









Agentic Framework for Automated Security Analysis in Multi-User Extended Reality

Mutahar Ali, Habiba Farrukh

{mutahara, habibaf}@uci.edu

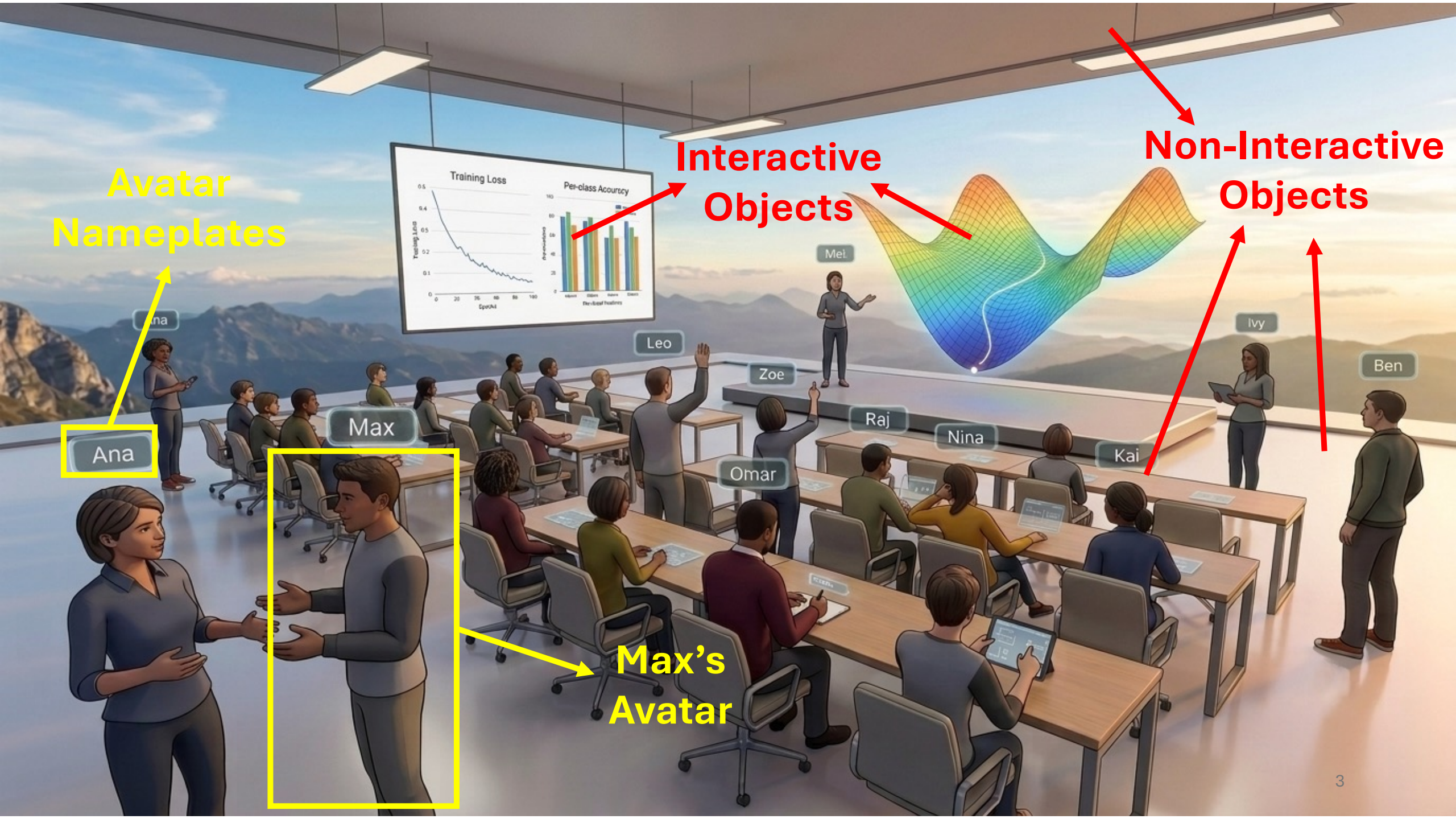
USENIX Security Symposium 2026

XR Security Workshop @ MobiCom 2026

UCI University of California, Irvine

SPECT Lab

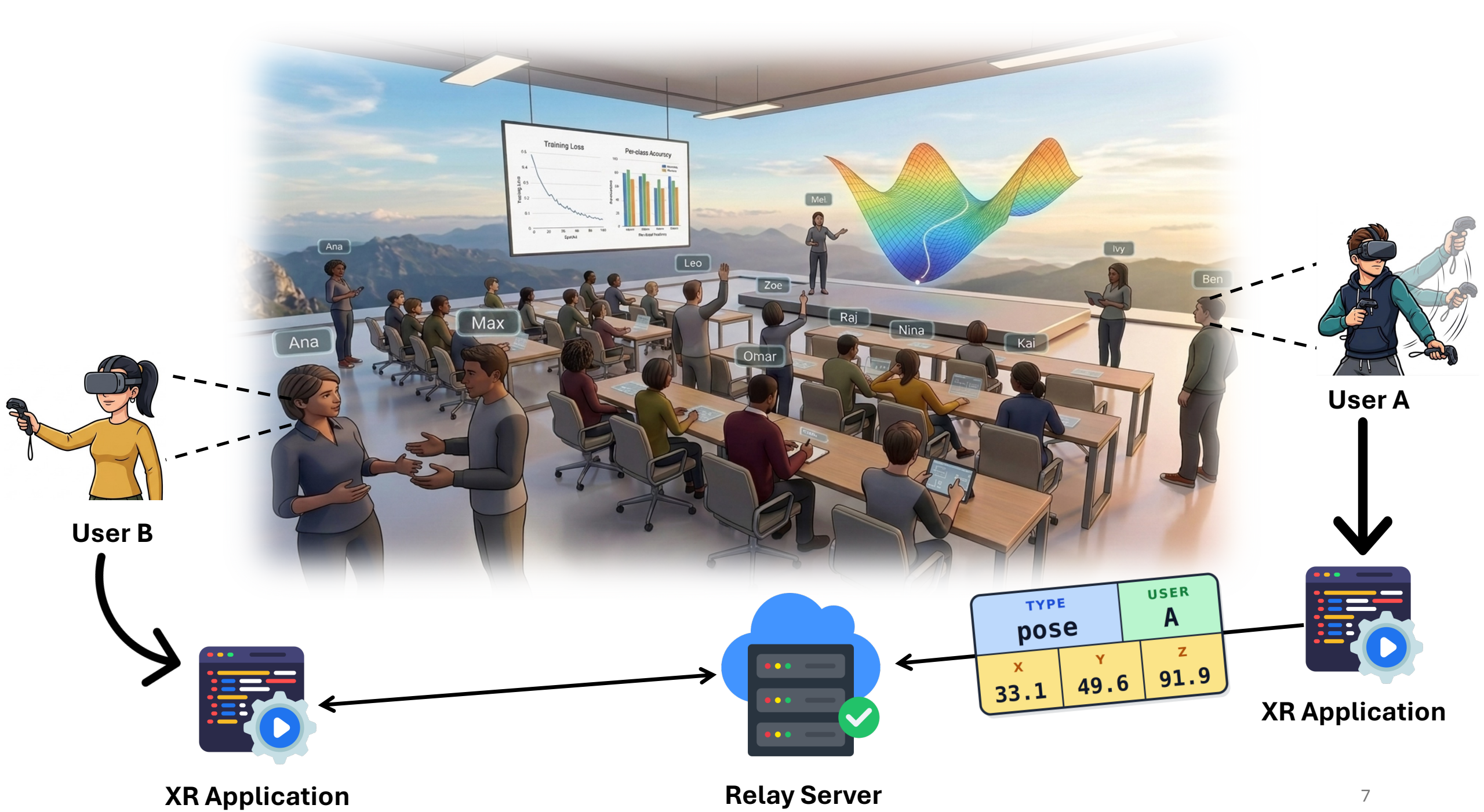

Training Loss
Per-class Acourtcy
Mel
Ana
Leo
Zoe
Ivy
Ben
Max
Raj
Nina
Kai
Ana
Omar

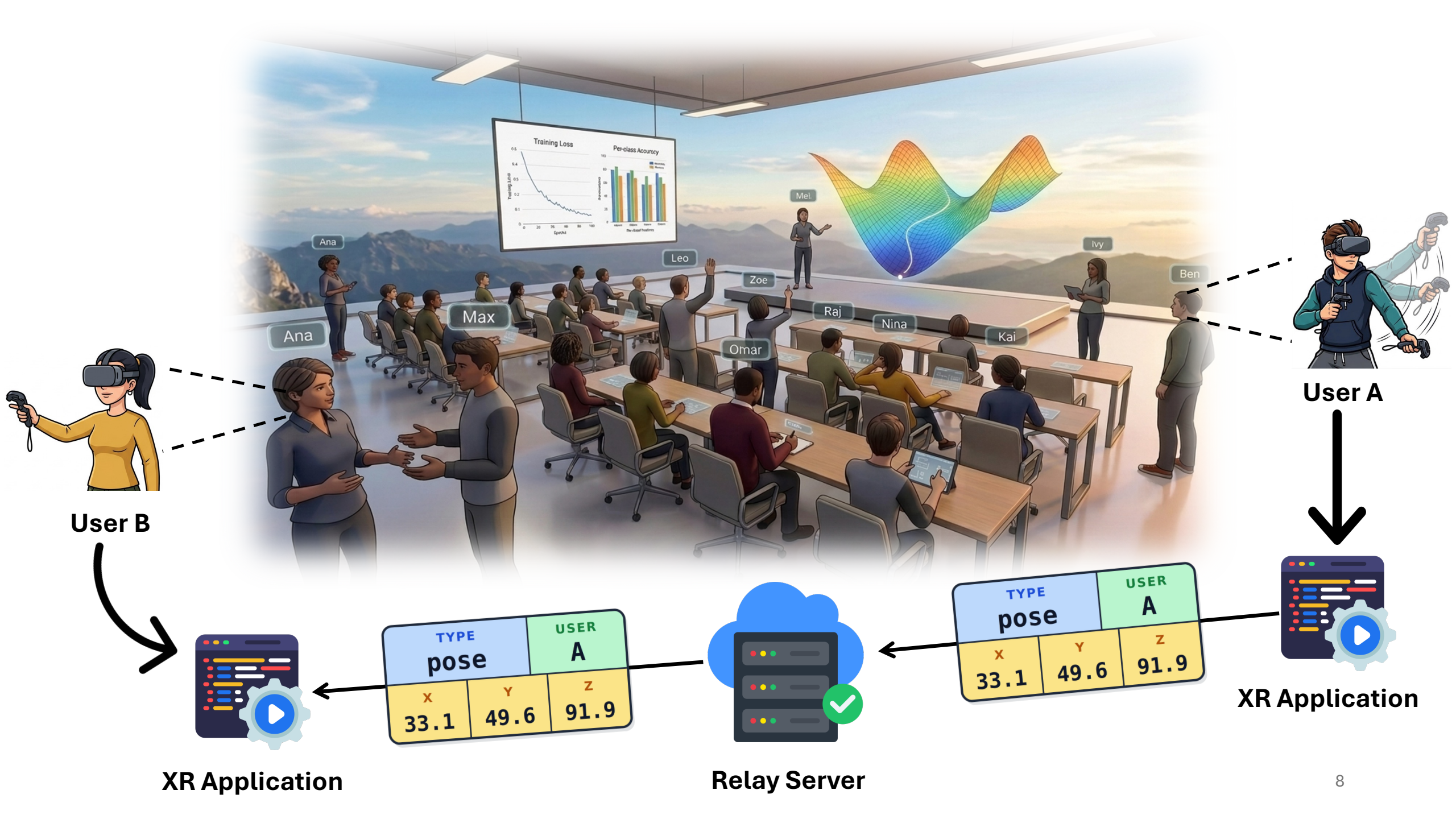

Avatar Nameplates
Interactive Objects
Non-Interactive Objects
Max's Avatar
Training Loss
Per-class Acuracy
Ana
Max
Leo
Mel
Zoe
Raj
Nina
Omar
Kai
Ivy
Ben

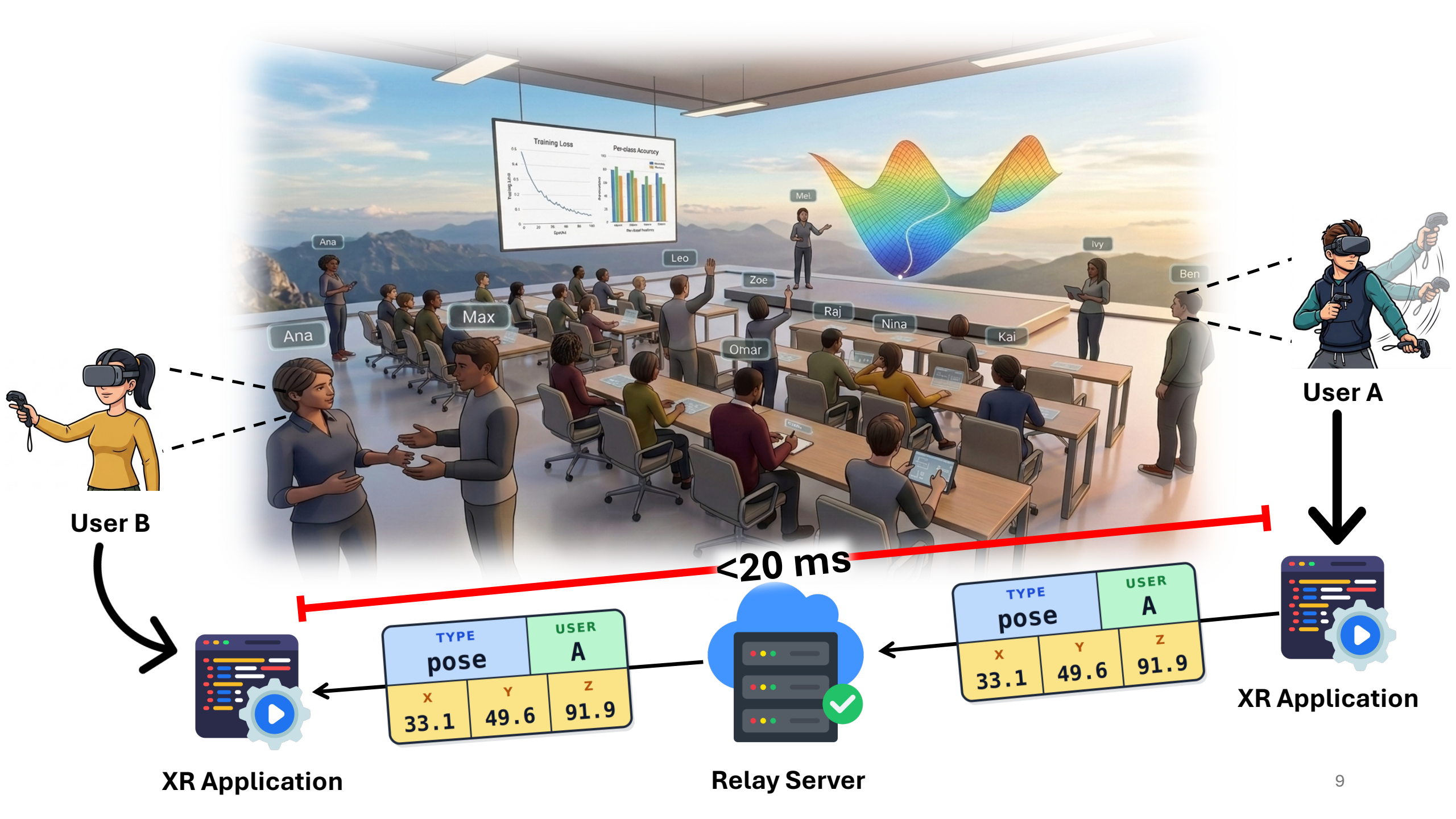

Training Loss
Per-class Accuracy
Mei
Ana
Leo
Zoe
Ivy
Ben
Max
Raj
Nina
Ana
Omar
Kai

