MAGNeT: Multimodal Adaptive Gaussian Networks for Intent Inference in Moving Target Selection across Complex Scenarios

Xiangxian Li
Shandong University
Weihai, China
xiangxianli@sdu.edu.cn

Yawen Zheng*
Institute of Software Chinese Academy of Sciences
Beijing, China
School of Software, Shandong University
Jinan, China
zhengyawen@iscas.ac.cn

Baiqiao Zhang
Shandong University
Weihai, China
The Hong Kong University of Science and Technology
Hong Kong, China
baiqiao.zhang@connect.ust.hk

Yijia Ma
Shandong University
Weihai, China
mayijia@mail.sdu.edu.cn

Xianhui Cao
AiLF Instruments
Weihai, China
hans@ailf.com.cn

Juan Liu
Shandong University
Weihai, China
Shandong Key Laboratory of Intelligent Electronic Packaging Testing and Application
Weihai, China
zzzliujuan@sdu.edu.cn

Yulong Bian
Shandong University
Weihai, China
Shandong Key Laboratory of Intelligent Electronic Packaging Testing and Application
Weihai, China
bianyulong@sdu.edu.cn

Jin Huang
Institute of Software Chinese academy of sciences
Beijing, China
huangjin@iscas.ac.cn

Chenglei Yang
School of Software, Shandong University
Jinan, China
chl_yang@sdu.edu.cn

Abstract

Moving target selection in multimedia interactive systems faces unprecedented challenges as users increasingly interact across diverse, dynamic contexts—from live streaming in moving vehicles to VR gaming in varying environments. Existing approaches rely on probabilistic models that relate endpoint distribution to target properties (size, speed). However, these methods require substantial training data for each new context and lack transferability across scenarios, limiting their practical deployment in diverse multimedia environments where rich multimodal contextual information is readily available. This paper introduces MAGNeT (Multimodal Adaptive Gaussian Networks), which addresses these problems by combining classical statistical modeling with context-aware multimodal method. MAGNeT dynamically fuses pre-fitted Ternary-Gaussian models from various scenarios based on real-time contextual cues, enabling effective adaptation with minimal training data while preserving model interpretability. We take experiments on self-constructed 2D and 3D moving target selection datasets under in-vehicle vibration conditions. Extensive experiments demonstrate that MAGNeT achieves lower error rates with few-shot samples, by applying context-aware fusion of Gaussian experts from multi-factor conditions.

CCS Concepts

• **Computing methodologies → Machine learning**.

Keywords

Moving Target Selection, Gaussian Mixture Model, Multi-modal Fusion, Few-shot Learning

ACM Reference Format:
Xiangxian Li, Yawen Zheng, Baiqiao Zhang, Yijia Ma, Xianhui Cao, Juan Liu, Yulong Bian, Jin Huang, and Chenglei Yang. 2025. MAGNeT: Multimodal Adaptive Gaussian Networks for Intent Inference in Moving Target Selection across Complex Scenarios. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Moving target selection has become increasingly prevalent in multimedia interactive content, such as interactive live streaming and gaming [13], and is studied as a fundamental task for generalized interfaces (e.g., 2D touchscreens [10, 12] and 3D virtual reality spaces [8, 28]). This task poses significant challenges to users' perception-action systems [11, 22], especially when performed under complex scenarios involving human factors and environmental perturbations [3, 14, 19]. For instance, Figure 1 illustrates a scenario where a user watches a live stream on a tablet while traveling in a car. The vehicle's movement makes it challenging to accurately tap on a real-time comment when it scrolls. This leads to the user selecting an unintended comment rather than the intended one. Such errors significantly degrade both user experience and interaction efficiency. Therefore, improving the efficacy of moving target selection amid a combination of human, environmental, and device-related factors.

