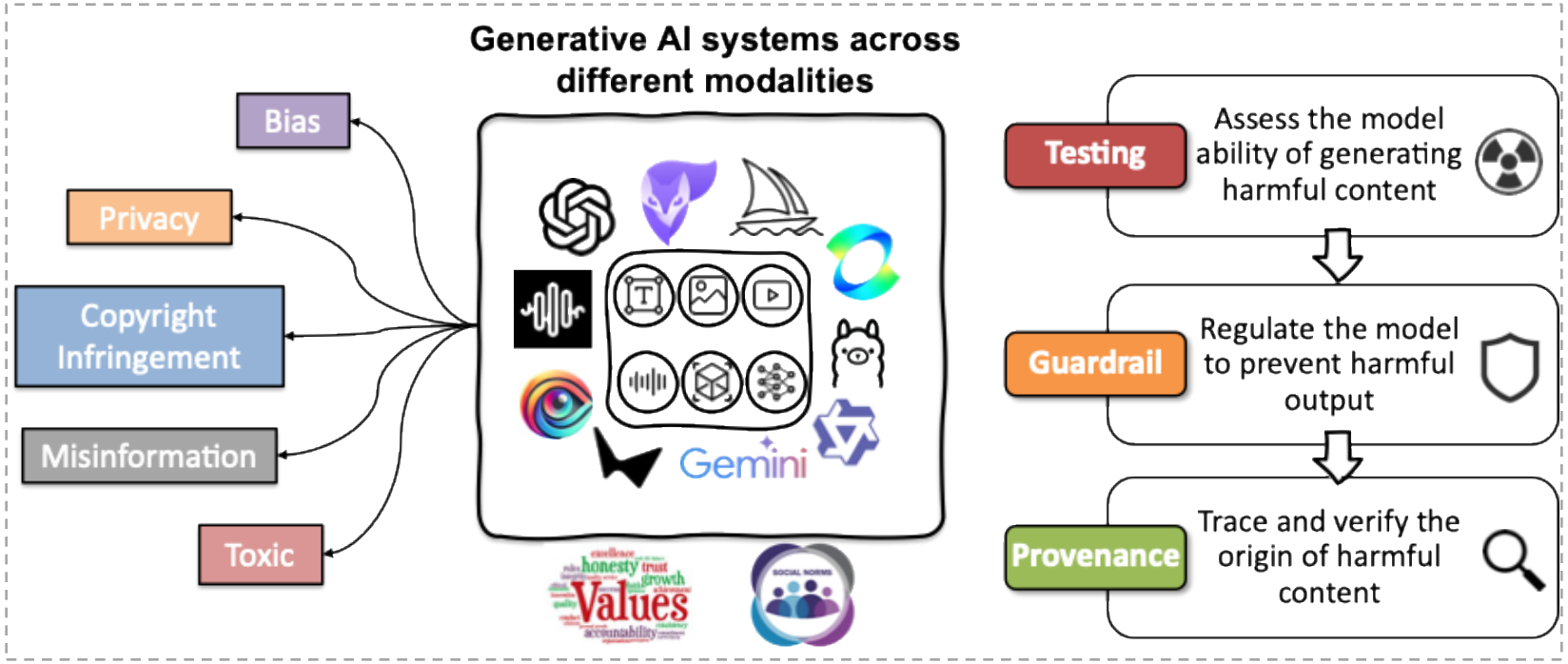
- Build secure and robust AI systems for safety-critical scenarios by proactively **identifying** threats and **mitigating** vulnerabilities.
- From the **digital layer** to the **physical world**, and from **model-level** to **system-level.**



Build powerful AI that is safe, secure, and truly trustworthy for people and society

Ethics & Value Alignment

- Ensure AI systems **reflect human values and social norms** by incorporating ethical principles into model design and deployment.
- Promote alignment through **testing**, **guardrails**, and **provenance** across **different modalities** like video, image, 3D, text, audio, etc.