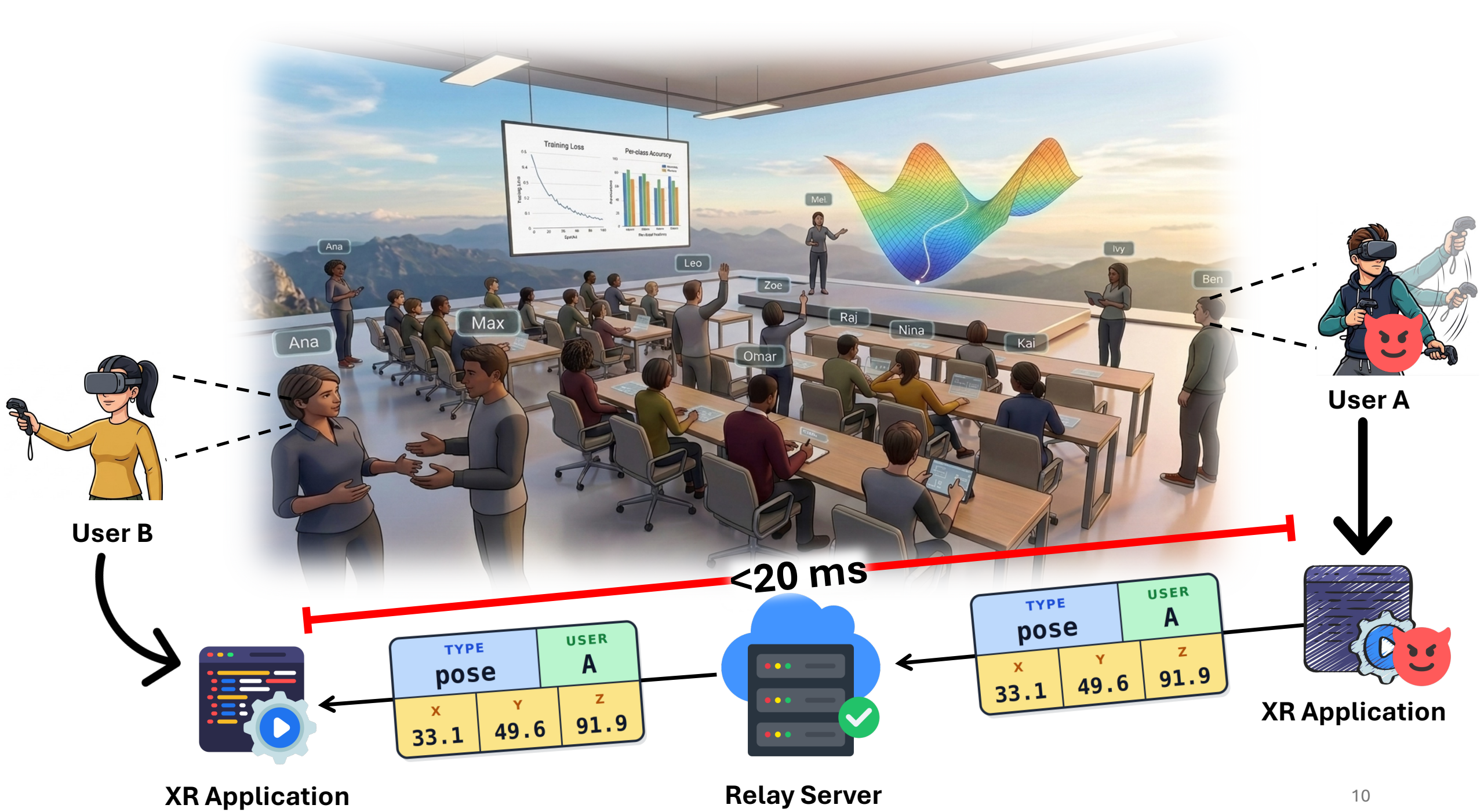

Training Loss
Per-class Accuracy
Mei
Ana
Leo
Zoe
Ivy
Ben
Max
Raj
Nina
Kai
Omar
Ana
User B
User A
XR Application
Relay Server
XR Application

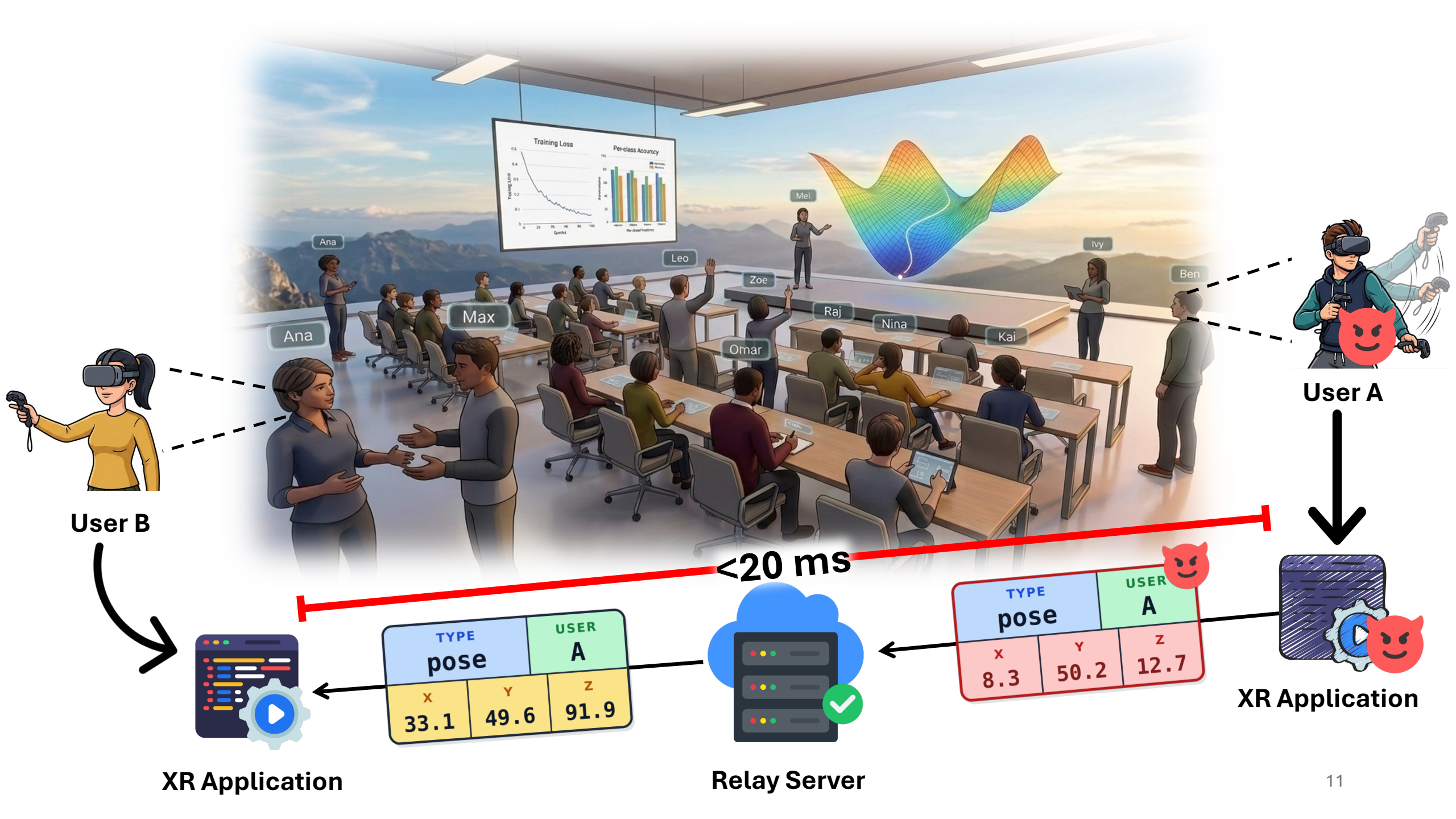

Training Loss
Per-class Accuracy
Mei
Ana
Leo
Zoe
Ivy
Ben
Max
Raj
Nina
Kai
Omar
Ana
User B
User A
XR Application
Relay Server
XR Application

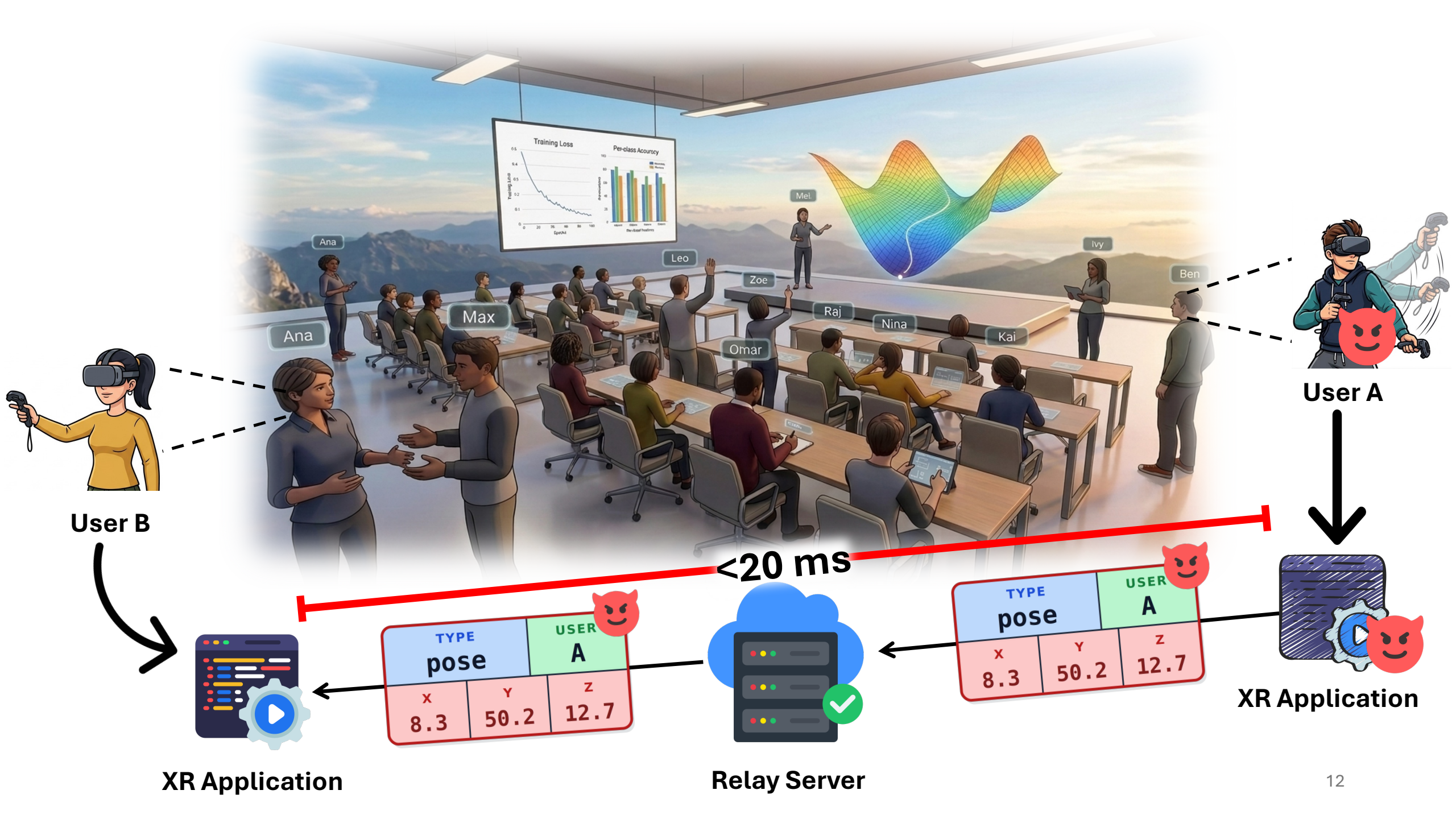

Training Loss
Per-class Accuracy
Mei
Ana
Leo
Zoe
Ivy
Ben
Max
Raj
Nina
Kai
Ana
Omar
User B
User A
TYPE
pose
USER
A
X
33.1
Y
49.6
Z
91.9
XR Application
XR Application
Relay Server

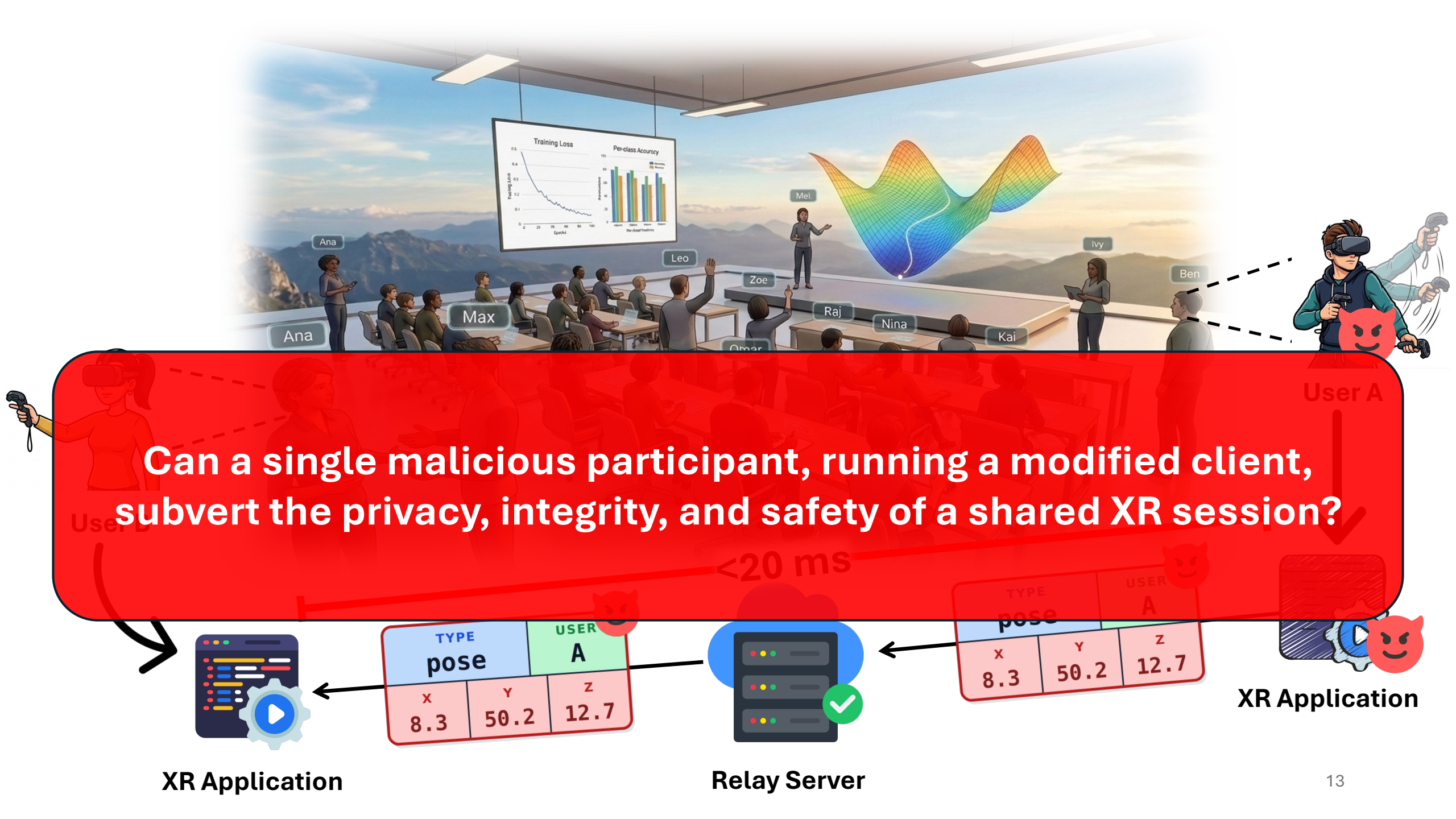

Training Loss
Per-class Accuracy
Mei
Ana
Leo
Zoe
Ivy
Ben
Max
Raj
Nina
Kai
Ana
Omar
User B
User A
TYPE
pose
USER
A
X
33.1
Y
49.6
Z
91.9
TYPE
pose
USER
A
X
33.1
Y
49.6
Z
91.9
XR Application
XR Application
Relay Server


