


# A Gaze Prediction Model for Task-Oriented Virtual Reality

K. Mammou<sup>1</sup> and K. Mania<sup>1</sup> 

<sup>1</sup>School of Electrical and Computer Engineering, Technical University of Crete, Greece

## Abstract

*In this work, we present a gaze prediction model for Virtual Reality task-oriented environments. Unlike past work which focuses on gaze prediction for specific tasks, we investigate the role and potential of temporal continuity in enabling accurate predictions in diverse task categories. The model reduces input complexity while maintaining high prediction accuracy. Evaluated on the OpenNEEDS dataset, it significantly outperforms baseline methods. The model demonstrates strong potential for integration into gaze-based VR interactions and foveated rendering pipelines. Future work will focus on runtime optimization and expanding evaluation across diverse VR scenarios.*

## CCS Concepts

• **Human-centered computing** → **Virtual reality**; • **Computing methodologies** → **Neural networks**; **Rendering**;

## 1. Introduction

Gaze prediction in Virtual Reality (VR) can replace eye trackers in techniques such as foveated rendering, thus aiding in solving the latency issues they suffer from [ATSD23]. However, the dynamic and immersive nature of VR makes the prediction challenging, especially in real-time, task-oriented applications such as games [KDCM16]. Recent gaze prediction models show promising results. DGaze [HLZ\*20] achieves real-time CNN-based gaze prediction in dynamic scenes under free-viewing conditions, but its performance deteriorates in task-oriented game scenarios. Fixation-Net [HBLW21] focuses on forecasting eye fixations during specific visual search tasks in VR and therefore cannot be directly applied to different tasks. Our work aims to develop a model capable of accurate gaze prediction in VR across diverse task categories while simplifying input requirements to ensure compatibility with existing systems. Ideally, the model should be lightweight and fast, making it suitable for gaze-contingent rendering.

## 2. Implementation

Our approach focuses mainly on the predictive power of temporal continuity [HLG20], since it is particularly evident in task-oriented VR environments, where gaze is frequently guided by specific goals [KDCM16]. Current models [HLZ\*20, HBLW21] usually require a complex input for their prediction, i.e. sequences of frames and gaze, head velocity sequences, as well as object positions sequences. Our model uses only sequences of past frames and gaze points to learn gaze behaviour patterns over time. The proposed architecture consists of three modules: (1) the Image Sequence Module (ISM) to capture temporal motion features from consecutive frames, (2) the Gaze Sequence Module (GSM) to learn

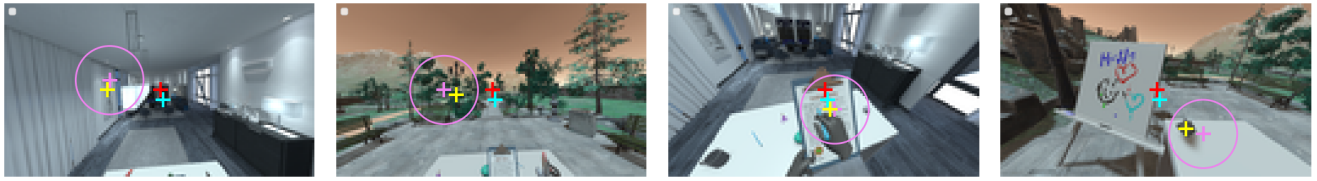
temporal gaze patterns, and (3) the Gaze Fusion Module (FM) that integrates both outputs to predict a single gaze point.

### 2.1. Dataset and Pre-processing

We used the OpenNEEDS dataset [EZW\*21], which includes data from 44 users performing tasks such as reading, manipulating objects, drawing, aiming and shooting in two virtual environments. We utilized the 8-bit sRGB frames down-sampled to 128×71 resolution and the 3D gaze vectors. Frames were normalized to the [0, 1] range, and gaze points were converted to 2D visual angles and similarly normalized. Outliers were removed using the Interquartile Range Method (IQR) to improve model robustness and performance. Input sequences were then created, with each sequence consisting of consecutive frames and their corresponding gaze points, ensuring data continuity from the same user and scene. The final dataset was split into training (80%), validation (10%) and testing sets (10%).

