Manual-Free Gaze Interaction via Bayesian-Based Implicit Intention Prediction

Taewoo Jo, Ho Jung Lee, Sulim Chun, and In-Kwon Lee

***Abstract*—Eye gaze is regarded as a promising interaction modality in extended reality (XR) environments. However, to address the challenges posed by the Midas touch problem, the determination of selection intention frequently relies on the implementation of additional manual selection techniques, such as explicit gestures (e.g., controller/hand inputs or dwell), which are inherently limited in their functionality. We hereby present a machine learning (ML) model based on the Bayesian framework, which is employed to predict user selection intention in real-time, with the unique distinction that all data used for training and prediction are obtained from gaze data alone. The model utilizes a Bayesian approach to transform gaze data into selection probabilities, which are subsequently fed into an ML model to discern selection intentions. In Study 1, a high-performance model was constructed, enabling real-time inference using solely gaze data. This approach was found to enhance performance, thereby validating the efficacy of the proposed methodology. In Study 2, a user study was conducted to validate a manual-free technique based on the prediction model. The advantages of eliminating explicit gestures and potential applications were also discussed.**

***Index Terms*—Extended reality (XR), eye tracking, interaction technique, intent prediction, implicit gaze.**

I. Introduction

INTERACTING with a target constitutes a fundamental task in the field of human-computer interaction (HCI), as it enables users to perform a variety of tasks, including text input, object manipulation, and locomotion. Given the substantial impact of rapid, precise, and ergonomic pointing and selection techniques on enhancing the interaction experience, the development of novel interaction techniques remains a vibrant research domain within the field of HCI. Hand-held controllers and hand-tracking-based ray casting techniques are frequently employed for pointing tasks. However, the prolonged use of these methods often results in physical fatigue, known as the "gorilla-arm effect" [1], [2]. Furthermore, the efficacy of these approaches is diminished when the hands are occupied with other physical objects, thereby limiting their effectiveness for pointing. The advent of head-mounted displays (HMDs) capable of eye-tracking has led to a surge of interest in gaze-based pointing. Gaze serves as a pointing modality, offering several advantages over freehand pointing. First, it requires minimal physical effort [3]. Second, it facilitates rapid pointing by allowing for simultaneous observation and pointing [3], [4], [5], [6], [7]. However, the implementation of gaze as a standalone input modality poses significant challenges. A significant concern pertains to the "Midas touch problem", which arises from the challenge of discerning the user's selection intention solely through eye movement. This limitation impedes the efficacy of gaze-based unimodal interactions [8].

Several methods have been proposed to address this issue. These methods allow users to manually express their selection intent through explicit gestures. These methods include multimodal techniques (e.g., controller [9], hand [10], [11], [12], [13]), dwelling [4], [7], [14], [15], [16], eye blink [17], eyelid movement [18], and smooth pursuit [19], [20]. These methodologies have been shown to accurately acquire three-dimensional targets and efficiently prevent unwanted selection. This capacity significantly mitigates the occurrence of the "Midas touch". However, manual techniques that require explicit gestures have notable limitations. Controller- and hand-gesture-based techniques are impractical when the hands are occupied. They can also be less accessible for users with upper-body impairments or older adults [21], [22]. Dwell-based selection introduces a delay in expressing intention, slowing work and disrupting flow [23]. In summary, the necessity of an explicit gesture creates modality-specific disadvantages. Furthermore, a variety of methodologies have been developed to predict users' selection intentions [12], [14], [15], [24], [25]. These methodologies have been used to automate selection processes [26]. This approach has the potential to facilitate rapid and intuitive interaction by eliminating the need for manual selection. However, many previous methods have not demonstrated real-time implementation and usually only acknowledge the potential to extend their approach to manual-free interaction techniques [24], [25]. Despite studies devising manual-free interaction techniques, quantitative comparisons with established manual techniques remain under-explored [26]. This paucity of research hinders a comprehensive understanding of the merits of truly manual-free interaction.

Received 13 January 2025; revised 23 September 2025; accepted 24 September 2025. Date of publication 29 September 2025; date of current version 10 November 2025. This work was supported in part by the National Research Foundation of Korea under Grant RS-2024-00348094 and in part by Korea Radio Promotion Association under Grant RNIX20230200, grant funded by the Republic of Korea government (MSIT). Recommended for acceptance by D. Iwai. *(Corresponding author: In-Kwon Lee.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Institutional Review Board of Yonsei University under Application No. 7001988-202508-HR-2398-03.

The authors are with the Department of Computer Science, Yonsei University, Seoul 03722, Republic of Korea (e-mail: twj5349@yonsei.ac.kr; dearshawn@yonsei.ac.kr; slchun@yonsei.ac.kr; iklee@yonsei.ac.kr).

Digital Object Identifier 10.1109/TVCG.2025.3615198

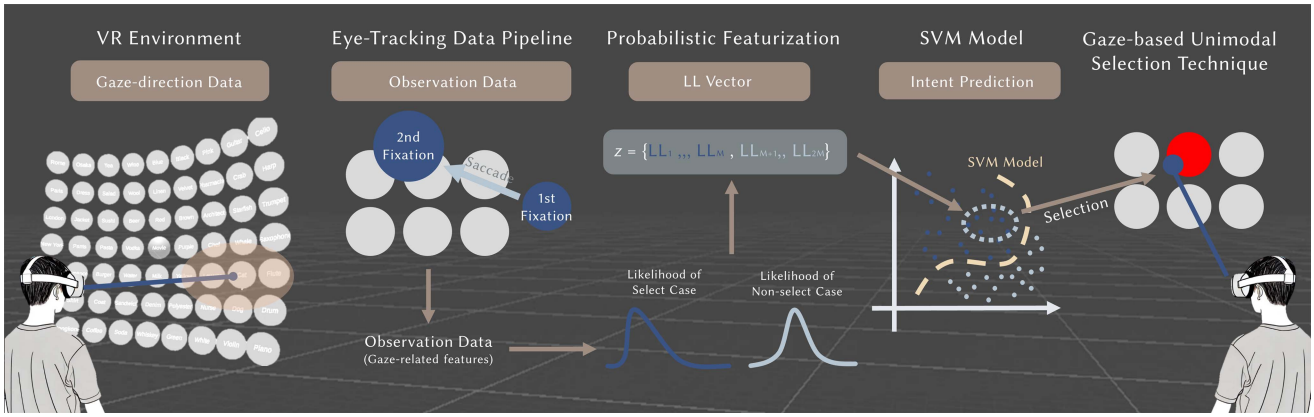

Fig. 1. We hereby propose a model for predicting selection intentions and a technique for manual-free interaction, both of which use only eye-gaze data. Initially, the user's eye gaze movement features were collected while observing and acquiring a three-dimensional target. Subsequently, we identified features that exhibited a substantial discrepancy based on the intention class. Subsequently, a Bayesian strategy was employed to convert the observation data into a log-likelihood vector. The vector is subsequently fed into the machine learning model, which infers the user's intention to select the target. The interaction technique utilizes this inference to automatically select the target, thereby eliminating the need for manual selection.

In this paper, we propose an ML model that uses Bayesian statistics to predict user selection intentions. The model calculates selection probabilities using the Bayesian approach, offering a sophisticated statistical framework for understanding user behavior. The method uses the Bayesian approach to transform gaze-related features into likelihood vectors, providing a rational basis for predicting selection intentions (see Fig. 1, Probabilistic Featurization). Consequently, the model's ability to predict user selection intentions improves. Additionally, the model enables real-time inference and gaze-based, manual-free interaction. A gaze-based, manual-free interaction technique was designed, and a user study was conducted to quantitatively compare it with traditional manual techniques. This highlights its strengths. The manual-free technique exhibited faster selection speeds than dwell and comparable selection speeds to controller- and hand-gesture-based methods. Furthermore, the study revealed that the proposed technique had lower workload and physical demand scores than conventional manual methods.

In summary, this study's contributions are as follows:

- *Method for predicting selection intention in real-time using only gaze data:* We propose a probabilistic feature extraction method that enables high-accuracy prediction of selection intention. We also constructed a high-performance, lightweight ML model.
- *Insights on a gaze-based manual-free technique:* Building on our high-performance intention prediction model, we designed a manual-free technique that eliminates the need for explicit gestures. Through a user study, we quantitatively compared this technique with established manual techniques, analyzed its strengths, and discussed potential applications.

