Journal of
Eye Movement Research

MDPI

Article

Far-Field Non-Contact Eye Tracking for Visual Attention Analysis in Runway Viewing Environments

Hannuo Tian [1], Ding-Bang Luh [1], Mingyang Guo [2], Qianli Lin [3] and Jun Ma [2,3,*]

[1] School of Art and Design, Guangdong University of Technology, Guangzhou 510090, China; tianhannuopro@163.com (H.T.); luhdingbang@126.com (D.-B.L.)
[2] College of Automotive Studies, Tongji University, Shanghai 201804, China; gmingyang1998@163.com
[3] College of Design and Innovation, Tongji University, Shanghai 200092, China; luliqianli@gmail.com
* Correspondence: majun.tongji@foxmail.com

Abstract

Conventional eye trackers are difficult to deploy in fashion-runway settings because they require head stabilization, wearable devices, close-range cameras, or infrared illumination. This study develops a far-field, non-contact eye-tracking framework and prototype system for measuring target-directed visual attention in runway-viewing settings using standard cameras. The framework integrates eye-region normalization, 3D gaze estimation, individualized homography-based point-of-regard mapping onto a scene plane, adaptive filtering, DBSCAN-based outlier removal, and kernel density estimation. In Experiment 1, 16 participants completed a controlled calibration-grid benchmark with the proposed system and EyeLink 1000 Plus. Compared with the EyeLink 1000 Plus, the proposed system exhibited higher positional error (A_{pos}: $2.50 \pm 0.25^{\circ}$ vs. $1.74 \pm 0.55^{\circ}$) and a higher value on the processed-output spatial-precision measure (A_{spa}: $2.31 \pm 0.24^{\circ}$ vs. $1.02 \pm 0.37^{\circ}$). These results characterize end-to-end localization under the calibrated benchmark geometry and provide a controlled reference for the subsequent AOI-level evaluation. In Experiment 2, an independent sample of 16 participants completed an instructed right chest targeting task at 2 m, 4 m and 6 m. AOI-retained sample percentages were 80.4%, 97.3%, and 96.0%, respectively, and gaze-density maps concentrated within the predefined upper-body target region. These findings support the feasibility of target-directed AOI-level localization at far-field distances without chin-rest stabilization under the tested simulated runway-observation protocol.

Keywords: non-contact eye tracking; runway viewing; remote gaze estimation; visual attention; stacked hourglass network; homography mapping

Check for updates

Academic Editor: Michael Burch

Received: 11 August 2026
Revised: 25 September 2026
Accepted: 1 October 2026
Published: 8 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Fashion shows are among the most concentrated forms of visual communication in contemporary creative industries [1]. Through meticulously orchestrated set designs, model performances, and spatial configurations, these events serve as platforms for conveying brand identity, cultural values, and evolving aesthetic sensibilities [2]. Key Opinion Leaders (KOLs) positioned in runway observation areas play a central role in aesthetic evaluation and content dissemination. Their patterns of visual attention are therefore commercially and culturally relevant to understanding fashion-show information delivery and garment presentation [3,4]. Examining how KOLs allocate attention across garment and body regions during runway performances is consequently of practical significance.

Conventional approaches to studying KOL preferences have relied primarily on post-hoc subjective methods such as questionnaires and semi-structured interviews [3,5], which are fundamentally limited in their capacity to reconstruct the fine-grained, moment-by-moment visual selection process during live viewing. Eye-tracking offers a more direct approach by recording time-resolved gaze behavior, enabling objective characterization of overt visual attention and information-sampling patterns [6–8]. Although eye-tracking has been applied across human–computer interaction, marketing, and fashion retail contexts to quantify attentional allocation [9–11], these studies have predominantly relied on screen-based or simulated environments whereas direct empirical investigation in live runway-observation settings remains limited.

In parallel, appearance-based gaze estimation using deep learning has attracted considerable research attention as a flexible alternative to hardware-intensive approaches [12,13]. Unlike corneal-reflection systems that depend on active infrared illumination [14], appearance-based gaze-estimation methods infer gaze from facial or ocular appearance captured by conventional cameras, reducing reliance on specialized active-infrared hardware [12,13]. Recent advances have addressed a range of challenges including semi-supervised landmark detection [15], iris geometry modeling [16], cross-subject appearance debiasing [17], cross-domain generalization [18], and calibration-free webcam-based estimation [19]. Park et al. [20] demonstrated that structurally interpretable ocular keypoints estimated from ordinary images provide a robust basis for model-based gaze estimation. At the same time, previous evaluations of remote eye trackers have shown that data quality can vary with head position and acquisition geometry [21], while recent reviews have identified substantial heterogeneity in working distance, measurement procedures, and reported oculomotor parameters across studies [22]. These findings indicate that gaze-estimation performance should be interpreted with respect to the specific recording geometry and application context in which it is measured.

Despite this collective progress, applying gaze estimation to visual attention analysis in live runway scenarios remains substantially underexplored. First, existing studies of KOL and influencer preferences frequently rely on retrospective self-report methods such as surveys and interviews [3,5], which cannot capture time-resolved visual selection during live viewing; second, existing eye-tracking technologies involve trade-offs among measurement accuracy, working distance, participant mobility, hardware requirements, and deployment complexity, which can make their use difficult in runway observation settings [23–25]; third, much of the appearance-based gaze-estimation literature focuses on gaze-direction or gaze-point estimation and generalization, whereas application-specific mapping of these estimates to interpretable scene-level AOIs remains a separate methodological step [12]; fourth, runway observation requires incorporating changes in the distance between the observer and the target, and more importantly, it requires interpreting gaze estimation based on clothing or body regions within the scene space.

Motivated by these limitations, this paper proposes a far-field non-contact eye-tracking framework for target-directed visual-attention measurement in runway-viewing environments. The framework integrates four core modules: (1) camera-based eye-image acquisition and preprocessing; (2) 3D gaze estimation via a Park et al. [20]-based stacked hourglass network with multi-stage heatmap refinement; (3) gaze-to-point-of-regard (POR) mapping through a homography scene coordinate transformation; and (4) attention analysis and visualization using adaptive filtering, density-based clustering, and kernel density estimation (KDE). A prototype system implementing this framework was developed and evaluated through two experiments: a controlled post-calibration benchmark comparison against the EyeLink 1000 Plus (SR Research Ltd., Ottawa, Canada), and an applied experiment in a simulated runway observation scenario. The principal contributions are as follows.

(1) We propose a far-field, non-contact eye-tracking framework that combines appearance-based 3D gaze-direction estimation with individualized homography-based planar gaze-to-scene mapping for scene-level AOI analysis using standard cameras.
(2) We design a dedicated attention-analysis pipeline that combines adaptive One Euro filtering, Density-based spatial clustering of applications with noise (DBSCAN)-based outlier removal, and KDE to convert noisy POR sequences into stable, interpretable area-of-interest (AOI)-level gaze representations suitable for garment- and design-level analysis.
(3) We implement the framework as a prototype system with real-time visualization, parameter configuration, and structured data export, supporting operation by researchers without specialized programming expertise.
(4) We provide an application-specific empirical evaluation of the framework: a controlled post-calibration benchmark against a commercial infrared eye tracker at predefined calibration-grid locations, and a target-directed runway-observation experiment with an independent participant sample at viewing distances of 2 m, 4 m and 6 m. These experiments separately characterize controlled positional performance and application-oriented AOI-level localization.

2. Materials and Methods

2.1. Overall Framework

Figure 1 illustrates the overall architecture of the proposed framework, which consists of four main functional modules. The Data Acquisition and Eye-Region Preprocessing module acquires facial and ocular signals from standard RGB camera video and converts each frame into normalized eye-region images. The 3D Gaze Estimation and Analysis module estimates ocular structural landmarks using a stacked hourglass network and calculates gaze direction from a 3D eyeball geometry model. The Gaze-to-POR Mapping module converts estimated gaze directions into scene-level points of regard through an individualized calibration procedure and homography transformation. The Attention Analysis and Visualization module generates AOI-based attention representations for subsequent visual behavior analysis.

2.2. Data Acquisition and Eye-Region Processing

2.2.1. Face and Eye Region Detection

Face region detection is the foundation of eye-tracking analysis. The camera acquires input image frames, which are mirrored and converted to grayscale. Face detection employs the maximum-margin object detection (MMOD)-based convolutional neural network detector provided by the dlib library [26,27], offering robust localization under varying poses and lighting conditions. To reduce computational overhead, input images are downsampled to half resolution prior to inference, with resulting bounding boxes scaled back by a factor of 2. Following face detection, dlib's five-point shape predictor was used to localize the inner and outer corners of both eyes together with a nose landmark [26,28], providing the landmarks used for eye-region extraction.

2.2.2. Landmark-Based 2D Affine Alignment of the Eye Region

To reduce image-plane variation in eye position, in-plane rotation, and scale during viewing, a composite 2D affine transformation was constructed for each eye to align the detected eye region to a standardized input image of fixed size (36×60 px in this work). The transformation consists of translation, in-plane rotation, isotropic scaling, and recentering. This preprocessing standardizes the local eye crop in the image plane and does not perform

virtual-camera normalization or explicitly compensate for general 3D head-to-camera pose variation [13].

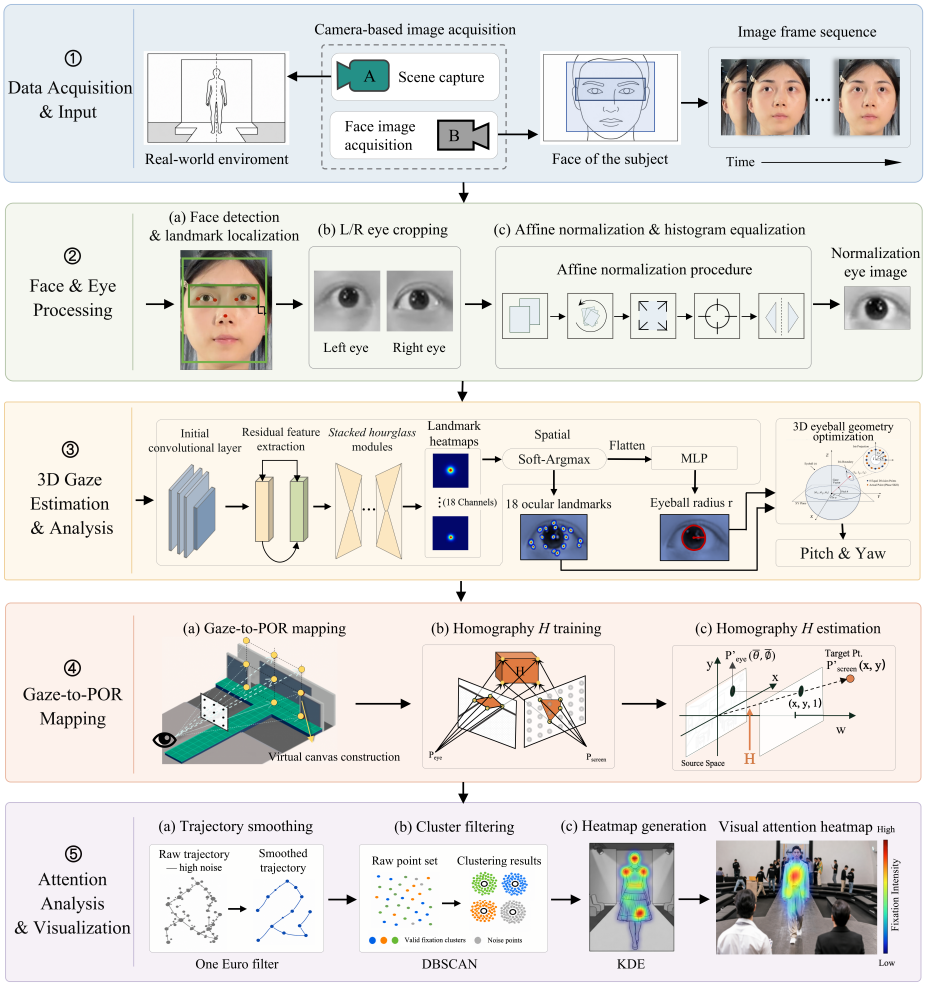


Figure 1. Overall framework of the proposed far-field non-contact eye-tracking system. Note. Arrows between processing blocks indicate the workflow. A and B denote scene-image and face-image acquisition, respectively, and H denotes the homography matrix. Colored backgrounds distinguish the processing stages, whose headings correspond to the relevant methodological subsections.