Previous work has primarily focused on inferring user intent by modeling uncertainties arising from device characteristics, target properties, and human factors. Huang et al. proposed the Ternary-Gaussian model [11], which relates endpoint distributions to a target's spatial and motion properties. This model provides a statistical criterion for the Bayesian framework, enabling the inference of user intent in moving target selection tasks [12, 29]. The model has been expanded to be applied in other scenarios by either incorporating additional impactful factors into the Ternary-Gaussian framework or embedding the Ternary-Gaussian as a kernel within another mechanism [8–10, 12, 21, 25–28] . However, probabilistic models require substantial data for model fitting, and existing models—often tailored to single contexts—lack transferability; each new scenario demands collection of new data and model parameters.

To address these limitations, this paper formulates the problem of intent inference for moving target selection under complex scenarios, and proposes a Multimodal Adaptive Gaussian Network for Target selection (MAGNeT)[1]. The proposed MAGNeT effectively senses multimodal environmental data and enables few-shot model adaptation by dynamically fusing pre-fitted Ternary-Gaussian models from various previous scenarios. MAGNeT incorporates user profiling, sensor data, and contextual information of selection tasks, uses a Gaussian mixture model to combine prior models, and applies a self-adaptive learning mechanism to adjust expert model parameters for each context. This significantly reduces prediction error even with limited training samples.

Extensive experiments using self-collected datasets of moving target selections under in-vehicle vibration conditions shows that, MAGNeT achieves lower error rate with only a few-shot samples per user in each condition. Further ablation experiments confirm the effectiveness of combining prior experts. In case studies, MAGNeT adaptively adjusts the weights for environmental factors, and demonstrates strong adaptability. Our key contributions include:

- We formulate the problem of intent inference for moving target selection in complex scenarios, and propose MAGNeT, a framework leveraging multimodal context-aware weighting and multi-expert Gaussian modeling.
- We introduce a self-adaptive update mechanism based on multi-modal context information and multi-expert models, adapting prior knowledge while maintaining interpretability.
- We construct datasets for moving target selection in complex contexts and demonstrate that our approach can dynamically adjust weights to reduce errors even in few-shot settings.

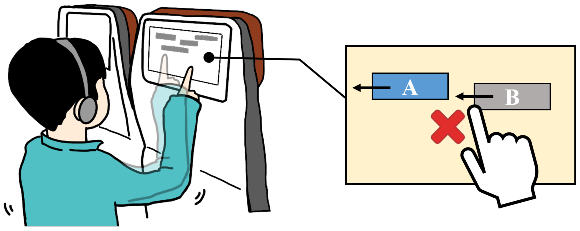

Figure 1: A user is performing touch interaction in a bumpy vehicle cabin. During this process, the car's bumping causes their arm to sway, leading them to accidentally select target B instead of the intended target A.

2 Related Works

2.1 Modeling Moving Target Selection Behavior

To enhance user performance in moving target selection, researchers have, on one hand, conducted in-depth studies on explicit assistance methods with visual cues [7, 17, 23], and on the other hand, focused on implicit approaches—such as inferring user intent through probabilistic computational models—which leverage interaction behavior uncertainty[29]. Thus, characterizing the uncertainty of moving target selection has become a key concern in the human-computer interaction (HCI) field. One of the most widely used models is the Ternary-Gaussian model [11]. It assumes that the selection endpoint follows a Gaussian distribution composed of three Gaussian components associated with target size, speed, and absolute pointing accuracy. The Ternary-Gaussian model provides statistical criteria for understanding user interaction intentions by using Bayes' theorem to determine touch selection targets [12]. The Ternary-Gaussian model has been extended from 1D to 2D and 3D space [12, 28] and in various contexts including modeling path steering and pursuit, crossing-based moving target selection [10], endpoint distribution of arbitrary shapes [25, 26], touching in different scenarios [27], and spatiotemporal selection [9, 21].