Training Loss
Per-class Accuracy
Mei
Ana
Leo
Zoe
Ivy
Ben
Max
Raj
Nina
Kai
Ana
Omar
User A
User B
<20 ms
TYPE
pose
USER
A
X
33.1
Y
49.6
Z
91.9
TYPE
pose
USER
A
X
33.1
Y
49.6
Z
91.9
XR Application
XR Application
Relay Server


Training Loss
Per-class Accuracy
Ana
Max
Ana
Leo
Mei
Zoe
Raj
Nina
Omar
Kai
Ivy
Ben
User B
User A
<20 ms
TYPE
pose
USER
A
X
33.1
Y
49.6
Z
91.9
XR Application
Relay Server
TYPE
pose
USER
A
X
33.1
Y
49.6
Z
91.9
XR Application


Training Loss
Per-class Accuracy
Ana
Max
Ana
Leo
Mei
Zoe
Raj
Nina
Omar
Kai
Ivy
Ben
User B
User A
<20 ms
TYPE
pose
USER
A
X
33.1
Y
49.6
Z
91.9
TYPE
pose
USER
A
X
8.3
Y
50.2
Z
12.7
XR Application
XR Application
Relay Server


Training Loss
Per-class Accuracy
Ana
Max
Ana
Leo
Mei
Zoe
Raj
Nina
Omar
Kai
Ivy
Ben
User A
User B
<20 ms
TYPE
pose
USER
A
X
8.3
Y
50.2
Z
12.7
XR Application
Relay Server
TYPE
pose
USER
A
X
8.3
Y
50.2
Z
12.7
XR Application



Can a single malicious participant, running a modified client, subvert the privacy, integrity, and safety of a shared XR session?

Challenges

Goal: Given a multi-user XR app, we can identify the security violations resulting from a single malicious participant.

Challenges

Goal: Given a multi-user XR app, we can identify the security violations resulting from a single malicious participant.

1. There is no unified model of how state is represented and synchronized in multi-user XR apps.

Challenges

Goal: Given a multi-user XR app, we can identify the security violations resulting from a single malicious participant.

1. There is no unified model of how state is represented and synchronized in multi-user XR apps.
2. There is no formal definition of what safe interaction requires.

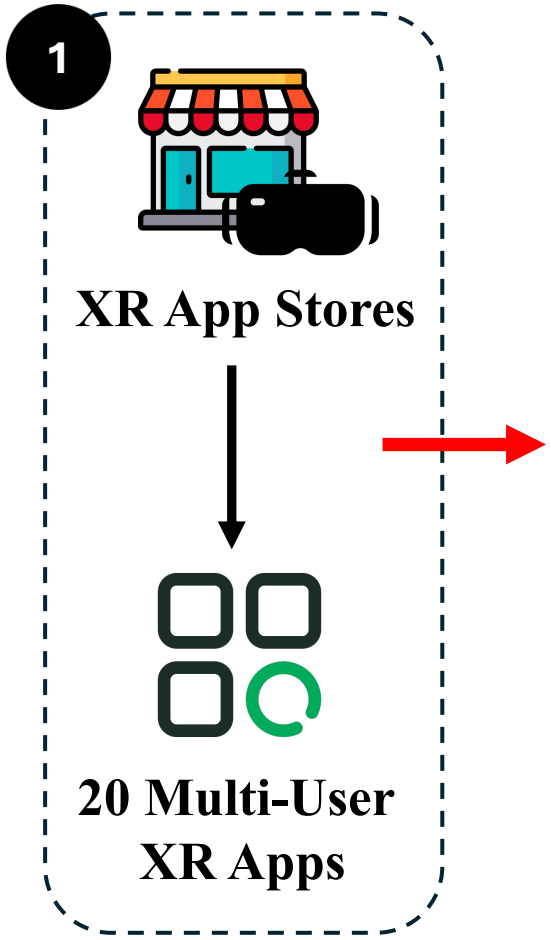
Challenges

Goal: Given a multi-user XR app, we can identify the security violations resulting from a single malicious participant.

1. There is no unified model of how state is represented and synchronized in multi-user XR apps.
2. There is no formal definition of what safe interaction requires.
3. There is no fixed program signature for the security-relevant code because violations are semantic, and enforcement differs across apps.

Methodology

Methodology



Methodology

1



XR App Stores

20 Multi-User XR Apps

2



Static Analysis

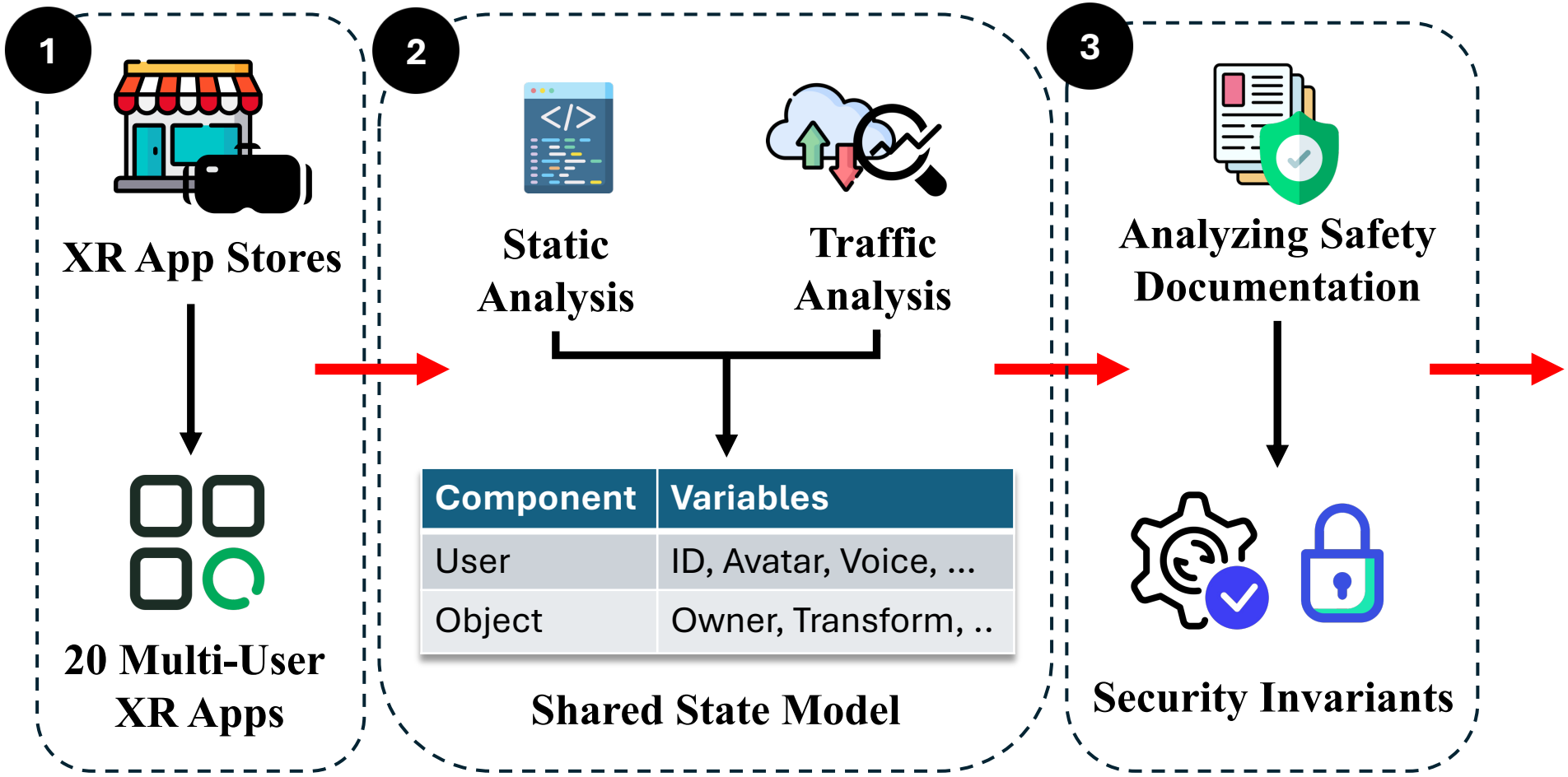


Traffic Analysis

Component	Variables
User	ID, Avatar, Voice, …
Object	Owner, Transform, ..

Shared State Model

Methodology



Methodology

1



XR App Stores

20 Multi-User XR Apps

2



Static Analysis



Traffic Analysis

Component	Variables
User	ID, Avatar, Voice, …
Object	Owner, Transform, ..

Shared State Model

3



Analyzing Safety Documentation





Security Invariants

4





Dynamic Analysis and Runtime Instrumentation



Surveillance



Impersonation



Tampering

Attacks

Methodology

1



XR App Stores

20 Multi-User XR Apps

2

Static Analysis



Traffic Analysis

Component	Variables
User	ID, Avatar, Voice, ...
Object	Owner, Transform, ..

