SHORT-PAPER

Demo: Perception Graph for Cognitive Attack Reasoning in Augmented Reality

RONGQIAN CHEN, The George Washington University, Washington, D.C., United States

SHU HONG, The George Washington University, Washington, D.C., United States

RIFATUL ISLAM, Kennesaw State University, Kennesaw, GA, United States

MAHDI IMANI, Northeastern University, Boston, MA, United States

GANG TAN, Pennsylvania State University, University Park, PA, United States

TIAN LAN, The George Washington University, Washington, D.C., United States

Open Access Support provided by:

The George Washington University

Northeastern University

Kennesaw State University

Pennsylvania State University

Published: 27 October 2025

Citation in BibTeX format

MobiHoc '25: Twenty-sixth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing
October 27 - 30, 2025
TX, Houston, USA

Conference Sponsors:
SIGMOBILE

Demo: Perception Graph for Cognitive Attack Reasoning in Augmented Reality

Rongqian Chen
George Washington University
Washington D.C., USA
rongqianc@gwu.edu

Shu Hong
George Washington University
Washington D.C., USA
shu.hong@gwu.edu

Rifatul Islam
Kennesaw State University
Kennesaw, GA, USA
rislam11@kennesaw.edu

Mahdi Imani
Northeastern University
Boston, MA, USA
m.imani@northeastern.edu

Gang Tan
Pennsylvania State University
University Park, PA, USA
gtan@psu.edu

Tian Lan
George Washington University
Washington D.C., USA
tlan@gwu.edu

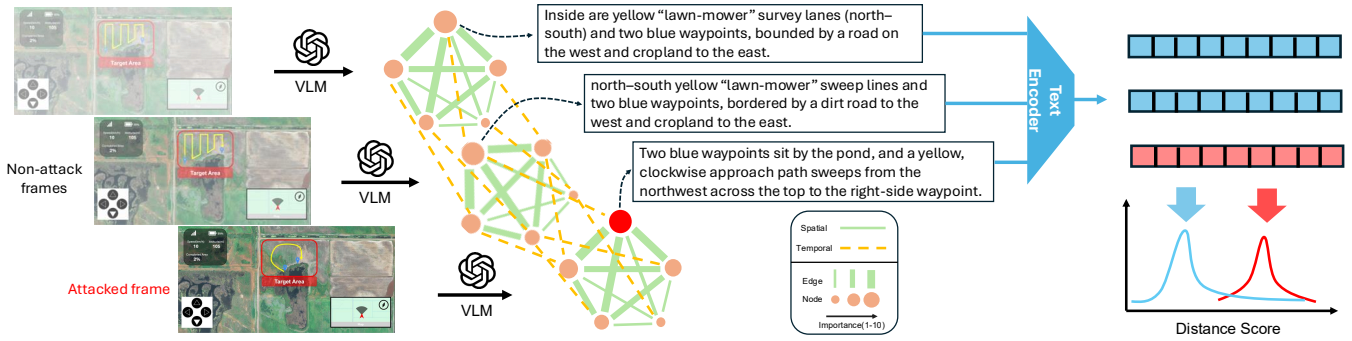

Figure 1: Perception Graph overview — cognitive knowledge is spatially and temporally stored, encoded, and reasoned upon.

Abstract

Augmented reality (AR) systems are increasingly deployed in tactical environments, but their reliance on seamless human-computer interaction makes them vulnerable to cognitive attacks that manipulate a user's perception and severely compromise user decision-making. To address this challenge, we introduce the Perception Graph, a novel model designed to reason about human perception within these systems. Our model operates by first mimicking the human process of interpreting key information from an MR environment and then representing the outcomes using a semantically meaningful structure. We demonstrate how the model can compute a quantitative score that reflects the level of perception distortion, providing a robust and measurable method for detecting and analyzing the effects of such cognitive attacks.

ACM Reference Format:
Rongqian Chen, Shu Hong, Rifatul Islam, Mahdi Imani, Gang Tan, and Tian Lan. 2025. Demo: Perception Graph for Cognitive Attack Reasoning in Augmented Reality. In *The Twenty-sixth International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '25), October 27–30, 2025, Houston, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3704413.3765312

MobiHoc '25, Houston, TX, USA

ACM ISBN 979-8-4007-1353-8/2025/10
https://doi.org/10.1145/3704413.3765312

1 Introduction

In augmented reality (AR) systems, the seamless blending of digital and physical worlds introduces new vulnerabilities to cognitive attacks that manipulate human perception. Such attacks—like inserting fake objects or removing real ones—can severely impair a user's ability to interpret critical information. Traditional computer vision models, which typically operate at the pixel level, often fail to detect these semantically meaningful alterations. Supervised learning approaches also face limitations, as they require large amounts of training data, which is difficult to collect in real AR environments and across diverse attack patterns, especially in safety-critical scenarios.

To address these limitations, we developed the Perception Graph model, a novel approach for modeling human perceptions in a few shots in AR systems. Our model uses pre-trained vision language models (VLMs) to mimic human interpretation and understanding [1]. It transforms AR visuals into a context-aware, semantically rich representation that detects perception changes and computes a distortion score, quantifying cognitive attack impact. This yields human-like understanding and a robust, interpretable foundation for mission-critical perception security.