2.2 Gaussian Mixture Model for Modeling Uncertainty

Existing models analyze individual factors like target size, speed, and depth influencing moving target selection. However, these factors combine non-linearly, making it hard for single models to capture their coupled effects. Gaussian Mixture Models (GMMs), widely used for interaction uncertainty modeling [4, 20], approximate complex probability distributions. For example, GMMs improve robot behavior prediction by integrating environmental disturbances and user intent. Yet, current GMMs rely on static weights, struggling with real-time changes in dynamic interactions. This points to a need for context-aware dynamic weight calibration. This would

[1] The project page is https://yibuxulong.github.io/MAGNeT_project/.

involve real-time sensing of environmental fluctuations, user operations, and target motion (e.g., speed and size) to adaptively adjust expert model weights, overcoming GMM limitations in dynamic coupling. In dynamic systems, Mixture of Experts (MoE) models use gating networks to adjust sub-model weights, providing a framework for multi-factor coupling [15, 24]. However, traditional MoE models depend on offline data, limiting real-time adaptability to contextual changes [2, 5]. Bayesian online adaptive methods can update weights but face high computational costs for high-dimensional features and lack explicit modeling of factor coupling [6].

3 Problem Formulation

In the scenario of moving target selection based on the Ternary-Gaussian model, taking 2D interface as an example, each target $t_i \in T$ is defined by spatial attributes (screen coordinates (x_i,y_i), target size $w_i \in W$) and motion attributes (velocity $v_i \in V$, moving direction α_i). Using the coordinates (x_s, y_s) of a user's interaction endpoint s and its relative state to targets as input, the method leverages pre-fitted parameters θ_{init} derived from large-scale empirical datasets to initialize the model, requiring at least 9 predefined pairs of Gaussian parameters (μ and Σ) corresponding to distinct $W \times V$ conditions and a minimum of 100 endpoint samples per condition for robust bivariate distribution estimation. For each target t_i, the model predicts its endpoint distribution parameters (μ_i and Σ_i) based on its attributes, computes the likelihood p_i for an observed endpoint s, and applies Bayesian inference to determine the posterior probability, ultimately selecting the target with the highest posterior probability as the user's intended choice.

However, existing methods not only require large-scale datasets for initial model construction but also employs fixed parameters, thereby limiting adaptability to novel interaction contexts. Considering that multiple Ternary-Gaussian expert models (θ_1, ..., θ_k) derived from diverse acquisition devices or user populations—exhibit inherent variability in parameter ranges and inter-parameter relationships, their generalization abilities in new environments are uncertain. Different from the previous learning paradigm, we propose MAGNeT as an adaptive parameter adjustment strategy that tailors model parameters to current settings. This is complemented by context-aware feature extraction, which captures intricate situational characteristics to facilitate fusion of Ternary-Gaussian expert models within a Gaussian mixture framework. Our approach enhances selection performance under small sample conditions while retaining the interpretability advantages of the original Ternary-Gaussian model framework.

4 Methodology

To address the issues of intent inference under complex interaction scenario, this paper proposes a **Multimodal Adaptive Gaussian Network (MAGNeT)**. As illustrated in Figure 2, MAGNeT consists of three components: (1) Multimodal Context-Aware Weighting module captures and fuses heterogeneous contextual information, (2) Gaussian Parameter Adaptive Adjustment module dynamically adapts Gaussian parameters based on target-specific contexts, and (3) Multi-expert Fusion module generates probabilistic touch predictions through Gaussian mixture modeling.

4.1 Multimodal Context-Aware Weighting

The multimodal context-aware weighting module serves as the foundation of our framework, employing specialized encoders to extract features from different modalities and generating adaptive fusion weights for expert selection.

4.1.1 User Feature Encoder. User characteristics including gesture type, age, and gender are first normalized and then processed through a dedicated user encoder f_u:

$$\mathbf{h}_u = f_u(\mathrm{Norm}(\mathbf{u})) \tag{1}$$

where $\mathbf{u}$ represents the raw user features and f_u is the user encoder that transforms normalized user characteristics into a latent representation $\mathbf{h}_u \in \mathbb{R}^{d_u}$, where d_u is the dimension of user characters.