Translation T_i: the coordinate origin is shifted to the geometric center of the eye's inner and outer corners:

$$T_i = \begin{bmatrix} 1 & 0 & -c_x \\ 0 & 1 & -c_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{1}$$

where (c_x, c_y) denotes the geometric center of the two eye corners.

Rotation T_r: let the inner and outer corner coordinates of either eye be $P_l(x_l, y_l)$ and $P_r(x_r, y_r)$ respectively. The tilt angle θ between the inter-corner axis and the image horizontal is:

$$\theta = \arctan\left(\frac{y_r - y_l}{x_r - x_l}\right). \tag{2}$$

A reverse rotation of $-\theta$ corrects this tilt:

$$T_r = \begin{bmatrix} \cos(-\theta) & -\sin(-\theta) & 0 \\ \sin(-\theta) & \cos(-\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{3}$$

Scaling T_s: the current eye width is defined as $w_e = 1.5\sqrt{(x_r - x_l)^2 + (y_r - y_l)^2}$, where 1.5 preserves a contextual margin. The resulting eye width is then normalized to the preset output width $W_0 = 60$ px:

$$s = \frac{W_0}{w_e}, \qquad T_s = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4}$$

Recentering T_c: the scaled coordinate origin is translated to the center of the output image $\left(\frac{W_0}{2}, \frac{H_0}{2}\right)$ where $W_0 = 60$ px and $H_0 = 36$ px:

$$T_c = \begin{bmatrix} 1 & 0 & W_0/2 \\ 0 & 1 & H_0/2 \\ 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

The composite transformation $T = T_c \cdot T_s \cdot T_r \cdot T_i$ resamples the local eye region into a standardized image with uniform orientation and fixed resolution. To reduce distributional discrepancy from left–right mirror symmetry, the left-eye crop is horizontally flipped after extraction [29], enabling a single set of network weights to perform binocular inference. After normalization, Contrast-Limited Adaptive Histogram Equalization was applied to the cropped grayscale eye image to suppress local brightness inhomogeneity while avoiding overamplification of noise in near-uniform regions, followed by intensity normalization to the range $[0, 1]$ [30]. These operations reduce the influence of illumination variation on local texture distribution and improve the recognizability of key structures such as iris edges, eye corners, and eyelid contours [31]. The resulting image provides a reliable, standardized input for subsequent hourglass-based structural reconstruction.

2.3. 3D Gaze Estimation and Analysis

2.3.1. Feature Extraction of Eye Images

The trained network [20] was obtained using synthetic UnityEyes eye images and annotations, with extensive data augmentation and curriculum learning. Their reference model was optimized using Adam with a learning rate of 5×10^{-4}, a batch size of 16, and was trained for 6.8 million steps. In the present study, the pretrained network weights were retained without parameter-level retraining, fine-tuning, or domain adaptation using participant images or runway recordings. As shown in Figure 2, the implementation comprises an initial convolutional stage, residual-block feature extraction, and a multi-stage stacked-hourglass module for localization of 18 ocular landmarks.

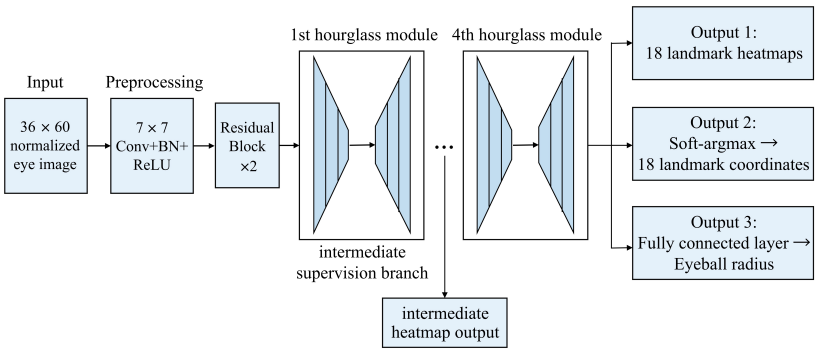


Figure 2. Stacked hourglass network architecture.

(1) Initial Convolutional Layer Setup

In the initial stage, the network performs preliminary feature extraction and resolution adjustment through a pre-convolutional module. Specifically: a 7×7 convolutional layer with stride $s_0 = 1$ and output channel count $n = 32$ extracts low-level edge and texture features with a large receptive field; batch normalization (BN) and a rectified linear unit (ReLU) activation are then applied; finally, two residual blocks sequentially expand the feature-map channel count from n to $2n$ and then compress it back to n, while limiting feature-channel width and computational complexity. With $s_0 = 1$, the initial stage preserves the spatial resolution of the normalized eye crop, yielding an output feature map of $H \times W$ ($H = 36$ px, $W = 60$ px).

(2) Residual Block Design

Each residual block follows a bottleneck structure: a 1×1 convolution compresses channels from n_{in} to $\frac{n_{\text{out}}}{2}$; a 3×3 convolution performs spatial feature fusion; a final 1×1 convolution restores channels to n_{out}. BN and ReLU precede each convolution. Skip connections add the input directly when $n_{\text{in}} = n_{\text{out}}$, or via a 1×1 projection convolution otherwise, preventing gradient vanishing and ensuring trainability of the deep network.

(3) Hourglass Module Design

In the present configuration, each hourglass module is constructed through recursive nesting with depth $D_{\max} = 4$, enabling feature integration across multiple spatial scales. At each recursion level d, the input feature map $x^{(d)}$ is divided into two parallel branches. The uplink path preserves spatial resolution and extracts fine local features through a residual block:

$$u^{(d)} = R_{\text{up}}^{(d)}\left(x^{(d)}\right), \tag{6}$$

where $R_{\text{up}}^{(d)}$ denotes the independent residual block on the uplink path. The features retained here are rich in fine-grained edge-contour and local-texture information, forming the basis for precise localization. The downlink path first halves spatial resolution via max-pooling, then extracts global semantic features through a residual block yielding $L_1^{(d)}$. If $d < D_{\max}$, $L_1^{(d)}$ serves as input to the next recursive level; at the recursion bottom ($d = D_{\max}$):

$$L_2^{(D_{\max})} = R_{\text{low}}^{(D_{\max})}\left(L_1^{(D_{\max})}\right), \tag{7}$$

where $R_{\text{low}}^{(d)}$ denotes the independent residual block on the downlink path. After each level obtains $L_2^{(d)}$, a further residual block yields $L_3^{(d)}$, which is upsampled via bilinear interpolation to match $u^{(d)}$; element-wise summation fuses high-resolution local features with low-resolution global semantics, preserving a constant channel count throughout.

The landmark-learning formulation uses intermediate supervision after each of the first three hourglass modules, following standard stacked-hourglass practice [32]. During model training, a 1×1 convolution projects each module output onto the heatmap space, with mean squared error (MSE) used as the landmark-heatmap objective:

$$\mathcal{L}_{\text{inter}} = \sum_{k=1}^{3} \sum_{j=1}^{18} \left\| \hat{H}_{k,j} - H_j^* \right\|_F^2, \tag{8}$$

where H_j^* is the ground-truth Gaussian heatmap for the j-th keypoint and $\hat{H}_{k,j}$ is the corresponding intermediate prediction from the k-th module. The final module's output is supervised by the same loss, enhancing gradient flow to early layers and accelerating convergence. The final heatmap tensor has dimension $N \times 18 \times H' \times W'$, representing 8 inner-eyelid boundary points, 8 iris-edge points, 1 iris center, and 1 eyeball center.

In the present implementation, the available stage-wise landmark heatmaps are averaged before Soft-Argmax coordinate extraction. This inference-time aggregation was adopted to combine the multiple supervised stage-wise predictions already produced by the stacked-hourglass network, rather than relying on a single stage output when processing the low-resolution eye crops used in the present system. The operation introduces no additional trainable parameters and does not modify the inherited pretrained weights. Because no ablation comparison was performed, this procedure is treated as an implementation design choice rather than as a demonstrated performance improvement.

(4) Keypoint Coordinate Extraction

Accurate 2D keypoint coordinates are extracted using the differentiable soft-Argmax operation, circumventing the resolution ceiling of hard-Argmax. A normalized coordinate grid matching the heatmap spatial size is pre-established, with horizontal grid $\{x_j\}_{j=0}^{W'-1}$ and vertical grid $\{y_i\}_{i=0}^{H'-1}$ expanded to matrices $\mathbf{X}$ and $\mathbf{Y}$ via meshgrid. For the flattened heatmap vector $H_k \in \mathbb{R}^{H'W'}$, a temperature-scaled Softmax yields the normalized spatial probability distribution:

$$p_m^{(k)} = \frac{\exp(\beta H_{k,m})}{\sum_{m'=1}^{H'W'} \exp(\beta H_{k,m'})}, \qquad m = 1, \ldots, H'W'. \tag{9}$$

The temperature parameter $\beta = 100$ sharpens the distribution toward the maximum-response location while preserving differentiability. The expected normalized coordinates are:

$$\begin{cases} \hat{x}_k = \sum_{m=1}^{H'W'} p_m^{(k)} \cdot x_{\text{ref},m}, \\ \hat{y}_k = \sum_{m=1}^{H'W'} p_m^{(k)} \cdot y_{\text{ref},m}. \end{cases} \tag{10}$$

The coordinates are mapped back to the pixel domain with a sub-pixel center offset:

$$\begin{cases} x_k = \hat{x}_k(W' - 1) + 0.5, \\ y_k = \hat{y}_k(H' - 1) + 0.5. \end{cases} \tag{11}$$

This process is repeated for all 18 keypoints, yielding a coordinate tensor of shape $N \times 18 \times 2$. Since the heatmap resolution is $\frac{1}{s}$ of the input image resolution, all coordinates are multiplied by the stride factor s to restore the original eye-image coordinate system.

(5) Eye Radius Regression

All 18 keypoint coordinates are flattened into a 36-dimensional structure vector and passed through a three-layer multilayer perceptron (MLP) (100 neurons per hidden layer, BN, ReLU), outputting a scalar eyeball radius estimate $\hat{r}$ through a final single-neuron fully connected layer. The eye-radius regression head uses mean absolute error (MAE) as its training objective, $\mathcal{L}_r = |\hat{r} - r^*|$, where r^* denotes the ground-truth radius label.

2.3.2. Gaze Direction Calculation Based on the 3D Eyeball Model

After obtaining the 18 keypoint coordinates and estimated eyeball radius $\hat{r}$, 3D gaze direction is estimated through geometric optimization of a 3D eyeball model [16,20], as shown in Figure 3. Let (e_x, e_y, e_z) be the eyeball center coordinates, (i_x, i_y, i_z) the iris center coordinates, and r the estimated radius. Gaze direction is parameterized by pitch angle θ (vertical deflection, positive upward) and yaw angle ϕ (horizontal deflection).

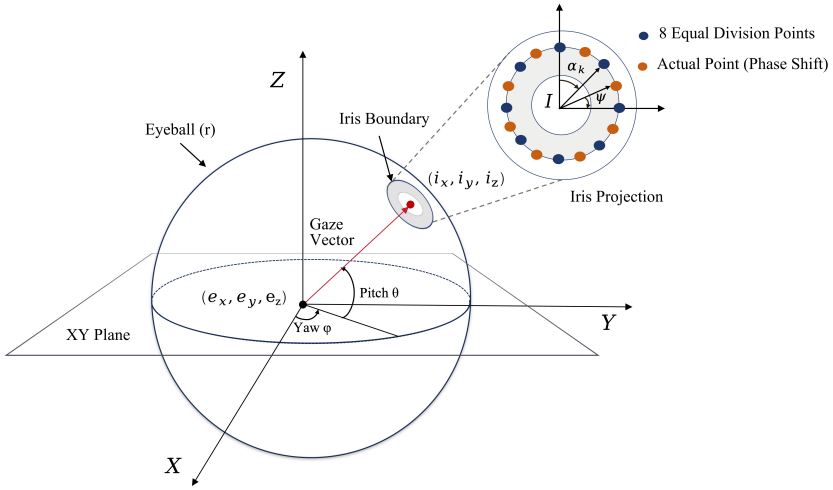


Figure 3. 3D eyeball projection model. Note. The red dot marks the iris center (i_x, i_y, i_z). The red line connects the eyeball center (e_x, e_y, e_z) to the iris center and represents the gaze direction in the geometric model.

