Published: 27 October 2025

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

SHORT-PAPER

Poster: Time-Aware LSTM for Gaze Prediction in Mixed Reality Under Latency Perturbations

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

RONGQIAN CHEN, 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

Pennsylvania State University

Kennesaw State University

Northeastern University

Poster: Time-Aware LSTM for Gaze Prediction in Mixed Reality Under Latency Perturbations

Shu Hong
George Washington University
shu.hong@gwu.edu

Rongqian Chen
George Washington University
rongqianc@gwu.edu

Rifatul Islam
Kennesaw State University
rislam11@kennesaw.edu

Mahdi Imani
Northeastern University
m.imani@northeastern.edu

Gang Tan
Pennsylvania State University
gtan@psu.edu

Tian Lan
George Washington University
tlan@gwu.edu

ABSTRACT

Cognitive attacks in mixed reality (MR), e.g., latency perturbations that induce frame-time jitter, can divert visual attention and degrade task performance. We study 2D gaze prediction under such disturbances and propose a time-aware sequence model that handles irregular sampling by supplying elapsed times Δt between observations and conditions on sparse event/object context available at prediction time via learned token embeddings. Using time-based windows, we evaluate within-user and cross-user temporal generalization on MR recordings spanning multiple attack intensities. Results indicate accurate, time-robust gaze regression under latency perturbations, supporting adaptive MR interfaces in adversarial settings.

CCS CONCEPTS

• **Human-centered computing** → **Interaction design**.

KEYWORDS

Gaze Prediction, Cognitive Attack, LSTM, Mixed Reality, Latency Perturbation

ACM Reference Format:
Shu Hong, Rongqian Chen, Rifatul Islam, Mahdi Imani, Gang Tan, and Tian Lan. 2025. Poster: Time-Aware LSTM for Gaze Prediction in Mixed Reality Under Latency Perturbations. 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/10.1145/3704413.3765313

1 INTRODUCTION

Mixed reality (MR) environments present unique challenges for human attention modeling due to the complex interplay between virtual and physical elements in the user's field of view. Cognitive attacks in MR, e.g., induced latency spikes, intentionally burden perception and decision-making, diverting user attention away from task goals. In cognitive psychology, gaze is a reliable behavioral correlate of the locus of visual attention [1–3]; hence, predicting a user's gaze point in real time offers a practical control signal for adaptive interfaces and safety.

MobiHoc '25, October 27–30, 2025, Houston, TX, USA

ACM ISBN 979-8-4007-1353-8/2025/10
https://doi.org/10.1145/3704413.3765313

2 DATASET AND PRE-PROCESSING

2.1 Dataset Description

The gaze prediction model was developed based on data from Penn State's Tactical Search and Rescue (TSR) testbed. We utilize 8 datasets from three participants with six different latency levels, each completing ten-minute MR training sessions involving dynamic latency conditions representing cognitive attacks. During each one-minute interval, participants experience continuous cognitive attacks at a fixed intensity level. Multimodal data were captured, including eye-tracking metrics (e.g., eye position, gaze direction, gaze targets), inertial head-tracking data (head position and rotation), real-time scene video features (e.g., brightness, luminance, and spectral entropy), event logs (e.g., LaserGunFired, MedBoxUsed), and system telemetry parameters such as latency levels. Specifically, we obtain time-synchronized numeric streams and auxiliary signals as:

- **Eye-Tracking (10 dims, averaged across eyes):** per-frame 3D eye position (x, y, z), eye rotation quaternion (q_x, q_y, q_z, q_w), and 3D gaze direction (d_x, d_y, d_z), averaged from left/right into `AvgEye*` and `AvgGazeDirection*`.
- **Head-Tracking (7 dims):** 3D head position and head rotation quaternion (3+4).
- **Attack Intensity (1 dim):** a per-frame scalar joined from the per-interval attack schedule/logs (FPS/latency level) and forward-filled within each interval.
- **Object Tokens:** lists of object names/IDs aligned to the nearest sensor frame.
- **Event Tokens:** lists of event names/IDs aligned to the nearest sensor frame within a small tolerance.
- **Ground-Truth Gaze:** 2D gaze (g_x, g_y) from the device in pixel.

2.2 Feature Engineering Pipeline

Numeric features and scaling. All numeric features are z-scored with `StandardScaler` fit on the *training* split and reused for validation/test.

Token embedding. Event/object sets are mapped to integer IDs using a vocabulary built from training data (with `PAD` and `UNK`). Within a batch, tokens are padded, embedded (16-d `nn.Embedding`),

and *mean-pooled* over non-pad tokens to form a frame-level embedding; we use the tokens from the prediction frame.

Time-based windows. For each anchor time t_i, we form a window $[t_i - W,\ t_i]$ in seconds and include all frames in this interval, preserving irregular timestamps. We *do not* resample to a fixed length; per-step gaps $\Delta t_j = t_j - t_{j-1}$ are carried to the model. Targets are future screen-normalized gaze $(g_x, g_y) \in [0,1]^2$ at horizon Δ_{pred} (e.g., 100 ms), aligned to the nearest label within $\pm\epsilon$ (we use $\epsilon = 40$ ms) using only observations up to t_i.

3 MODEL

We predict future gaze $\hat{\mathbf{g}} = (\hat{g}_x, \hat{g}_y) \in [0,1]^2$ at a fixed horizon Δ_{pred} from multimodal telemetry observed up to t_i. Ground-truth $\mathbf{g}_{t_i+\Delta_{\text{pred}}}$ is matched by nearest-neighbor in time within $\pm\epsilon$; inputs strictly exclude future information.