4.1.2 Environmental Feature Encoders. Environmental data from vibration and acceleration sensors are processed through specialized encoders with temporal attention mechanisms:

For vibration signals $\mathbf{e}_{vib} \in \mathbb{R}^{T \times d_v}$:

$$\mathbf{h}_v = f_v(\mathrm{TemporalAttention}(\mathbf{e}_{vib})) \tag{2}$$

For acceleration signals $\mathbf{e}_{acc} \in \mathbb{R}^{T \times d_a}$:

$$\mathbf{h}_a = f_a(\mathrm{TemporalAttention}(\mathbf{e}_{acc})) \tag{3}$$

where f_v, f_a, d_v and $d_a$are the vibration encoder, acceleration encoder, vibration dimension, and acceleration dimension, respectively. The temporal attention mechanism allows the model to focus on the most relevant time steps for each environmental modality before feature encoding.

4.1.3 Target Feature Encoder. The target-specific information from the interaction scene is processed through a target encoder f_t:

$$\mathbf{h}_t = f_t(\mathbf{t}_i) \tag{4}$$

where $\mathbf{t}_i$ contains the target information including spatial coordinates and geometric properties, and f_t is the target feature encoder.

4.1.4 Context-Aware Weighting. All encoded representations are integrated through a context-aware weighting (CAW) model to generate expert fusion weights:

$$\mathbf{h}_{con} = \mathrm{Concat}(\mathbf{h}_u, \mathbf{h}_v, \mathbf{h}_a, \mathbf{h}_t) \tag{5}$$

$$\mathbf{w} = [w_1, w_2, \ldots, w_k] = \mathrm{CAW}(\mathbf{h}_{con}) \tag{6}$$

where $\mathbf{w} \in \mathbb{R}^k$ represents the fusion weights for k experts, and $\sum_{i=1}^{k} w_i = 1$. The CAW architecture consists of three linear layers with LeakyReLU activations, where batch normalization and dropout are applied after the first layer.

4.2 Gaussian Parameter Adaptive Adjustment

This module performs target-specific Gaussian parameter adaptation, enabling the model to dynamically adjust to different touch targets and interaction contexts.

4.2.1 Target-Specific Coordinate System. For each target t_i, we establish a coordinate system with normal and tangent axes based on the target's geometric properties:

$$\mathbf{t}_i = \{(x'_i, y'_i), \mathbf{v}_i, \mathbf{w}_i\} \tag{7}$$

where (x'_i, y'_i) are the transformed coordinates, $\mathbf{v}_i$ is the tangent axis, and $\mathbf{w}_i$ is the normal axis.