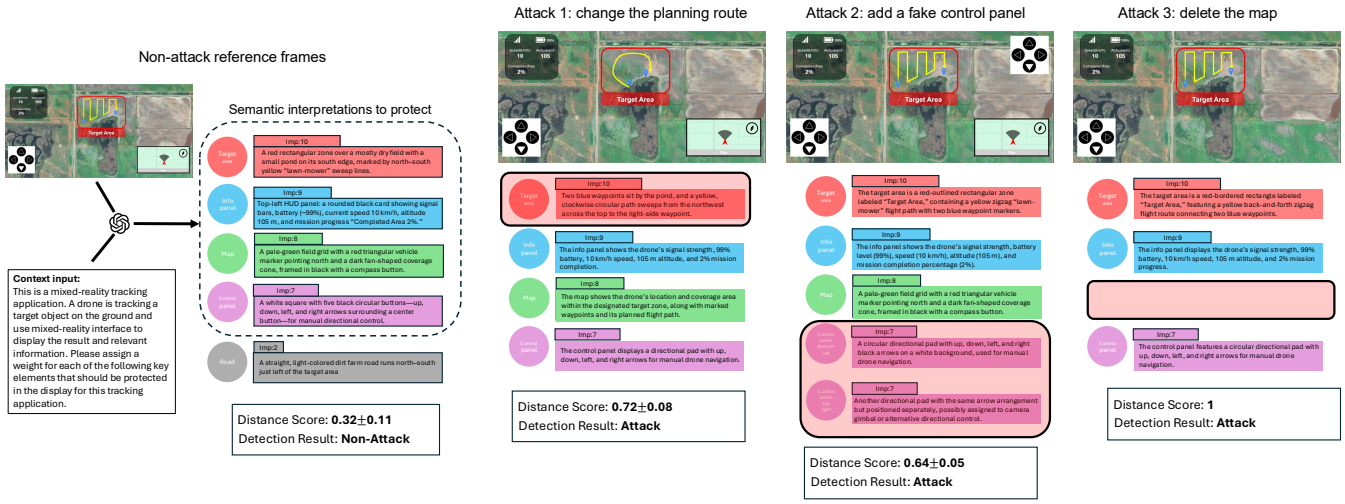

Figure 2: Attack reasoning examples in an agricultural drone scene — alterations in Perception Graph information or structural changes result in higher distance scores.

2 Methodology

Our Perception Graph model is designed to detect cognitive attacks in AR environments through a two-phase process.

Graph Construction Phase. As illustrated in Fig.1 and Fig.2, the construction phase begins by generating reference graphs that encode the ground-truth semantic interpretation of a scene. These graphs are derived from the semantic outputs of Vision–Language Models (VLMs). While VLMs capture the underlying meaning of objects and relationships, their natural language descriptions often vary in word choice and phrasing. To resolve this variability, each description is passed through a text encoder, which projects it into a latent embedding space. In this space, semantic meaning is represented by the direction of the embedding vector, and similarities between descriptions can be consistently measured with cosine similarity.

In addition to encoding meaning, the construction phase also generate contextual weights for each object. These weights quantify the object's relative importance in the scene, allowing the model to focus protection efforts on critical objects. For example, traffic signs, hazard warnings, or navigation markers—while treating less relevant objects with lower priority. This selective emphasis ensures that detection resources are concentrated where cognitive attacks would cause the greatest harm.

Detection phase. During the detection phase, the model processes new AR frames to generate a perception graph and aligns it with the stored reference graphs. Scene objects are represented as nodes carrying both semantic embeddings and contextual weights. Semantic changes—such as addition, removal, or modification of nodes—are detected by comparing embeddings of corresponding nodes. To quantify differences, we define a distance function: $Distance = \sqrt{1 - Sim(E_1, E_2)}$, where $Sim(E_1, E_2)$ denotes the cosine similarity between embeddings E_1 and E_2. Smaller distances indicate semantic consistency, while larger distances reveal deviations. A distance of 1 corresponds to a missing node (i.e., no semantic match), and distances exceeding a defined threshold trigger a potential cognitive attack alert.

Table 1: Attack detection based on distance scores and Z-scores relative to the normal distribution ($\mu = 0.32$, $\sigma = 0.11$).

Attack Type	Distance	Z-score	Detection
Route Modification	0.72 ± 0.08	3.6σ	Attack
Fake Control Panel	0.64 ± 0.05	2.9σ	Attack
Map Deletion	1.00	6.2σ	Attack

3 Demonstration

In Fig. 2, we show three representative cognitive attacks and their impact on graph distance scores over 10 reference frames. Under normal conditions, distances follow a distribution with mean $\mu = 0.32$ and standard deviation $\sigma = 0.11$, capturing natural variation in VLM-generated embeddings.

To quantify deviations, we compute the *Z-score* for each observed distance d:

$$Z = \frac{d - \mu}{\sigma}. \quad (1)$$

Table 1 reports the results. Route modification yields 3.6σ, fake control panel 2.9σ, and map deletion 6.2σ—all well outside normal variation. Frames exceeding a threshold (e.g., $Z > 2$) are flagged as potential cognitive attacks.

This statistical mapping transforms raw distances into interpretable evidence of semantic deviation, enabling robust detection of AR cognitive attacks.

References

[1] Rongqian Chen, Allison Andreyev, Yanming Xiu, Mahdi Imani, Bin Li, Maria Gorlatova, Gang Tan, and Tian Lan. 2025. A Neurosymbolic Framework for Interpretable Cognitive Attack Detection in Augmented Reality. *arXiv preprint arXiv:2508.09185* (2025).