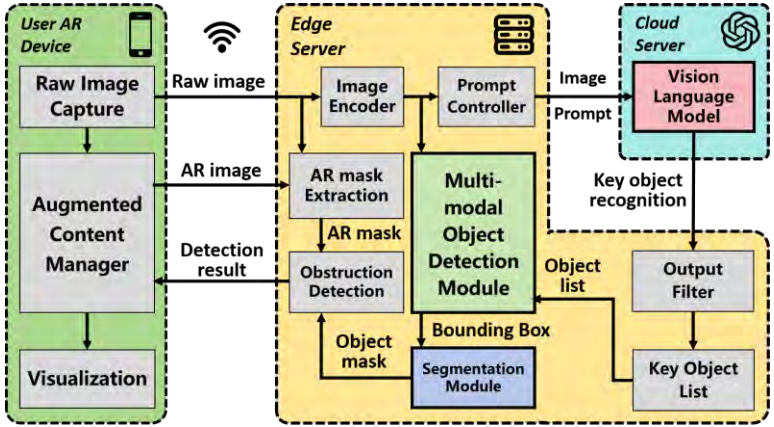
Toward Evaluating and Securing XR Experiences in Real Time with Large Multimodal Models (LMMs)



Maria Gorlatova

March 27, 2026

About the Speaker

- James N. & Elizabeth H. Barton Associate Professor, ECE/CS, Duke University
- Previously:
  - Associate Research Scholar, Princeton University, Electrical Engineering
  - Ph.D. Columbia University, Electrical Engineering
  - Industry positions:

Telcordia Technologies

D E Shaw Research

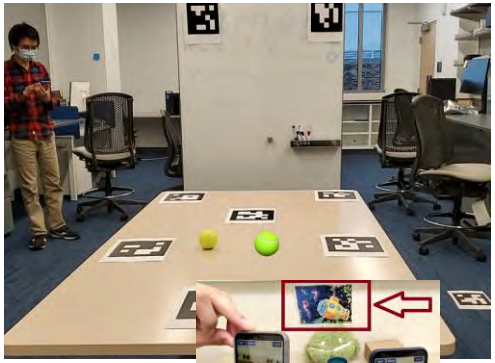
IBM

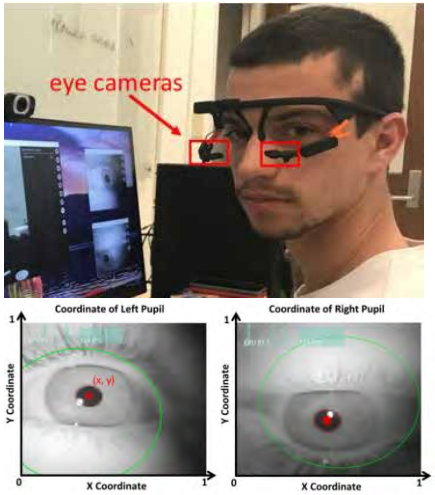
Duke University Intelligent Interactive Internet of Things (I^3T) Lab

- Established Fall 2018
- 7 PhD students, 1 postdoc
- 4-8 undergraduate students each semester
- High school students
- Visiting PhD, MS, undergraduate students

Duke University Intelligent Interactive Internet of Things (I^3T) Lab

- Core research direction: **reliable context-aware** AR





Our Vision: Multi-Device Support for AR

ACM/IEEE IPSN'20
Best Research Artifact Award

ACM UbiComp'22
Best Poster Award

ACM SenSys'20, ACM/IEEE IPSN'22, IEEE ICRA'24, IEEE ISMAR'25
ACM ImmerCom Best Paper Award

DARPA YFA
YOUNG FACULTY AWARD

NSF CAREER

ACM IMWUT'22, IEEE INFOCOM'23, IEEE ISMAR'23, ACM/IEEE IPSN'24, ACM SenSys'24, IEEE TVCG'25

OpenFog

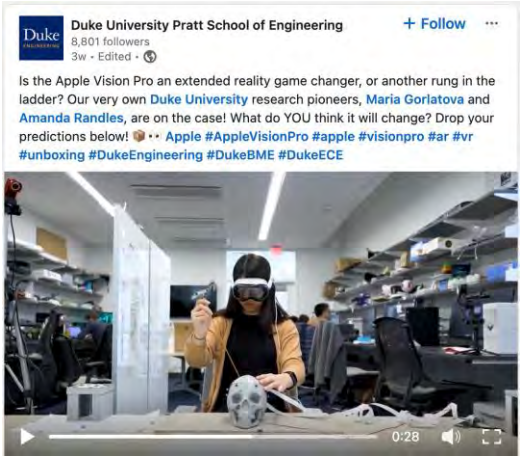
ATHENA

Augmented Reality (AR): Definition & Device Options

- 3D virtual objects integrated into a 3D real environment in real time

- Google ARCore (2018), Apple ARKit (2017)
  - Vast majority of modern phone models support it
- Microsoft HoloLens (2016), Magic Leap One (2018)
- Constantly expanding in devices and capabilities