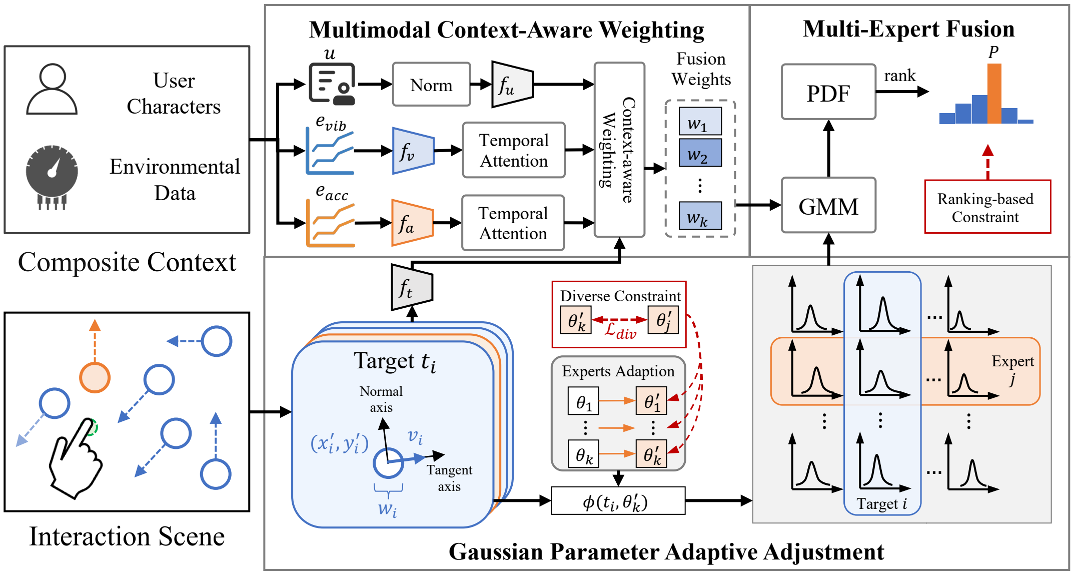

Figure 2: Illustrative diagram of MAGNeT, which features a Multimodal Context-Aware Weighting module that deciphers how prior models collaborate in a given scene. And its Gaussian Parameter Adaptive Adjustment module refines prior parameters specifically for the current context. The resulting ensemble of Gaussian distributions predicts outcomes via Gaussian mixture.

4.2.2 *Expert Parameter Adaptation.*

We maintain k expert parameter sets $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k\}$. Each expert's parameters are adapted based on the target-specific context and encoded features:

$$\boldsymbol{\theta}'_k = \text{ExpertAdaptation}(\boldsymbol{\theta}_k, \mathbf{h}_t, \mathbf{h}_{con}) \tag{8}$$

The expert adaptation process leverages both the target-specific features $\mathbf{h}_t$ from encoder f_t and the global context $\mathbf{h}_{con}$.

4.2.3 *Diversity Constraint.*

To ensure expert diversity and prevent parameter collapse, we introduce a diversity constraint $\mathcal{L}_{div}$:

$$\mathcal{L}_{div} = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j \neq i} \text{sim}(\boldsymbol{\theta}'_i, \boldsymbol{\theta}'_j) \tag{9}$$

where $\text{sim}(\cdot, \cdot)$ measures the similarity between expert parameters, encouraging diverse specialization across experts.

4.2.4 *Gaussian Distribution Generation.*

For each expert k and target t_i, we generate the corresponding Gaussian distribution:

$$\mathcal{G}_{k,i} = \phi(t_i, \boldsymbol{\theta}'_k) \tag{10}$$

where $\phi(\cdot, \cdot)$ is the Gaussian parameterization that computes the mean and covariance matrix based on the adapted parameters.

4.3 Multi-expert Fusion

The final module combines predictions from multiple experts through Gaussian mixture to generate the final probability distribution.

4.3.1 *Gaussian Mixture Model Construction.*

The Gaussian mixture model combines expert predictions weighted by the context-aware fusion weights derived from the encoded features:

$$p(\mathbf{y}|t_i, \mathbf{h}_{con}) = \sum_{k=1}^{K} w_k \mathcal{G}_{k,i}(\mathbf{y}) \tag{11}$$

where $\mathbf{y} \in \mathbb{R}^2$ represents the 2D touch coordinates, and the weights w_k are computed from the concatenated encoded features $\mathbf{h}_{con}$.

4.3.2 *Probability Density Function.*

The GMM outputs a probability density function (PDF) that captures the uncertainty of touches:

$$\text{PDF}(t_i) = p(\mathbf{y}|t_i, \mathbf{h}_{con}) \tag{12}$$

4.3.3 *Ranking-based Prediction.*

The final prediction is obtained by ranking targets based on their probability values:

$$\text{rank}(t_i) = \arg\max_{\mathbf{y}} \text{PDF}(t_i) \tag{13}$$

4.4 Loss Function and Training Strategy

Our training objective combines two constraints:

$$\mathcal{L}_{total} = \mathcal{L}_{rank} + \lambda_{div}\mathcal{L}_{div} \tag{14}$$

4.4.1 *Ranking-based Constraint.*

The ranking loss ensures that the ground truth target has higher probability than negative samples:

$$\mathcal{L}_{rank} = \max(0, \text{margin} + \log p(\mathbf{y}_{neg}|t_{neg}) - \log p(\mathbf{y}_{pos}|t_{pos})) \tag{15}$$

where t_{pos} is the ground truth target and t_{neg} represents negative targets. The complete framework effectively integrates multimodal contextual information through specialized encoders (f_u, f_v, f_a, f_t) and adaptive parameter adjustment.