The 2D displacement of the iris center relative to the eyeball center, $\Delta x = i_x - e_x$ and $\Delta y = i_y - e_y$, yields closed-form initial estimates:

$$\begin{cases} \theta_0 = \arcsin\left(\frac{\Delta y}{r}\right), \\ \phi_0 = \arcsin\left(\frac{-\Delta x}{r\cos\theta_0}\right). \end{cases} \tag{12}$$

These initial values provide a physically grounded starting point for the subsequent nonlinear optimization and substantially reduce the number of iterations required for convergence.

Relying solely on the iris-center displacement ignores the full 3D shape of the iris, limiting accuracy. Therefore, a nonlinear optimization step [16,20] is employed that incorporates all eight equally spaced iris-edge keypoints (azimuth angles $\alpha_k = \frac{k\pi}{4}, \quad k = 0, \ldots, 7$) as geometric constraints. The parameter vector $(\theta, \phi, \delta, \psi)$ is jointly optimized, where δ is the angular displacement of the iris relative to the eyeball, and ψ is the initial phase shift of the iris circumference.

$$\tilde{\alpha}_k = \alpha_k + \psi, \qquad \alpha_k = \frac{k\pi}{4}, \quad k = 0, \ldots, 7. \tag{13}$$

The image-plane predicted coordinates of the k-th iris-edge point follow the 3D rotational projection relationship:

$$\begin{cases} \hat{x}_k = e_x - r\cos(\theta + \delta\sin\tilde{\alpha}_k)\cdot\sin(\phi + \delta\cos\tilde{\alpha}_k), \\ \hat{y}_k = e_y + r\sin(\theta + \delta\sin\tilde{\alpha}_k). \end{cases} \tag{14}$$

The predicted iris-center pixel coordinates at the current gaze angle are:

$$\begin{cases} \hat{i}_x = e_x + r(-\cos\theta\cdot\sin\phi), \\ \hat{i}_y = e_y + r\sin\theta. \end{cases} \tag{15}$$

The iris-center reprojection error is thus:

$$\begin{cases} \Delta x_c = \hat{i}_x - i_x = e_x - r\cos\theta\sin\phi - i_x, \\ \Delta y_c = \hat{i}_y - i_y = e_y + r\sin\theta - i_y. \end{cases} \tag{16}$$

The composite objective function minimizes both iris-edge and iris-center reprojection errors:

$$\mathcal{L}(\theta, \phi, \delta, \psi) = \frac{1}{8}\sum_{k=0}^{7}\left[\left(\hat{x}_k - i_x^{(k)}\right)^2 + \left(\hat{y}_k - i_y^{(k)}\right)^2\right] + \Delta x_c^2 + \Delta y_c^2, \tag{17}$$

where $\left(i_x^{(k)}, i_y^{(k)}\right)$ are the network detection coordinates of the k-th iris edge keypoint. The optimization process calls the truncated Newton method with physical constraint boundaries for each parameter: $\theta \in [-0.4\pi, 0.4\pi]$, $\phi \in [-0.4\pi, 0.4\pi]$, $\delta \in [-0.01\pi, 0.5\pi]$, and $\psi \in [-\pi, \pi]$.

After convergence, the final estimated gaze angles $(-\theta, \phi)$ are output. The sign negation on θ adapts to the image coordinate convention in which the y-axis points downward, ensuring that the output θ is positive for upward gaze.

For benchmark context, Park et al. [20] reported mean angular gaze errors of 7.1° on Columbia Gaze and 11.9° on EYEDIAP using their model-based gaze-estimation approach with 20 calibration images per participant.

2.4. Gaze-to-POR Mapping

3D gaze estimation is defined in this paper as estimating the 3D gaze direction parameterized by pitch and yaw, whereas the subsequent gaze-to-POR mapping represents the resulting POR in the 2D coordinate system of the calibrated scene plane. To map estimated 3D gaze direction to an attention position in the runway scene, a homography transformation is adopted as the core mapping model. As shown in Figure 4, given the estimated gaze angles, the corresponding normalized 2D gaze point is obtained by projecting the gaze direction onto a unit virtual plane: $[\tan\phi, \tan\theta]$. This is equivalent to the projection of a unit gaze vector onto the image plane of a pinhole camera with unit focal length, consistent with standard gaze normalization conventions [13].

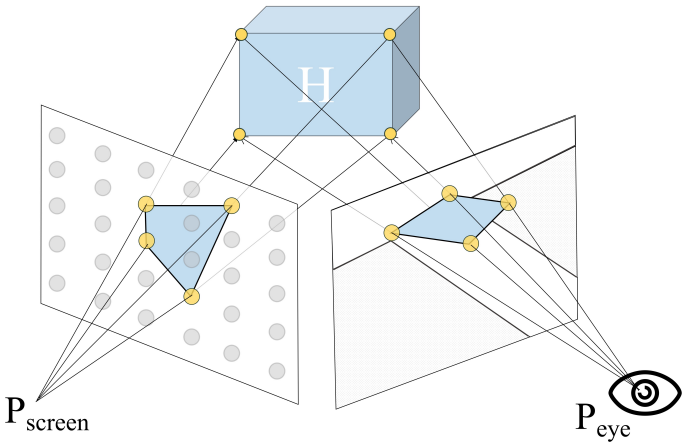


Figure 4. Schematic diagram of POR mapping based on homography transformation. Note. Yellow dots indicate corresponding points in the schematic. H denotes the homography matrix mapping normalized 2D gaze coordinates to scene-plane point-of-regard (POR) coordinates.

To establish the correspondence between the normalized gaze point and the runway scene coordinates, an individualized calibration procedure is performed prior to each session. $N = 16$ calibration points with known scene coordinates $\{p_i\}_{i=1}^{N}$ are arranged on

the runway reference plane, covering the full spatial extent of the viewing area. Participants sequentially fixate on each calibration point for 5 s; the stable central 3 s window is used to extract the corresponding normalized gaze coordinates (u_i, v_i), represented in homogeneous form as $q_i = [u_i, v_i, 1]^T$. This produces N matched pairs $\{(q_i, p_i)\}_{i=1}^{N}$, where $p_i = [x_i, y_i, 1]^T$ denotes the homogeneous scene-level POR. The runway reference plane is modelled as a virtual canvas defined by four corner points on the physical floor or backdrop, establishing the pixel-coordinate system in which all POR outputs are expressed. The mapping between q_i and p_i is modelled as:

$$\lambda_i p_i = H q_i, \tag{18}$$

where H is a 3×3 homography matrix and λ_i is a non-zero homogeneous scale factor for the i-th correspondence. The homography matrix H was estimated from the calibration correspondences using the standard Direct Linear Transform formulation for planar projective transformations [33]. After calibration, H is applied online to convert each estimated gaze angle into the corresponding 2D scene gaze point.

2.5. Attention Analysis and Visualization

2.5.1. Gaze Trajectory Smoothing via Adaptive Filtering

To suppress trajectory jitter and remove unreliable samples, the raw POR sequence is further processed through adaptive filtering, density clustering, and KDE. The parameter settings used in these processing stages were established through repeated preliminary pilot trials before the formal experiments and were subsequently held fixed across participants and experimental conditions.

The One Euro filter dynamically adjusts filtering intensity according to instantaneous gaze movement speed [34]. When the estimated POR velocity is low, the cutoff approaches $f_{c,\min}$, producing stronger smoothing of small fluctuations; as POR velocity increases, the cutoff increases so that the filter follows faster positional changes with reduced lag. Let the raw observation at time t be x_t, the filtered output be $\hat{x}_t$, the sampling interval be $\Delta t_t = t - t_{t-1}$, and the cutoff frequency be f_c. The smoothing coefficient α is determined by:

$$\tau = \frac{1}{2\pi f_c}, \tag{19}$$

$$\alpha(f_c, \Delta t_t) = \frac{1}{1 + \frac{\tau}{\Delta t_t}} = \frac{2\pi f_c \Delta t_t}{2\pi f_c \Delta t_t + 1}. \tag{20}$$

The recursive filtering equation is:

$$\hat{x}_t = \alpha(f_c, \Delta t_t) \cdot x_t + [1 - \alpha(f_c, \Delta t_t)] \cdot \hat{x}_{t-1}. \tag{21}$$

A high f_c drives $\alpha \to 1$ (rapid response); a low f_c drives $\alpha \to 0$ (strong smoothing). An independent low-pass filter applied to the first-order finite difference of the raw signal yields a smooth instantaneous rate estimate $\hat{\dot{x}}_t$:

$$\hat{\dot{x}}_t = \alpha(f_{c,d}, \Delta t_t) \cdot \frac{x_t - \hat{x}_{t-1}}{\Delta t_t} + [1 - \alpha(f_{c,d}, \Delta t_t)] \cdot \hat{\dot{x}}_{t-1}, \tag{22}$$

where $f_{c,d} = 1$ Hz is the fixed derivative cutoff frequency. The main-signal cutoff frequency is then adapted linearly:

$$f_c = f_{c,\min} + \beta \left| \hat{\dot{x}}_t \right|, \tag{23}$$

where $\hat{x}_t$ is the filtered rate-of-change estimate; its magnitude is used so that the adaptive cutoff depends on the speed of signal variation rather than its direction. $f_{c,\min} = 0.2\,\text{Hz}$ is the minimum cutoff governing static smoothing strength and $\beta = 0.1$ is the rate sensitivity coefficient. When $\|\hat{x}_t\|$ is small, f_c approaches $f_{c,\min}$, providing stronger smoothing of small-amplitude fluctuations; as $\|\hat{x}_t\|$ increases, f_c increases and the filter responds more rapidly to changes in the POR signal.

2.5.2. DBSCAN Attention Point Clustering

Even after filtering, the POR sequence may contain outliers from blinks, sudden head movements, or model inference failures, which distort the gaze density distribution and create spurious attention hotspots. Therefore, DBSCAN was applied to the 2D POR coordinates to identify spatially dense clusters and classify isolated samples as noise [35]. The neighborhood radius and minimum neighborhood size were fixed at $\varepsilon = 40\,\text{px}$ and MinPts $= 30$, respectively. All viewing-distance conditions were processed in the same scene-image coordinate system without distance-specific resizing or rescaling before DBSCAN or KDE. Thus, $\varepsilon = 40\,\text{px}$ represented the same neighborhood radius in image coordinates across conditions, whereas MinPts $= 30$ was a sample-count criterion and therefore had no spatial or angular unit. Samples not assigned to a sufficiently dense cluster were treated as noise and excluded from subsequent analysis.

2.5.3. Attention Heatmap Generation Based on KDE

To obtain a continuous probability density estimate of the spatial gaze distribution, KDE was applied to the post-clustering gaze samples retained by the processing pipeline $\{p_i\}_{i=1}^{n}$, where $p_i = (x_i, y_i)$, to generate a smooth 2D probability density field. The attention density function at any scene location $p = (x, y)$ is:

$$\hat{f}(p) = \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{p - p_i}{h}\right), \tag{24}$$

where n is the number of gaze samples retained after filtering and clustering, h is the bandwidth parameter, and K is a 2D isotropic Gaussian kernel [36]. A fixed bandwidth of $h = 30\,\text{px}$ was used in the scene-image coordinate system. KDE was used only to visualize the spatial distribution of the processed gaze samples and did not enter AOI assignment or the cross-distance statistical analyses. The resulting density maps were normalized to [0, 1], mapped to a cool-to-warm pseudocolor scale, and overlaid on the corresponding runway images, with warmer colors indicating higher relative gaze density.

2.6. Prototype System Development

As shown in Figure 5, the proposed framework is implemented as a prototype eye-tracking system for practical visual attention analysis in runway viewing scenarios. The system is developed in Python 3.9 and structured into three functional layers: a data input layer, a core computation layer, and a result output layer. The data input layer receives real-time camera video streams alongside auxiliary inputs including participant metadata, scene reference images, calibration grid configurations, and mapping parameters. The core computation layer implements the full algorithmic pipeline described in the preceding sections, comprising three parallel processing units responsible for 3D gaze estimation, gaze-to-POR mapping, and visual attention analysis. The result output layer supports both real-time visualization and post-session data export. During real-time operation, a live scene view is overlaid with the current POR position and gaze trajectory. After each session, the system exports POR sample sequences in CSV format, DBSCAN-based clustering results, and attention heatmaps.