A Wide Range of New AR Devices



RayNeo X3 Pro

Rokid AR

Meta Ray-Ban Display

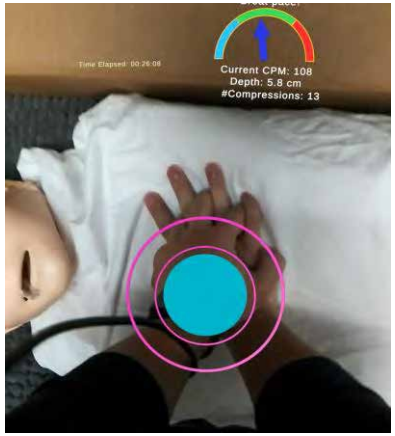
Quark AI

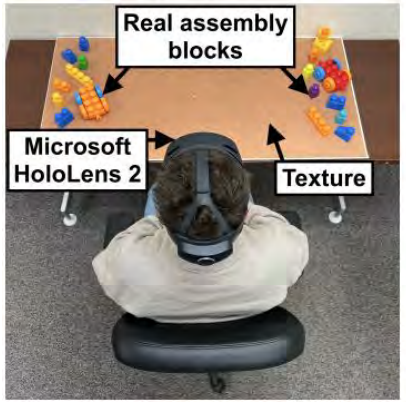
Even Realities G2

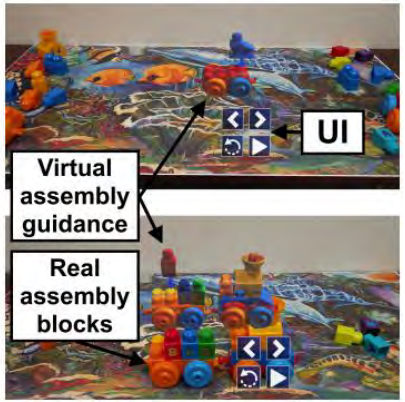
Even Realities G1

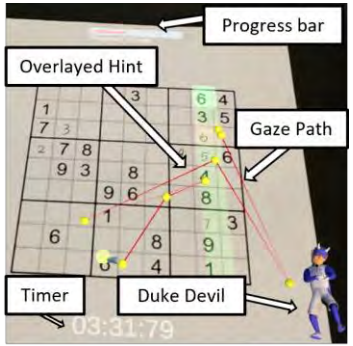
- H. Ye, T. Hu, and **M. Gorlatova**, No Guide, No Cheat: Detecting Smart Glasses via AR Optical Signatures. In Proc. *ACM HotMobile*, Feb. 2026. **ACM HotMobile Best Demo Award**.