Shared State Model

3



Analyzing Safety Documentation



Security Invariants

4





Dynamic Analysis and Runtime Instrumentation



Surveillance



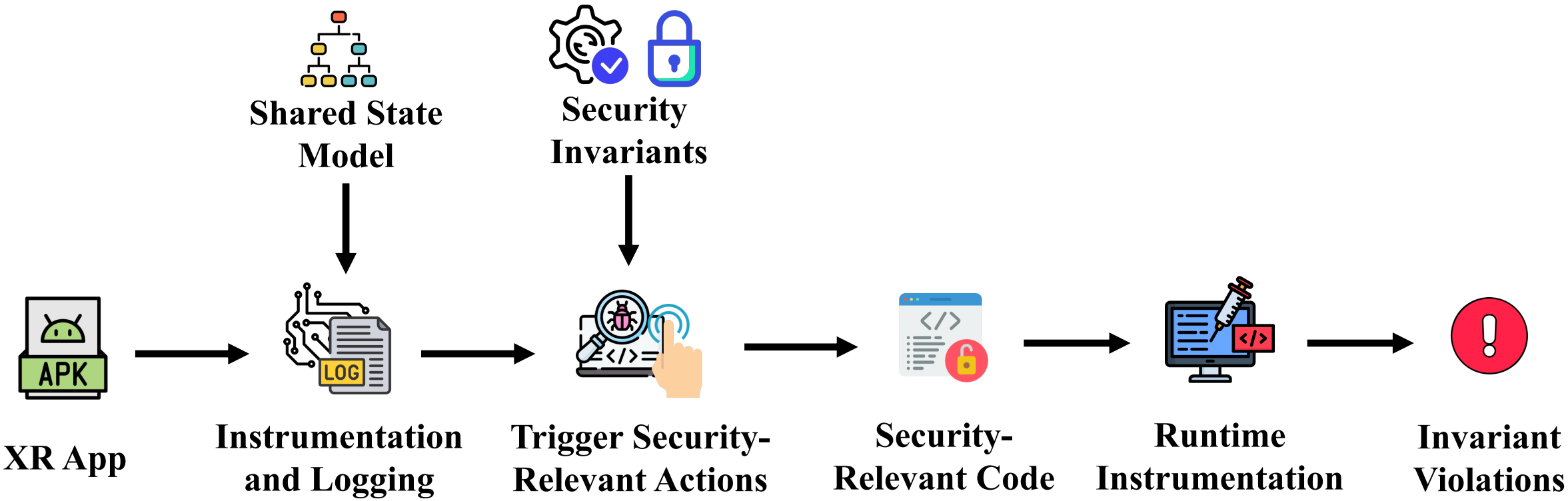
Impersonation



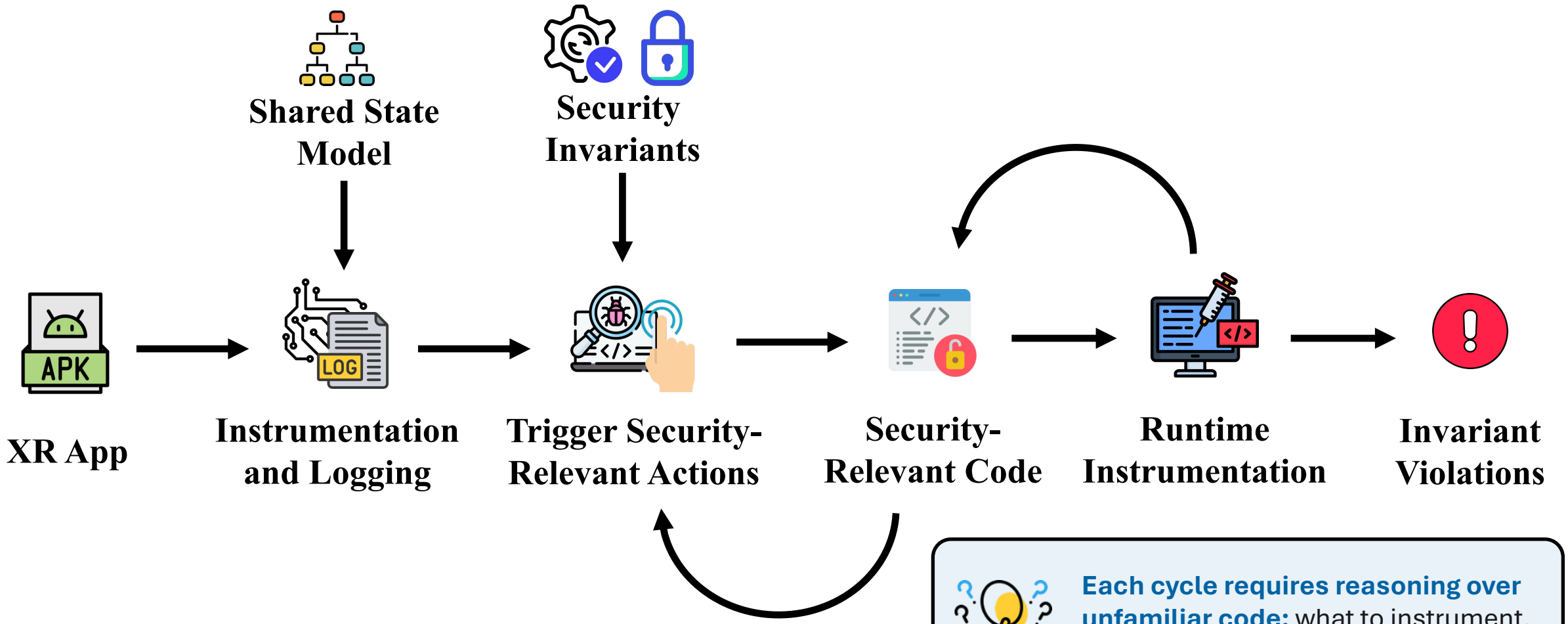
Tampering

Attacks

Methodology

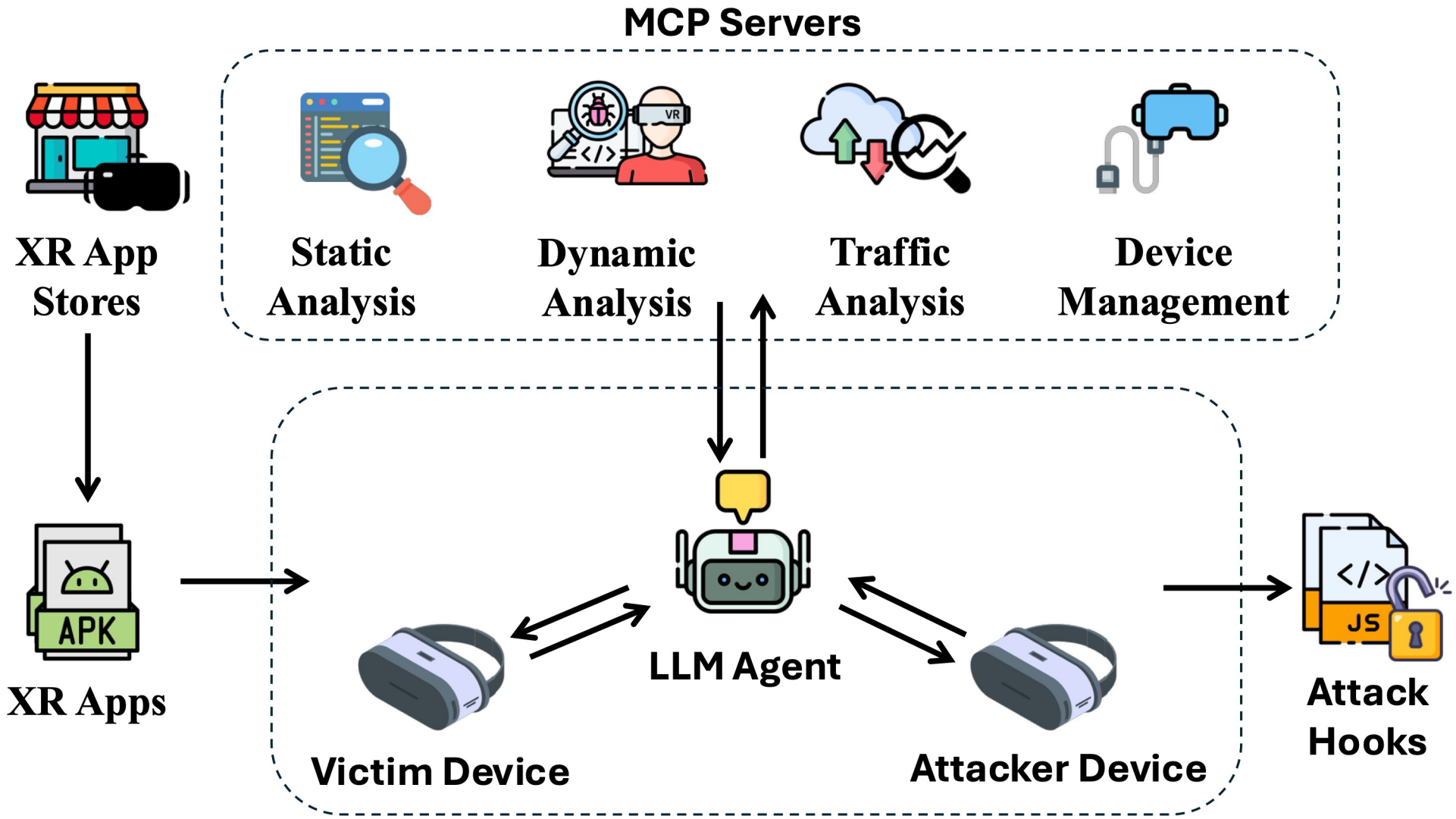


Methodology

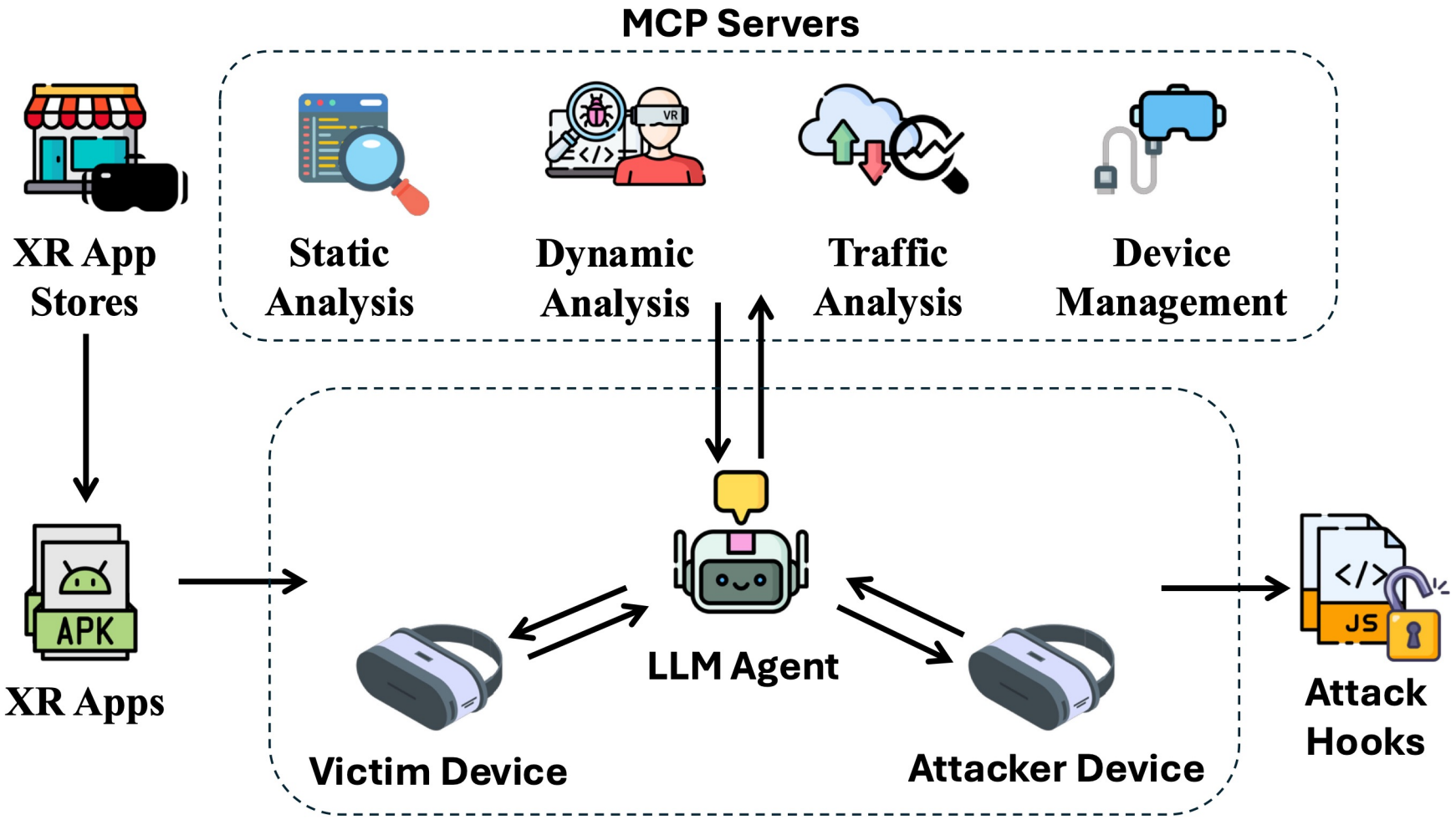


Each cycle requires reasoning over unfamiliar code: what to instrument, which symbols to search next, and what the runtime effect means.

Methodology

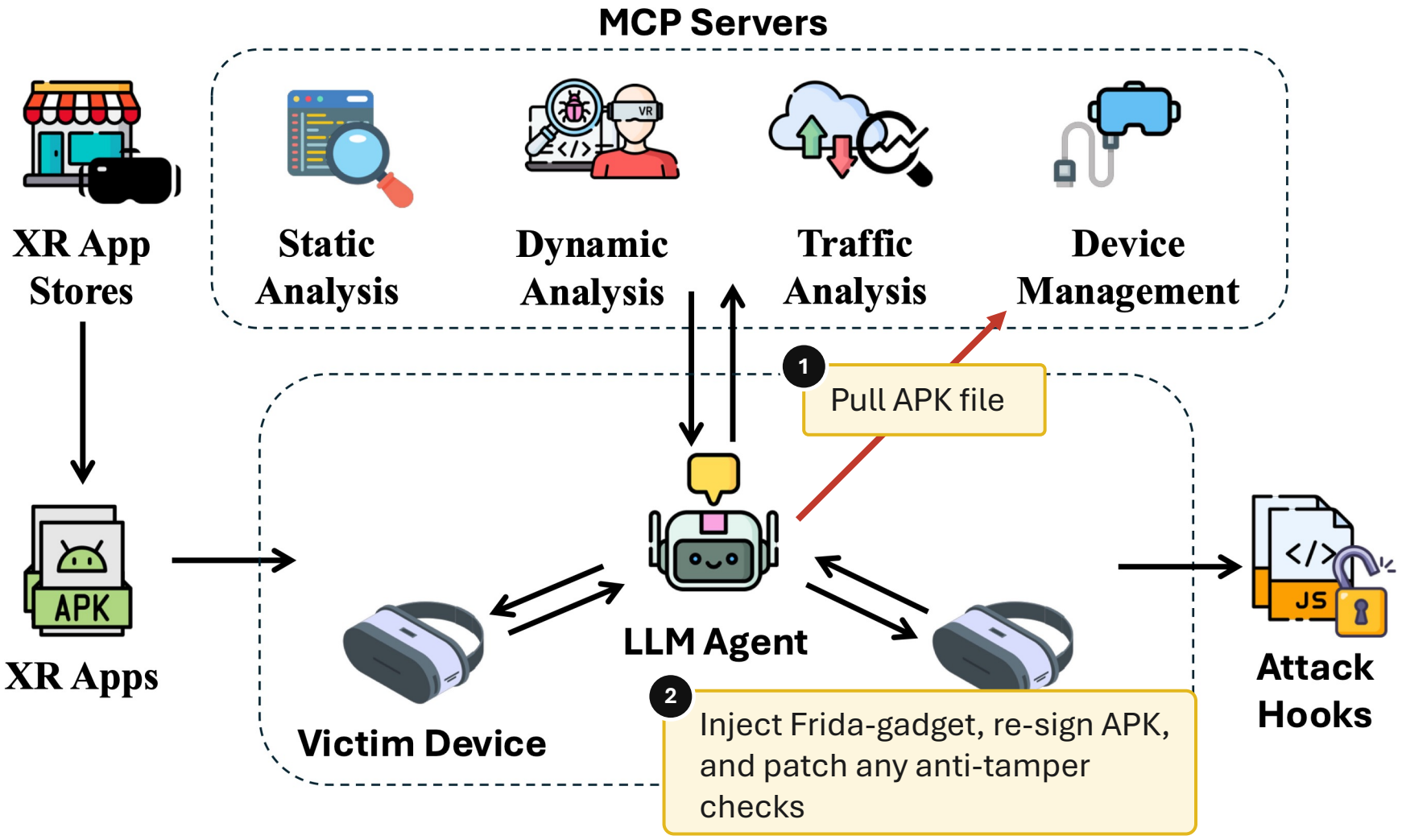


Methodology



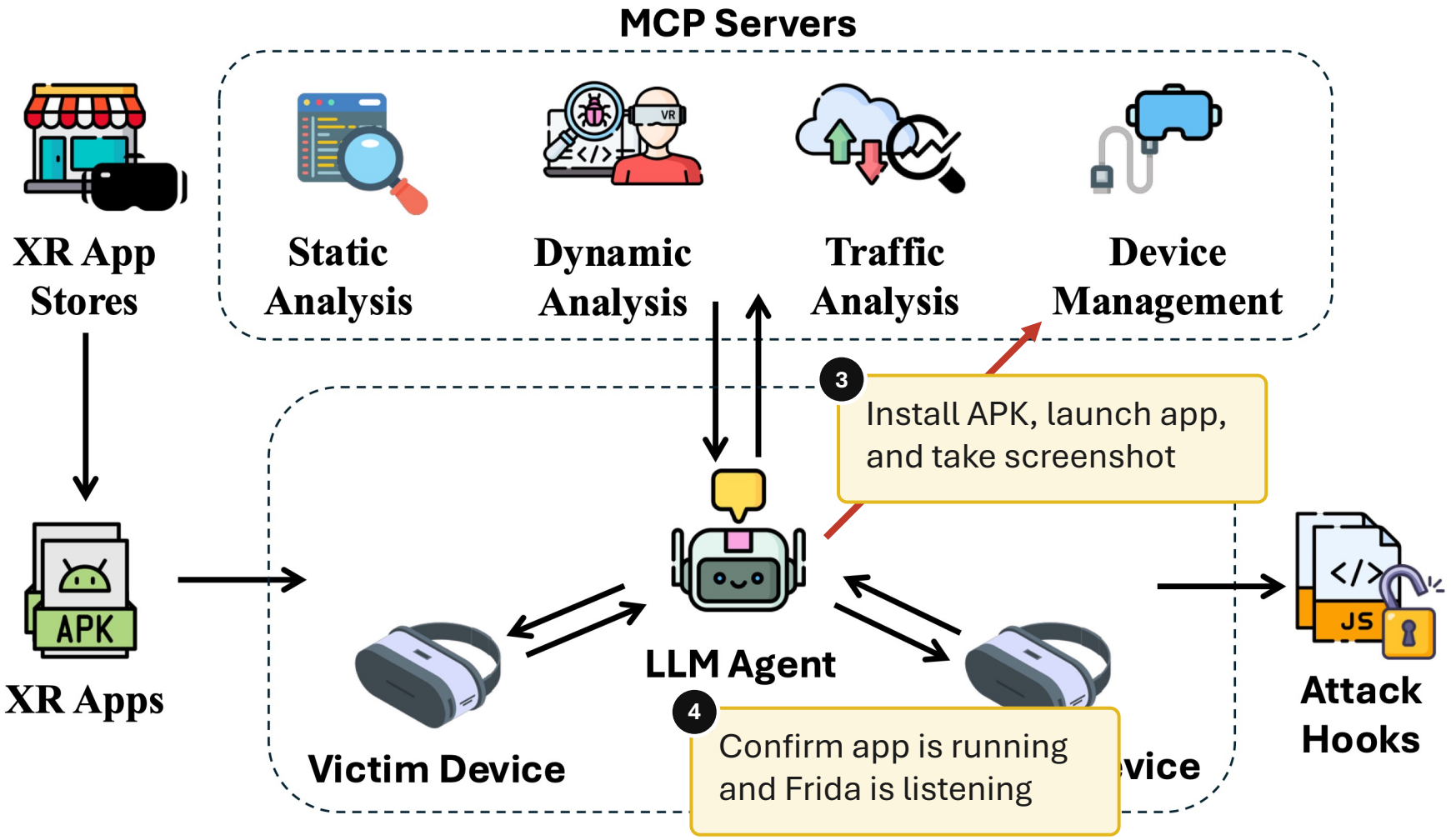
1. **Provisioning for Instrumentation**
2. Identify security-relevant code
3. Create hook, instrument, run
4. Verify attack success