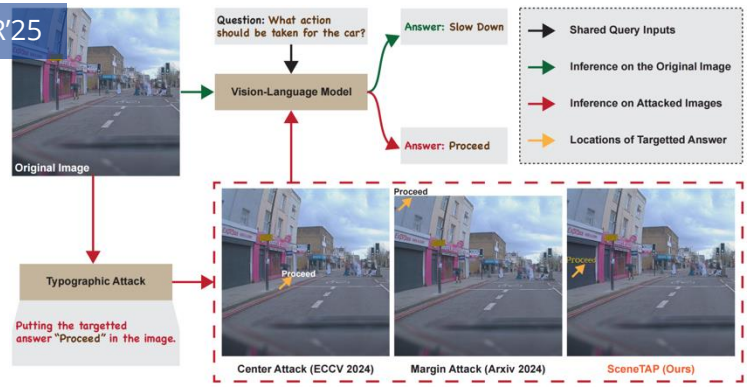
CCS'24, NeurIPS'24, AAAI'24, NAACL'25

IJCAI'24, TDSC'25, TOSEM'25, ICML'25, NDSS'25, ICML'25

AAAI'20, NeurIPS'20, MM'20,TPAMI'21, 25 AAAI'21, AAAI'23, AAAI' 23, TKDE'23, ECCV'24, TPAMI'24, ICML'24, NDSS'24, TIFS'25, NDSS'24, ICLR'25, TDSC'25

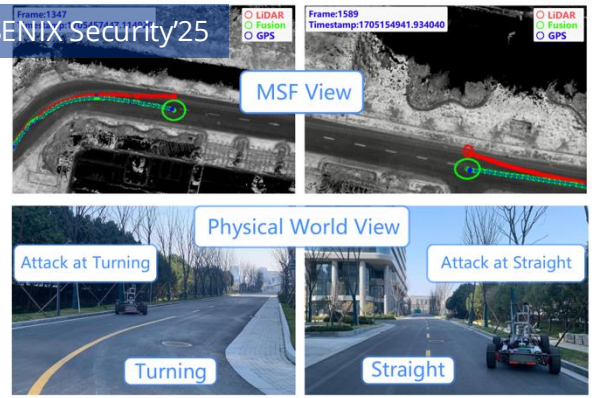
Selected Highlights

CVPR'25



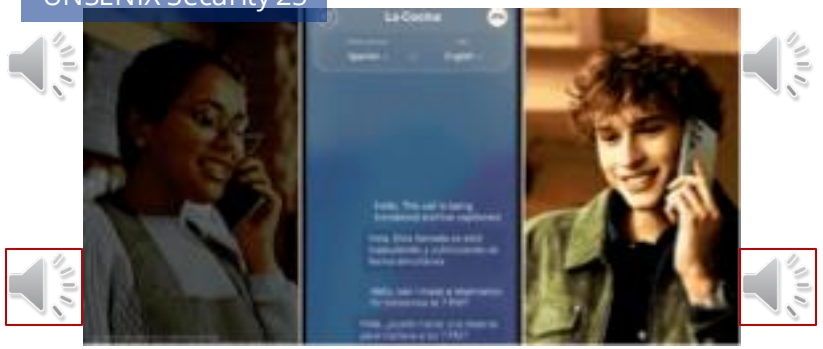
Typographic Attack (First practical attack)

UNSENIX Security'25



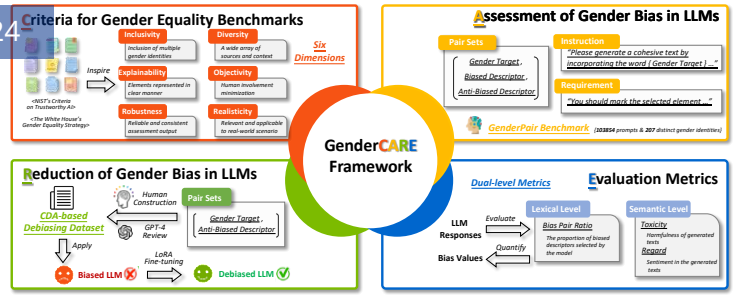
GPS Spoofing Attack (First practical attack)