II. Related Works

A. Pointing Technique

Natural and efficient pointing and selection techniques are essential for providing optimal XR user experiences. The pointing task, which precedes selection, requires a technique that enables precise and fast targeting with minimal physical exertion [5], [27]. The HCI community has developed methods to evaluate pointing performance based on Fitts' law [28], [29] and has proposed various pointing techniques applicable in 2D screens and 3D environments [5], [12], [27], [29]. The controller-based pointing method is one of the most widely used techniques in XR environments. Additionally, hand-tracking pointing techniques have been proposed [1], [30], [31]. However, these pointing methods require significant physical effort during prolonged use [1], [5], a phenomenon referred to as the "gorilla arm" effect [2]. Recently, gaze-based pointing techniques have gained attention for their ability to facilitate fast and natural pointing movements with minimal physical effort [3], [5], [32], [33].

B. Manual Selection Technique With Gaze-Based Pointing

Although gaze offers superior pointing performance, unlike controllers or hands, it has not been widely adopted as a selection technique. This is largely due to the "*Midas touch problem*," where unintended selections occur when gaze is used as a selection [8].

To prevent this, researchers have proposed interaction techniques that rely on manual confirmation via explicit gestures. One approach is multimodal gaze interaction, which expresses selection intention using other modalities, such as hand-held controllers [9], and hand gestures [10], [12], [13], [34]. The most common technique is Gaze + Pinch, enabled by on-device hand-gesture tracking. It has been adopted as the default interaction technique on recent commercial XR HMDs (e.g., Apple Vision Pro). Another approach uses explicit eye movement gestures alone, such as dwell [14], [15], [23], [35], eye-blink [36], [37], and smooth pursuit [38]. These approaches have been found to effectively address the Midas Touch issue by requiring an explicit act to confirm selection. However, each modality has been observed to have distinct advantages and drawbacks. A comparative study by Mutasim et al. [23] reported that controller- and pinch-based multimodal methods result in faster and more natural selections. Conversely, dwell typically involves

a trade-off between speed and accuracy compared to controller- and pinch-based methods. Nevertheless, the necessity of an explicit gesture introduces modality-specific limitations. Multimodal techniques may experience timing discrepancies between gaze and confirming actions [39] and are limited by physical and ergonomic constraints (e.g., occupied hands being or user impairments). Furthermore, dwell time requires users to wait, which disrupts continuous work. With prolonged use, dwell time can contribute to eye fatigue [23]. These limitations underscore the need for approaches that simplify or eliminate explicit gestures.

C. Predicting User Intention With Gaze Dynamics

Recent work has investigated the prediction of implicit user intentions for interaction in XR [40]. These approaches infer intention from physiological and behavioral signals, including eye-gaze dynamics and fixation patterns [14], [15], [24], pupil responses [14], and brain signals such as EEG [25], [26], [41]. By leveraging these implicit signals, researchers have proposed methods that simplify or even eliminate the need for explicit gestures and correct erroneous inputs by aligning selections with the inferred intention. David-John et al. [24] have confirmed through their research that various gaze-related characteristics derived from the analysis of gaze data can predict selection intentions within XR environments. Furthermore, Narkar et al. [15] have demonstrated that using time-series gaze data and features as inputs in a long short-term (LSTM) memory model enables the real-time prediction of user selection intentions with high accuracy. They have also proposed a dwell-based technique that adjusts the dwell threshold adaptively based on predicted intentions, thereby simplifying the explicit gesture.

However, most prior systems either do not demonstrate real-time operation [24], [25], limiting their viability as practical interaction techniques, or retain a reliance on an explicit confirming gesture [14], [15]. While some studies have investigated fully manual-free interaction, quantitative comparisons with established manual techniques are rarely reported. Thus, the advantages of manual-free interaction remain insufficiently studied.

III. Method

A. Feature Selection

Previous studies have employed a strategy of selecting eye-gaze features and then feeding them into models to determine selection intention or attentional state. Hwang et al. [42] defined zone-in/out states, preprocessed gaze indices, and performed t-tests (and alternative tests when necessary) to use only statistically significant features in the classifier. David-John et al. [24] and Narkar et al. [15] also constructed event-based features (e.g., fixation and saccade) and continuous dynamic features (e.g., velocity and acceleration). Then, they selected a subset of features whose significance was verified to reduce complexity and overfitting. Our study follows this same principle, selecting features based on the independent-samples t-test as the key criterion, similar to Hwang et al. [42].

This study aims to identify the most effective observations for explaining selection intention, which is defined by two classes: click versus non-click. To this end, we first constructed a pool of candidate features from gaze signals. This set consists of event-based features based on fixation/saccade detection, including fixation duration, saccade amplitude/peak velocity, dispersion, and the K coefficient, which captures ambient-focal characteristics of gaze strategies [43]. These event-based features are defined based on the results of the I-VDT (Velocity and Dispersion Threshold Identification) algorithm. We used the event window employed by Bednarik et al. [44].

Class labeling was based on whether or not selection confirmation occurred. More specifically, when the system confirmed a selection while in the fixation state, the features belonging to that fixation interval were labeled as a click (class $= 1$). If the fixation ended without a selection, however, the features were labeled as a non-click (class $= 0$). Before the analysis, outliers were identified and removed using the interquartile range (IQR) for each combination of condition (target configuration and task type, which will be described later), feature, and class. The normality assumptions were then diagnosed for each condition.

Feature selection proceeded by performing independent-samples t-tests with class (1/0) as the factor for each condition and feature. We set the significance level to $p < 0.05$, and selected features that were statistically significant. Only the selected features were used as observations and subsequently fed to the intention classifier.

B. Bayes Score and SVM Inference Pipeline

This subsection describes a pipeline that begins with Bayes' rule to infer the user's intention, $Y \in \{0, 1\}$, which is a binary variable, from the posterior probability conditioned on the observation $\mathbf{x} = (x_1, \ldots, x_M)$. The pipeline then normalizes the score to Bayes' factor and log-likelihoods (LLs) in accordance with the binary label structure. Next, it approximates joint interactions with linear/kernel decision functions, and learns the coefficients with a support vector machine (SVM).

Our goal is to probabilistically infer the user's selection intention, Y (e.g., 1 for select and 0 for non-select), from the observation, $\mathbf{x} = (x_1, \ldots, x_M)$. According to Bayes' rule, the posterior probability $p(Y = y|x)$ can be written as follows:

$$p(Y = y|\mathbf{x}) = \frac{p(\mathbf{x}|Y = y)p(Y = y)}{p(\mathbf{x})}. \tag{1}$$

Since the possible values of Y are 0 or 1, we define the Bayes score, $s(\mathbf{x})$, using the posterior log-odds as follows:

$$\begin{aligned} s(\mathbf{x}) &= \log \frac{p(y = 1|\mathbf{x})}{p(y = 0|\mathbf{x})} \\ &= \log \frac{\pi_1}{\pi_0} + \log p(\mathbf{x}|y = 1) - \log p(\mathbf{x}|y = 0). \end{aligned} \tag{2}$$

In this case, a score of 0 or higher corresponds to $y = 1$, and scores below 0 correspond to $y = 0$. Here, π_i denotes the prior probability, and $\log p(\mathbf{x}|y = 1) - \log p(\mathbf{x}|y = 0))$ is the joint log-likelihood (joint LL), i.e., the log-odds of the class-conditional probability for the entire observation. The individual

LL for each feature is given by:

$$LL_i(x_i, y) = \log p(x_i|Y = y). \tag{3}$$

Assuming all features of the observation are independent, the ratio of joint densities factorizes into a product, yielding:

$$\begin{aligned}
&\log \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} \\
&= \sum_{i}^{M} \log p(x_i|y=1) - \sum_{i}^{M} \log p(x_i|y=0) \\
&= \sum_{i}^{2M} c_i LL_i(x_i, y_i),
\end{aligned} \tag{4}$$

where i is less than or equal to M, c_i and y_i are both 1. When i is greater than M, c_i is -1 and y_i is 0. Therefore,

$$s(\mathbf{x}) = \log \frac{\pi_1}{\pi_0} + \sum_{i}^{2M} c_i LL_i(x_i, y_i), \tag{5}$$

which allows a naive Bayes classifier. However, real gaze features usually have correlations and nonlinear interactions [12]. To account for this, we express the joint LL with a correction term $\kappa(\mathbf{x})$ as follows:

$$\log p(\mathbf{x}|y=1) - \log p(\mathbf{x}|y=0) = \sum_{i}^{2M} c_i LL_i(x_i, y_i) + \kappa(\mathbf{x}), \tag{6}$$

where $\sum c_i LL_i$ represents the main effects of each feature, and the term $\kappa(\mathbf{x})$ accommodates interaction effects (joint effects) among features.