Methodology



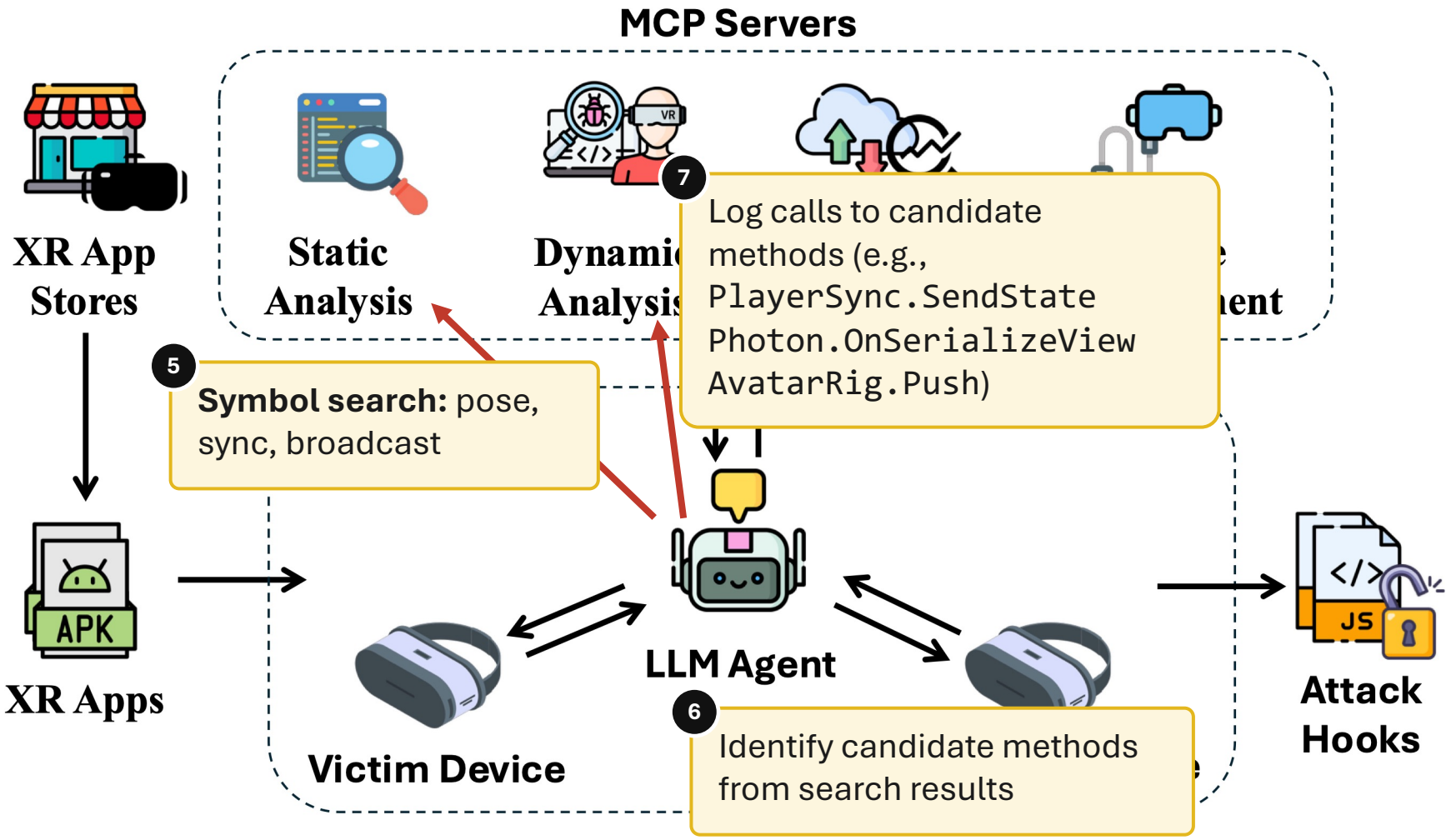
1. **Provisioning for Instrumentation**
2. Identify security-relevant code
3. Create hook, instrument, run
4. Verify attack success

Methodology



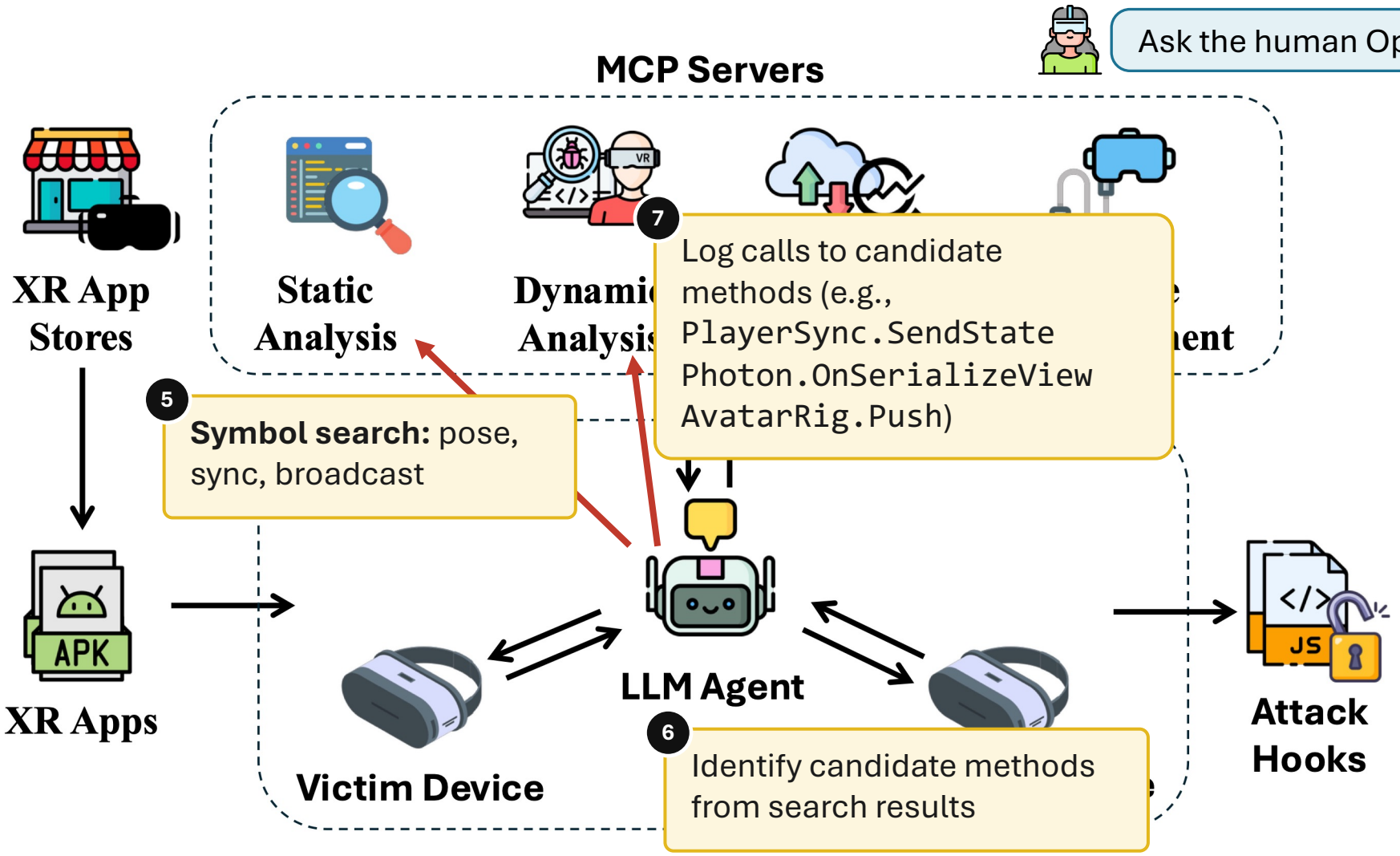
1. **Provisioning for Instrumentation**
2. Identify security-relevant code
3. Create hook, instrument, run
4. Verify attack success

Methodology



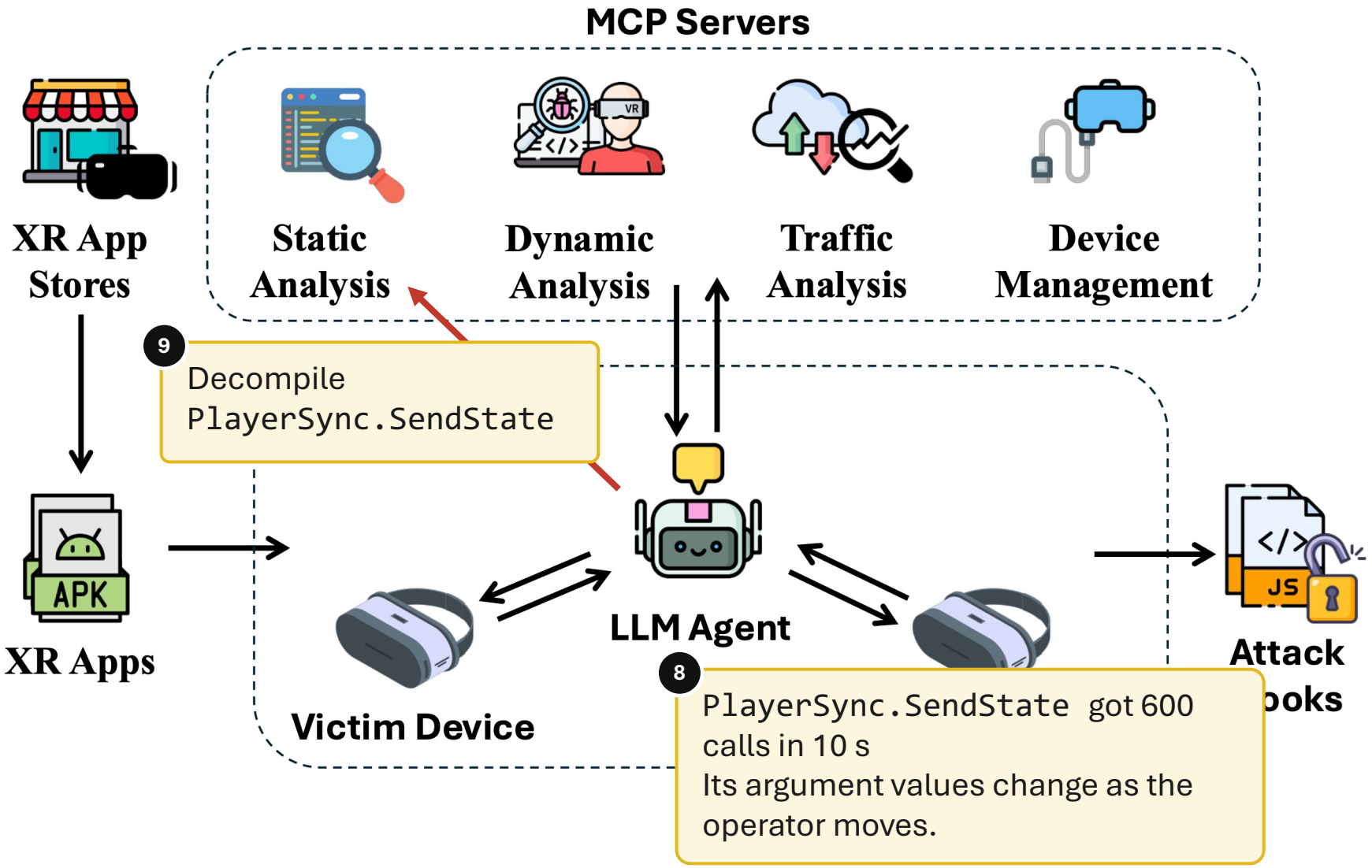
1. Instrumentation and logging
2. **Identify security-relevant code**
3. Create hook, instrument, run
4. Verify attack success

Methodology



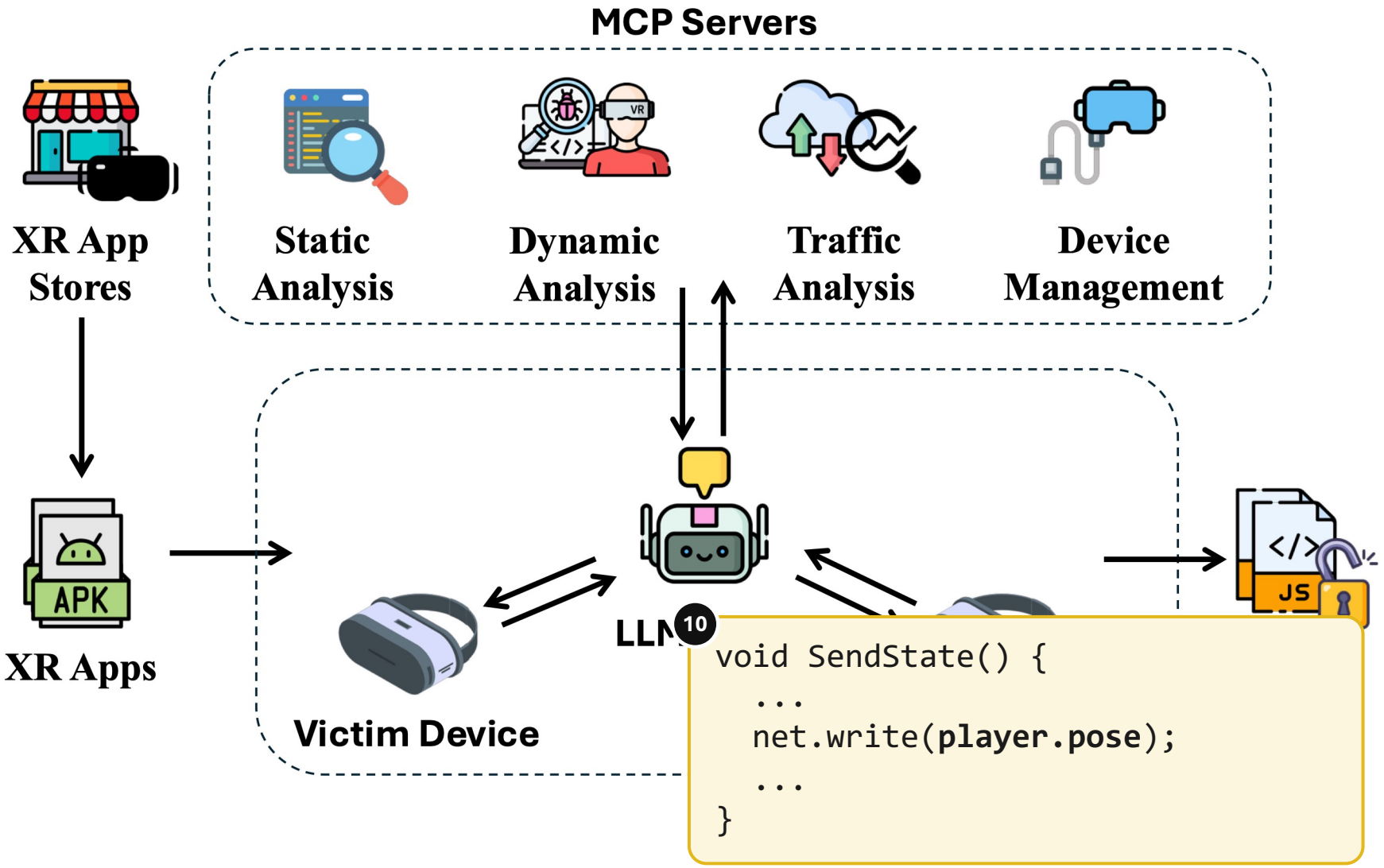
1. Instrumentation and logging
2. **Identify security-relevant code**
3. Create hook, instrument, run
4. Verify attack success