UNSENIX Security'25



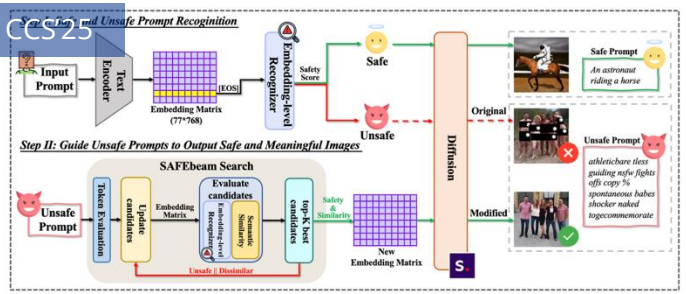
Denial of Translation Attack (First practical attack)

CCS'24



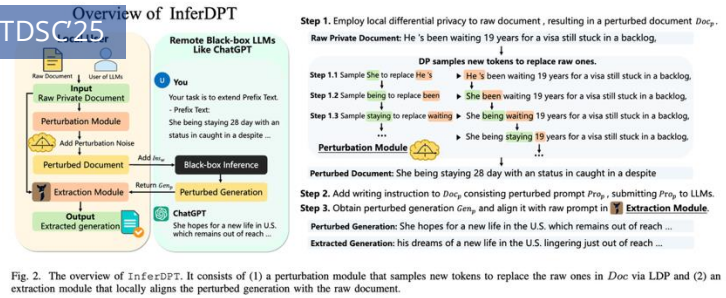
GenderCARE (Distinguished artifact award)

CCS'25



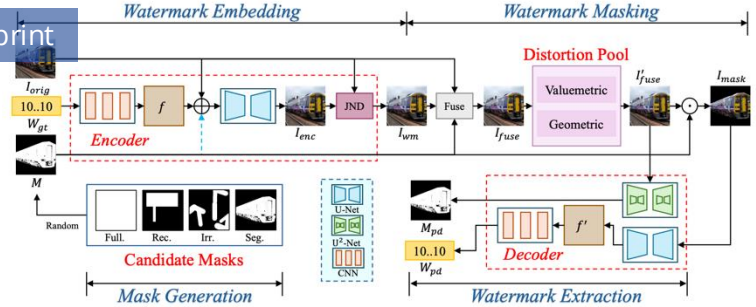
SafeGuider (SOTA defense for T2I models)

TDSC'25



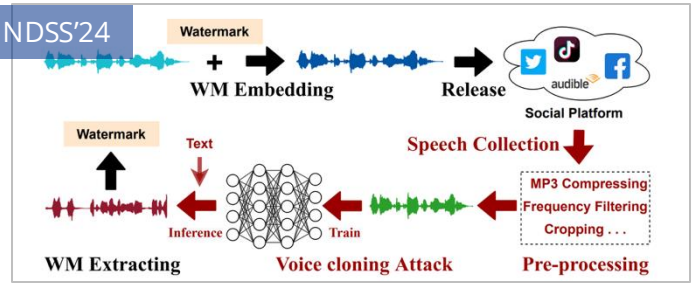
InferDPT (First solution for LLM inference-time privacy)

Preprint



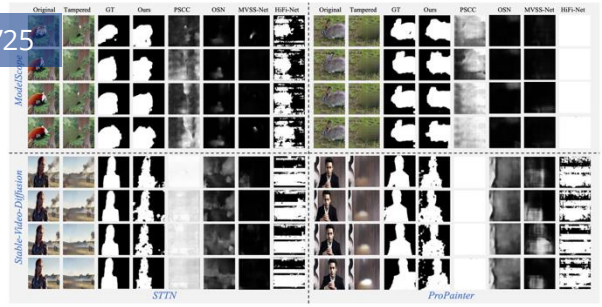
MaskWM (SOTA Gen-AI image WM)