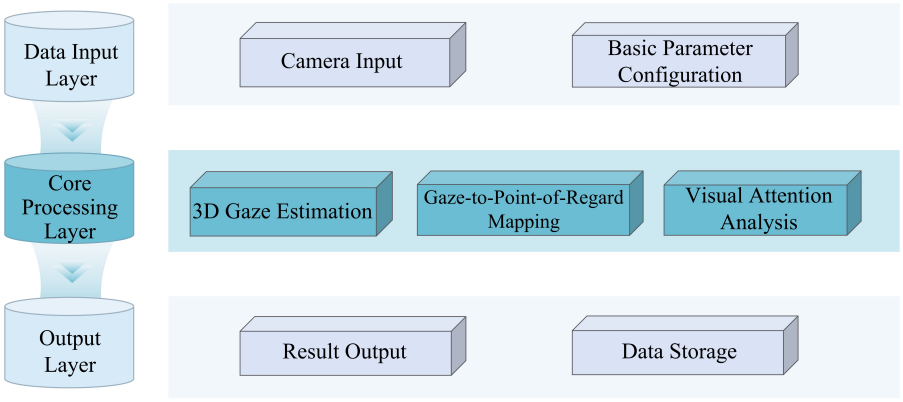


Figure 5. System overall architecture diagram.

As shown in Figure 6, the system additionally provides a graphical configuration panel enabling key parameter adjustment without requiring code modification, supported by a scrollable status log for monitoring processing state. This interface allows researchers to operate the prototype without specialized programming expertise and supports application-oriented evaluation in simulated runway-observation settings.

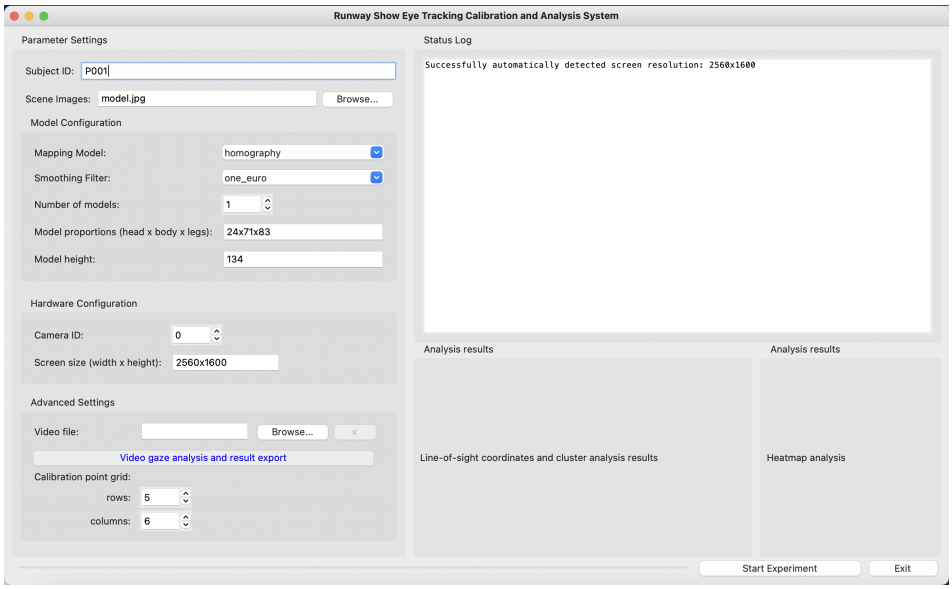


Figure 6. Interface diagram of the proposed system. Note. The displayed parameter values are illustrative; the experimental settings are specified in the Methods.

2.7. Experimental Evaluation

To evaluate the proposed framework, two experiments were conducted. Experiment 1 benchmarked the proposed system against a commercial infrared eye tracker under controlled laboratory conditions to characterize post-calibration positional accuracy and spatial precision. Experiment 2 evaluated application-oriented AOI-localization feasibility in a simulated runway-observation scenario. Both experiments were executed on a Mac computer equipped with an Apple M2 chip (Apple Inc., Cupertino, CA, USA). For both experiments, the fixed pretrained gaze-estimation network described in Section 2.3.1 was used, while participant-specific calibration was performed only at the subsequent gaze-to-scene mapping stage, where the homography was estimated from calibration samples collected prior to each session.

2.7.1. Experiment 1: Controlled Calibration-Grid Benchmark—Methods

(1) Participants and Benchmark setup

Experiment 1 involved 16 participants (8 males and 8 females; mean age ± SD, 24.8 ±3.5 years). Participants were recruited through targeted online notices circulated within university and research-institute networks. Before enrollment, interested volunteers provided demographic and visual-status information together with information on their academic, research, or professional background relevant to the study context. Eligibility required prior involvement in at least one of the following activities: runway or fashion observation, aesthetic or design evaluation, fashion-content interpretation or dissemination, or related professional review and feedback. Eligibility was assessed from the information provided during this pre-enrollment screening according to the study-defined criteria; no separate occupational credential-verification procedure was used.

Eligibility also required a myopia refractive error not exceeding 3.00 D in magnitude and the ability to resolve the experimental targets without corrective eyeglasses. To avoid reflections and partial occlusion in the RGB eye images, no participant wore eyeglasses during the experiment. Written informed consent was obtained from all participants.

The benchmark device was the EyeLink 1000 Plus, a desktop-mounted video-based pupil-corneal-reflection eye tracker used as the laboratory reference comparator for the RGB-webcam system [37,38]. Video of the participant's face and eye regions was acquired using a USB RGB webcam labeled "HD WEBCAM". During experimental data acquisition, the camera operated at 1920 × 1080 px and an achieved frame rate of 30 fps. As shown in Figure 7, the participant sat at a table (height: 75 cm), with head position stabilized using a chin rest (height: 25 cm). The chin rest was used only in Experiment 1 to minimize head-motion-related variability and to provide a controlled common geometry for the within-protocol comparison with the EyeLink 1000 Plus; it is not a hardware requirement of the proposed system.

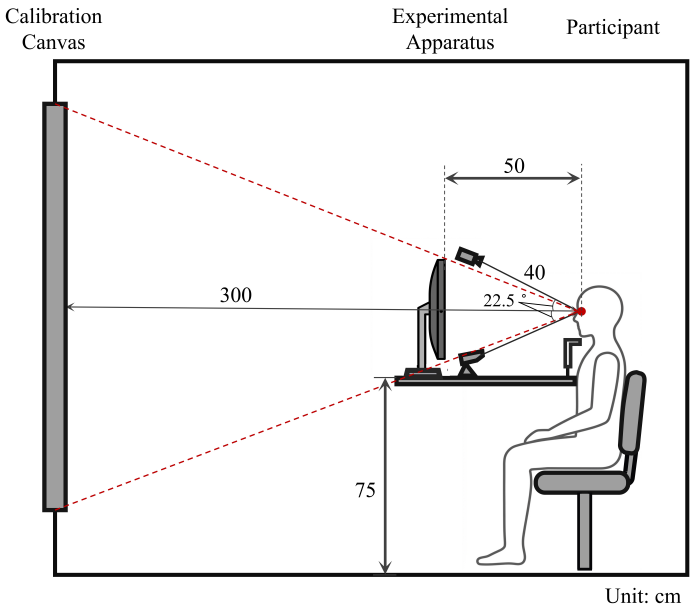
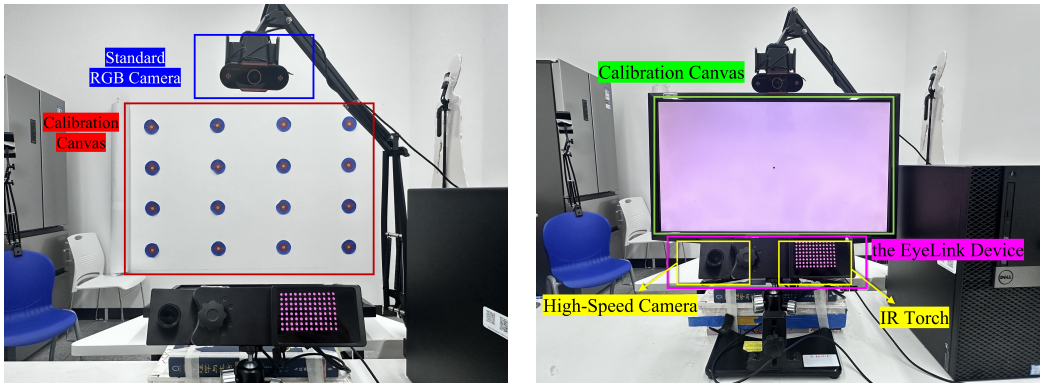


Figure 7. Spatial layout of Experiment 1.

A 23-inch monitor (50.9 × 28.6 cm) was positioned 50 cm from the eyes. The EyeLink infrared unit, equipped with a 35-mm lens for head-stabilized tracking, was positioned 22.5° below the participant's line of sight at a 40 cm camera-to-eye distance. The experimental task required fixation only at the 16 inset target locations, with maximum target

eccentricities of approximately 19.0° horizontally and 12.1° vertically. This configuration was used as the study-specific benchmark geometry for the EyeLink system. The proposed system's cameras were symmetrically mounted 22.5° above at the same distance. This symmetric placement was used to reduce differences in nominal camera-angle offset between the two acquisition configurations. Figure 8 shows the experimental setup of the two systems.



(**a**) (**b**)

Figure 8. Experimental setup of the two systems: (**a**) Proposed system; (**b**) EyeLink 1000 Plus.

(2) Calibration and Benchmark Tasks

Because the EyeLink 1000 Plus performs gaze prediction on the monitor plane while the proposed system maps gaze to a calibrated scene plane, a unified comparison framework was established. A rectangular calibration canvas (305.4 × 171.6 cm) was mounted on a wall 3 m from the participant's eyes, with its center aligned with the monitor center along the participant's horizontal line of sight (Figure 7). The physical canvas was represented by a 1920 × 1080 virtual coordinate system, corresponding to approximately 0.159 cm/px in both horizontal and vertical directions. Canvas dimensions were derived from monitor dimensions by a proportional scaling factor of 6, ensuring that corresponding points on the monitor and canvas subtended identical visual angles and thus corresponded to the same 3D gaze direction. Sixteen evenly distributed target points were marked on both surfaces, as shown in Figure 9. The same 16 spatial target locations were used for participant-specific calibration and subsequent evaluation. Although the spatial target locations themselves were not held out, the evaluation samples were acquired in a separate post-calibration fixation recording after H had been estimated and were not used to fit the homography. Accordingly, A_{pos} quantifies target-referenced post-calibration localization error at previously calibrated locations. During evaluation, participants again fixated each target for 5 s; the central 3 s window was used for metric computation. Figure 10 shows the experimental procedure. For each participant, the proposed system and EyeLink were evaluated in separate recordings, with the order of the two system measurements randomized across participants. The two systems used the same metric definitions while retaining their native processing pipelines. EyeLink recorded right-eye gaze at 2000 Hz with filter setting 2, and was analyzed from its native output. For the proposed system, A_{pos} and A_{spa} were calculated from the POR output after the operational One Euro filtering and DBSCAN stages. The tracker-recorded validation summaries yielded a mean validation error of $0.67 \pm 0.19°$ (range, 0.29–0.97°) and a mean maximum validation error of $3.59 \pm 1.51°$ (range, 1.55–7.37°). For both systems, A_{pos} and A_{spa} were calculated separately at each target and equally averaged across the 16 targets to obtain participant-level values.