When interaction effects are weak, around a representative point, $\bar{\mathbf{z}}$, where the data are concentrated (where $\mathbf{z}(\mathbf{x}) = [LL_1(x_1), \ldots, LL_{2M}(x_{2M})]$), a first-order Taylor expansion allows for a linear approximation:

$$\begin{aligned}
s(\mathbf{x}) &\approx s_{lin}(\mathbf{x}) \\
&= \log \frac{\pi_1}{\pi_0} + b + \mathbf{w}^\intercal \mathbf{z}(\mathbf{x}),
\end{aligned} \tag{7}$$

where $\mathbf{w}$ and b are learned from the data, and $b + \mathbf{w}^\intercal \mathbf{z}(\mathbf{x})$ plays the role of a first-order approximation to $\sum c_i LL_i + \kappa(\mathbf{x})$.

In contrast, when interaction effects are strong and nonlinear, higher-order interactions must be considered. In this case, we use the kernel trick with a Gaussian kernel. With the kernel function defined as $K(\bar{\mathbf{z}}(\mathbf{x}), \mathbf{z}(\mathbf{x})) = e^{-\gamma||\bar{\mathbf{z}}-\mathbf{z}||^2}$, the Representer Theorem enables us to express the interaction effects as a linear combination of basis functions. This results in:

$$\begin{aligned}
s(\mathbf{x}) &\approx s_{ker}(\mathbf{x}) \\
&= \log \frac{\pi_1}{\pi_0} + b + \sum_{i} \alpha_j K(\bar{\mathbf{z}}(\mathbf{x}), \mathbf{z}(\mathbf{x})),
\end{aligned} \tag{8}$$

where the nonlinear kernel, K, flexibly absorbs joint interactions.

Accordingly, since the Bayes score $s(\mathbf{x})$ becomes similar to the SVM problem, we trained the coefficients of the decision functions ($(\mathbf{w}, b)$ or (α, b)) with SVM. This ensures that $s_{lin}(\mathbf{x})$ and $s_{ker}(\mathbf{x})$ make decisions similar to the true $s(\mathbf{x})$. During inference, features extracted in real time are transformed into LLs to construct $\mathbf{z}(\mathbf{x})$, and the sign of the learned $s(\mathbf{z})$ determines click/non-click.

C. Observation to LL Conversion

The objective of this stage is to transform the observation vector $\mathbf{x} = (x_1, \ldots, x_M)$ into an LL vector $\mathbf{z}(\mathbf{x}) = [LL_1(x_1), \ldots, LL_{2M}(x_{2M})]$ that can be directly fed into the Bayes score $s(\mathbf{x})$ defined in Section III-B. To this end, for class $y \in \{0, 1\}$ and task context c, we fit the class-conditional pdf $p(x|Y = y, C)$ for each feature x_i and compute the individual LLs using the fitted distributions. Since distributional properties can vary with task context (e.g., target configuration or task difficulty), all fittings were performed per condition and per class.

We employed maximum likelihood estimation (MLE) for distribution fitting, and the candidate family consisted of 15 representative PDFs, including Gaussian, Weibull, and log-normal distributions. For each (x_i, y, C) combination, we estimated the candidate distributions using MLE and evaluated the goodness-of-fit using the Kolmogorov–Smirnov (KS) test. In this study, we considered a distribution fitted only when the KS p-value was 0.05 or less. When multiple candidates met this criterion for the same combination, we selected the distribution with the smaller p-value as the final model. Using the selected $p(x_i|Y = y, C)$, the individual LL at the online inference time under condition C is computed as:

$$LL_i(x_i, y_i, C) = \log p(x_i|Y = y_i, C). \tag{9}$$

Computing LL_i for all dimensions yields $\mathbf{z}(\mathbf{x})$, which is then fed to the decision function $s(\mathbf{x})$ (with linear or kernel approximation), as defined in Section III-B, to determine the final intention (click versus non-click).

IV. Study 1: Construction and Evaluation of the Selection Intention Prediction Model

The primary goal of our study is to construct a real-time selection intention prediction model using only gaze data. Accordingly, this section describes three stages: 1) collecting gaze data when selection intention emerges during a target acquisition task, 2) statistical feature selection and linear transform (LLT), and 3) training and evaluating a model. Finally, we select the model to be used for a manual-free selection technique.

A. Data Collection

1) Participants and Apparatus: We recruited 20 participants (M = 24.35, SD = 2.03, 11 male, 9 female). All had normal or corrected-to-normal vision and did not wear glasses. People with color-vision deficiency were also excluded. We implemented the VR environment in Unity 2022.3.16f1 and ran it on a Meta Quest Pro HMD. We recorded eye-tracking data at 60–66 Hz.

2) Task Design: We implemented a modified version of the target acquisition task used in previous research [45], [46] (see Figs. 2 and 3). In this task, spheres were arranged in a 9×7 grid.

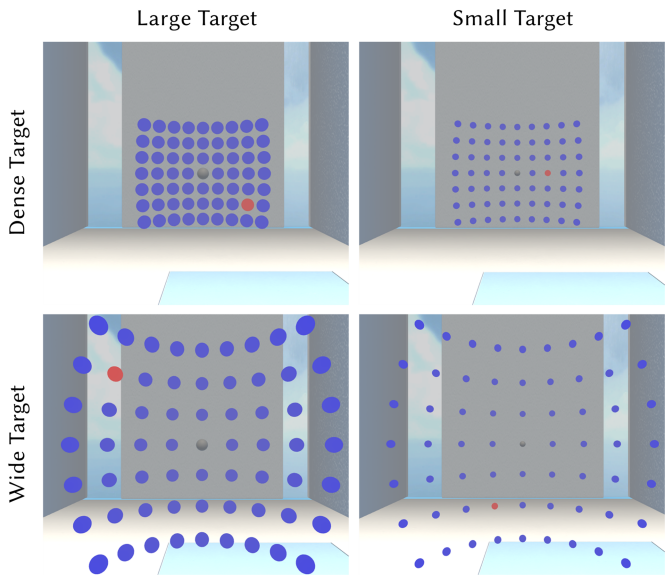

Fig. 2. We created four target configurations by combining two factors: target size (*Large* or *Small*) and target density (*Wide* or *Dense*).

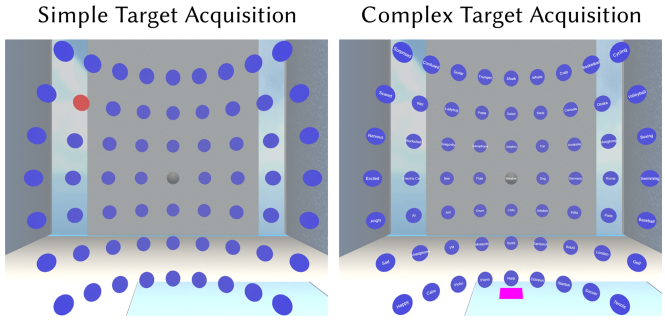

Fig. 3. We categorized acquisition complexity into two levels based on the difficulty of identifying the target. In the *Simple* acquisition task (left), a red target object was placed among blue distractors, and participants were instructed to click on the target. In the *Complex* acquisition task (right), each distractor had a random word written on it. Participants had to find and click the target object that had a word belonging to the same category as the word written on the starting object.

A central start object (colored gray) was surrounded by distractors (colored blue). The objects located between the outermost layer and the starting object served as potential target objects (26 in total). These distractors were randomly activated as task targets. Four configurations were created by varying two levels of target density (*Dense* and *Wide*) and two target sizes (*Small* and *Large*) (see Fig. 2). In the *Dense* target condition, objects were spaced 7.5 degrees apart. In the *Wide* condition, the spacing was 15 degrees. The *Small* and *Large* target widths were 3 and 6 degrees, respectively.

Two different task types were designed to present a variety of task difficulties. The first is a simple acquisition task (or *Simple*), which is similar to the task used in previous studies [45], [46]. When the user points to the starting object, one of the potential target objects is randomly activated and turns red. After accurately pointing to and confirming the selection of the target object, it turns blue again, accompanied by auditory feedback. The user should then reset by pointing to the starting object again, which activates a new target object. The task is completed after interacting with all potential target objects.

The second task was a complex acquisition task (or *Complex*) in which a category word was displayed on the starting object. The potential target objects were still colored blue and contained words belonging to that category. The distractor objects showed unrelated words (see Fig. 3). At the start of the task, participants are asked to locate and select the object containing the correct category word. After accurately pointing to the object and confirming their selection, the target object turns red and a sound feedback is played. After pointing again at the starting object again, a new category word appears, and the words on all objects are shuffled. Then, one potential target is randomly selected as the new target. The task ends when all potential targets have been interacted with. The words used in the *Complex* task were chosen from the language commonly used by the participants.

3) Study Design and Procedure: The data collection study followed a $2 \times 2 \times 2$ within-subjects factorial design: *Target Density* (*Dense* and *Wide*) $\times$ *Target Size* (*Small* and *Large*) $\times$ *Task Complexity* (*Simple* and *Complex*). Participants performed the task in eight different conditions, completing two tasks per condition, with each task consisting of 26 trials. In all conditions, gaze pointing was used as the pointing technique, and selection was confirmed by clicking the trigger button on the controller.