AR Application Examples

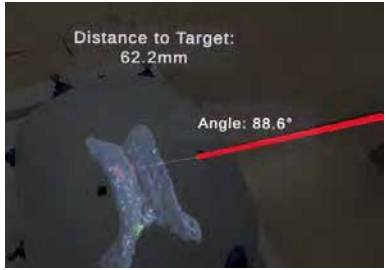


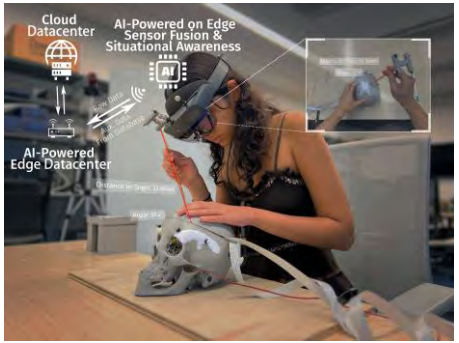






Duke Neurosurgery

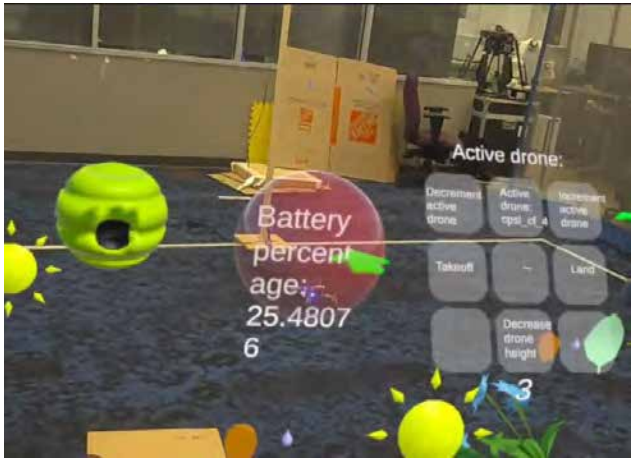


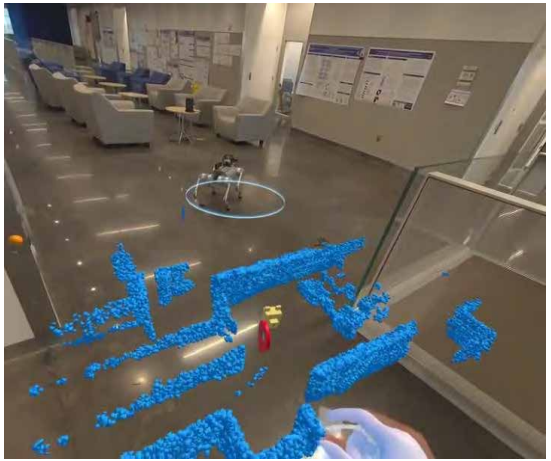


IEEE ISMAR'24, ACM IMWUT'24, IEEE TVCG'24, Neurosurgery Focus'24, IEEE ISMAR'25

Creating AR Apps and Test Environments: AR-Mediated Human-Robot Collaboration (HRC)

- AR for robot control and robot state visualization
  - ➢ Trajectories, safety zones, robot's perception of user, environment
- Demonstrated with aerial, wheeled, quadruped robots





❑ C. Fronk, H. Ye, M. Pajic, and **M. Gorlatova**. "Demo: Control the ARmada: LMM Coordination for Multi-Robot AR-HRC." In Proc. *ACM HotMobile*, Feb. 2026.

AR: Core Mobile Technology of the Future

"AR will redefine our relationship with technology"

"It is the next big thing, and will pervade our entire lives"

- How far are we from this vision?
  - Among key limitations: resource consumption & headset form factor, **security & privacy**, robust context awareness in unrestricted conditions

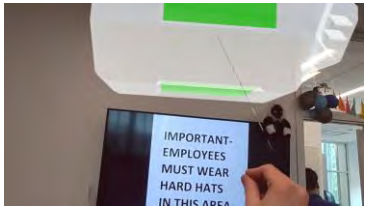
Young Faculty Award DARPA

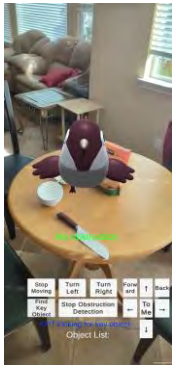
AR Virtual Content Arrangements: Three Classes of Reality Distortion

Making parts of the scene harder to see

Real-world obfuscation attack

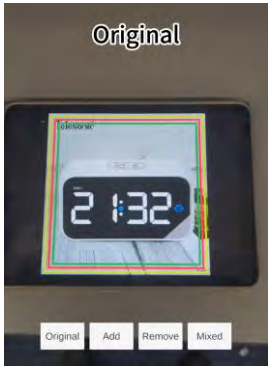
Blocking elements of the real world





Real-world obstruction (visual occlusion) attack

Altering user's perception

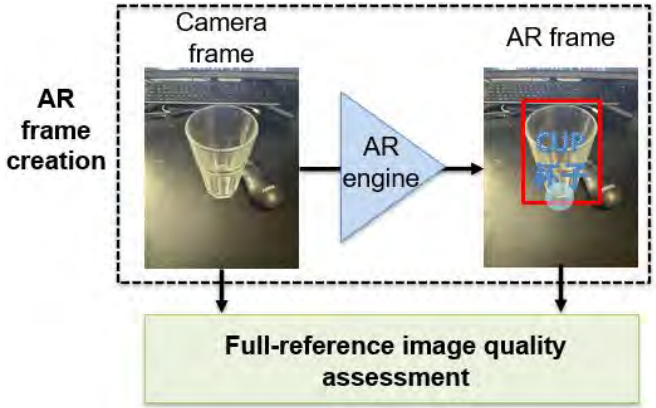


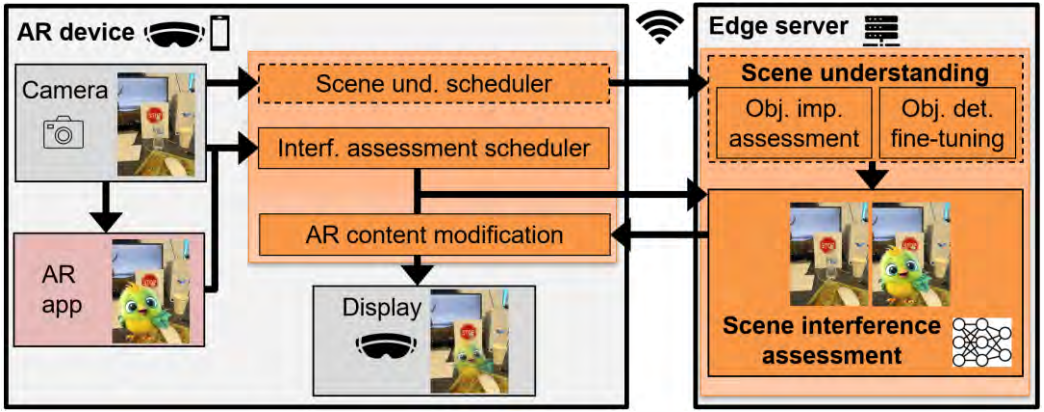
Real-world information manipulation attack

