Gaze prediction in virtual reality using visual and head motion cues

Predicting where you'll look next in VR, without an eye tracker

Springer Best Paper Award — ISVC 2025

C. Petrou[1], Eleanna Kouletou[1], H. Partaourides[2], A. Balomenos[1], Y. Kopsinis[1], S. Chatzis[3]

1 Libra AI Technologies, Athens | 2 Ethical AI Novelties, Limassol | 3 Cyprus University of Technology, Limassol

01 Introduction

Virtual reality headsets work best when they know where the user is looking. This powers techniques like foveated rendering, which sharpens only the area you're focused on to save computing power. But many headsets, including the Apple Vision Pro, don't expose raw eye-tracking data due to hardware limits or privacy rules.

Without knowing where the eyes are pointed, VR systems can lag or waste resources on regions the user isn't even looking at. We built a system that predicts where a user will look next using only head movement and on-screen visual content, no eye tracker required.

02 Our approach

Our model fuses two signals: a lightweight AI (UniSal) that scans video frames for visually striking regions, and the headset's own motion data. A time-series model then combines both to forecast where the eyes will move next, up to a second ahead.

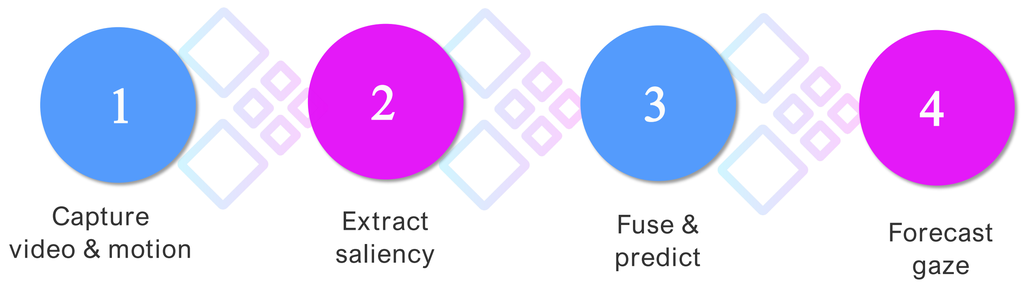

Fig. 1. Multi-modal gaze prediction pipeline: saliency features and head motion are fused and passed through a time-series model to forecast future gaze direction.

03 Dataset at a glance

Evaluated on the EHTask benchmark of 360° VR recordings:

15 360° videos	**30** Participants	**4** Viewing conditions
69,120 Training samples	**30 Hz** Sampling rate	**80/20** Train / val split

04 Key results

Our models cut gaze-prediction error roughly in half versus simple head-position baselines:

5.17° LSTM prediction error (Spherical RMSE)

~49% lower error vs. head-position baseline

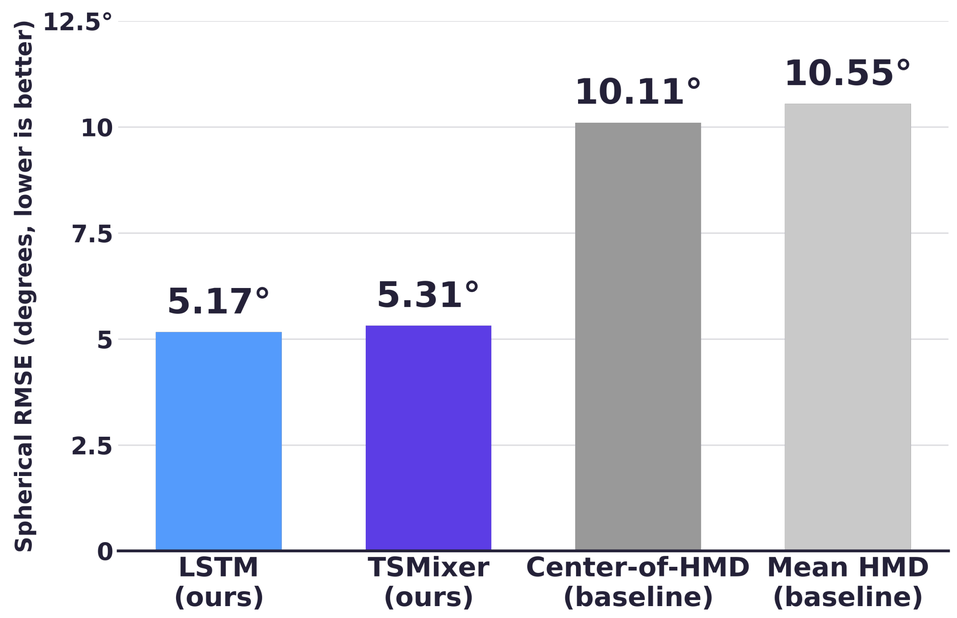

Fig. 2. Gaze prediction error (Spherical RMSE, lower is better): our models vs. simple head-position baselines.

05 System performance

Predictions stayed reliable up to about 333 ms into the future, after which accuracy gradually approached the baseline. LSTM held its advantage better than TSMixer at longer horizons, and the full pipeline ran in real time on both a Linux workstation and Apple Silicon hardware.

333 ms Reliable prediction horizon

Real-time On Linux & Apple Silicon (M1)

06 Why it matters

- Enables foveated rendering and similar techniques without dedicated eye-tracking hardware.
- Respects strict privacy limits on headsets like the Apple Vision Pro, which withhold raw gaze data.
- Runs in real time on both Linux and Apple Silicon, ready for next-generation VR headsets.

Scan for the full paper

References & data

[1] Hu et al. "EHTask: Recognizing User Tasks from Eye and Head Movements in Immersive VR." IEEE TVCG, 2023.

[2] Chen et al. "TSMixer: An All-MLP Architecture for Time Series Forecasting." 2023.

Acknowledgement

This work was supported by the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Actions (MSCA), Grant Agreement No. 101129865 (ACCESS).

Funded by the European Union