NDSS'24



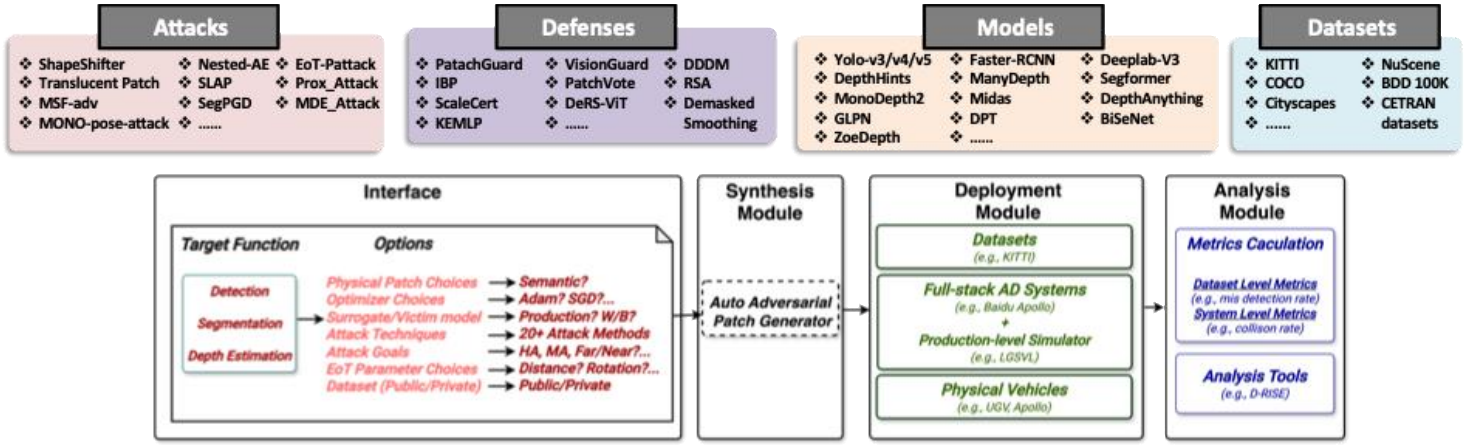
Timbre Watermarking (First semantic audio WM)

ICLR'25



VideoShield(First Gen-AI video WM)

The SG's Largest & Comprehensive Platform for Evaluating AV Robustness



A fake person causes car to slow-down or brake on the bridge.

A handhold patch cross the crosswalk misleading the depth estimation.

- Organized by **Ministry of Defence of Singapore** (MINDEF)
- High coverage: **19+** attacks, **13+** defenses, **14+** models, **6+** datasets
- Physical test in the national center
- Enhance Singapore **national security** and technological independence
- Support critical infrastructure and **defense** readiness