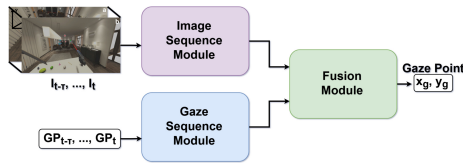
### 2.2. Model Architecture

The ISM consists of 5 ConvLSTM2D layers (ReLU activation), 4 MaxPooling layers for dimensionality reduction, and 4 fully connected (FC) layers (sizes 64, 32, 32, and 16). A dropout layer (rate = 0.5) prevents overfitting. The input is a sequence of 10 continuous frames. The GSM processes a sequence of 10 continuous gaze points with 4 LSTM layers (ReLU), 2 FC layers (sizes 32 and 16), and a dropout layer. Finally, the FM merges the ISM and GSM outputs using a maximum operation, followed by a FC layer (size 16, ReLU) and a dropout layer. The final FC layer (size 2, sigmoid) outputs the predicted gaze point  $(x_g, y_g)$ . The model is trained using Mean Absolute Error as the loss function and the Adam opti-



**Figure 1:** The purple cross denotes the ground truth gaze position, with the purple circle illustrating the foveal region with radius 15°. The yellow cross represents the prediction of our model, the red cross shows the center baseline and the blue the mean baseline.

mizer with an initial learning rate of 0.001. Training was performed on Google Colab using the NVIDIA L4 Tensor Core GPU, with a batch size of 64 for 10 epochs.



**Figure 2:** Architecture of the proposed model.

### 3. Evaluation

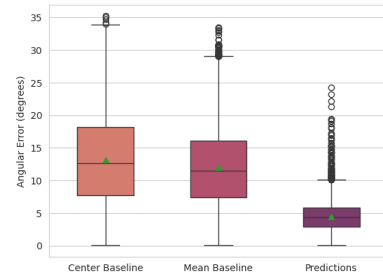
Following the approach of Hu et al. [HLZ\*20], the model was evaluated based on its prediction error, recall rate (to assess its potential for integration into foveated rendering pipelines), and runtime performance. Two baselines (center and mean) were defined for comparison. For the recall rate, a foveal radius of 15 degrees, centered at the ground truth gaze point, was applied. The model achieved a low median error with a narrow IQR, indicating robustness, accuracy, and consistency (Figure 3). It demonstrated a 66.43% and 63.08% improvement over the center and mean baselines, respectively. Additionally, the model significantly outperformed the baselines in terms of recall rate (Table 1), achieving values suitable for practical applications. Observing the visualised results (Figure 1), we notice that the predictions follow the pattern of the ground truth closely. However, the average runtime of approximately 150ms remains a notable limitation, affecting its viability for real-time use.

	Center	Mean	Ours
Mean Recall Rate	60.8%	69.46%	99.87%

**Table 1:** Recall rates of our model and the baselines.

### 4. Conclusion and Future Work

We propose a gaze prediction model for task-oriented VR environments that primarily leverages temporal information. Our main contribution is a model that achieves strong accuracy and consistency with data from diverse tasks and users, demonstrating adaptability and flexibility, while requiring a relatively simple input. Future work will focus on optimizing runtime performance to enable integration into real-time rendering pipelines. Additionally, we aim to enhance the model's accuracy and evaluate its effectiveness across diverse scenarios, datasets, and real-time applications.



**Figure 3:** Angular error comparison between our model and the baselines.

### Acknowledgments

This work has been supported by the Horizon Europe Research & Innovation Programme under Grant agreement N. 101092612 (Social and hUman ceNtered XR - SUN project). Views and opinions expressed in this work are those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the European Commission can be held responsible for them.

### References

- [ATSD23] ARABADZHIYSKA E., TURSUN C., SEIDEL H.-P., DIDYK P.: Practical saccade prediction for head-mounted displays: Towards a comprehensive model. *ACM Trans. Appl. Percept.* 20, 1 (Jan. 2023). URL: <https://doi.org/10.1145/3568311>, doi:10.1145/3568311. 1
- [EZW\*21] EMERY K. J., ZANNOLI M., WARREN J., XIAO L., TALATHI S. S.: Openneeds: A dataset of gaze, head, hand, and scene signals during exploration in open-ended vr environments. In *ACM Symposium on Eye Tracking Research and Applications* (2021), pp. 1–7. 1
- [HBLW21] HU Z., BULLING A., LI S., WANG G.: Fixationnet: Forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2681–2690. 1
- [HLG20] HU Z., LI S., GAI M.: Temporal continuity of visual attention for future gaze prediction in immersive virtual reality. *Virtual Reality & Intelligent Hardware* 2, 2 (2020), 142–152. 1
- [HLZ\*20] HU Z., LI S., ZHANG C., YI K., WANG G., MANOCHA D.: Dgaze: Cnn-based gaze prediction in dynamic scenes. *IEEE transactions on visualization and computer graphics* 26, 5 (2020), 1902–1911. 1, 2
- [KDCM16] KOULIERIS G. A., DRETTAKIS G., CUNNINGHAM D., MANIA K.: Gaze prediction using machine learning for dynamic stereo manipulation in games. In *2016 IEEE Virtual Reality (VR)* (2016), pp. 113–120. doi:10.1109/VR.2016.7504694. 1