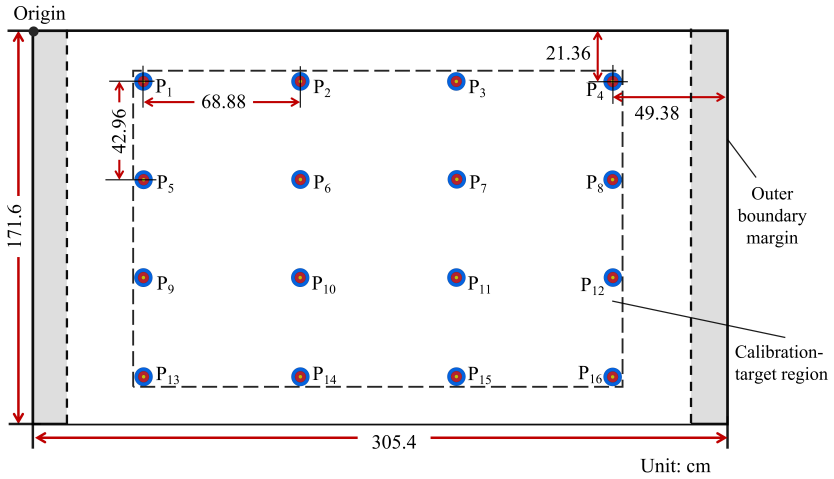
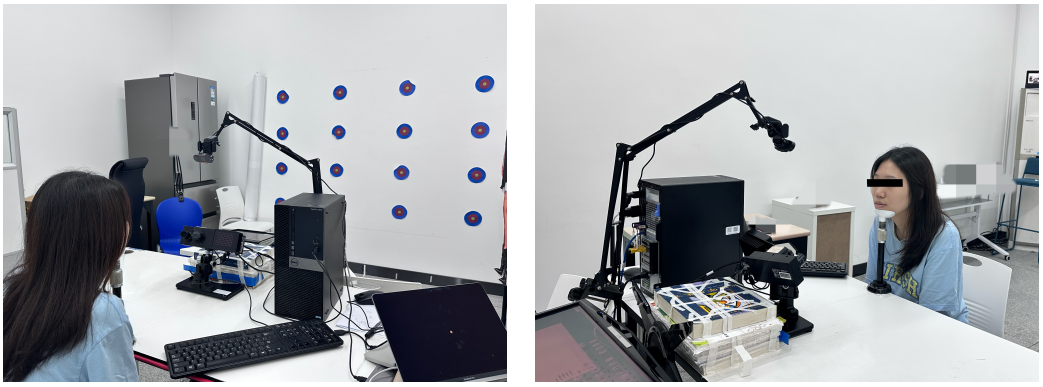


Figure 9. Wall canvas and 16-point distribution map for Experiment 1.



(**a**) (**b**)

Figure 10. Procedure record of Experiment 1: (**a**) Back view; (**b**) Front view.

(3) Evaluation Indicators

Target-referenced localization error was quantified by A_{pos} (px), defined as the mean Euclidean distance between each estimated gaze sample and the corresponding target position:

$$A_{\text{pos}} = \frac{1}{N}\sum_{i=1}^{N}\|g_i - t_i\|_2. \tag{25}$$

A_{pos} is a combined localization-error measure that reflects both systematic offset and sample dispersion. It is therefore not directly comparable to validation or manufacturer-specified accuracy figures defined as the offset of the mean gaze position from the target.

Spatial precision was quantified by the root-mean-square (RMS) spatial-dispersion metric A_{spa} (px), defined as the two-dimensional RMS distance of the gaze samples from their centroid:

$$A_{\text{spa}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\|g_i - \bar{g}\|_2^2}, \tag{26}$$

Here, g_i and t_i denote the 2D gaze and target positions, respectively, and $\bar{g}$ denotes the centroid of the gaze samples within the same target interval. For the proposed system, A_{spa} was calculated from POR samples available after the fixed operational One Euro filtering and DBSCAN stages. Accordingly, A_{spa} characterizes the spatial precision of the processed end-to-end POR output. Lower A_{pos} indicates lower combined localization error, whereas lower A_{spa} indicates higher spatial precision.

Because both metrics are expressed in pixels and therefore depend on the coordinate scale, participant-level A_{pos} and A_{spa} values were additionally converted to physical distance and degrees of visual angle. All cross-system metrics were computed in a common 1920×1080 coordinate system. Proposed-system coordinates, originally represented in a 2560×1600 scene frame, were transformed according to $x_c = x\left(\frac{1920}{2560}\right)$ and $y_c = y\left(\frac{1080}{1600}\right)$. EyeLink coordinates were retained in their native 1920×1080 frame. The 1920×1080 virtual canvas represented a physical area of 305.4×171.6 cm, yielding an approximately isotropic scale of 0.159 cm/px. Pixel distances were therefore converted as

$$d_{\text{cm}} = d_{\text{px}} \frac{305.4}{1920}. \tag{27}$$

Given the $D = 300$ cm eye-to-canvas distance, the corresponding visual angle was calculated as

$$\theta_{\text{deg}} = 2 \arctan\left(\frac{d_{\text{cm}}}{2D}\right) \frac{180}{\pi}, \qquad D = 300 \text{ cm}. \tag{28}$$

The angular values are reported to facilitate interpretation relative to conventional eye-tracking literature. The Experiment 1 proposed-system metrics reported in this study were calculated from the POR sequences after the operational filtering and DBSCAN stages and therefore characterize the processed end-to-end system output rather than the unfiltered gaze-estimation signal.

2.7.2. Experiment 2: Simulated Runway Observation

(1) Participants and Runway-Observation Setup

Experiment 2 applied the proposed system to a simulated runway observation scenario. A separate group of 16 participants (8 males and 8 females; mean age ± SD, 26.4 ±3.0 years) who had not participated in Experiment 1 was recruited using the same targeted online recruitment, pre-enrollment screening procedure, and eligibility and exclusion criteria as in Experiment 1. Video of the participants' face and eye regions was acquired using the same USB RGB webcam labeled "HD WEBCAM" as in Experiment 1, at a resolution of 1920×1080 px and an achieved frame rate of 30 fps. Unlike Experiment 1, no chin rest or other mechanical head-stabilization device was used in Experiment 2. Participants were allowed to maintain their natural head posture and make small spontaneous head movements during observation, while the frame-wise face and eye localization and eye-region normalization stages continuously updated the eye regions used for gaze estimation. The experimental layout is shown in Figure 11.

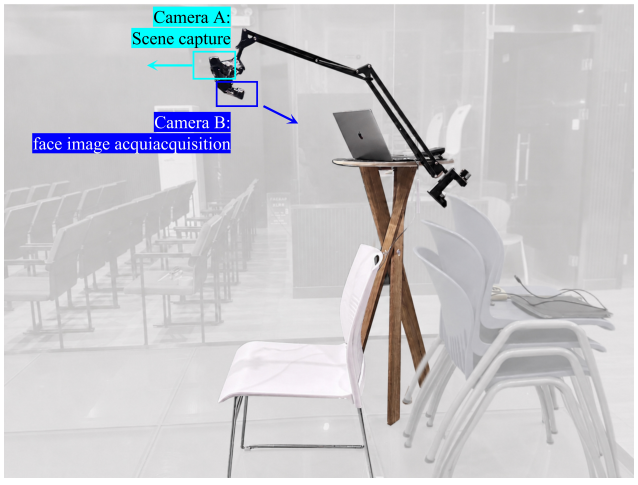


Figure 11. Device layout diagram for Experiment 2.

(2) Calibration and Observation Tasks

For Experiment 2, the observation distances were determined according to the spatial characteristics of runway viewing. Considering that the minimum social distance between runway models and KOLs is typically 2 m during a performance, the nearest observation point was set at a distance of 2 m from the participants' eyes. Based on this reference distance, multiples of 2 m were defined as two additional observation nodes in the direction of the line of sight along the runway: 4 m and 6 m, corresponding to the medium-distance and long-distance observation phases, respectively. A calibration canvas was positioned at 7 m, 1 m beyond the farthest observation node, so that the physical calibration surface did not obstruct the subsequent 6 m, 4 m and 2 m approach path. During calibration, each participant fixated the 16 evenly distributed calibration points sequentially for 5 s, with the central 3 s window used for homography estimation. The resulting participant-specific homography was retained across all three observation distances to maintain a single-calibration condition during the continuous approach. Because the target subsequently occupied depths of 6 m, 4 m and 2 m, its increasing departure from the calibration plane could introduce depth-dependent POR misregistration, particularly at the 2 m condition. This changing-depth configuration was retained to approximate the approach trajectory of a runway model.

In the formal observation task, to simulate the walking trajectory of a runway model towards the audience, a trained experimental assistant played the role of a model, continuously walking towards the participants from in front of the canvas. Along the walking path, locations 6 m, 4 m and 2 m from the participants' eyes were designated as Points 1, 2, and 3, respectively. The assistant walked through these nodes without pausing. Each 2 m interval took approximately 5–8 s, corresponding to an overall approach duration of approximately 10–16 s from 6 m to 2 m. Because the distance between the experimental assistant and the participants changed continuously during the walking process, the size of the model's projection in the scene image also changed accordingly. As shown in Figure 12, keyframes were extracted from the video recording when the experimental assistant was at points 1, 2, and 3, allowing for a comparison of the gaze results under different distance conditions with a unified spatial reference. POR samples corresponding to the three predefined observation conditions were analyzed separately for AOI localization and visualization. Participants were instructed to attend to the model's right-chest region. No unique point-level ground-truth coordinate was specified or recorded. The right-chest region therefore served as the predefined target AOI for evaluating whether the estimated gaze remained spatially concentrated as viewing distance changed. This region was selected because it contains salient garment details such as breast pockets and lapel stitching. Figure 13 shows a representative experimental record.

(3) AOI Definition and Evaluation Measures

For AOI analysis, the visible body of the model in each distance-specific scene frame was partitioned into three mutually exclusive semantic body regions (Figure 14): AOI-1 (head: face and hairstyle), AOI-2 (upper body: torso and garment details), and AOI-3 (lower body: trousers and footwear). The AOIs followed the visible anatomical and garment boundaries in each scene; their image extents therefore varied with the projected size and position of the model. Samples falling outside all three AOIs were classified as off-target.

To quantify the distance-dependent change in AOI extent, the body-region templates comprised a 25 × 36 cm head-and-neck region, a 44 × 64 cm upper-body region, and a 44 × 77 cm lower-body region. The upper- and lower-body widths corresponded to the measured shoulder width of 42 cm plus a 1 cm margin on each side, and the three regional

heights summed to the model's height of 177 cm. For parallel body and reference planes, the projected dimensions on the 7 m reference plane scaled with viewing distance. The corresponding projected AOI extents in the 2560 × 1600 reference coordinate system and their angular extents at each viewing distance are reported in Table 1. As expected from perspective projection, all three AOIs increased in both image extent and visual angle as viewing distance decreased, with the largest extents occurring at 2 m.

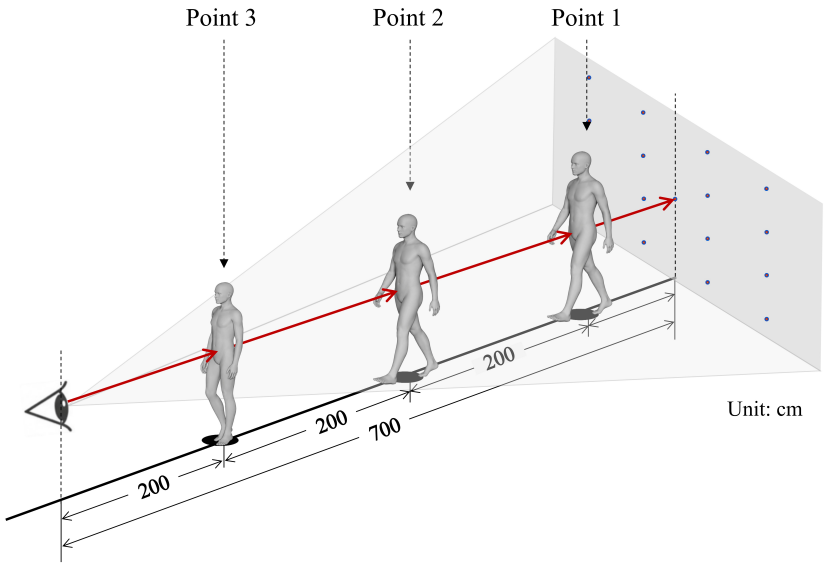
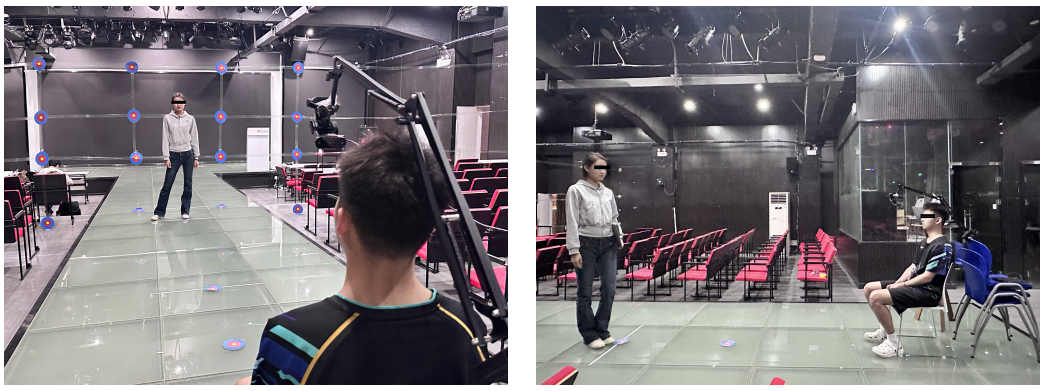


Figure 12. Schematic diagram of the observation task in Experiment 2. Note. The red arrows indicate the schematic viewing direction along the observation axis. The colored dots on the reference canvas mark the calibration points. Points 1, 2, and 3 correspond to viewing distances of 6, 4, and 2 m, respectively.