Before the experiment began, participants completed a demographic questionnaire. Then, they received instructions on the tasks, the VR HMD, and the controller use, followed by an explanation of the eye-tracking calibration protocol. After calibration, participants underwent a practice session in which they completed five trials for each of the eight conditions to familiarize themselves with the task. Then, they participated in the main session, in which the order of the conditions was counterbalanced using a Latin square [47]. After each task, participants completed a one-question survey about discomfort on a 10-point Likert scale to assess their level of fatigue [48]. If a score of 7 or higher was reported, a rest period was provided. If a score of 10 was reported, the experiment was terminated. Additionally, the experiment was terminated if the total duration exceeded 90 minutes, in order to prevent the degradation of the quality of the gaze data due to accumulated fatigue. Participants who completed the entire study finished 416 trials. The study was conducted under institutional IRB approval.

4) Eye Tracking Data Pipeline: This subsection explains how the eye-gaze data collected in Section IV-A were processed and recorded. Our data pipeline converts eye and head tracking data into gaze direction data and computes gaze-related features when a fixation event is detected.

First, we collected eye (eye-in-head frame) direction data and head (head-in-world frame) direction data using the HMD's tracking function. Then, we computed gaze direction data (eye-in-world frame) by quaternion multiplication. Next, the head and gaze direction data are converted to angular units to compute angular dispersion and velocity for each frame. The I-VDT algorithm uses these computed dispersion and velocity values to detect fixation and saccade events [24], [49]. A velocity greater than 70 degrees/second is classified as a saccade, while events with a velocity less than 30 degrees/second, a dispersion of less

than one degree, and a duration of at least 100 ms are classified as fixations.

The I-VDT algorithm detects fixation events and computes gaze-related features within a data window, as shown in Fig. 1 [44]. When a fixation is detected, features are computed for the previous fixation (first fixation) and the preceding saccade. Features are also computed for the currently detected fixation (second fixation). When the fixation ends, the calculation of gaze-related features stops. The list of computed features can be found in the Supplementary Material.

B. Feature Selection and LL Transform

1) Feature Selection: For feature selection, we first split the dataset by context (8 types) and conducted statistical tests for each gaze-related feature by class (two types: click or non-click). Class labeling was based on whether manual selection was confirmed. If the selection was confirmed during a fixation interval, the features from that interval were labeled as class 1 (click); if the fixation ended without a selection, the features were labeled as class 0 (non-click). We then removed outliers using IQR. After confirming the normality assumption, we conducted independent-samples t-tests similar to prior work. Only features that were significant ($p < 0.05$) in at least 6 of the 8 contexts were selected. A summary of the selected features and detailed statistical analyses are provided in the Supplementary Material.

2) LL Transformation With Context and Class-Wise PDF: This subsection explains how observations are transformed into LL. For each dimension, x_i of the selected observation, $\mathbf{x} = (x_1, \ldots, x_M)$, we fit the class-conditional likelihoods, $p(x_i|Y = y_i, C)$, by class, $y \in \{0, 1\}$, and task context, C, using maximum likelihood estimation. We considered candidate distributions to include 15 pdfs, such as Gaussian, Weibull, and Log-Normal. We considered a fit to be valid only when the Kolmogorov-Smirnov (KS) test yielded $p < 0.05$. If multiple candidates satisfied the criterion, we chose the one with the smaller p-value). Next, we computed $LL_i(x_i, y_i) = \log p(x_i|Y = y_i, C)$ and constructed the LL vector, $\mathbf{z}(\mathbf{x}) = [LL_1, \ldots, LL_{2M}]$. Detailed fitting outcomes and KS test results are available in the Supplementary Material.

C. Training and Evaluating ML Model

This subsection describes how we trained and evaluated the prediction model. Due to class imbalance (e.g., approximately 1:4 in Simple and 1:80 in Complex), we re-sampled to a ratio of 1:3 and applied a train/validation/test split of 6:2:2. We used a Gaussian kernel-based SVM and optimized the hyperparameters gamma and C with a grid search. To prevent overfitting, we conducted 5-fold cross-validation. We report the following metrics: accuracy, F1-score, and AUC-ROC.

1) All-Context versus Context-Specific Model: We first fixed the LL input and compared an All-context model, which was trained on all 8 contexts combined, with Context-specific models, which were trained separately for each of the 8 contexts. We evaluated both the All-context and the Context-specific models on each individual context dataset. For a given context, the Context-specific model exhibited slightly superior performance (see Table I). We attribute this to the advantage of learning the interaction effect from data restricted to a specific context. In contrast, the All-context model must accommodate heterogeneous distributions across contexts. Additionally, the Complex task models showed higher classification performance than the Simple task models. The inference time for all models was less than 1 ms.

2) Ablation Study: We verified the contribution of input representations for the Complex × Small × Wide condition through an ablation study with the following conditions:

- *All & Obs:* All raw observations without feature selection
- *Selected & Obs:* All raw observations only with selected features
- *Selected & LLs:* LLs only with selected features

The metrics for each condition are reported in Table II. The evaluation results show that the Selected Features condition outperformed the All Features condition, aligning with prior research [42]. Furthermore, the highest performance was achieved by using the proposed LL transform.

V. Study 2: Evaluation of the Gaze-Based Interaction Techniques

The objective of Study 2 was to evaluate the performance and user experience of the gaze-based Manual-free selection technique in realistic interaction contexts and to quantitatively compare it with existing manual selection techniques. Our objective was to provide insights that underscore the potential of manualfree selection. To minimize the Midas touch effect, we first selected the two environments with the highest F1 scores identified in Study 1. These corresponded to the two configurations (Small × Dense, Small × Wide) with the complex target-acquisition scenario adopted from Study 1. Manualfree then employed the intentionprediction pipeline established in Study 1 and operated with a contextspecific model tailored to the selected configurations. Thus, the study was designed to minimize unnecessary automatic selections by Manual-free while clearly demonstrating its strengths in representative usage scenarios.

A. User Study

1) Participants: We recruited 25 participants (M = 23.36, SD = 2.21, Male = 10, Female = 15) for Study 2. None of the users had participated in Study 1, and all had normal or corrected-to-normal vision. Of those with corrected-to-normal vision, only those using non-glasses corrective devices were eligible. Individuals with color blindness were also excluded from participation. 16 of the participants reported previous experience with VR environment and HMDs. We used the same VR environment and HMD as in Study 1.

2) Task and Interaction Techniques: The task implemented in this user study was the *Complex* target acquisition task used in Study 1 (see Fig. 3. Participants had to read the category word displayed on the starting object and find the target object labeled with a word belonging to that category among the distractors. In Study 2, only two types of target configurations were used: *Small* × *Dense* and *Small* × *Wide*.

TABLE I
PERFORMANCE COMPARISON BETWEEN ALL-CONTEXT AND CONTEXT-SPECIFIC MODELS

Task Type	Target Cofiguration	All-Context Model			Context-Specific Model		
		Accuracy	F1	AUC-ROC	Accuracy	F1	AUC-ROC
Simple	Large × Dense	0.744	0.565	0.751	0.756	0.584	0.787
	Large × Wide	0.808	0.514	0.814	0.827	0.607	0.816
	Small × Dense	0.753	0.594	0.769	0.769	0.598	0.792
	Small × Wide	0.794	0.584	0.788	0.781	0.552	0.817
Complex	Large × Dense	0.928	0.882	0.967	0.929	0.883	0.969
	Large × Wide	0.933	0.875	0.977	0.952	0.913	0.982
	Small × Dense	0.927	0.884	0.981	0.947	0.921	0.976
	Small × Wide	0.943	0.904	0.953	0.972	0.957	0.987

TABLE II
COMPARISON OF MODEL PERFORMANCE BASED ON FEATURE SELECTION AND LL CONVERSION AT COMPLEX × SMALL × WIDE CONDITION

	Accuracy	F1	AUC-ROC
All & Obs	0.751	0.321	0.649
Selected & Obs	0.838	0.455	0.685
Selected & LLs	**0.972**	**0.957**	**0.987**

Four interaction techniques were used, each with different pointing and selection methods as follows:

1) *Cont:* Gaze-based pointing and controller-based selection
2) *Hand:* Gaze-based pointing and hand gesture-based selection
3) *Dwell:* Gaze-based pointing and dwell-based selection
4) *Manual-free* (Ours): Gaze-based pointing and selection using an intention prediction model

Gaze-based pointing uses gaze direction data, computed using a raycasting method, to identify the object at which the user is pointing.