5 Data Collection

5.1 Participants

10 participants (3 females) were recruited in this study. The average age of the participants was 23.4 years old (± 2.84), and all of them were right-handed. Every participant had experience in using touch-based devices such as smartphones and tablets. None of the participants reported having any perceptual or motor impairments.

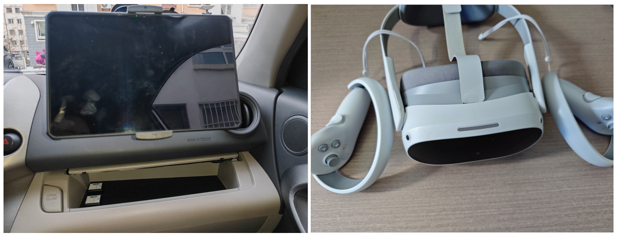

Figure 3: Apparatus used in the experiment.

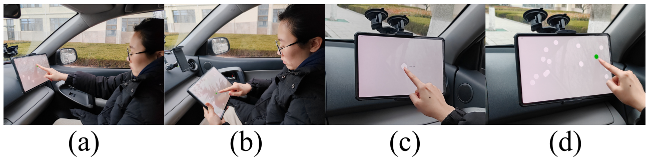

Figure 4: Two interactive gestures and task interface of 2D moving target selection.

5.2 Apparatus

This study collected environmental vibration and user interaction data within a real vehicle on a 2-km closed campus loop to minimize external interference and ensure safety. The route featured 6 right-angle turns and 8 speed bumps. The weighted root mean square acceleration (RMSA) is used to characterize the intensity of the vibration, in accordance with the ISO 2631–1 standard (International Organization for Standardization, 1997 [1]). Due to non-regular, aperiodic vibrations, RMSA was calculated over 3-second windows preceding each target selection.

During the experiment, the experimenter drove a car (Toyota RAV4) along the specified experimental route. Environmental vibration was measured using two inertial sensors (Wit BWT901BLECL5.0, Wit WTVB01-BT50) mounted on the passenger-side platform. They were used to collect acceleration data and vibration data respectively. The acceleration data included the acceleration of the x, y, and z axes, and the vibration data included the vibration velocity, vibration angle, vibration displacement, and vibration frequency of the x, y, and z axes. As shown in Figure 3, A Huawei MatePad Pro tablet (2560×1600, 12.6", 240 PPI) displayed 2D target selection tasks and recorded touch points, mounted near-vertially on a bracket. A Pico 4 headset (4320×2160, 2.56" per eye, 1200 PPI, 105° FOV) presented 3D selection tasks and captured spatial pointing data.

5.3 Design

5.3.1 2D moving target selection. The experiment adopted a within-subjects design, involving a cross-combination of 4 target size levels, 4 target speed levels, and 2 interaction gestures:

- Target size (W): 65, 95, 125, and 155 px
- Target speed (V): 300, 550, 800, and 1050 px/s
- Interaction gesture (P): tablet fixed, tablet handheld

Each participant completed 12 repeated trials under each experimental condition, totaling $W(4) \times V(4) \times P(2) \times 10$ participants × 12 repetitions = 3840 trials. During the experiment, participants adopted two interaction gestures: the tablet fixed in the car (Figure 4(a)) and the tablet handheld by the participant (Figure 4(b)). The order of interaction gestures and the presentation sequence of target conditions were counterbalanced across participants, with rest periods inserted between different conditions. Additionally, the test order within each condition was randomized. Each participant took approximately 25 minutes to complete the entire test.