(**a**) (**b**)

Figure 13. Experiment record of Experiment 2: (**a**) Front view; (**b**) Side view.

For each participant and viewing distance, N_{proc} denotes the valid POR samples available after the gaze-processing pipeline and before AOI assignment. Samples assigned to the three non-overlapping AOIs constitute $N_{\text{retained}} = \sum_{j=1}^{3} N_{\text{AOI},j}$, whereas samples outside all AOIs constitute N_{off}; therefore, $N_{\text{proc}} = N_{\text{retained}} + N_{\text{off}}$. The AOI-retained sample percentage was calculated as

$$R_{\text{retained}} = \frac{N_{\text{retained}}}{N_{\text{proc}}} \times 100\%. \quad (29)$$

For AOI j,

$$P_{\mathrm{AOI},j} = \frac{N_{\mathrm{AOI},j}}{N_{\mathrm{proc}}} \times 100\%. \tag{30}$$

Table 1. Projected AOI extents across viewing distances.

Distance (m)	Head AOI	Upper-Body AOI	Lower-Body AOI
6	140 × 201 px (2.39° × 3.44°)	246 × 358 px (4.20° × 6.11°)	246 × 431 px (4.20° × 7.34°)
4	210 × 302 px (3.58° × 5.15°)	369 × 537 px (6.30° × 9.15°)	369 × 646 px (6.30° × 11.00°)
2	419 × 604 px (7.15° × 10.29°)	737 × 1074 px (12.55° × 18.18°)	737 × 1292 px (12.55° × 21.79°)

Note. All dimensions are reported as width × height. Pixel extents represent the projected body-region dimensions in the 2560 × 1600 reference coordinate system; angular extents (in parentheses) were calculated from the corresponding physical body-region dimensions and viewing distances. The head AOI corresponds to the head-and-neck template.

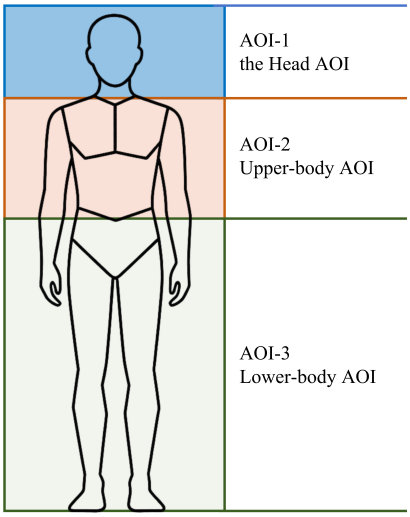


Figure 14. Body-region AOI definitions for the frontal-view runway target.

2.8. Statistical Analysis

Participant-level paired data were analyzed using nonparametric methods to avoid relying on normality assumptions with the sample size available in each experiment. All tests were two-sided, and the significance level was set at $\alpha = 0.05$. Descriptive statistics are reported as mean ± standard deviation together with participant-level values.

For Experiment 1, between-system comparisons were performed using two-sided exact Wilcoxon signed-rank tests on participant-level paired measurements. Rank-biserial correlation was reported as the paired effect-size measure. Mean paired differences were calculated as proposed-system minus EyeLink values, and 95% confidence intervals were estimated using participant-level percentile bootstrap resampling with 200,000 resamples while preserving the paired structure. For target-level analysis, Bland–Altman agreement was assessed across the 16 paired target-position means, and target dependence of the participant-level between-system differences was tested using Friedman tests, with Kendall's W reported as the effect-size measure.

For Experiment 2, Friedman tests were used to examine the effect of viewing distance on (1) AOI-retained sample percentage and (2) the percentage of post-processing gaze samples assigned to the upper-body AOI. When a Friedman omnibus test was statistically significant, exact pairwise Wilcoxon signed-rank tests were conducted, and the resulting

p-values were adjusted using the Holm procedure. Kendall's *W* was reported as the effect-size measure for Friedman tests, and rank-biserial correlation was reported for pairwise Wilcoxon comparisons. Statistical analyses were conducted in Python 3.9.

3. Results

3.1. Experiment 1: Controlled Calibration-Grid Benchmark—Results

Per-participant benchmark results for the EyeLink 1000 Plus and the proposed system are presented in Table 2. After coordinate harmonization, all pixel-based measurements were expressed in a common 1920 × 1080 coordinate system, and the corresponding visual angles were calculated using the same experimental geometry. For positional error, the proposed system yielded a mean A_{pos} of 82.19 ± 8.37 px (2.50 ± 0.25°), compared with 57.18 ± 18.10 px (1.74 ± 0.55°) for EyeLink. The mean paired difference was 25.01 px (0.76°), with a 95% bootstrap confidence interval of 17.25–32.66 px (0.52°–0.99°). The difference was significant ($W = 2$, $p < 0.001$, $r_{rb} = 0.971$). For the processed-output spatial-precision measure A_{spa}, the proposed system yielded a larger value than EyeLink under the present benchmark. Mean A_{spa} was 76.21 ± 7.90 px (2.31 ± 0.24°) for the proposed system and 33.64 ± 12.21 px (1.02 ± 0.37°) for EyeLink. The mean paired difference was 42.57 px (1.29°), with a 95% bootstrap confidence interval of 35.86–49.64 px (1.09°–1.51°), and the difference was significant ($W = 0$, $p < 0.001$, $r_{rb} = 1.000$).

Table 2. Results of Experiment 1.

Participant ID	EyeLink A_{pos}, px (°)	EyeLink A_{spa}, px (°)	Proposed A_{pos}, px (°)	Proposed A_{spa}, px (°)
1	44.37 (1.35°)	44.39 (1.35°)	83.04 (2.52°)	76.47 (2.32°)
2	55.75 (1.69°)	45.30 (1.38°)	83.08 (2.52°)	77.40 (2.35°)
3	60.65 (1.84°)	34.47 (1.05°)	71.30 (2.17°)	70.16 (2.13°)
4	46.80 (1.42°)	14.76 (0.45°)	82.78 (2.51°)	76.99 (2.34°)
5	34.79 (1.06°)	14.67 (0.45°)	92.92 (2.82°)	85.93 (2.61°)
6	61.51 (1.87°)	34.29 (1.04°)	79.64 (2.42°)	68.63 (2.08°)
7	60.94 (1.85°)	29.98 (0.91°)	88.10 (2.68°)	75.70 (2.30°)
8	66.82 (2.03°)	42.10 (1.28°)	80.86 (2.46°)	67.23 (2.04°)
9	114.18 (3.47°)	31.46 (0.96°)	101.96 (3.10°)	91.07 (2.77°)
10	45.11 (1.37°)	25.73 (0.78°)	84.01 (2.55°)	72.00 (2.19°)
11	57.93 (1.76°)	43.69 (1.33°)	86.34 (2.62°)	91.24 (2.77°)
12	62.28 (1.89°)	18.27 (0.55°)	75.85 (2.30°)	78.21 (2.38°)
13	42.70 (1.30°)	19.77 (0.60°)	65.03 (1.98°)	62.85 (1.91°)
14	39.23 (1.19°)	43.27 (1.31°)	82.60 (2.51°)	78.10 (2.37°)
15	57.84 (1.76°)	54.33 (1.65°)	77.43 (2.35°)	74.48 (2.26°)
16	63.93 (1.94°)	41.74 (1.27°)	80.08 (2.43°)	72.85 (2.21°)
Mean ± SD	57.18 ± 18.10 (1.74 ± 0.55°)	33.64 ± 12.21 (1.02 ± 0.37°)	82.19 ± 8.37 (2.50 ± 0.25°)	76.21 ± 7.90 (2.31 ± 0.24°)

Note. A_{pos} = combined target-referenced localization error, calculated as the mean sample-to-target Euclidean distance; A_{spa} = spatial precision operationalized as RMS spatial dispersion of the analyzed POR output. Pixel values are expressed in the common 1920 × 1080 coordinate frame, corresponding to 0.159 cm/px. Visual-angle values were derived using the 3-m eye-to-canvas distance. Lower values indicate better performance.

Target-level analysis further showed that the positional-error difference was not spatially uniform. Across the 16 target positions, the mean proposed-minus-EyeLink bias in A_{pos} was +25.01 px (+0.76°), with Bland–Altman limits of agreement from −40.29 to +90.31 px (−1.22° to +2.74°). The between-system difference varied significantly across target locations ($\chi^2(15) = 84.40$, $p < 0.001$, Kendall's $W = 0.352$), indicating a target-dependent component in positional accuracy. In contrast, the between-system difference in A_{spa} was more consistent across the target grid. The mean A_{spa} bias was +42.57 px (+1.29°),

with limits of agreement from +23.19 to +61.94 px (+0.70° to +1.88°), and no significant target-position effect was detected ($\chi^2(15) = 20.06$, $p = 0.170$, Kendall's $W = 0.084$).

Although the proposed system showed lower positional accuracy and spatial precision than the laboratory-grade EyeLink reference, its mean A_{pos} and A_{spa} remained within approximately 2.5° and 2.3°, respectively. Moreover, the relatively small between-participant variability of the proposed system (SD = 0.25° for A_{pos} and SD = 0.24° for A_{spa}) indicates stable performance across participants. These results suggest that the system retains useful spatial localization capability under controlled conditions, particularly for applications in which gaze is interpreted at the level of broader regions of interest.

3.2. Experiment 2: Performance Across Viewing Distances

Per-participant results for the 6 m, 4 m and 2 m viewing conditions are reported in Tables 3–5. The mean AOI-retained sample percentages were 95.96 ± 4.10% at 6 m, 97.31 ± 3.50% at 4 m, and 80.43 ± 14.95% at 2 m. Participant-level AOI-retained sample percentages ranged from 84.9% to 100.0% at 6 m, from 88.1% to 100.0% at 4 m, and from 54.6% to 100.0% at 2 m. The sample counts further show that the reduction at 2 m was associated with a markedly larger number of post-processing POR samples classified as off-target. Across the 16 participants, N_{off} totaled 746 of 3681 processed samples at 2 m, compared with 99 of 3618 at 4 m and 142 of 3524 at 6 m.

Table 3. Experimental results under the 6 m condition.

Participant	N_{proc}	Head AOI (% of N_{proc})	Upper-Body AOI (% of N_{proc})	Lower-Body AOI (% of N_{proc})	N_{off}	AOI-Retained Samples (%)
1	256	45.7	52.0	0.0	6	97.7
2	219	13.2	71.7	0.0	33	84.9
3	205	1.0	98.0	0.0	2	99.0
4	237	0.0	94.9	0.0	12	94.9
5	191	15.2	77.5	0.0	14	92.7
6	200	16.0	81.5	0.0	5	97.5
7	273	19.4	76.2	0.0	12	95.6
8	220	6.4	93.6	0.0	0	100.0
9	222	0.0	96.8	0.0	7	96.8
10	219	26.5	62.1	0.0	25	88.6
11	212	21.7	74.1	0.0	9	95.8
12	229	9.2	88.6	0.0	5	97.8
13	200	11.0	87.0	0.0	4	98.0
14	192	4.7	93.2	0.0	4	97.9
15	225	0.0	100.0	0.0	0	100.0
16	224	2.7	95.5	0.0	4	98.2
Mean	–	12.0	83.9	0.0	–	95.96 ± 4.10

The small nonzero lower-body percentages may reflect brief gaze shifts away from the instructed right-chest target; however, the present data cannot determine whether these samples represent actual gaze shifts or occasional localization error.