The selection mechanism for the *Manual-free* condition relied on gaze data processed through the pipeline. The computed observation data was then fed into the Bayesian-based ML model. Based on the model's inference, the system automatically decided whether to select the pointed object. For the other three manual selection methods, the initial option is the *Dwell* condition. In this condition, an object is selected when the user holds a fixed gaze on it for over 600 ms. In the *Cont* and *Hand* conditions, the pointed object was selected when the trigger button on the handheld controller was pressed and a pinch gesture was observed. The manual-free technique takes up to 12 ms from the time it takes to infer intention from eye-tracking data to the time it takes to perform the selection.

3) Study Design and Procedure: The user study followed a 2×4 within-subjects factorial design: *Target Configuration* (*Small* × *Dense* and *Small* × *Wide*) by *Interaction Technique* (*Cont*, *Hand*, *Dwell* and *Manual-free*). Participants performed the Complex target acquisition task under eight different conditions, completing one task per condition. Each task consisted of 26 selection trials.

Prior to the start of the experiment, participants completed a demographic questionnaire. Then, they were then instructed on the task to be performed and how to use the VR HMD and controller. Instructions for the eye tracking calibration protocol were also provided. After completing the calibration, participants underwent a practice session in which they experienced five trials of each of the eight condition to familiarize themselves with the task. This was followed by the main session, in which the order of conditions was counterbalanced using a counterbalanced Latin square design [47].

After each task, participants completed a discomfort survey [48], the System Usability Scale (SUS) [50], and the NASA-TLX [51] survey. Participants were given a break if they rated discomfort as 7 or higher, and the experiment was terminated if they rated discomfort as 10. The study was IRB approved.

B. Results

We conducted a statistical analysis to examine the effects of the *Target Configuration* and *Interaction Technique* factors. First, we assessed the normality of the data using the Shapiro-Wilk test. If the data met the normality assumption, we proceeded with a two-way repeated measures analysis of variance (RM ANOVA) to test for interaction effects and main effects. Additionally, RM-ANOVAs conducted separately for the *Small* × *Dense* and *Small* × *Wide* configurations to observe the effect of the *Interaction Technique* factor.

For all RM-ANOVA analyses, if Mauchly's test indicated a violation of the sphericity assumption, we adjusted the degrees of freedom with Greenhouse-Geisse correction. If the data violated the normality assumption, we used the Friedman test instead. When a significant interaction or main effect was observed by RM-ANOVA or the Friedman test, post-hoc pairwise comparisons were performed using Bonferroni correction for multiple comparisons. The following section highlights the results where statistical significance was found; detailed statistical analysis results can be found in the Supplementary Material.

1) Time to Completion (TTC): TTC was measured as the time from when the participant pointed to the starting object until they selected target object. The mean TTC for each task was calculated by averaging the TTCs across the 26 trials. A smaller TTC indicates faster target acquisition.

The two-way RM ANOVA revealed no significant interaction effect. However, main effects were observed for both the *Target Configuration* and *Interaction Technique* factors (see Fig. 4 left). To further investigate, a one-way RM ANOVA was performed for each configuration type. In the *Small* × *Dense* condition, the *Interaction Technique* factor significantly influenced the Time to Completion (TTC). Post-hoc tests showed that the *Cont*, *Hand*, and *Manual-free* conditions resulted in significantly shorter TTC than the *Dwell* condition. Similarly, for the *Small* × *Wide* condition, significant differences in TTC were also found among *Interaction Technique*, with the *Dwell* condition

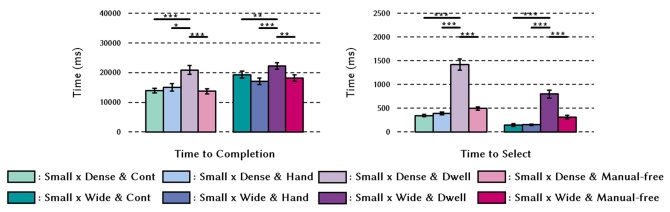

Fig. 4. TTC (Left) and TTS (Right) based on *Target Configuration* and *Interaction Technique*. Statistically significant differences identified through post-hoc tests are indicated by connecting lines, where ∗, ∗∗, and ∗∗∗ represent p-values less than 0.05, 0.01, and 0.001, respectively.

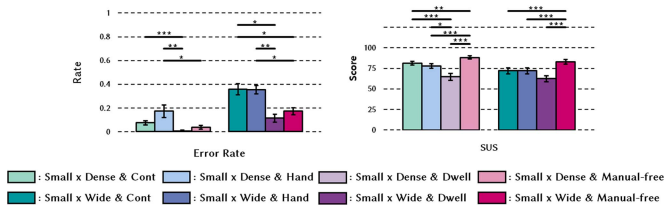

Fig. 5. The error rate (Left) and the SUS survey score (Right), which was collected through interviews, based on *Target Configuration* and *Interaction Technique*. Statistically significant differences identified through post-hoc tests are indicated by connecting lines, where ∗, ∗∗, and ∗∗∗ represent p-values less than 0.05, 0.01, and 0.001, respectively.

exhibiting the longest TTC. These findings demonstrate that the *Dwell* condition consistently results in the slowest target acquisition time across all configurations.

2) Time to Select (TTS): TTS refers to the time between the participant's perception of a target and the completion of the selection by clicking. Target perception is defined as the moment when the participant fixates their gaze on the target prior to selection confirmation. A shorter TTS indicates faster selection confirmation.

The analysis using a two-way RM ANOVA revealed significant interaction effects as well as main effects of the two factors on TTS (see Fig. 4 right). Additionally, one-way ANOVA for each configuration type showed significant differences in TTS based on the *Interaction Technique* factor. Post-hoc tests for the *Small* × *Dense* indicated that the *Dwell* condition had significantly longer TTS compared to the *Cont*, *Hand*, and *Manual-free* conditions. Similarly, in the *Small* × *Wide* condition, the *Dwell* condition also exhibited significantly longer TTS compared to the other conditions. However, no statistically significant differences were observed between the remaining conditions. These results suggest that the *Dwell* condition consistently leads to the longest selection times.

3) Error Rate: The error rate was calculated by dividing the number of unsuccessful clicks by the total number of clicks during the task. A higher error rate indicates that the participants had more difficulty making selections due to configuration or technique.

The analysis of error rate using a two-way RM ANOVA revealed significant interaction effects and main effects of both factors (see Fig. 5 left). Furthermore, one-way RM ANOVA for each configuration type indicated that the *Interaction Technique* factor significantly influenced the error rate. In the *Small* × *Dense* condition, post-hoc tests showed that the *Hand* condition had a significantly higher error rate compared to the *Dwell* and *Manual-free* conditions. For the *Small* × *Wide* condition, the *Cont* condition exhibited a higher error rate than the *Dwell* and *Manual-free* conditions, while the *Hand* condition also showed a higher error rate compared to the *Dwell* and *Manual-free* conditions. These findings suggest that the *Hand* condition consistently led to more frequent error trials compared to the gaze-based unimodal methods, such as *Dwell* and *Manual-free*, across all configurations. Moreover, in the *Small* × *Wide* configuration, the *Cont* condition also resulted in more errors compared to the *Dwell* and *Manual-free* conditions.

4) System Usability Scale (SUS): The SUS survey scores, which reflect the usability of the system, indicate that higher scores correspond to better usability.

The two-way RM ANOVA analysis revealed significant main effects of both the *Interaction Technique* and *Target Configuration* factors (see Fig. 5 right). Furthermore, one-way ANOVA for each configuration demonstrated that the SUS scores differed significantly based on the *Interaction Technique*. In the *Small* × *Dense* configuration, post-hoc comparisons indicated that the *Manual-free* condition yielded significantly higher SUS scores than all other conditions, whereas the Dwell condition scored significantly lower than the *Cont* and *Hand* conditions. In the *Small* × *Wide* configuration, the *Manual-free* condition again achieved significantly higher SUS scores than all other conditions. These results indicate that the *Manual-free* condition provided superior usability irrespective of the configuration.

5) NASA-TLX: Perceived workload was assessed using the total score of the NASA TLX questionnaire, with a lower score indicating less workload. Furthermore, analyses were conducted for the six sub-dimensions of NASA-TLX.

Overall Score: A two-way RM ANOVA revealed significant main effects for each factor; however, no interaction effect was observed (see Fig. 6 Overall). Specifically, the *Interaction Technique* factor demonstrated statistically significant differences in overall scores for each configuration. Further post-hoc analysis showed that the *Manual-free* condition produced significantly lower overall scores compared to all other conditions. This result indicates that, regardless of the configuration, participants perceived the workload to be the lowest when using the *Manual-free* technique.