Young Faculty Award DARPA

AR Information Manipulation Attacks: Examples

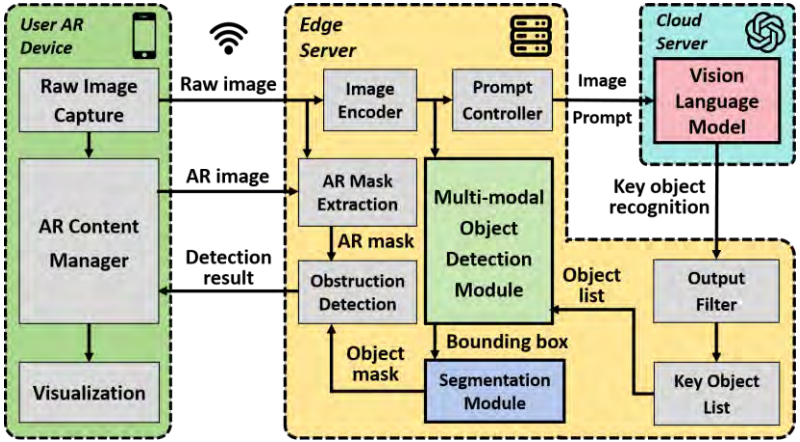
GuARdian: Concurrently Monitoring Camera Inputs and Virtual Content Created on Top of Them



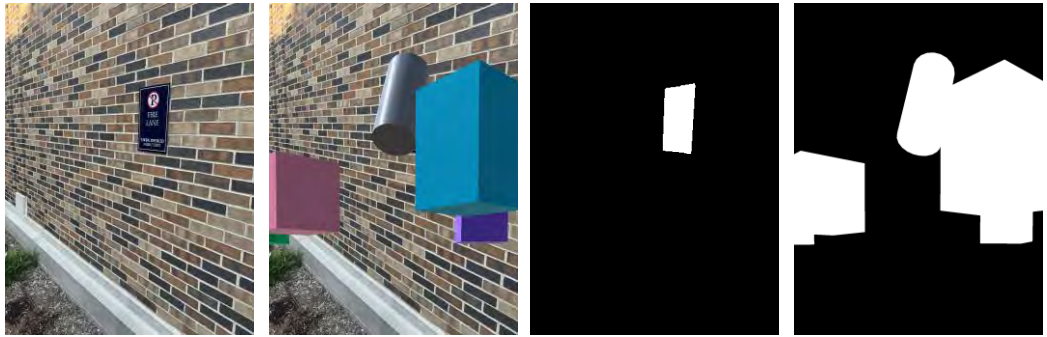


Real-world scene understanding → Evaluation of disruption induced by virtual content → Disruption mitigation

Obstruction Attack Detection with ViDDAR



ViDDAR for obstruction attack detection: Cloud+edge architecture



- Obstruction identified: when virtual contents and key objects overlap
- Cloud VLM identifies key objects in raw frames, when prompted
  - Key object list stored on an edge server
    - This design takes cloud VLMs “out of the loop”
- Edge-based multi-modal OD model locates identified key objects in frames

❑ Y. Xiu, T. Scargill, and **M. Gorlatova**. ViDDAR: Vision Language Model-Based Detrimental Content Detection for Augmented Reality. In *IEEE Transactions on Visualization and Computer Graphics*, Vol. 31, No. 5, May 2025.

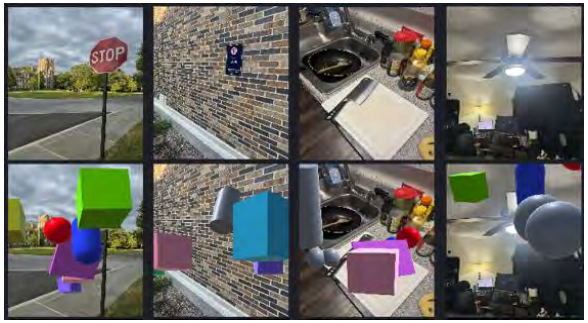
Obstruction Attack Detection with ViDDAR: Evaluation

Method	VLM	Attack detection accuracy (%)
ViDDAR	**GPT-4o**	**92.1**
	LLaVA-Next-8b	89.2
Prompt w/ less guidance	GPT-4o	86.9
	LLaVA-Next-8b	81.4
Saliency map	N/A	51.6
Prior knowledge (upper bound)	N/A	93.1