A Friedman test indicated a significant effect of viewing distance on the AOI-retained sample percentage ($\chi^2_F(2) = 9.50$, $p = 0.009$, Kendall's $W = 0.30$). Holm-adjusted pairwise Wilcoxon signed-rank tests showed higher AOI-retained sample percentages at 6 m than at 2 m ($W = 8$, $p_{adj} = 0.002$, $r_{rb} = 0.88$) and at 4 m than at 2 m ($W = 10$, $p_{adj} = 0.003$, $r_{rb} = 0.85$). No statistically significant difference was detected between the 6 m and 4 m conditions ($W = 55$, $p_{adj} = 0.528$, $r_{rb} = -0.19$).

Table 4. Experimental results under the 4 m condition.

Participant	N_{proc}	Head AOI (% of N_{proc})	Upper-Body AOI (% of N_{proc})	Lower-Body AOI (% of N_{proc})	N_{off}	AOI-Retained Samples (%)
1	202	9.9	87.1	0.0	6	97.0
2	276	17.4	82.6	0.0	0	100.0
3	294	16.7	71.4	0.0	35	88.1
4	184	41.3	56.0	0.0	5	97.3
5	208	38.0	62.0	0.0	0	100.0
6	188	35.6	63.3	0.0	2	98.9
7	141	7.1	84.4	0.0	12	91.5
8	226	34.1	60.6	0.0	12	94.7
9	251	24.7	71.7	0.0	9	96.4
10	242	17.4	82.6	0.0	0	100.0
11	246	35.4	64.6	0.0	0	100.0
12	250	31.2	63.2	0.8	12	95.2
13	162	28.4	71.6	0.0	0	100.0
14	265	39.2	60.8	0.0	0	100.0
15	274	52.2	45.6	0.0	6	97.8
16	209	57.4	42.6	0.0	0	100.0
Mean	–	30.4	66.9	0.05	–	97.31 ± 3.50

Table 5. Experimental results under the 2 m condition.

Participant	N_{proc}	Head AOI (% of N_{proc})	Upper-Body AOI (% of N_{proc})	Lower-Body AOI (% of N_{proc})	N_{off}	AOI-Retained Samples (%)
1	182	0.0	61.0	0.0	71	61.0
2	226	0.0	89.8	0.0	23	89.8
3	206	0.0	100.0	0.0	0	100.0
4	201	0.0	98.0	0.0	4	98.0
5	203	0.0	71.9	0.0	57	71.9
6	295	0.0	54.6	0.0	134	54.6
7	228	0.0	60.1	3.1	84	63.2
8	251	0.0	70.5	0.0	74	70.5
9	271	0.0	80.1	0.0	54	80.1
10	289	0.0	77.5	5.5	49	83.0
11	248	0.0	89.9	0.0	25	89.9
12	236	0.0	75.0	10.6	34	85.6
13	228	0.0	60.5	0.0	90	60.5
14	246	38.6	48.4	0.0	32	87.0
15	187	14.4	84.5	0.0	2	98.9
16	184	0.0	92.9	0.0	13	92.9
Mean	–	3.3	75.9	1.2	–	80.43 ± 14.95

The mean percentages of post-processing gaze samples N_{proc} assigned to the upper-body AOI were 83.9%, 66.9%, and 75.9% at 6 m, 4 m, and 2 m, respectively. A Friedman test did not detect a statistically significant difference in upper-body AOI assignment percentage across the three viewing distances ($\chi^2_F(2) = 4.50$, $p = 0.105$, Kendall's $W = 0.14$). Thus, although the proportion of post-processing gaze samples falling within the predefined AOIs varied significantly with viewing distance, the upper-body AOI assignment percentage did not show a statistically detectable cross-distance difference.

The KDE gaze-density maps are shown in Figures 15–17. Across participants, the highest-density regions were predominantly located within the upper-body AOI and were concentrated around the predefined right-chest observation target. This spatial concentration was observed at 6 m, 4 m, and 2 m, although the distribution and AOI-retained sample percentages varied among participants and across viewing distances.

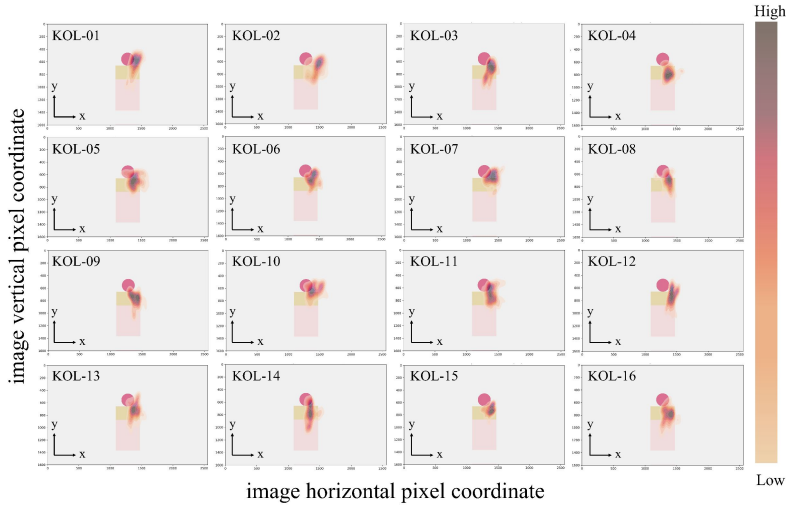


Figure 15. Gaze heatmap under the 6 m condition. Note. The body-region representation reflects the scene view at this distance; only the portions within the displayed image area are shown.

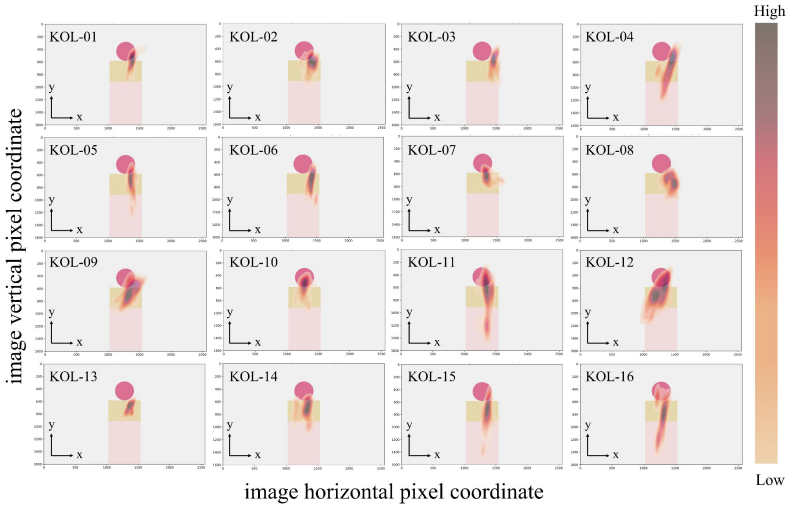


Figure 16. Gaze heatmap under the 4 m condition. Note. The body-region representation reflects the scene view at this distance; only the portions within the displayed image area are shown.

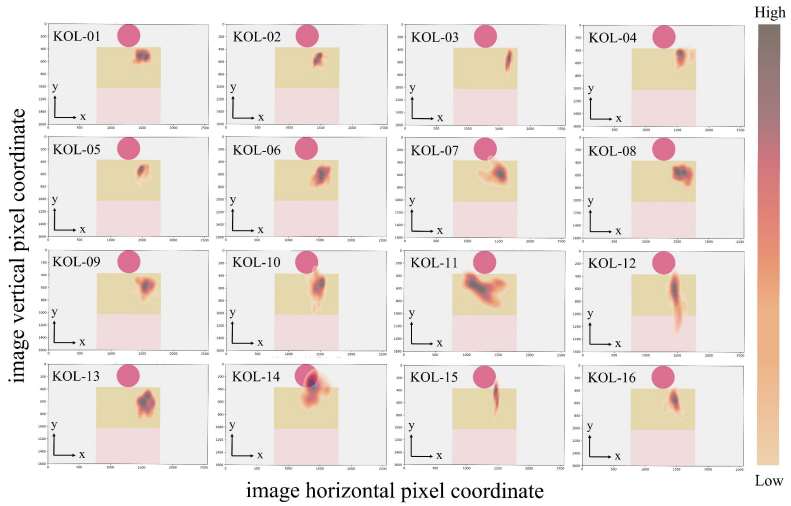


Figure 17. Gaze heatmap under the 2 m condition. Note. The body-region representation reflects the scene view at this distance; only the portions within the displayed image area are shown.

4. Discussion

4.1. General Findings

Experiment 1 provided a controlled characterization of the post-calibration target-localization performance of the complete RGB-to-scene gaze-mapping pipeline under the fixed 3 m benchmark geometry. Although the proposed system showed higher positional error than the EyeLink 1000 Plus, its mean error remained within a low-single-digit visual-angle range (2.50° vs. 1.74°). This benchmark therefore provided a study-specific performance reference for the subsequent application-oriented AOI-level evaluation. Because Experiment 2 used different 2, 4, and 6 m viewing geometries and did not obtain independent point-level ground truth, the Experiment 1 results should not be interpreted as positional-accuracy estimates for those conditions.

The observed performance gap is consistent with the different sensing and estimation architectures of the two systems. The proposed framework derives scene-level POR from passive RGB imagery through eye-region normalization, gaze-direction estimation, individualized scene mapping, and downstream POR processing, allowing uncertainty from multiple stages to propagate to the final gaze estimate. This limits fine point-of-gaze discrimination; accordingly, Experiment 2 focused on localization within predefined anatomical AOIs rather than fine point locations.

Experiment 2 further examined the system through an application-oriented AOI-localization feasibility test under viewing conditions designed to approximate runway observation. AOI-retained sample percentages remained high at 6 m and 4 m but decreased at 2 m, indicating reduced AOI-level localization performance at the closest viewing distance. Importantly, this reduction occurred despite the AOIs having their largest projected and angular extents at 2 m. For example, the upper-body AOI increased from approximately 246 × 358 px (4.20° × 6.11°) at 6 m to 737 × 1074 px (12.55° × 18.18°) at 2 m. Thus, the lower AOI-retained sample percentage at 2 m cannot be explained by a smaller projected target or a smaller AOI acceptance region.

At 2 m, the reduction in AOI-retained sample percentage was accompanied by a marked increase in N_{off}, indicating that more post-processing POR samples fell outside the predefined AOIs. A plausible contributor is the changing-depth geometry relative to the single-plane calibration. As the target moved progressively farther from the calibration plane, depth-dependent mismatch between the mapped POR coordinates and the distance-specific AOI boundaries may have increased, resulting in more samples being classified as off-target at the closest viewing distance. A further plausible contributor is target motion relative to the static keyframe AOIs. For an approximately constant-speed approach along the line of sight, the rate of change in projected linear size increases approximately with the inverse square of viewing distance. The model's projected boundaries may therefore diverge more rapidly from the keyframe AOIs within the analysis interval at 2 m, potentially increasing N_{off}. Neither this motion-related AOI mismatch nor the depth-dependent mapping error was independently quantified; the observed reduction therefore cannot be attributed uniquely to either mechanism.

Across the three viewing distances, the upper-body AOI accounted for the largest percentage of the post-processing gaze samples (N_{proc}), and the KDE maps showed that high-density gaze regions were generally concentrated around the predefined right-chest target. The upper-body AOI assignment percentage did not show a statistically detectable difference across viewing distances ($p = 0.105$). Because participants were explicitly instructed to attend to the right-chest region, these distributions primarily reflect the system's ability to localize instructed visual attention rather than spontaneous attentional preference. The consistency of the gaze-density maps with the predefined target nevertheless supports

the feasibility of using the proposed framework for AOI-level attention measurement across changing observer-to-target distances.

4.2. Limitations and Future Work

First, Experiment 1 evaluated newly acquired post-calibration samples at the same spatial locations used for participant-specific calibration. The reported positional errors therefore characterize same-location post-calibration performance rather than held-out spatial generalization. Because the point-level calibration data required for reliable refitting were not retained, future work should evaluate independent spatial target sets or pre-specified held-out calibration procedures. In addition, the Experiment 1 proposed-system metrics reflect the fixed operational preprocessing configuration. Intermediate POR outputs were not retained at every processing stage, precluding retrospective calculation of unfiltered or stage-specific precision measures, the exact Experiment 1 DBSCAN rejection rate, and systematic parameter-sensitivity analyses. The retained Experiment 1 outputs for the proposed system also did not include per-target mean gaze coordinates or sample-level POR coordinates, preventing retrospective calculation of the conventional target-referenced mean-position offset. This limits direct comparison of the proposed system with manufacturer specifications and studies that report accuracy using that definition. Future evaluations should retain sample-level outputs and report a metric-matched mean-position accuracy measure alongside A_{spa}. Prospective studies should therefore retain stage-specific outputs and assess filtering and clustering sensitivity.