Methodology



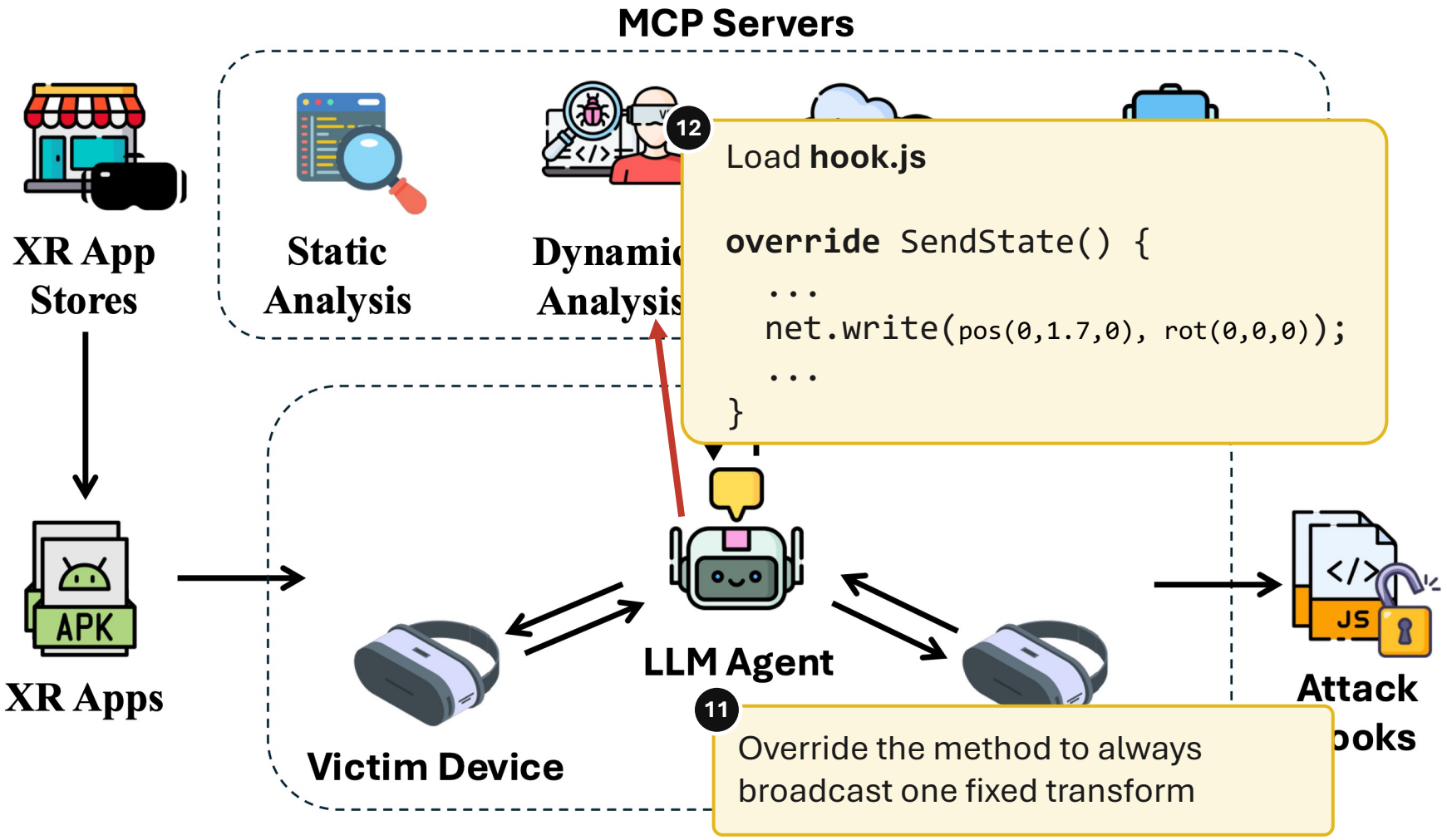
1. Instrumentation and logging
2. **Identify security-relevant code**
3. Create hook, instrument, run
4. Verify attack success

Methodology



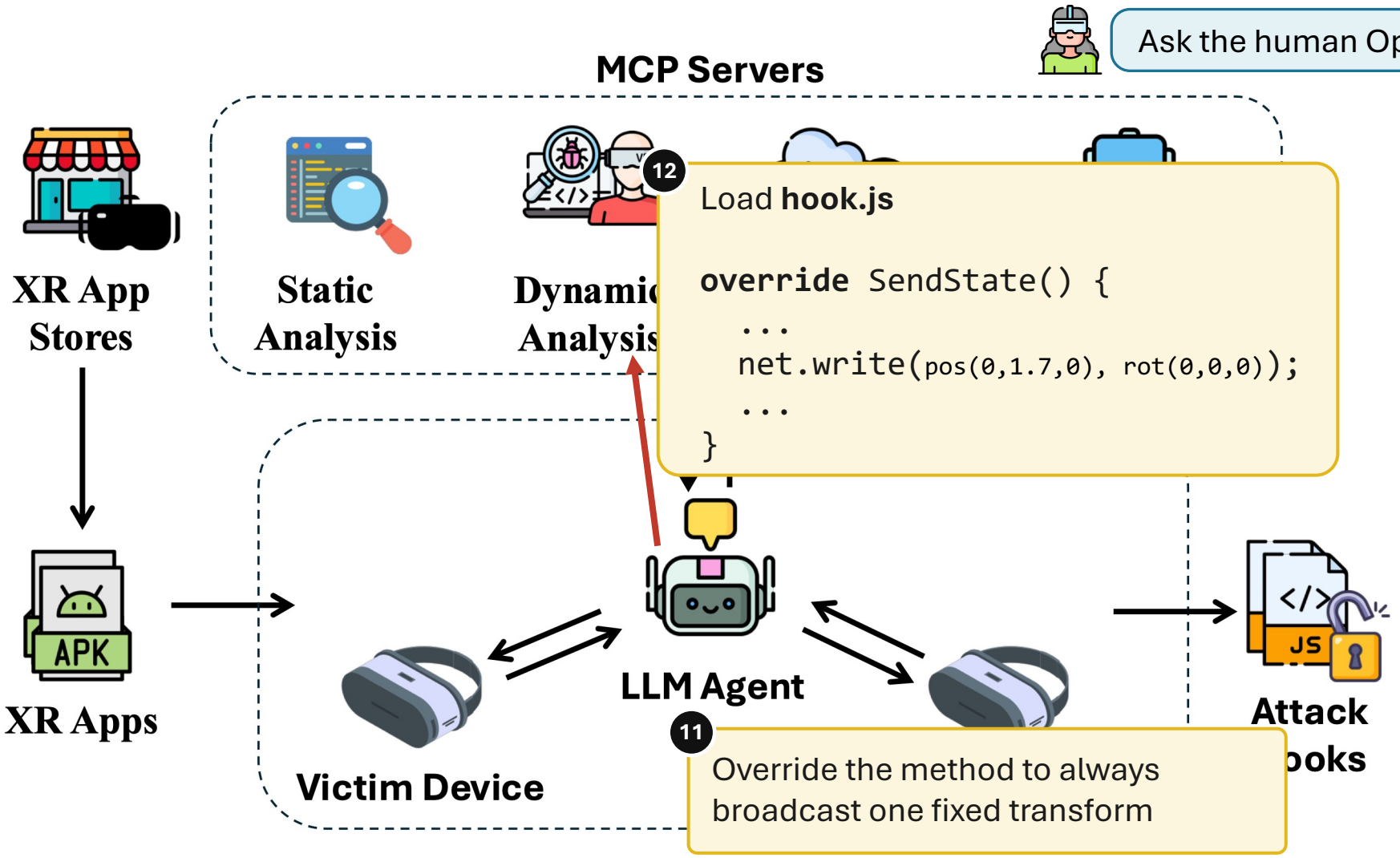
1. Instrumentation and logging
2. **Identify security-relevant code**
3. Create hook, instrument, run
4. Verify attack success

Methodology



1. Instrumentation and logging
2. Identify security-relevant code
3. **Create hook, instrument, run**
4. Verify attack success

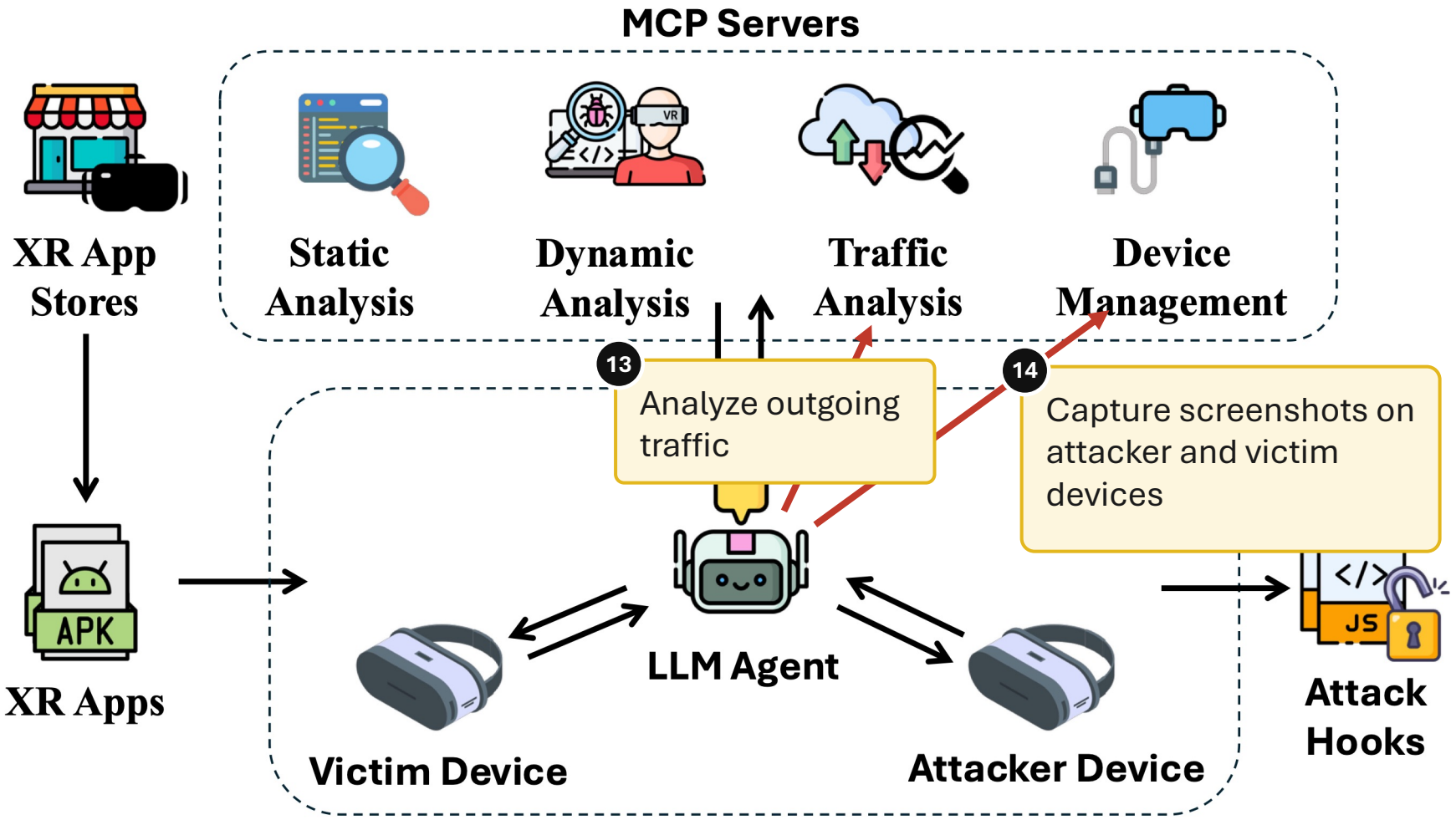
Methodology



Ask the human Operator to walk around the room

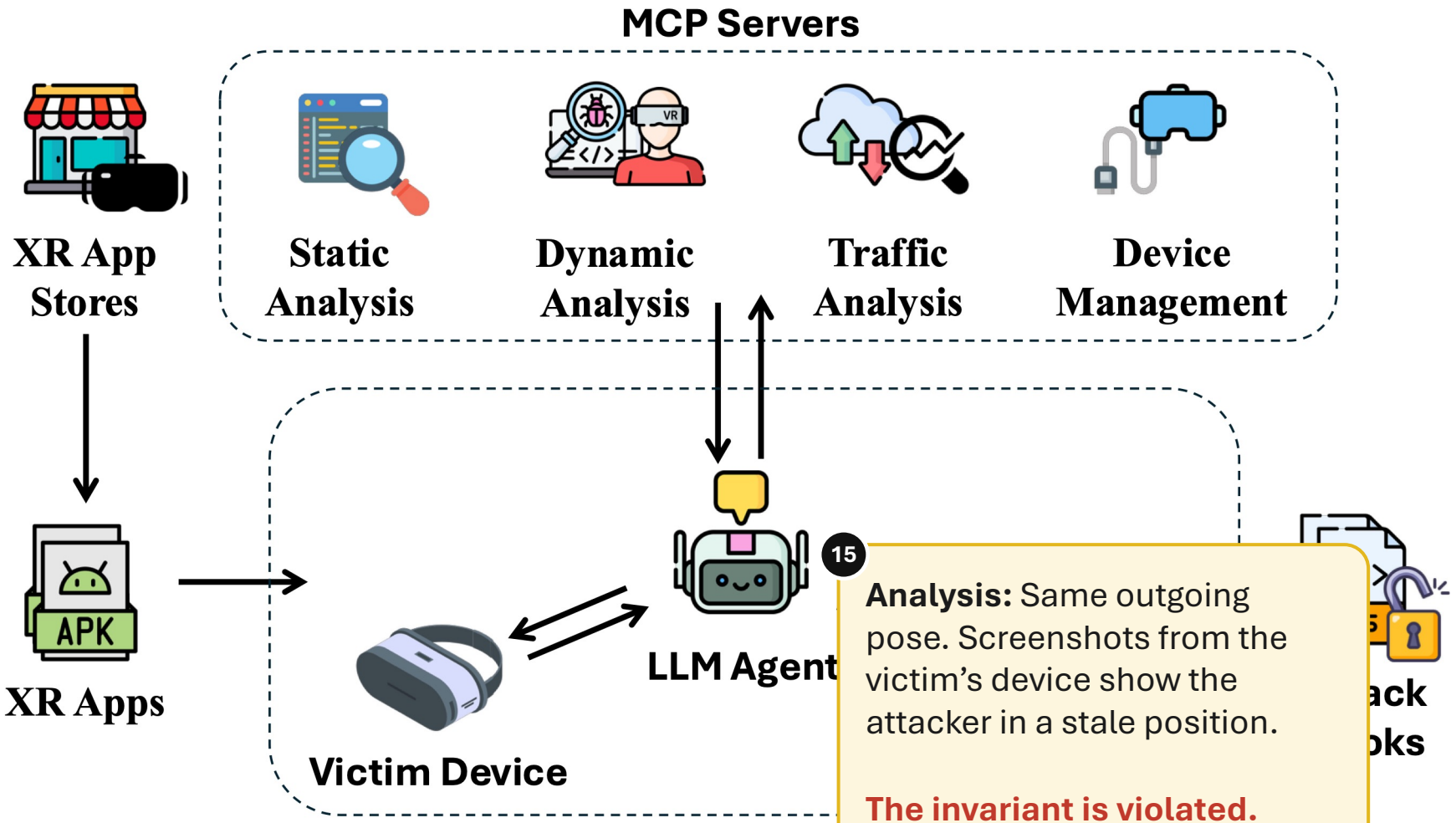
1. Instrumentation and logging
2. Identify security-relevant code
3. **Create hook, instrument, run**
4. Verify attack success

Methodology



1. Instrumentation and logging
2. Identify security-relevant code
3. Create hook, instrument, run
4. **Verify attack success**

Methodology



1. Instrumentation and logging
2. Identify security-relevant code
3. Create hook, instrument, run
4. **Verify attack success**