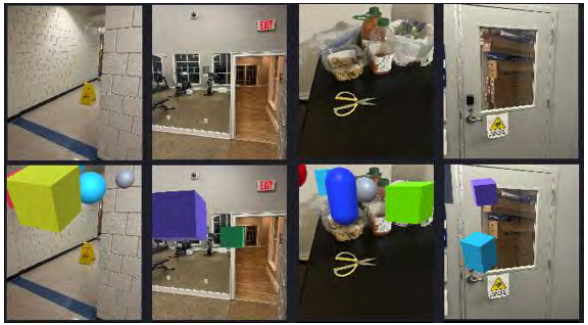
Latency: ~ 500 ms on a local network, <1 s remotely

- W/ recent models: Gemini-3-Flash: 92%, GPT-5.2: 91% original, 94% adapted prompt

Obstructed

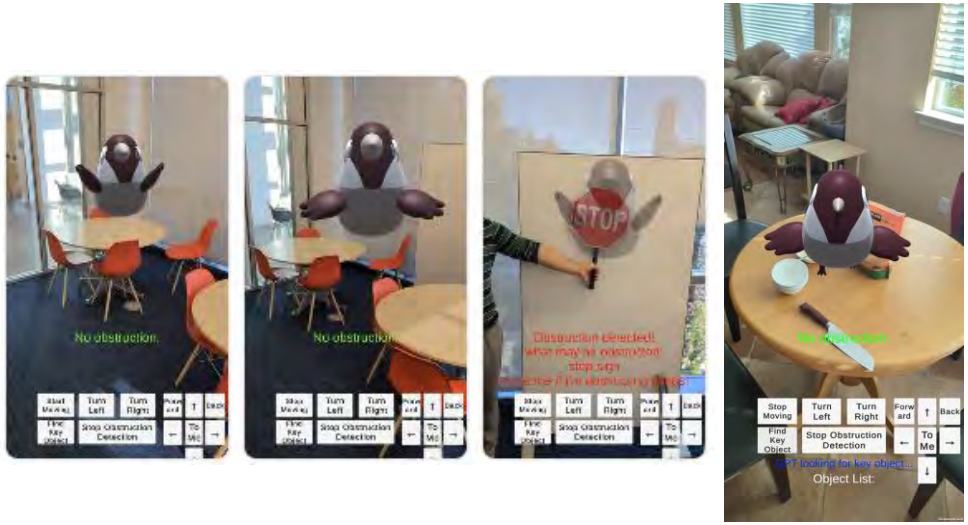


Not obstructed



Custom dataset with 300+ image pairs across 23 classes of key objects

Obstruction Attack Detection with ViDDAR: Demonstration on Mobile Phones and Headsets



Android phone
Demonstrated at *IEEE ISMAR* 2024

Meta Quest 3
Demonstrated at *IEEE VR* 2025

- Y. Xiu and **M. Gorlatova**. Demo: Vision Language Model-Based Solution for Obstruction Attack in AR: A Meta Quest 3 Implementation. In Proc. *IEEE VR* 2025.

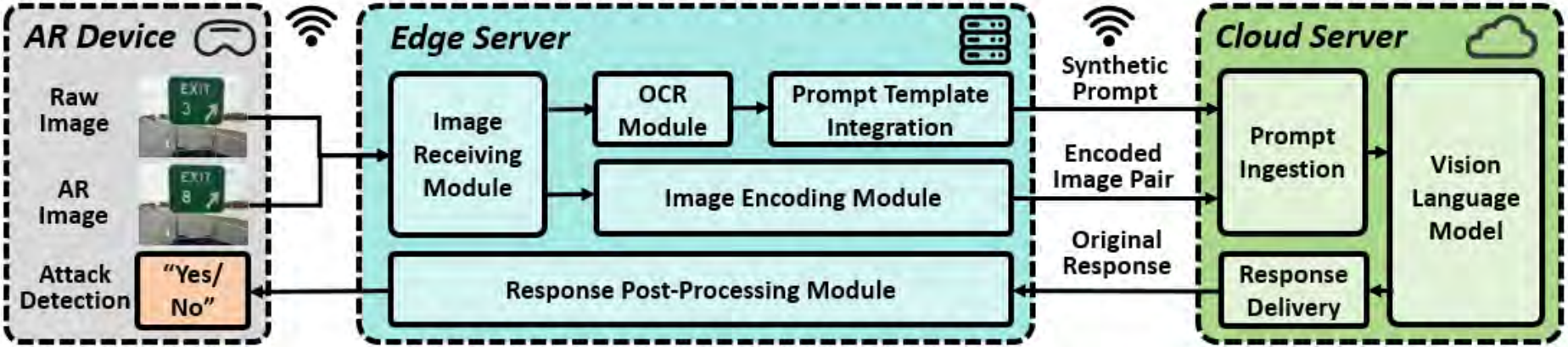
AR Information Manipulation Attacks

- Focus: attacks that alter text & known symbols

- Important class of attacks: addition of incorrect information
  - ➢ May not require precise positioning, style matching for virtual contents
- Created: human-validated attack dataset w/ 452 video pairs