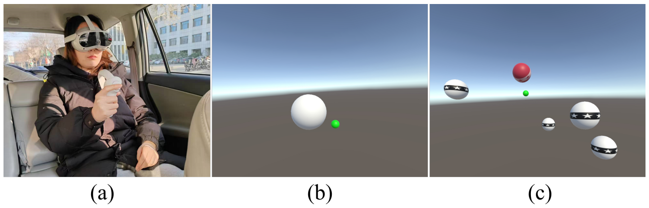

Figure 5: Interactive gesture and task interface of 3D scenario.

5.3.2 3D moving target selection. Within-subjects design was employed in the experiment, involving crossed combinations of 4 target size levels and 4 target speed levels:

- Target size (W): 0.04, 0.08, 0.12, and 0.16 m
- Target speed (V): 0.22, 0.34, 0.45, and 0.56 m/s

Each participant completed 6 repeated trials under each experimental condition, totaling $W(4) \times V(4) \times 10$ participants × 6 repetitions = 960 trials. The presentation sequence of target conditions was counterbalanced across participants, with rest intervals inserted between different conditions. Additionally, the test order within each condition was randomized. One participant took approximately 15 minutes to complete the entire experiment.

5.4 Procedure

5.4.1 2D moving target selection. Participants sat in the passenger seat, adjusted the seat to a comfortable position, and fastened their seatbelt. Once ready, the experimenter started the vehicle and drove along the predetermined route. In the 2D moving target selection task, participants initiated the test by tapping the "Start" button at the center of the screen, as shown in Figure 4 (c). At the beginning of each trial, 15 circular targets appeared at random positions on the screen and moved in random directions at a fixed speed. If a target reached the screen edge, it rebounded according to the law of specular reflection while maintaining its speed. Among these 15 circular targets, 14 were white, and only one was specially marked green, as shown in Figure 4 (d). Participants were instructed to tap the green target on the screen with their finger as quickly and accurately as possible. Each participant had only one attempt per task, and the system recorded the specific coordinates of their touch point on the screen regardless of whether the tap successfully hit the target's interior. The screen would then automatically clear, and the "Start" button would reappear at the center. Participants needed to tap this button again to initiate the next trial. When the vehicle completed one lap along the predetermined route and returned to the starting point, participants took a full rest before switching interaction gestures for further testing.

5.4.2 3D moving target selection. Participants made selection operations by squeezing the trigger of a handle, with the handle real-time mapped to a cursor in the 3D scene. The 3D cursor, a green

Table 1: Statistics of datasets, # means the number of item.

Dataset	# Samples	# Targets	# Test	# Validation
MTS-2D	3,840	15	1,536	384
MTS-3D	960	5	384	96

sphere with a diameter of 1 cm, was dynamically aligned with the handle's physical position in the real world, allowing participants to determine the handle's location by observing the green sphere. Participants initiated the test by selecting a white sphere at the center of their field of view (Figure 5 (b)).

At the start of each trial, 5 spherical targets appeared at random positions in the 3D space in front of the participant, moving in random directions at a fixed speed. The targets' movement range had a depth variation interval of 0.25 m to 0.6 m. Among the 5 spherical targets, only one was marked red (Figure 5 (c)). Participants were instructed to use the handle to select the red target as quickly and accurately as possible.

Each participant had only one attempt per task, and the system recorded the specific spatial coordinates of the 3D cursor regardless of whether the selection successfully hit the target's interior. Subsequently, all targets were automatically cleared, and a white sphere reappeared at the center of the field of view. Participants could initiate the next trial by selecting this sphere again.

6 Experiments

6.1 Experimental Settings

6.1.1 Dataset. The experiments are conducted on two datasets. Statistics of datasets are shown in Table 1.

Moving Target Selection 2D (MTS-2D) is a dataset containing 3,840 samples of 2D moving target selection touch data. Of these, 1,536 samples are used for model testing and 384 for validation.

Moving Target Selection 3D (MTS-3D) is a dataset containing 960 samples of 3D moving target selection data. Of these, 384 samples are used for model testing and 96 for validation.