Findings

20 Apps

VRCHAT
REC ROOM
ENGAGE
Arkio
GORILLA TAG
VICTORYXR
CHILLOUTVR
Spatial

4 Platforms

STEAMVR™
VIVE

8 Attacks



Surveillance
- Video
- Audio



Tampering



Impersonation

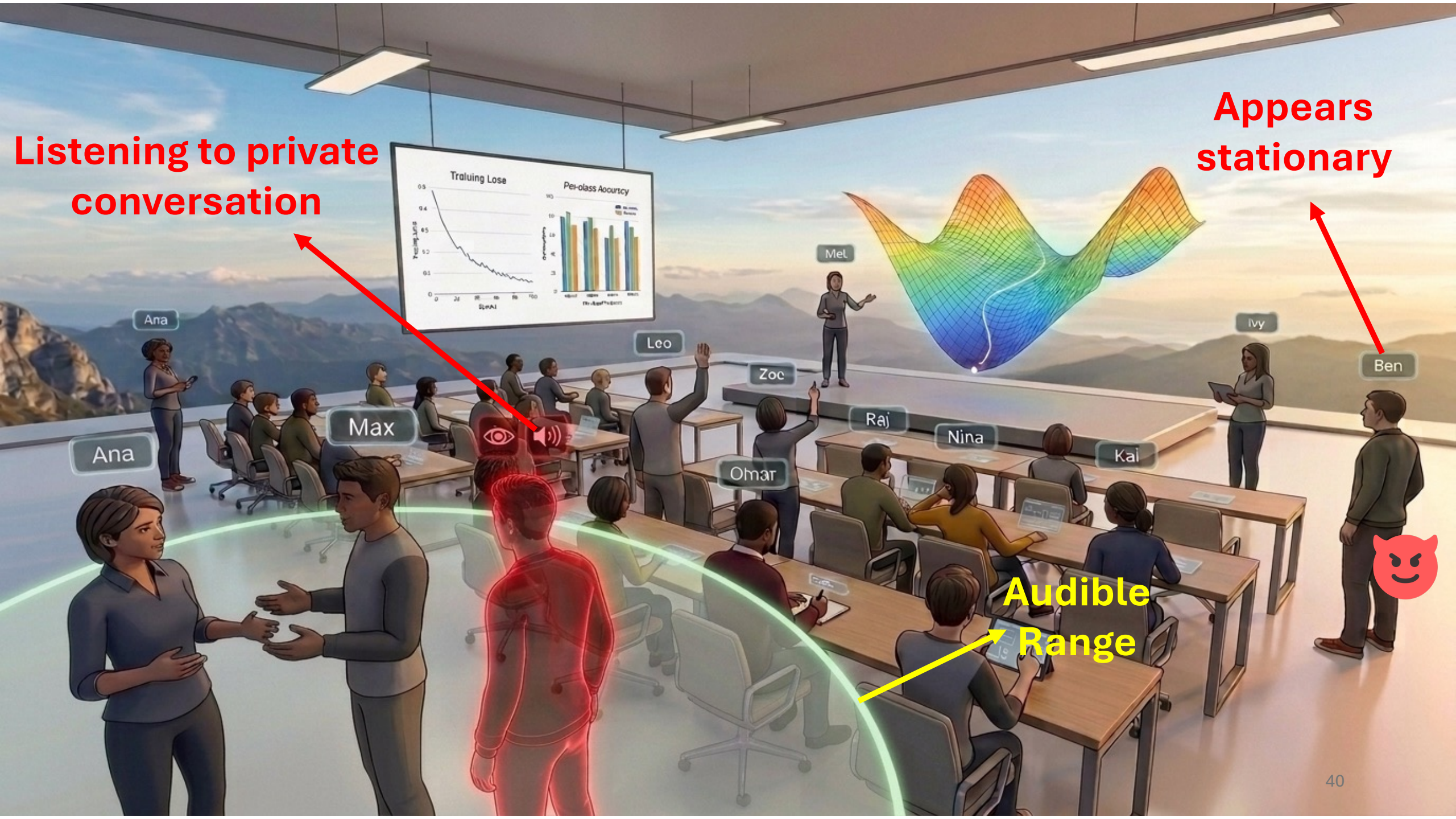


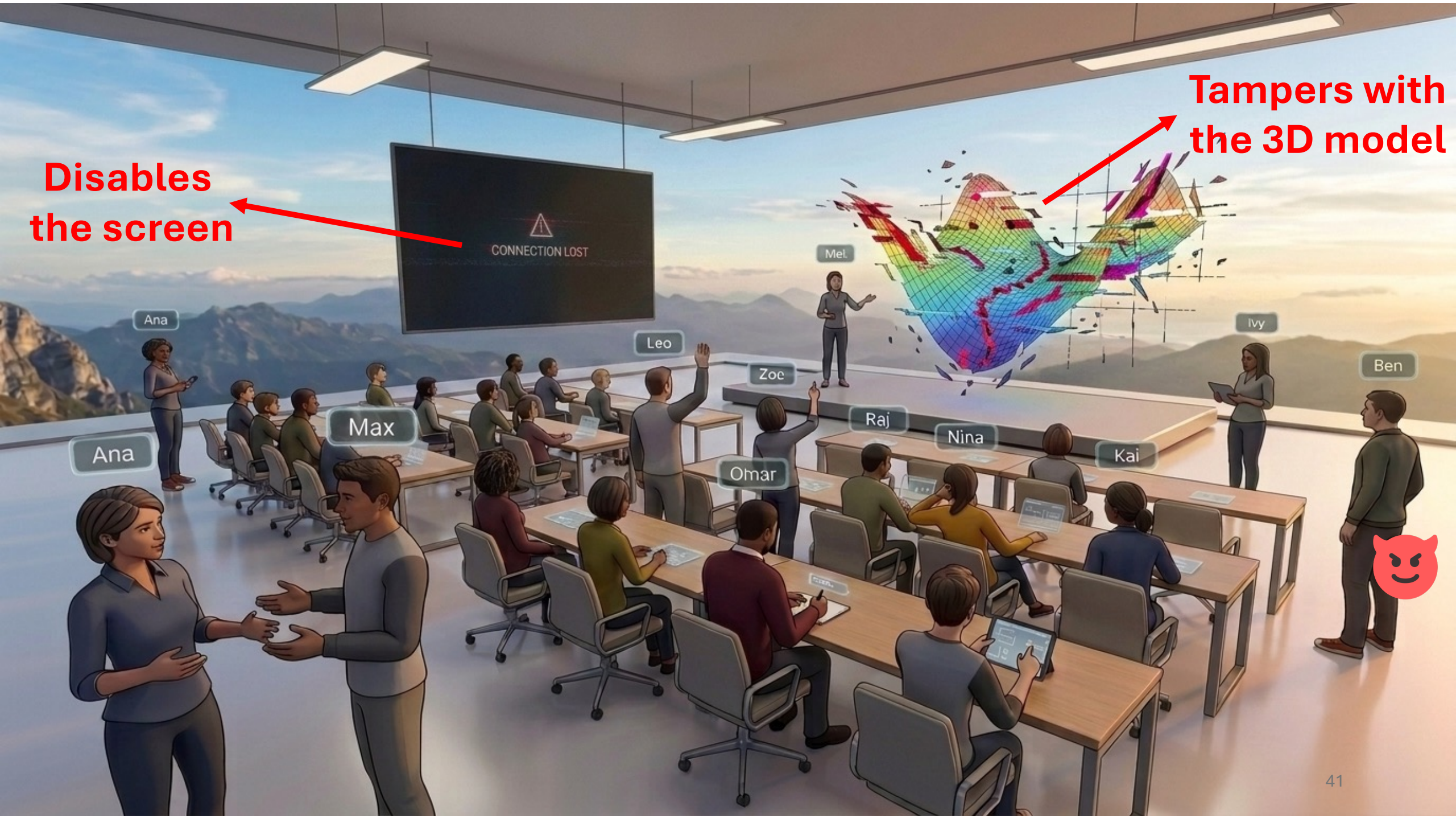
Safety Bypass
- Block
- Host Controls



Disruption
- Avatar Crashing
- Forced Disconnection


Training Loss
Per-class Acourtcy
Mel
Ana
Leo
Zoe
Ivy
Ben
Max
Raj
Nina
Kai
Ana
Omar


Listening to private conversation
Appears stationary
Audible Range
Ana
Max
Ana
Leo
Zoe
Omar
Mel
Raj
Nina
Kai
Ivy
Ben


Disables the screen
Tampers with the 3D model
CONNECTION LOST
Mel
Ana
Leo
Zoe
Ivy
Ben
Max
Ana
Raj
Nina
Kai
Omar

Findings

Which attacks are possible in which apps across all 20 apps?

APP	ATTACKS					PLATFORMS			
Application	Video surveillance	Object tampering	Block evasion	Forced disconn.	Avatar cloning	Quest	Vive	Android	PC VR
Rec Room	✓	✓	✓	✓	—	✓	—	✓	✓
Spatial	✓	✓	—	✓	—	✓	✓	✓	—
ChilloutVR	✓	✓	✓	✗	✓	—	—	—	✓
EngageVR	✓	✓	—	✗	—	✓	✓	✓	✓
3D Organon XR	✓	✓	—	✓	—	✓	✓	✓	✓
Banter	✓	✓	✓	✗	✓	✓	—	—	✓

• • •

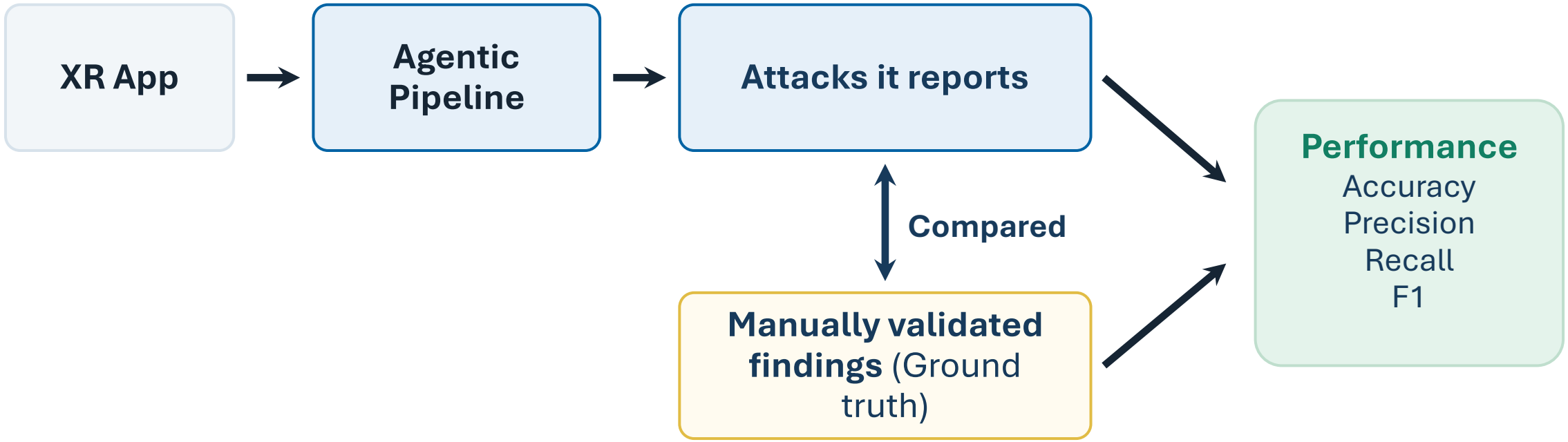
Findings

Which attacks are possible in which apps across all 20 apps?

APP	ATTACKS					PLATFORMS			
Application	Video surveillance	Object tampering	Block evasion	Forced disconn.	Avatar cloning	Quest	Vive	Android	PC VR
Rec Room	✓	✓	✓	✓	—	✓	—	✓	✓
Spatial	✓	✓	—	✓	—	✓	✓	✓	—
ChilloutVR	✓	✓	✓	✗	✓	—	—	—	✓
EngageVR	✓	✓	—	✗	—	✓	✓	✓	✓
3D Organon XR	✓	✓	—	✓	—	✓	✓	✓	✓
Banter	✓	✓	✓	✗	✓	✓	—	—	✓

• • •

Evaluation


XR App
Agentic Pipeline
Attacks it reports
Compared
Manually validated findings (Ground truth)
Performance
Accuracy
Precision
Recall
F1

Results

Attack	Accuracy	Precision	Recall	F1
Video surveillance	1.00	1.00	1.00	1.00
Audio surveillance	0.93	1.00	0.93	0.96
Object manipulation	0.79	1.00	0.79	0.88
Block evasion	0.79	0.57	1.00	0.73
Host-control evasion	0.93	1.00	0.86	0.92
Forced disconnection	0.71	0.57	0.80	0.67
Overall (macro-average)	**0.86**	**0.86**	**0.90**	**0.86**

Results

Attack	Accuracy	Precision	Recall	F1
Video surveillance	1.00	1.00	1.00	1.00
Audio surveillance	0.93	1.00	0.93	0.96
Object manipulation	0.79	1.00	0.79	0.88
Block evasion	0.79	0.57	1.00	0.73
Host-control evasion	0.93	1.00	0.86	0.92
Forced disconnection	0.71	0.57	0.80	0.67
Overall (macro-average)	**0.86**	**0.86**	**0.90**	**0.86**

Results

Attack	Accuracy	Precision	Recall	F1
Video surveillance	1.00	1.00	1.00	1.00
Audio surveillance	0.93	1.00	0.93	0.96
Object manipulation	0.79	1.00	0.79	0.88
Block evasion	0.79	0.57	1.00	0.73
Host-control evasion	0.93	1.00	0.86	0.92
Forced disconnection	0.71	0.57	0.80	0.67
Overall (macro-average)	**0.86**	**0.86**	**0.90**	**0.86**

Results

Attack	Accuracy	Precision	Recall	F1
Video surveillance	1.00	1.00	1.00	1.00
Audio surveillance	0.93	1.00	0.93	0.96
Object manipulation	0.79	1.00	0.79	0.88
Block evasion	0.79	0.57	1.00	0.73
Host-control evasion	0.93	1.00	0.86	0.92
Forced disconnection	0.71	0.57	0.80	0.67
Overall (macro-average)	**0.86**	**0.86**	**0.90**	**0.86**