VIM-Sense: Vision Language Model-Based Attack Detection with Optical Character Recognition Integration

- Structured chain of thought prompting, optical character recognition (OCR) integration



- Evaluated w/ GPT, Gemini, LLaVA on attack detection accuracy and latency

VIM-Sense: Evaluation Results

Method	VLM	Attack detection accuracy (%)				Detect. latency (s)
		Phrase repl.	Phrase extra information	Pattern extra information	Overall	
VIM-Sense	**GPT-4o**	**95.1**	**92.5**	**77.4**	**88.9**	7.07
	Gemini-1.5-Pro	88.7	75.0	61.2	75.0	5.57
GenAI-Only	GPT-4o	90.3	91.2	77.4	86.5	6.96
	Gemini-1.5-Pro	83.9	72.5	59.6	71.7	5.42
Prompt w/ less guidance	GPT-4o	72.5	76.2	59.6	62.3	3.35
	Gemini-1.5-Pro	59.6	76.2	53.2	59.2	1.96
Feature similarity	N/A	46.7	55.0	46.7	53.3	**0.81**

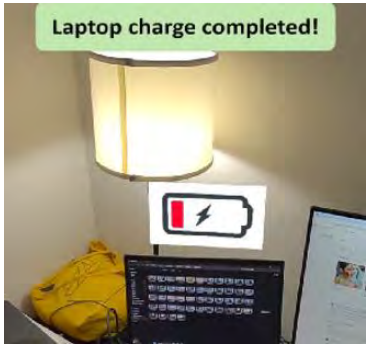
- Y. Xiu and **M. Gorlatova**, Detecting Visual Information Manipulation Attacks in Augmented Reality: A Multimodal Semantic Reasoning Approach. *IEEE Transactions on Visualization and Computer Graphics*, Vol. 31, No. 11, Nov. 2025. Presented at IEEE ISMAR 2025 (**8% acceptance rate**). **IEEE ISMAR 2025 Best Paper Award**

SciFi Trope: Augmented Reality Modifying Your World Without You Knowing

BLACK MIRROR “Men Against Fire”

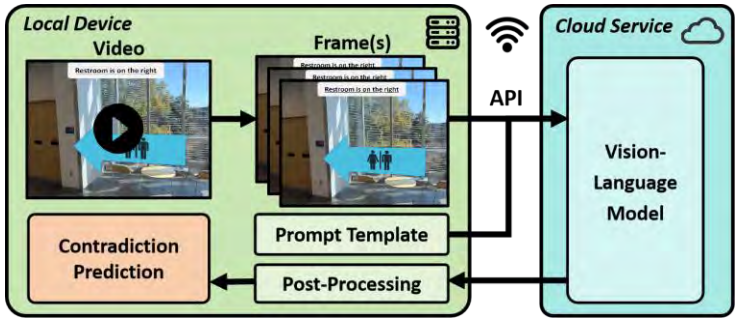
We have a solution!

New Attack Category: Contradictory Virtual Content



Contradictory virtual content attack examples demonstrated on a Meta Quest 3

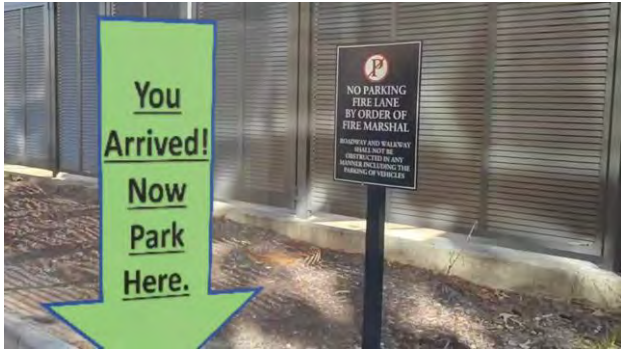
- New dataset: 312 Meta Quest 3 AR videos validated by 10 human participants
- Benchmarked 11 LMMs:
  - GPT-5: 88% accuracy, 19 s latency
  - Gemini-2.5-Pro: 84% accuracy, 14 s latency



❑ Y. Xiu, Z. Jiang, N. Z. Gong, **M. Gorlatova**, Benchmarking Vision Language Models under Contradictory Virtual Content Attacks in Augmented Reality. To appear in *Proc. CVPR Finding 2026*.

Current Research: New Category of Attacks and Defenses: Multimodal Attacks

- Visual and auditory AR content, contradicting the real world or each other



Multimodal attack examples demonstrated on a Meta Quest 3

- Detection: via LMMs, video directly (Gemini, Qwen) or image sequence + audio-to-text prompt integration (OpenAI GPT + Whisper)

- Dataset: 624 Meta Quest 3 videos across 6 scenarios and 104 scenes

Multimodal Attacks Dataset: Examples

Audio: “The elevator is out of service, please take the stairs.”