Second, Experiment 2 used a single planar calibration at 7 m for targets subsequently observed at 6, 4, and 2 m, and the resulting depth-dependent mapping mismatch was not independently quantified. In addition, the fixed DBSCAN radius ($\varepsilon = 40$ px) and KDE bandwidth ($h = 30$ px) represented different proportions of the projected body size across viewing distances. Consequently, cross-distance comparisons retain a processing-parameter component: DBSCAN may affect the composition of the processed sample set, whereas KDE affects only the visualized density distribution and does not enter AOI assignment or the statistical analyses. Experiment 2 also lacked independent frame-by-frame point-level gaze ground truth; the AOI-retained sample percentage should therefore be interpreted as a post-processing AOI-inclusion measure. Future work should evaluate multi-depth calibration and depth-aware scene mapping under controlled target-depth conditions.

Third, each experiment included 16 relatively young participants. Spectacle wearers and participants with myopia greater than 3.00 D in magnitude were not evaluated. Future work should include larger and more diverse participant samples.

Fourth, the inherited eye-region landmark network was trained on synthetic UnityEyes imagery and was not fine-tuned or domain-adapted using the far-field RGB recordings collected in the present study. A domain mismatch therefore exists between the original training imagery and the present acquisition conditions, but its contribution to the observed end-to-end localization error was not independently quantified.

Fifth, Experiment 2 was conducted without mechanical head stabilization, but head position and orientation were not independently measured. Accordingly, the present study does not quantify head-motion tolerance or POR robustness as a function of head translation or rotation.

5. Conclusions

This study developed a far-field, non-contact eye-tracking framework for visual-attention measurement in fashion-runway settings. By integrating appearance-based 3D gaze-direction estimation, individualized scene mapping, and AOI-oriented attention

analysis, the framework connects RGB-based gaze estimation with interpretable scene-level visual-attention measurement.

Experiment 1 characterized the operating performance of the integrated RGB-to-scene gaze-mapping pipeline under the controlled benchmark condition.

Although the proposed system showed greater point-level positional error and a higher processed-output A_{spa} value than the EyeLink 1000 Plus under the controlled benchmark, the proposed system maintained scene-level localization within a low-single-digit visual-angle range, supporting its intended use for broader AOI-level analysis. Experiment 2 further supported the feasibility of target-directed AOI localization across the tested viewing distances, although the AOI-retained sample percentage decreased at 2 m.

Overall, these findings support the feasibility of the proposed framework for far-field, non-contact AOI-level visual-attention measurement under the present simulated runway-observation conditions. Future work should evaluate the framework's performance in unrestricted live-runway environments, including continuous smooth-pursuit and synchronized tracking of multiple observers.

Author Contributions: Conceptualization, H.T., M.G. and Q.L.; Methodology, H.T., M.G. and Q.L.; Software, H.T.; Validation, H.T. and M.G.; Investigation, H.T., M.G. and Q.L.; Visualization, H.T.; Data Curation, H.T.; Writing—original draft preparation, H.T. and M.G.; Writing—review and editing, D.-B.L.; Supervision, D.-B.L. and J.M.; Funding acquisition, J.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Beijing Natural Science Foundation (grant number L247008) and the Guangdong Province Social Science Planning Project (grant number GD25YYS52).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Academic Integrity and Research Ethics Committee of Guangdong University of Technology (protocol code GDUTXS20250187, 3 July 2025).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Consent for publication was obtained from all identifiable individuals shown in the figures.

Data Availability Statement: The original data presented in this study are openly available in Zenodo (version v2) at https://doi.org/10.5281/zenodo.22706802.

Acknowledgments: The authors would like to thank all participants who volunteered for this study. The authors are also grateful to the Human-Centered Technology Laboratory and the New Media Exhibition and Communication Laboratory at Guangdong University of Technology for providing the testing venues and facilities, and to the Industrial Innovation Research Laboratory for its support and assistance.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mendes, S. The Instagrammability of the Runway: Architecture, Scenography, and the Spatial Turn in Fashion Communications. *Fash. Theory* **2021**, *25*, 311–338. [CrossRef]
2. SanMiguel, P.; Rus-Navas, A.; Sádaba, T. Fashion Shows: The Greatest Show on Earth. In *Proceedings of the Fashion Communication in the Digital Age*; Springer Nature: Cham, Switzerland, 2023; pp. 227–237. [CrossRef]
3. Sardar, S.; Tata, S.V.; Sarkar, S. Examining the influence of source factors and content characteristics of influencers' post on consumer engagement and purchase intention: A moderated analysis. *J. Retail. Consum. Serv.* **2024**, *79*, 103888. [CrossRef]
4. Spörl-Wang, K.; Krause, F.; Henkel, S. Predictors of social media influencer marketing effectiveness: A comprehensive literature review and meta-analysis. *J. Bus. Res.* **2025**, *186*, 114991. [CrossRef]
5. Wang, Y.; Lee, H.; Ko, E.; Milfeld, T. The impact of luxury brands' NFT fashion on brand attitude and electronic word-of-mouth. *Int. J. Advert.* **2026**, *45*, 1081–1116. [CrossRef]
6. Mahanama, B.; Jayawardana, Y.; Rengarajan, S.; Jayawardena, G.; Chukoskie, L.; Snider, J.; Jayarathna, S. Eye Movement and Pupil Measures: A Review. *Front. Comput. Sci.* **2022**, *3*, 733531. [CrossRef]

7. Spering, M. Eye Movements as a Window into Decision-Making. *Annu. Rev. Vis. Sci.* **2022**, *8*, 427–448. [CrossRef] [PubMed]
8. Laborde, Q.; Roques, A.; Armougum, A.; Vayatis, N.; Bargiotas, I.; Oudre, L. Vision toolkit part 2. Features and metrics for assessing oculomotor signal: A review. *Front. Physiol.* **2025**, *16*, 1661026. [CrossRef] [PubMed]
9. Jang, J.Y. Analyzing visual behavior of consumers in a virtual reality fashion store using eye tracking. *Fash. Text.* **2023**, *10*, 24. [CrossRef]
10. Nordfält, J.; Ahlbom, C.P. Utilising eye-tracking data in retailing field research: A practical guide. *J. Retail.* **2024**, *100*, 148–160. [CrossRef]
11. Novák, J.Š.; Masner, J.; Benda, P.; Šimek, P.; Merunka, V. Eye Tracking, Usability, and User Experience: A Systematic Review. *Int. J. Hum.–Comput. Interact.* **2024**, *40*, 4484–4500. [CrossRef]
12. Cheng, Y.; Wang, H.; Bao, Y.; Lu, F. Appearance-Based Gaze Estimation With Deep Learning: A Review and Benchmark. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 7509–7528. [CrossRef] [PubMed]
13. Zhang, X.; Sugano, Y.; Fritz, M.; Bulling, A. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 162–175. [CrossRef] [PubMed]
14. Wang, X.; Zhang, J.; Zhang, H.; Zhao, S.; Liu, H. Vision-Based Gaze Estimation: A Review. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *14*, 316–332. [CrossRef]
15. Sun, Y.; Zeng, J.; Shan, S. Gaze estimation with semi-supervised eye landmark detection as an auxiliary task. *Pattern Recognit.* **2024**, *146*, 109980. [CrossRef]
16. Nie, W.; Wang, Z.; Ren, W.; Zhang, H.; Liu, H. Iris Geometric Transformation Guided Deep Appearance-Based Gaze Estimation. *IEEE Trans. Image Process.* **2025**, *34*, 1616–1631. [CrossRef] [PubMed]
17. Kim, S.; Nam, W.J.; Lee, S.W. Appearance debiased gaze estimation via stochastic subject-wise adversarial learning. *Pattern Recognit.* **2024**, *152*, 110441. [CrossRef]
18. Xia, L.; Li, Y.; Cai, X.; Cui, Z.; Xu, C.; Chan, A.B. Collaborative contrastive learning for cross-domain gaze estimation. *Pattern Recognit.* **2025**, *161*, 111244. [CrossRef]
19. Kuric, E.; Demcak, P.; Majzel, J.; Nguyen, G. Democratizing eye-tracking? Appearance-based gaze estimation with improved attention branch. *Eng. Appl. Artif. Intell.* **2025**, *149*, 110494. [CrossRef]
20. Park, S.; Zhang, X.; Bulling, A.; Hilliges, O. Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*; ETRA '18; ACM: New York, NY, USA, 2018; pp. 1–10. [CrossRef]
21. Niehorster, D.C.; Cornelissen, T.H.W.; Holmqvist, K.; Hooge, I.T.C.; Hessels, R.S. What to expect from your remote eye-tracker when participants are unrestrained. *Behav. Res. Methods* **2018**, *50*, 213–227. [CrossRef] [PubMed]
22. Cantó-Cerdán, M.; Hernández-Rodríguez, C.J.; Martínez-Abad, A. Evidence on the parameters of oculomotor skills and normative values: A systematic review. *J. Optom.* **2025**, *18*, 100570. [CrossRef]
23. Fu, X.; Franchak, J.M.; MacNeill, L.A.; Gunther, K.E.; Borjon, J.I.; Yurkovic-Harding, J.; Harding, S.; Bradshaw, J.; Pérez-Edgar, K.E. Implementing mobile eye tracking in psychological research: A practical guide. *Behav. Res. Methods* **2024**, *56*, 8269–8288. [CrossRef] [PubMed]
24. Hooge, I.T.C.; Niehorster, D.C.; Hessels, R.S.; Benjamins, J.S.; Nyström, M. How robust are wearable eye trackers to slow and fast head and body movements? *Behav. Res. Methods* **2023**, *55*, 4128–4142. [CrossRef] [PubMed]
25. Niehorster, D.C.; Zemblys, R.; Beelders, T.; Holmqvist, K. Characterizing gaze position signals and synthesizing noise during fixations in eye-tracking data. *Behav. Res. Methods* **2020**, *52*, 2515–2534. [CrossRef] [PubMed]
26. King, D.E. Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* **2009**, *10*, 1755–1758.
27. King, D.E. Max-Margin Object Detection. *arXiv* **2015**, arXiv:1502.00046. [CrossRef]
28. Kazemi, V.; Sullivan, J. One Millisecond Face Alignment with an Ensemble of Regression Trees. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1867–1874. [CrossRef]
29. Kaur, H.; Jindal, S.; Manduchi, R. Rethinking Model-Based Gaze Estimation. In *Proceedings of the ACM on Computer Graphics and Interactive Techniques*; ACM: New York, NY, USA, 2022; Volume 5, pp. 1–17. [CrossRef] [PubMed]
30. Zuiderveld, K.J. Contrast Limited Adaptive Histogram Equalization. In *Graphics Gems*; Heckbert, P.S., Ed.; Elsevier: Amsterdam, The Netherlands, 1994; pp. 474–485. [CrossRef]
31. Zhang, X.; Sugano, Y.; Bulling, A. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*; ETRA '18; ACM: New York, NY, USA, 2018; pp. 1–9. [CrossRef]
32. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In *Proceedings of the Computer Vision—ECCV 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 483–499. [CrossRef]
33. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2004. [CrossRef]

34. Casiez, G.; Roussel, N.; Vogel, D. 1 € filter: A simple speed-based low-pass filter for noisy input in interactive systems. In Proceedings of the CHI '12: SIGCHI Conference on Human Factors in Computing Systems, Austin, TX, USA, 5–10 May 2012; pp. 2527–2530. [CrossRef]
35. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; Volume 96, pp. 226–231.
36. Angeles, D.; Kurtek, S.; Klein, E.; Brinkman, M.; Ferketich, A. Geometric framework for statistical analysis of eye tracking heat maps, with application to a tobacco waterpipe study. *J. Appl. Stat.* **2024**, *51*, 1191–1209. [CrossRef] [PubMed]
37. SR Research Ltd. *EyeLink 1000 Plus User Manual*, Version 1.0.12; SR Research Ltd.: Kanata, ON, Canada, 2017.
38. Kaduk, T.; Goeke, C.; Finger, H.; König, P. Webcam eye tracking close to laboratory standards: Comparing a new webcam-based system and the EyeLink 1000. *Behav. Res. Methods* **2024**, *56*, 5002–5022. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.