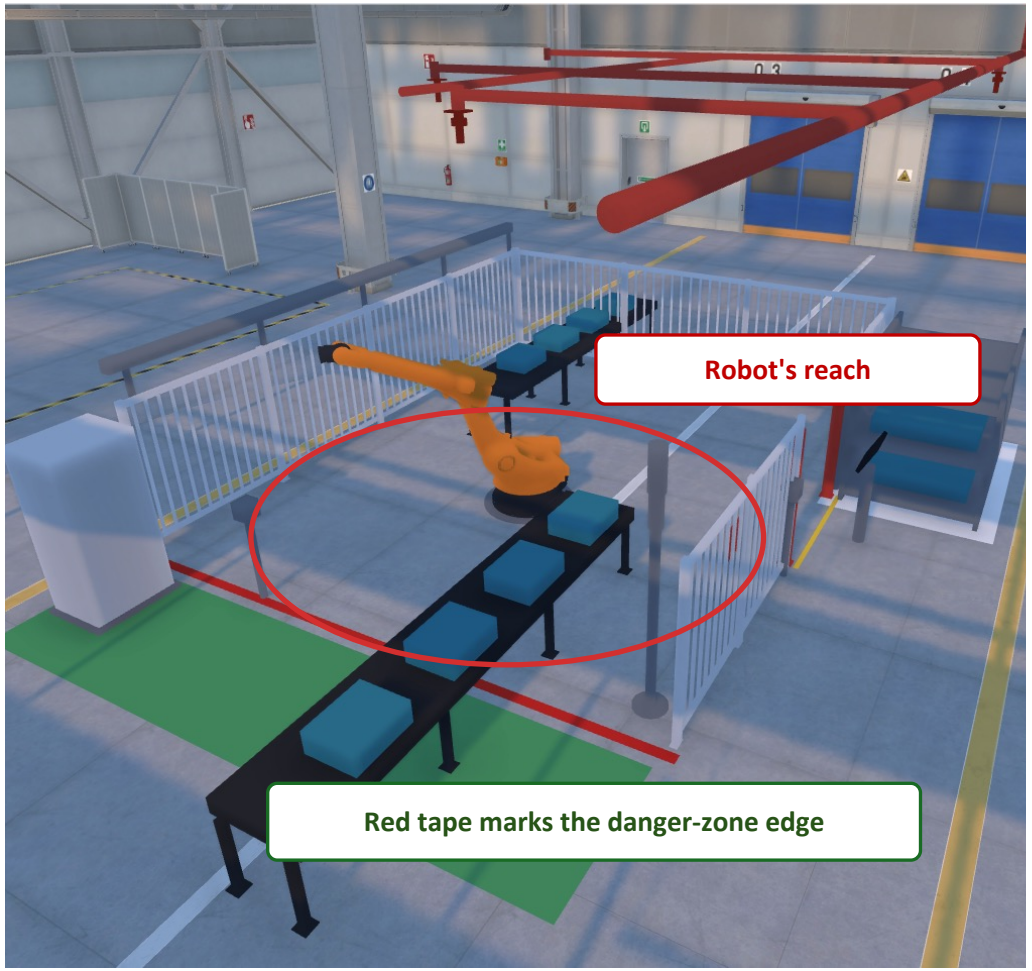
Cost Analysis

- **Manual Effort**
  - 60–90 minutes end-to-end per app (most of it autonomous)
  - Only 5–10 minutes of human operator time

- **Token Usage**
  - 1–2M input tokens, dominated by Ghidra decompilation returned through the static server

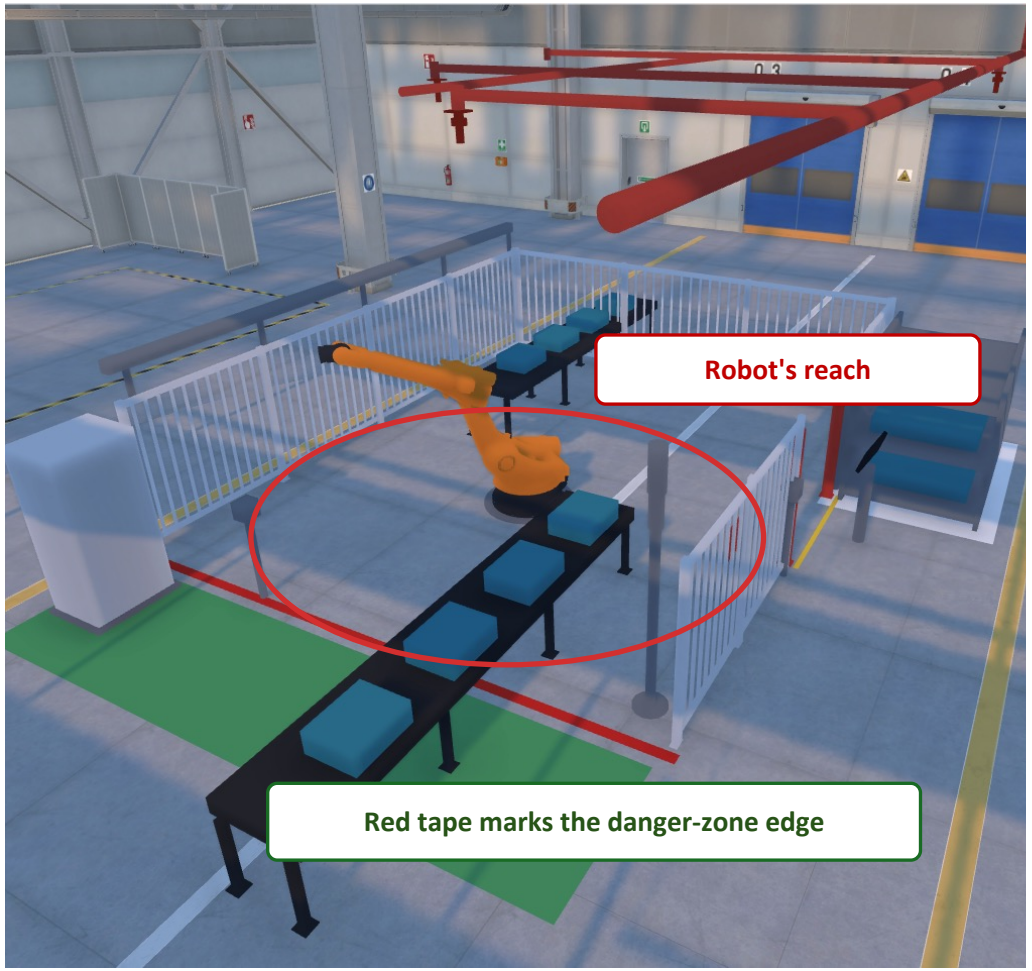
Collaborative Design and Review

As designed



Collaborative Design and Review

As designed

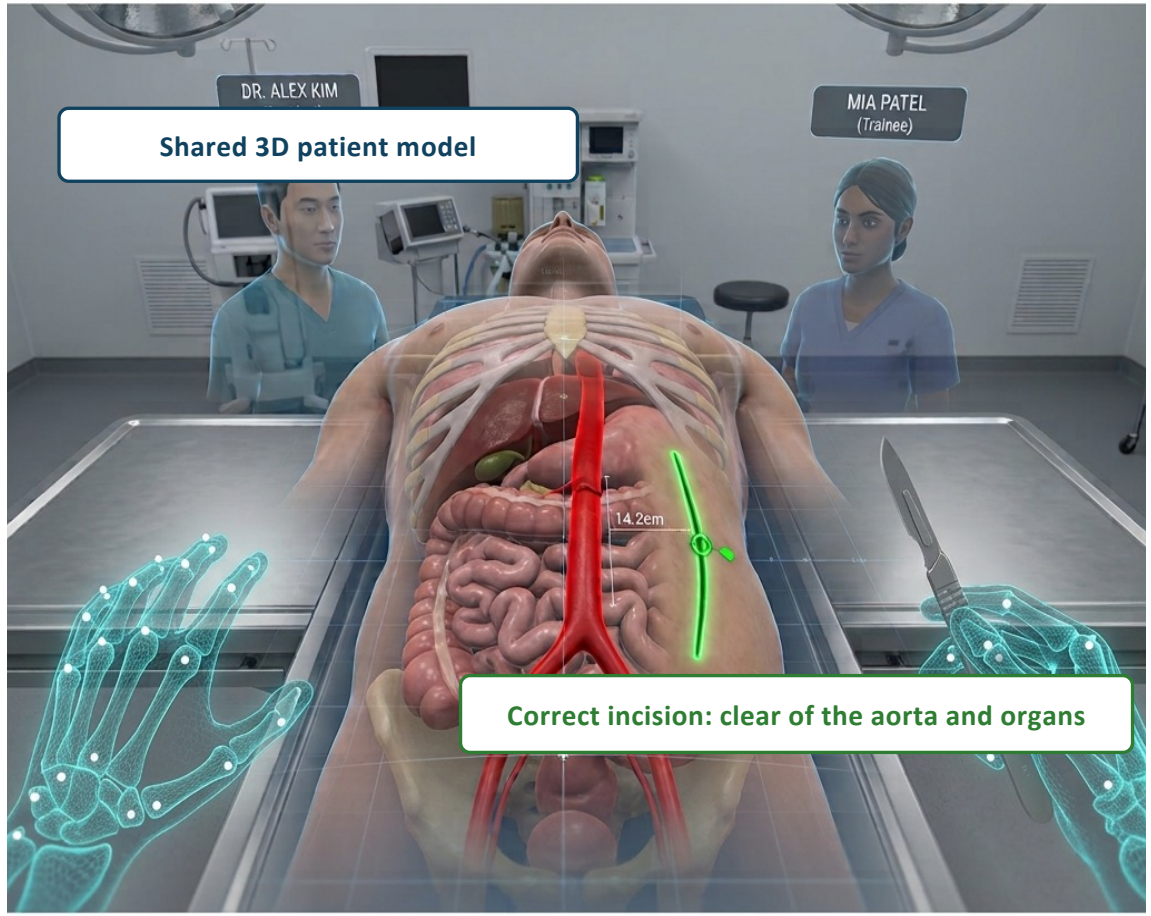


After tampering



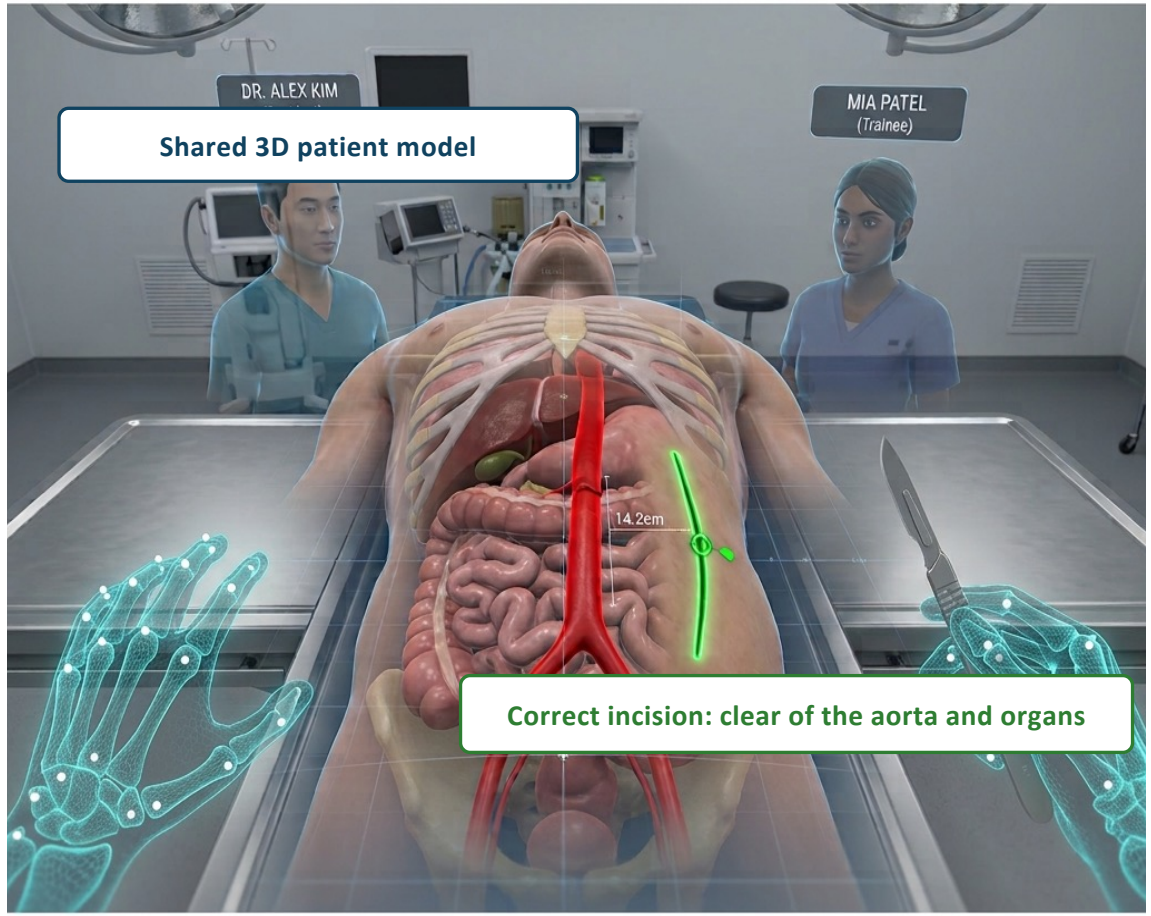
Medical Training

As designed

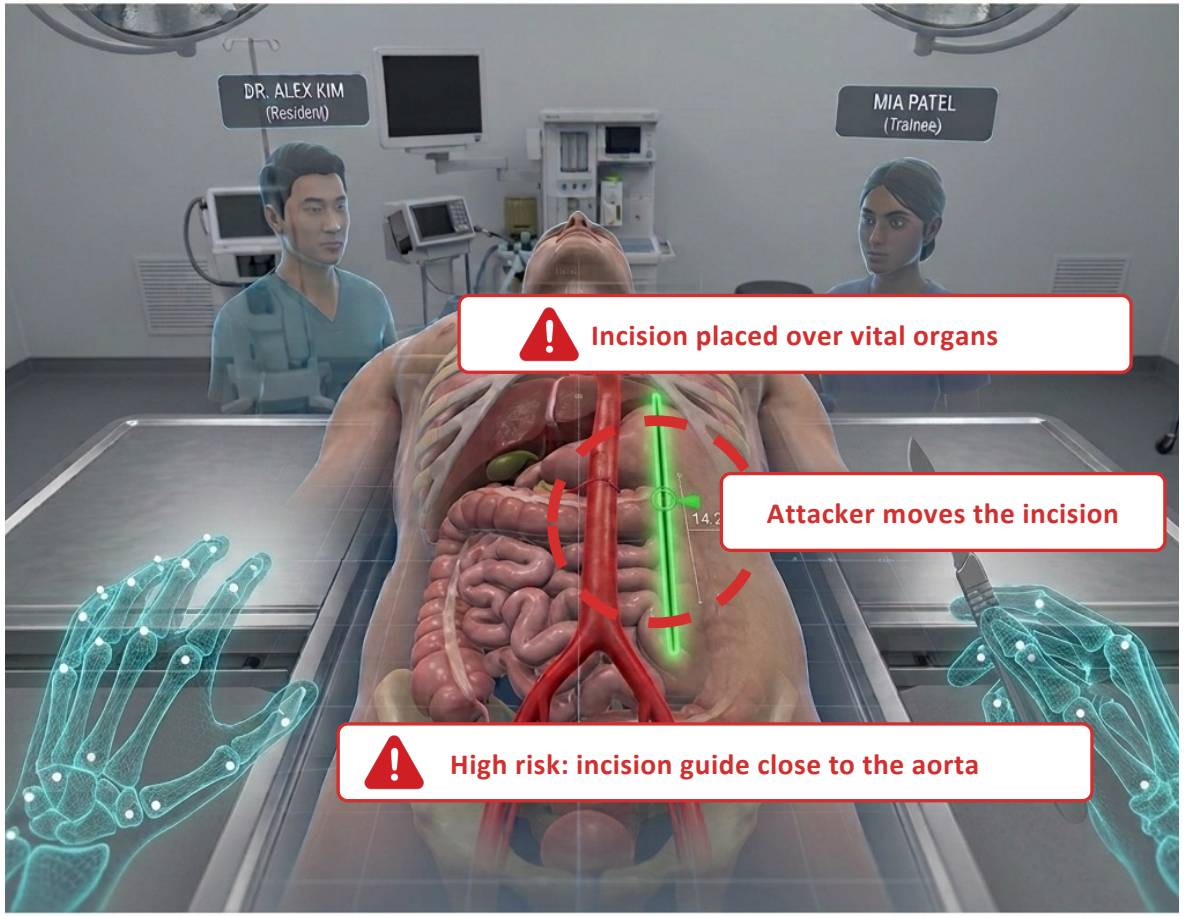


Medical Training

As designed



After tampering



Takeaways

Client-side trust introduces **security risks** in multi-user XR, and there is a trade-off between security and immersion.

Findings **disclosed** to vendors. Vendors can use this framework to perform **security regression tests** before releases.

Prompts, scripts, and MCP tooling released for reuse and reproducibility.

Defenses

- Selective server-side validation
- Client integrity and attestation
- State minimization

Video Demos of Attacks