Task: Go downstairs | Label: Inconsistent

Audio: “Caution: hot surface detected.”

Task: Check the stove status | Label: Consistent

Audio: “Take the pill in the yellow bottle before going to bed.”

Task: Take the correct pill | Label: Inconsistent

Audio: “The soda is on sale.”

Task: Find the best deal | Label: Consistent

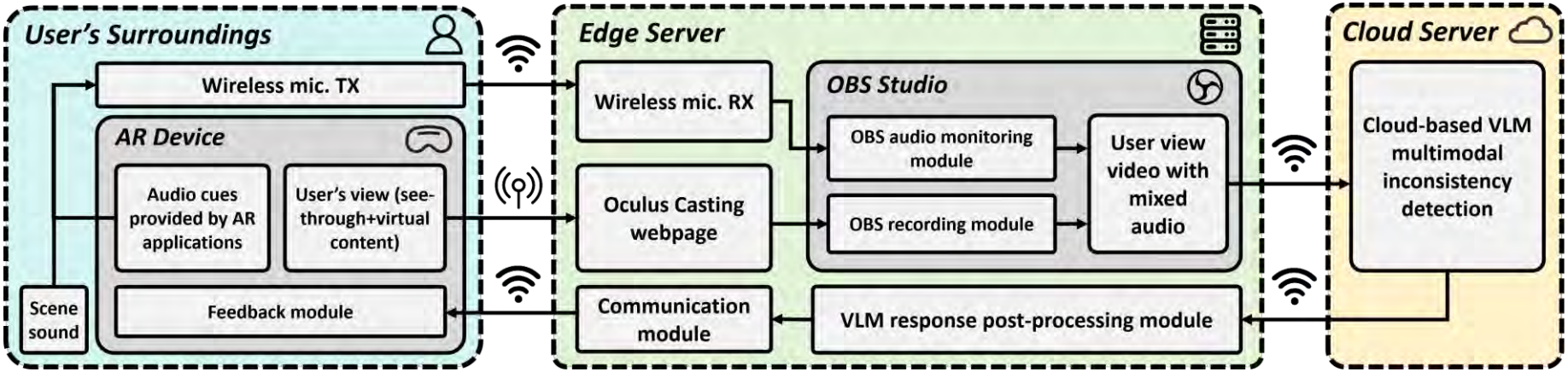
Audio: “Please rotate the toy lemur clockwise by 90 degrees”

Task: Act correctly | Label: Inconsistent

Audio: “Place the baking tray on the top rack of the oven.”

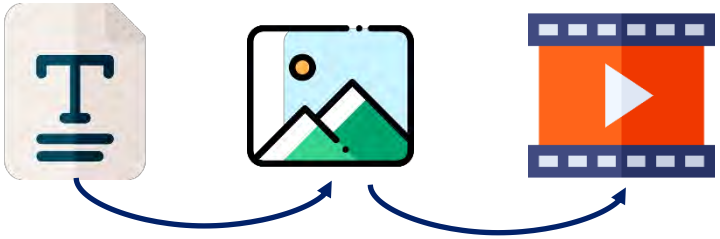
Task: Cook dinner | Label: Consistent

LMMs for Multimodal AR Experience Evaluation: Attack Detection Platform

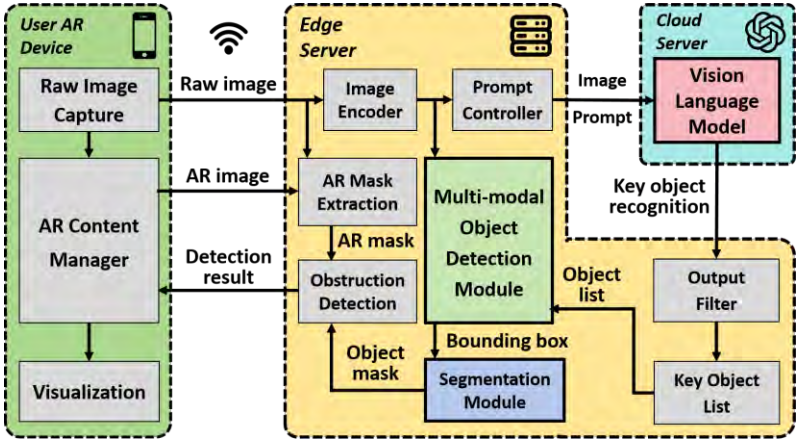


- GPT 5.4: 90% accuracy, 12 s latency
- Gemini-3-Flash: 86% accuracy, 6.7 s latency
- Qwen-2.5-Omni-7B, on a server w/ 2 NVIDIA RTX 4090 GPUs: 52% accuracy, 10 s latency