3.1 Overview

Our architecture has three components: (i) a time-aware LSTM encoder for numeric streams; (ii) Last-Observation-Carry-Forward (LOCF)-Decay encoders for sparse event and object signals; and (iii) a fusion MLP that outputs $\hat{\mathbf{g}}$.

3.2 Numeric stream encoder

A 3-layer unidirectional LSTM (hidden size 512) processes the standardized numeric sequence with packed batches. To reduce reliance on stale inputs under irregular sampling, we apply a simple time-decay gate to the final state:

$$\tilde{\mathbf{h}}_i \;=\; \exp(-\alpha\,\Delta t_i)\;\odot\;\mathbf{h}_i, \quad \alpha = \text{softplus}(\tilde{\alpha}) \geq 0,$$

where Δt_i is the gap since the last observed frame.

3.3 Event and object encoders (LOCF-Decay)

For events, we take the most recent token e^* at or before t_i, embed it (16-d), and apply exponential decay with staleness s_{evt}:

$$\bar{\mathbf{e}}_i \;=\; g_{\text{evt}}(s_{\text{evt}})\,\mathbf{U}_E[e^*] \;+\; \big(1 - g_{\text{evt}}(s_{\text{evt}})\big)\,\mathbf{U}_E[\text{NONE}],$$

$$g_{\text{evt}}(s) = \exp(-\beta_{\text{evt}} s),\ \beta_{\text{evt}} \geq 0.$$

We add compact auxiliaries (presence bit and capped staleness).

For objects, we locate the latest non-empty detection, mean-pool object embeddings with (x, y) coordinates through a small MLP ϕ, and apply decay with staleness s_{obj}. The decay rate can adapt to instantaneous head angular speed $\|\boldsymbol{\omega}_{\text{head}}(t_i)\|$:

$$\gamma_{\text{obj}} = \text{softplus}(\tilde{\gamma}_0 + \tilde{\gamma}_1 \|\boldsymbol{\omega}_{\text{head}}(t_i)\|), \quad g_{\text{obj}}(s) = \exp(-\gamma_{\text{obj}} s).$$

We also include presence, capped staleness, and capped object count.

3.4 Fusion and training

We concatenate $\tilde{\mathbf{h}}_i$, the decayed event/object representations, and the small auxiliary vector, and feed the result to an MLP with sigmoid output to ensure $\hat{\mathbf{g}} \in [0,1]^2$. The loss is mean squared error:

$$\mathcal{L}_{\text{MSE}} = \tfrac{1}{2}\big[(g_x - \hat{g}_x)^2 + (g_y - \hat{g}_y)^2\big].$$

4 EVALUATION AND RESULTS

4.1 Experimental Setup

We use time-based windows of $W = 5$ s ending at anchor t_i. The prediction horizon is $\Delta_{\text{pred}} = 100$ ms, so labels are taken at $t_{\text{label}} = t_i + \Delta_{\text{pred}}$. Training anchors are spaced by $S_{\text{train}} = 50$ ms; validation/test anchors use $S_{\text{val}} = S_{\text{test}} = 100$ ms. We train with Adam (lr $= 10^{-3}$), batch size 64, dropout 0.2, weight decay 10^{-5}, for up to 150 epochs with early stopping (patience 15, min_delta $= 10^{-5}$). The best-on-validation checkpoint is used for all reports.

Evaluation protocols.

Within-user: For each of 8 datasets, we perform a temporal split of 72% / 8% / 20% (train/val/test) with W-sized gaps between splits; anchors are restricted to their split.

LOUO: 8-fold cross-user evaluation holding out one user group G_k as test; validation comes from tail segments of the remaining groups with a W gap.

4.2 Results

Within-user training dynamics. Learning curves (Fig. 1) show stable convergence by epochs 30–50, with no overfitting under temporal splits with gaps.

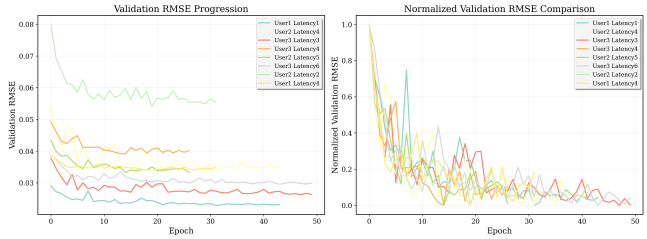

Figure 1: Within-user learning curves across 8 datasets.

Cross-user generalization (Leave-one-user-out, LOUO). Aggregate results across the 8 folds, our model attains $\text{RMSE}_{\text{norm}} = 0.0589$ with positive R^2 along both axes (CIs strictly above zero), indicating non-trivial cross-user generalization.

5 CONCLUSION

We presented a time-aware LSTM for 2D gaze prediction in MR, under cognitively induced latency perturbations (variable FPS). We evaluate both within-user and leave-one-user-out (LOUO) protocols. The approach achieves strong validation performance and robust gaze regression under latency perturbations, providing a practical signal for adaptive MR interfaces.

REFERENCES

[1] Michael I. Posner. 1980. Orienting of attention. *Quarterly Journal of Experimental Psychology* 32, 1 (1980), 3–25.
[2] Akanksha Saran, Ruohan Zhang, Elaine S. Short, and Scott Niekum. 2021. Efficiently Guiding Imitation Learning Agents with Human Gaze. In *Proc. AAMAS*. 1109–1117.
[3] Christopher D. Wickens. 2002. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science* 3, 2 (2002), 159–177.