Six Sub-dimensions: Results of the two-way RM ANOVA showed no interaction effects in all sub-dimensions. However, significant main effects of the *Interaction Technique* were observed in all sub-dimensions. Meanwhile, the configuration factor exhibited significant main effects in the Mental, Physical, and Effort sub-dimensions.

VI. Discussion

In Study 1, we evaluated a Bayesian approach that predicts selection intention using only gaze data. Furthermore, in Study 2, we designed and evaluated a manual-free technique that does not require explicit gestures. Based on the results of the two studies, our contributions are summarized as follows:

1) We propose a pipeline that predicts selection intention in real time and with high accuracy from gaze data, and we build the selection intention prediction model.

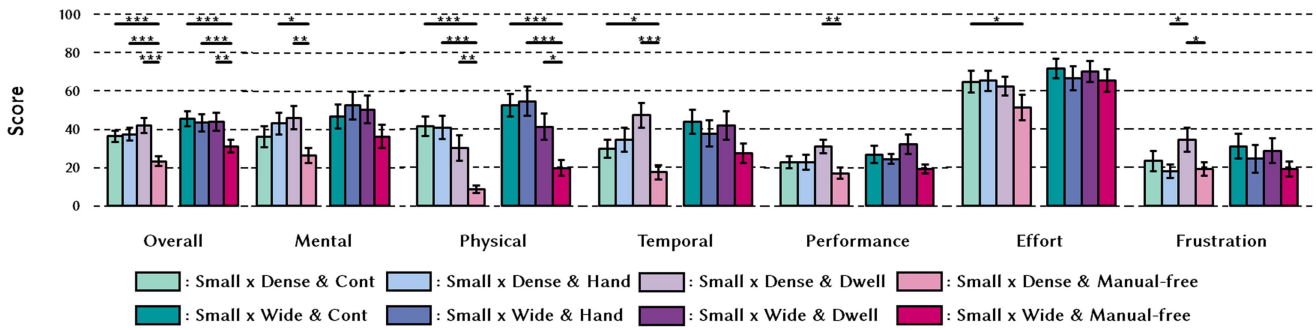

Fig. 6. Overall and 6 sub-dimension NASA-TLX scores based on *Target Configuration* and *Interaction Technique*. Statistically significant differences identified through post-hoc tests are indicated with connecting lines, where $*$, $**$, and $***$ represent p-values less than 0.05, 0.01, and 0.001, respectively.

2) We designed a *Manual-free* technique driven by the predicted intention and quantitatively evaluated its advantages over existing manual techniques.
3) We discuss how the proposed method and technique can be applied to the HCI and XR domains.

A. Bayesian Approach to Predicting Selection Intention

In this study, inspired by a Bayesian approach that successfully addressed problems in 2D screen touch [52], [53] and 3D target acquisition [12], [33], [46], [54], [55], we applied probabilistic featureization to convert gaze-related features into LL vector. We hypothesized that this LL vector would be more advantageous for ML models in predicting selection intention than traditional observation data formats, and Study 1 confirmed this hypothesis. The prediction performance was significantly higher with the LL vector compared to the raw data (observation data). This underscores the importance of modeling the intended decision and featurizing observations accordingly.

Building on prior work that used statistical feature selection, we selected only those gaze features that showed significant differences by intention for each task condition. Specifically, we used class-wise t-tests and included as inputs only the features that satisfied $p < 0.05$ [42]. Study 1 showed that this contributed to improved ML predictive performance, aligning with the prior literature. We attribute this to reduced redundancy from correlated features, thereby lowering model complexity and interaction effects ($\kappa(\mathbf{x})$).

Our model fits class-conditional distributions separately for each condition, because the distribution of gaze features varies with task context. For example, even the same gaze-related feature can shift with the target configuration or task difficulty [43]. Accordingly, we estimate distributions by maximum likelihood for each condition–class pair and compute LL. Comparing context-specific models with an all-context model showed higher performance for the context-specific variants. This suggests that mixing heterogeneous contexts makes it harder for the SVM to establish a stable decision boundary. Through this pipeline, we were able to build a high-performance intention prediction model, which in turn enabled the design of a highly accurate manual-free interaction technique.

B. Improving Interaction Experience in 3D Target Selection

Analyzing the results of Study 2, the *Manual-free* technique showed strengths across indicators and, depending on the condition, exhibited advantages over manual techniques in terms of speed, accuracy, usability and workload.

Participants demonstrated high selection accuracy when using *Dwell*, a traditional gaze-based manual selection technique, but provided feedback that their selection speed was slower than the other conditions and that they perceived it as inconvenient. Consistent with previous findings [23], [56], the *Dwell* technique resulted in less frequent error trials than *Hand* and *Cont*. However, TTC and TTS were significantly longer compared to multi-modal techniques and *Manual-free*. The characteristics of dwell-based techniques that require dwellings above a threshold were observed. However, the trade-off between high accuracy and slow performance was a drawback for system usability. Participants reported significantly lower SUS scores when using the dwell technique compared to the multimodal and *Manual-free* techniques, indicating that the system was the least user-friendly. The interview feedback reinforced these findings: P16 reported, "The slow click speed made it a bit of a pain to use," while P13 noted, "The delay in selection made it unclear whether the click was registered properly, which caused worry and stress."

Participants performed the task faster and reported higher usability scores when using the *Hand* technique, a multimodal technique, compared to *Dwell*. However, error trials were observed with a higher frequency compared to the other techniques. Specifically, the error rate was significantly higher compared to the *Dwell* and *Manual-free* techniques. Previous studies have reported that the *Hand* technique suffers from error trials due to the gesture being made too late or before the gaze is accurately fixated on the target [39], [57]. Such cases were also found in this study, such as P6's case, "Sometimes I kept holding the pinch gesture because it was annoying to keep my hand extended, and it was automatically clicked at this time," or P21's case, "Sometimes I couldn't click even if I pinched it, but if I think about it, it was when I already moved my gaze to find the next target."

Even when using the multimodal technique *Cont*, participants reported faster performance and higher usability scores compared to *Dwell*. In terms of error rate, it was higher than the *Dwell* in the *Small* $\times$ *Dense* configuration, and significantly

higher than the *Dwell* and *Manual-free* conditions in the *Small* $\times$ *Dense* configuration. However, the difference in familiarity compared to the *Hand* technique is revealed in the interview session, as P5's interview feedback "It felt like a mouse on the desktop, so it was more familiar and comfortable" and P12's response "Button clicks were more recognisable than hand gestures". However, for the *Small* $\times$ *Wide* configuration, which required more accurate selection, the disadvantages of the multimodal technique became apparent. Multimodal techniques, which require different modalities for pointing and selection, are more complex to use than gaze-based techniques, and appear to lead to significantly higher error rates in configurations requiring higher accuracy, as evidenced by P21's feedback that "Having to do it in two steps was more complicated than just using the eyes". This suggests that unimodal techniques are more appropriate in environments where accurate selection is required.

When using the *Manual-free* technique, participants demonstrated fast and accurate target selection and responded with a high usability score and low workload. The TTC and TTS were significantly shorter than the *Dwell* technique, while comparable to the two multimodal techniques, as no statistical significance was observed. In terms of error rates, the *Manual-free* technique outperformed the *Hand* condition in all configurations and the *Cont* condition in the *Small* $\times$ *Wide* configurations. Furthermore, the *Manual-free* technique achieved higher usability scores in all configurations. Regarding workload and physical demand, the *Manual-free* technique yielded significantly lower NASA-TLX scores in all configurations compared to the other techniques. P23 commented, "This (*Manual-free*) technique was convenient because it allowed accurate and quick clicks using only my eyes", and also P10 remarked, "*Dwell* technique caused fatigue because it required maintaining focus for a fixed period of time, which was not the case with natural clicking". This underlines the simplicity and intuitiveness of the *Manual-free* technique, which avoids the complexity of the two-step processes of multimodal techniques and the manual fixation required by *Dwell*. These advantages allow for faster and more accurate target acquisition.

In summary, the simplicity and low workload of the *Manual-free* technique makes it more suitable than multimodal techniques for environments that require precise interactions or prolonged use. In addition, its ability to support faster selection compared to unimodal dwelling suggests its effectiveness in high-frequency selection environments, such as text entry or menu selection tasks.

C. Applications of Manual-Free Technique

Leveraging the observed fast selection speed, low error rate, high usability, and low workload, the manualfree technique is well suited for tasks that involve frequent selections over long durations, such as text entry [33] and menu navigation [37], where it supports consecutive selections without the delay inherent to dwell and with lower workload than manual methods. These aspects are particularly relevant in situations where XR technologies are adopted, such as in clinical environments where hygiene is required, in manufacturing and maintenance tasks where both hands are occupied, in logistics and fieldwork, and in laboratory or educational settings where the conditions of the work inherently restrict the use of manual techniques for operating the XR interface. In these cases, enabling precise selection without using the hands makes the proposed approach particularly useful.