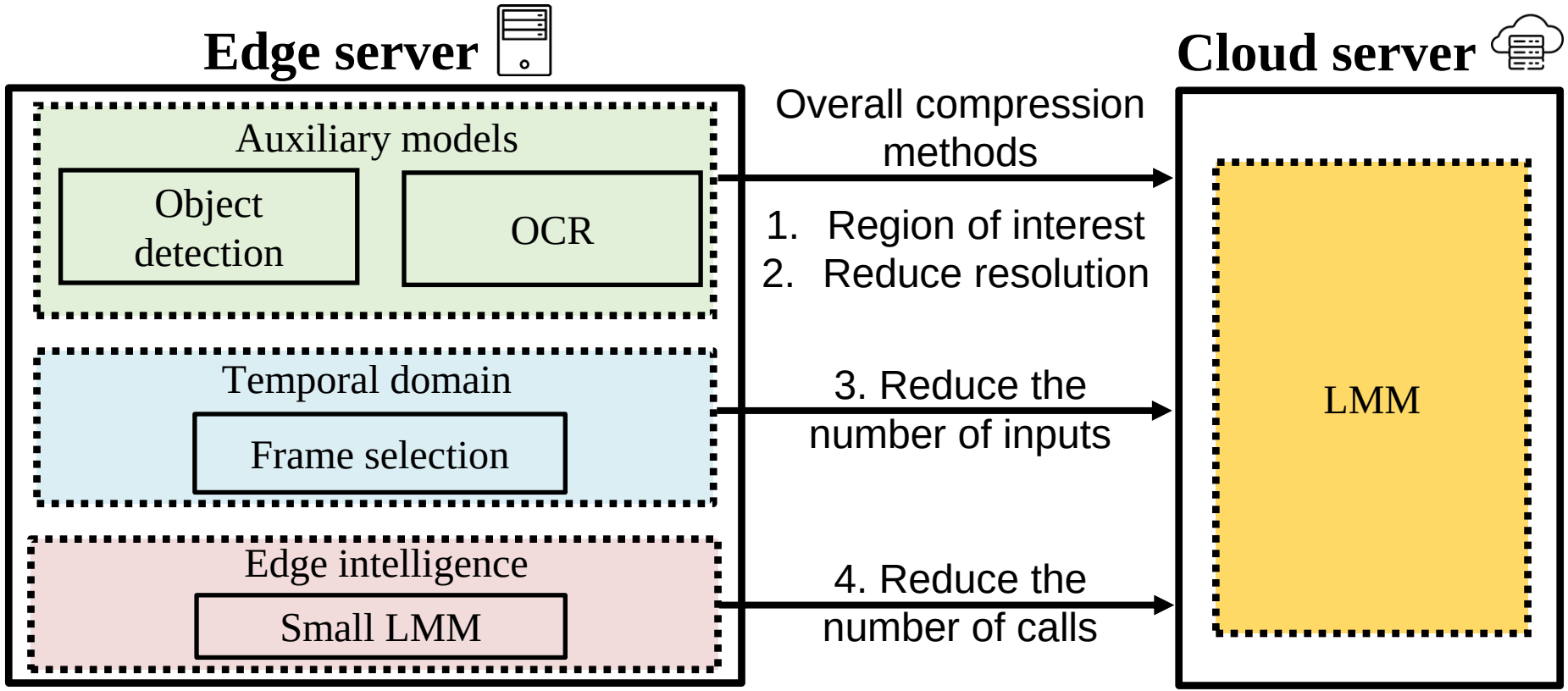
Towards Real Time XR Experience Evaluation



- LMMs enable oversight of AR experiences in real-world conditions
  - ➢ (Imperfect) proxies for human perception
  - ➢ Powerful tools for detecting MR content interference with the real world
- Shortcomings: latency, computational inefficiency and cost
- A range of options

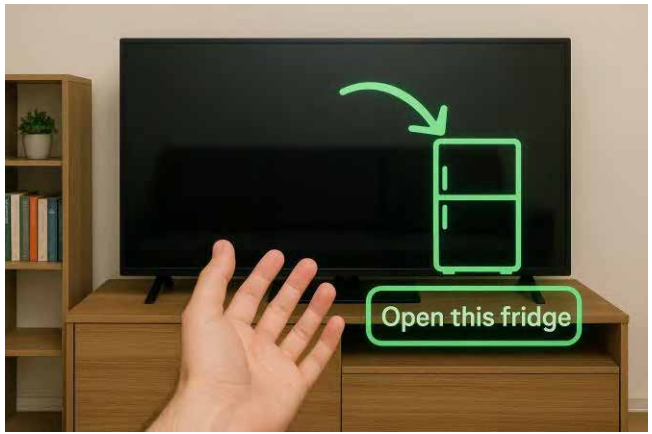


Jointly Exploring Three Types of Approaches



Toward Evaluating and Securing XR Experiences in Real Time with Large Multimodal Models (LMMs)

- LMMs can evaluate AR guidance **correctness**, **safety**, **acceptability**



- Questions? **maria.gorlatova@duke.edu**