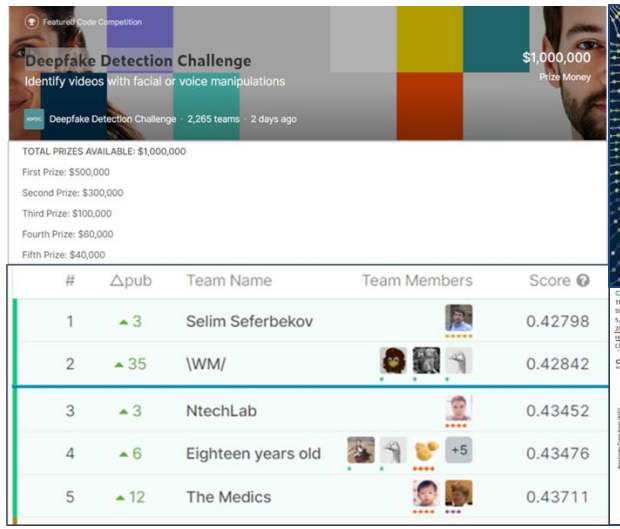
MINDEF SINGAPORE

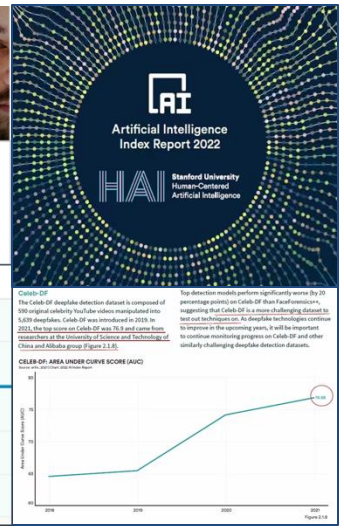
DSO NATIONAL LABORATORIES

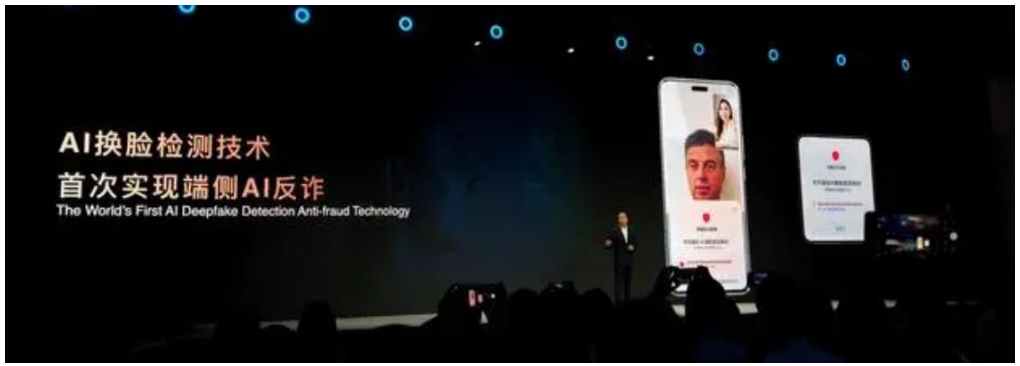
① Bus stop with bay
② Rain simulator
③ Slope
④ Signalled intersection
⑤ S-course
⑥ Signalled intersection
⑦ V2X communication
⑧ Charging station for vehicle and AutOnomous VehicLe MonItoring and EValuation SystEm (OLIVE)
⑨ Urban canyon
⑩ Pedestrian crossing
⑪ Crank course
⑫ Bus stop
⑬ Flash flood area

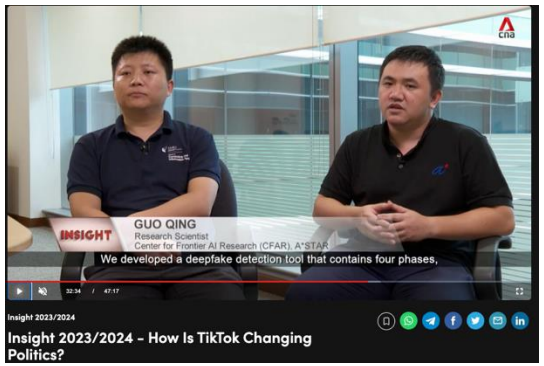
High-accuracy Deepfake Detection

- Ranked **2nd globally** in the Deepfake Detection Challenge (DFDC), organised by Facebook and MIT, with a $300,000 award.
- Recognized in the 2022 Stanford AI Index as **a key milestone** in deepfake detection.
- **3rd** rank in AISG Trust Media Challenge.
- The world's **first** on-device detection system.
- Partnering with **AGIL Trust**—Singapore's **national** digital trust infrastructure









Centre for Frontier AI Research

A*STAR CFAR

THANK YOU

ASTARSG

ASTAR-SG