From an accessibility perspective, it lowers the entry barrier for older adults or people with upper-body impairments by enabling precise interaction using gaze alone. Wu et al. [21] investigated the challenges older adults face when selecting and manipulating 3D objects in VR and found that interaction accuracy decreases during manual selection, leading to negative user experiences and highlighting the need for alternative modalities. The selection technique proposed in this study could address these issues by providing a more accessible solution. In addition, Franz et al. [22] compared different locomotion techniques for individuals with upper limb motor impairments in VR environments and found that methods that did not require a controller were the most preferred. By integrating the proposed selection technique with teleportation, it could serve as a valuable locomotion tool for users with motor impairments.

D. Limitations and Future Work

While we confirm the potential of manual-free interaction, several limitations and directions for future work remain. First, the model exhibits context dependence. Applying it to a new context requires collecting data and retraining the model. In our experiments, we fit condition-specific distributions by task type and target configuration. However, we cannot assume that the resulting model generalizes to other tasks or backgrounds. That is, a model optimized for a particular context may not transfer well to another. In practice, additional data collection and model tuning will be needed for the target environment. To reduce deployment cost, future work should explore contextrobust general models or domain adaptation methods that adapt quickly with limited data. For example, mechanical simulation data [46] or deep generative density estimation [58] could be used to predict likelihood distributions for new tasks without extra experiments, thereby reducing retraining overhead.

Second, our current commit mechanism uses a single threshold, which can lead to false positives depending on the situation. The system regards $s(x) > 0$ as intent to select. This fixed threshold risks inadvertent clicks when a user briefly glances at an object. Although such events were not problematic in our study, more complex environments could suffer from inaccurate autoselections that harm user experience. A practical remedy is to employ more conservative or dynamic thresholds. For example, allowing configuration of the commit criterion (e.g., $s(x) > 0.5 or > 1$) depending on application requirements can balance fast response and falsepositive suppression and improve the reliability of manualfree interaction.

Third, the current intention prediction model is limited to binary classification ("select" versus "nonselect") and does not yet distinguish what action the user intends to perform next (e.g., click, drag, open or close a menu, scroll). Extending the LLbased approach to intended action classification would

enable the system to proactively suggest and guide appropriate interactions based on context. When the system can infer what the user intends to do, it can support layered tool switching and predictive assistance that naturally progress, enabling a more intelligent gazebased interaction design.

Lastly, our user study was conducted under relatively static and simplified conditions. We validated the core capability with fixed virtual objects and without visual distractors. Real XR applications will involve moving targets, dynamic backgrounds, and a wide variety of situational changes. Future work should therefore test the proposed technique in more realistic and complex scenarios. For example, selecting moving objects, operating amid visual clutter, and using AR headsets that expose the real world would help validate robustness and practical utility.

VII. Conclusion

In VR/AR environments, 3D target acquisition typically relies on interaction techniques that require a manual selection step. These methods often increase the physical and cognitive load, which hinders the experience of natural interaction. To address this issue, we propose the *Manual-free* technique that uses gaze data to predict the user's selection intention, enabling interaction without the need for manual selection. Specifically, a probabilistic featureization is used to transform gaze-related features into LL vectors, which are then used to predict user intention through a Bayesian-based ML model. Furthermore, in user studies, by comparing the *Manual-free* technique with traditional manual selection methods, it demonstrated a more accurate selection performance and provided a more comfortable interaction experience. Finally, based on our results, we discuss how interaction techniques that do not require manual selection can be applied in practical scenarios.

References

[1] S. Boring, M. Jurmu, and A. Butz, "Scroll, tilt or move it: Using mobile phones to continuously control pointers on large public displays," in *Proc. 21st Annu. Conf. Australian Comput.-Hum. Interaction Special Int. Group*, 2009, pp. 161–168.

[2] J. D. Hincapié-Ramos, X. Guo, P. Moghadasian, and P. Irani, "Consumed endurance: A metric to quantify arm fatigue of mid-air interactions," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2014, pp. 1063–1072.

[3] A. S. Fernandes, T. S. Murdison, and M. J. Proulx, "Leveling the playing field: A comparative reevaluation of unmodified eye tracking as an input and interaction modality for VR," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 5, pp. 2269–2279, May 2023.

[4] V. Tanriverdi and R. J. Jacob, "Interacting with eye movements in virtual environments," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2000, pp. 265–272.

[5] L. Sidenmark, F. Prummer, J. Newn, and H. Gellersen, "Comparing gaze, head and controller selection of dynamically revealed targets in head-mounted displays," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 11, pp. 4740–4750, Nov. 2023.

[6] R. Vertegaal, "A fitts law comparison of eye tracking and manual input in the selection of visual targets," in *Proc. 10th Int. Conf. Multimodal Interfaces*, 2008, pp. 241–248.

[7] L. E. Sibert and R. J. Jacob, "Evaluation of eye gaze interaction," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2000, pp. 281–288.

[8] R. J. Jacob, "What you look at is what you get: Eye movement-based interaction techniques," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 1990, pp. 11–18.

[9] M. Kytö, B. Ens, T. Piumsomboon, G. A. Lee, and M. Billinghurst, "Pinpointing: Precise head-and eye-based target selection for augmented reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–14.

[10] K. Pfeuffer, B. Mayer, D. Mardanbegi, and H. Gellersen, "Gaze pinch interaction in virtual reality," in *Proc. 5th Symp. Spatial User Interaction*, 2017, pp. 99–108.

[11] R. Shi, Y. Wei, X. Qin, P. Hui, and H.-N. Liang, "Exploring gaze-assisted and hand-based region selection in augmented reality," *Proc. ACM Hum.-Comput. Interaction*, vol. 7, no. ETRA, pp. 1–19, 2023.

[12] Y. Wei et al., "Predicting gaze-based target selection in augmented reality headsets based on eye and head endpoint distributions," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–14.

[13] J. Kim, S. Park, Q. Zhou, M. Gonzalez-Franco, J. Lee, and K. Pfeuffer, "PinchCatcher: Enabling multi-selection for gaze pinch," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2025, pp. 1–16.

[14] T. Isomoto, S. Yamanaka, and B. Shizuki, "Dwell selection with ML-based intent prediction using only gaze data," *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 6, no. 3, pp. 1–21, 2022.

[15] A. S. Narkar, J. J. Michalak, C. E. Peacock, and B. David-John, "GazeIntent: Adapting dwell-time selection in VR interaction with real-time intent modeling," 2024, *arXiv:2404.13829*.

[16] P. Majaranta, U.-K. Ahola, and O. Špakov, "Fast gaze typing with an adjustable dwell time," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2009, pp. 357–360.

[17] X. Lu, D. Yu, H.-N. Liang, and J. Goncalves, "iText: Hands-free text entry on an imaginary keyboard for augmented reality systems," in *Proc. 34th Annu. ACM Symp. User Interface Softw. Technol.*, 2021, pp. 815–825.

[18] X. Yi, L. Qiu, W. Tang, Y. Fan, H. Li, and Y. Shi, "DEEP: 3D gaze pointing in virtual reality leveraging eyelid movement," in *Proc. 35th Annu. ACM Symp. User Interface Softw. Technol.*, 2022, pp. 1–14.

[19] A. Esteves, E. Velloso, A. Bulling, and H. Gellersen, "Orbits: Gaze interaction for smart watches using smooth pursuit eye movements," in *Proc. 28th Annu. ACM Symp. User Interface Softw. Technol.*, 2015, pp. 457–466.

[20] T. Piumsomboon, G. Lee, R. W. Lindeman, and M. Billinghurst, "Exploring natural eye-gaze-based interaction for immersive virtual reality," in *Proc. 2017 IEEE Symp. 3D User Interfaces*, 2017, pp. 36–39.

[21] Z. Wu, D. Wang, S. Zhang, Y. Huang, Z. Wang, and M. Fan, "Toward making virtual reality (VR) more inclusive for older adults: Investigating aging effect on target selection and manipulation tasks in VR," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2024, pp. 1–17.

[22] R. L. Franz, J. Yu, and J. O. Wobbrock, "Comparing locomotion techniques in virtual reality for people with upper-body motor impairments," in *Proc. 25th Int. ACM SIGACCESS Conf. Comput. Accessibility*, 2023, pp. 1–15.

[23] A. K. Mutasim, A. U. Batmaz, and W. Stuerzlinger, "Pinch, click, or dwell: Comparing different selection techniques for eye-gaze-based pointing in virtual reality," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, 2021, pp. 1–7.

[24] B. David-John, C. Peacock, T. Zhang, T. S. Murdison, H. Benko, and T. R. Jonker, "Towards gaze-based prediction of the intent to interact in virtual reality," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, 2021, pp. 1–7.

[25] G. R. Reddy, M. J. Proulx, L. Hirshfield, and A. Ries, "Towards an eye-brain-computer interface: Combining gaze with the stimulus-preceding negativity for target selections in XR," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2024, pp. 1–17.

[26] A. S. Yashin, Y. G. Shevtsova, E. P. Svirin, A. N. Vasilyev, and S. L. Shishkin, "Combining intuitive gaze-based control with EEG-based detection of motor imagery and quasi-movements," in *Proc. Symp. Eye Tracking Res. Appl.*, 2025, pp. 1–3.

[27] T. Grossman and R. Balakrishnan, "The bubble cursor: Enhancing target acquisition by dynamic resizing of the cursor's activation area," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2005, pp. 281–290.

[28] P. M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," *J. Exp. Psychol.*, vol. 47, no. 6, pp. 381–391, 1954.

[29] I. S. MacKenzie, "Fitts' law as a research and design tool in human-computer interaction," *Hum.- Comput. Interaction*, vol. 7, no. 1, pp. 91–139, 1992.

[30] F. Periverzov and H. Ilieş, "IDS: The intent driven selection method for natural user interfaces," in *Proc. 2015 IEEE Symp. 3D User Interfaces*, 2015, pp. 121–128.

[31] T. Luong, Y. F. Cheng, M. Möbus, A. Fender, and C. Holz, "Controllers or bare hands? A controlled evaluation of input techniques on interaction performance and exertion in virtual reality," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 11, pp. 4633–4643, Nov. 2023.

[32] M. Choi, D. Sakamoto, and T. Ono, "Kuiper belt: Utilizing the "out-of-natural angl," region in the eye-gaze interaction for virtual reality," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2022, pp. 1–17.
[33] Y. Ren, Y. Zhang, Z. Liu, Y. Li, L. Yuan, and N. Xie, "Eye-hand typing: Eye gaze assisted finger typing via Bayesian processes in ar," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 5, pp. 2496–2506, May 2024.
[34] H. Cho et al., "SonoHaptics: An audio-haptic cursor for gaze-based object selection in XR," in *Proc. 37th Annu. ACM Symp. User Interface Softw. Technol.*, 2024, pp. 1–19.
[35] T. Isomoto, T. Ando, B. Shizuki, and S. Takahashi, "Dwell time reduction technique using Fitts' law for gaze-based target acquisition," in *Proc. 2018 ACM Symp. Eye Tracking Res. Appl.*, 2018, pp. 1–7.
[36] A. R. Ramirez Gomez, C. Clarke, L. Sidenmark, and H. Gellersen, "Gaze hold: Eyes-only direct manipulation with continuous gaze modulated by closure of one eye," in *Proc. ACM Symp. Eye Tracking Res. Appl.*, 2021, pp. 1–12.
[37] T. Rolff et al., "A hands-free spatial selection and interaction technique using gaze and blink input with blink prediction for extended reality," 2025, *arXiv:2501.11540*.
[38] M. Vidal, A. Bulling, and H. Gellersen, "Pursuits: Spontaneous interaction with displays based on smooth pursuit eye movement and moving targets," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 439–448.
[39] Y. Park, J. Kim, and I. Oakley, "The impact of gaze and hand gesture complexity on gaze-pinch interaction performances," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2024, pp. 622–626.
[40] N. Sendhilnathan, A. S. Fernandes, M. J. Proulx, and T. R. Jonker, "Implicit gaze research for XR systems," 2024, *arXiv:2405.13878*.
[41] W. Nguyen, K. Gramann, and L. Gehrke, "Modeling the intent to interact with VR using physiological features," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 8, pp. 5893–5900, Aug. 2024.
[42] E. Hwang and J. Lee, "Looking but not focusing: Defining gaze-based indices of attention lapses and classifying attentional states," in *Proc. 2025 CHI Conf. Hum. Factors Comput. Syst.*, 2025, pp. 1–14.
[43] K. Krejtz, A. Duchowski, I. Krejtz, A. Szarkowska, and A. Kopacz, "Discerning ambient/focal attention with coefficient k," *ACM Trans. Appl. Percep.*, vol. 13, no. 3, pp. 1–20, 2016.
[44] R. Bednarik, H. Vrzakova, and M. Hradis, "What do you want to do next: A novel approach for intent prediction in gaze-based interaction," in *Proc. Symp. Eye Tracking Res. Appl.*, 2012, pp. 83–90.
[45] Y. Lu, C. Yu, and Y. Shi, "Investigating bubble mechanism for ray-casting to improve 3D target acquisition in virtual reality," in *Proc. IEEE Conf. Virtual Reality 3D User Interfaces*, 2020, pp. 35–43.
[46] H.-S. Moon, Y.-C. Liao, C. Li, B. Lee, and A. Oulasvirta, "Real-time 3D target inference via biomechanical simulation," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2024, pp. 1–18.
[47] J. V. Bradley, "Complete counterbalancing of immediate sequential effects in a latin square design," *J. Amer. Stat. Assoc.*, vol. 53, no. 282, pp. 525–528, 1958.
[48] A. S. Fernandes and S. K. Feiner, "Combating VR sickness through subtle dynamic field-of-view modification," in *Proc. IEEE Symp. 3D User Interfaces*, 2016, pp. 201–210.
[49] N. Sendhilnathan, T. Zhang, B. Lafreniere, T. Grossman, and T. R. Jonker, "Detecting input recognition errors and user errors using gaze dynamics in virtual reality," in *Proc. 35th Annu. ACM Symp. User Interface Softw. Technol.*, 2022, pp. 1–19.
[50] J. Brooke, "SUS: A quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.
[51] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, 1988.
[52] X. Bi and S. Zhai, "Bayesian touch: A statistical criterion of target selection with finger touch," in *Proc. 26th Annu. ACM Symp. User Interface Softw. Technol.*, 2013, pp. 51–60.
[53] Z. Li et al., "BayesGaze: A Bayesian approach to eye-gaze based target selection," in *Proc. Graph. Interface Conf.*, NIH Public Access, 2021, pp. 231–240.
[54] D. Yu, H.-N. Liang, X. Lu, K. Fan, and B. Ens, "Modeling endpoint distribution of pointing selection tasks in virtual reality environments," *ACM Trans. Graph.*, vol. 38, no. 6, pp. 1–13, 2019.
[55] H.-S. Moon, A. Oulasvirta, and B. Lee, "Amortized inference with user simulations," in *Proc. 2023 CHI Conf. Hum. Factors Comput. Syst.*, 2023, pp. 1–20.
[56] A. Esteves, Y. Shin, and I. Oakley, "Comparing selection mechanisms for gaze input techniques in head-mounted displays," *Int. J. Hum.- Comput. Stud.*, vol. 139, 2020, Art. no. 102414.
[57] M. Kumar, J. Klingner, R. Puranik, T. Winograd, and A. Paepcke, "Improving the accuracy of gaze input for interaction," in *Proc. 2008 Symp. Eye Tracking Res. Appl.*, 2008, pp. 65–68.
[58] S. T. Radev, U. K. Mertens, A. Voss, L. Ardizzone, and U. Köthe, "BayesFlow: Learning complex stochastic models with invertible neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 4, pp. 1452–1466, Apr. 2022.

Taewoo Jo received the BS degree in physics, and the MS degree in intelligent robotics from the Gwangju Institue of Science and Technology, Gwang-ju, Republic of Korea, in 2022 and 2024, respectively. He is currently working toward the PhD degree with the Department of Computer Science, Yonsei University. His research interests include extended reality and human-computer interaction.

Ho Jung Lee received the BS degree in computer science from Yonsei University, Seoul, South Korea, in 2023. He is currently working toward the PhD degree with the Integrated MS–PhD program with the Department of Computer Science, Yonsei University. His research interests include virtual reality, human–computer interaction, and reinforcement learning.

Sulim Chun received the BA degree in economics and applied statistics and the BS degree in computer science from Yonsei University, Seoul, South Korea, in 2024. She is currently working toward the MS degree in computer science with Yonsei University. Her research interests include virtual reality, human–computer interaction, and computer graphics.

In-Kwon Lee received the BS degree in computer science from Yonsei University, Seoul, South Korea, in 1989, and the MS and PhD degrees in computer science and engineering from the Pohang University of Science and Technology, Pohang, South Korea, in 1992 and 1997, respectively. He is a professor with the Department of Computer Science, Yonsei University. His research interests include computer graphics, human-computer interaction, and music technology.