The dataset partitioning is performed via uniform sampling across target conditions ($W \times V$), respectively. And the train set of each dataset is sampled from the remaining data according to the specific experimental settings described in Section 6.3. Both datasets include user characteristics (age, gender, gesture) and environmental vibration characteristics introduced in Section 5.2.

6.1.2 Expert Models. For the MTS-2D dataset, we have three expert models corresponding to sit-fixed, sit-handheld, and walk-handheld scenarios. We use **Expert (s-h)**, **Expert (s-f)**, and **Expert (w-h)** to denote the models derived from these scenarios, respectively:

- Expert (s-f): A 2D Ternary-Gaussian model with parameters fitted from experimental data of 25 participants. Participants remained seated and extended their arms to interact with a fixed tablet in the cabin for moving target selection.
- Expert (s-h): A 2D Ternary-Gaussian model with parameters fitted from experimental data of 21 participants. Participants sat while holding the tablet in one hand and using the other hand to tap the screen for moving target selection tasks.
- Expert (w-h): A 2D Ternary-Gaussian model with parameters fitted from experimental data of 21 participants. Participants walked while holding the tablet in one hand and using the other hand to tap the screen for moving target selection.

For the MTS-3D dataset, a 3D Ternary-Gaussian model (Expert (3D)) whose parameters were derived from [28].

6.1.3 Parameter Settings. For model training, we configure the batch size to 32 for the MTS-2D and 16 for the MTS-3D to ensure stable gradient estimation. The maximum number of training epochs is set to 50, with early stopping triggered if the validation loss fails to decrease for 10 consecutive epochs. We employ the AdamW optimizer for model optimization with a learning rate of 5×10^{-4} and weight decay of 1×10^{-4} to promote generalization. A cosine annealing schedule is applied for learning rate decay.

Regarding the specific architectural parameters of our MAGNeT framework, the feature encoders are configured as follows:

User and Target Encoders: Both the user encoder f_u and target encoder f_t are implemented as linear networks with hidden dimension $d_{hidden} = 64$, producing user representations $\mathbf{h}_u \in \mathbb{R}^{64}$ and target representations $\mathbf{h}_t \in \mathbb{R}^{64}$, respectively.

Environmental Encoders: The acceleration encoder f_a and vibration encoder f_v are implemented as bidirectional GRU networks, each with hidden dimension $d_{hidden} = 64$, generating environmental feature representations $\mathbf{h}_a, \mathbf{h}_v \in \mathbb{R}^{128}$.

Temporal Attention: The temporal attention mechanism employs multi-head attention with feature dimension 128 and 8 attention heads, enabling the model to capture diverse temporal dependencies in environmental signals.

Context-Aware Weighting: The softmax temperature parameter τ in the context-aware weighting module is set to 2.0 to control the sharpness of expert weight distributions.

6.2 Metrics

To analyze the model's ability to model interaction uncertainty in complex scenarios, this paper measures the error rate of the model's predictions. For the paradigm of secondary confirmation in interaction assistance [16, 18], we measured the Top-1 and Top-2 error rates, formed as $E@1$ and $E@2$. Additionally, to model the uncertainty caused by vibrations in an in-vehicle environment, we further measured the error rates under different levels of vibration.

- **Top-1 Error rates based on clustering (E_{clust})**: K-means clustering was performed on the acceleration and vibration data collected during the experiment. The number of clusters was determined by the silhouette coefficient. Using RMSA as a quantitative indicator, the midpoint of the clusters was calculated based on the cluster centers to serve as the dividing point. Statistical results show that in the 2D dataset, the data can be divided into two clusters. The mean RMSA of the first and second clusters are 0.7030 and 0.4516 separately. Therefore, for the 2D data, group 1 (G1) for E_{clust} is $RMSA < 0.5773$, and G2 is $RMSA \geq 0.5773$. The 3D dataset is also clustered into two groups: the mean RMSA of clusters are 0.3895 and 0.5768 separately. Therefore, for the 3D data, G1 for E_{clust} is $RMSA < 0.4831$, and G2 is $RMSA \geq 0.4831$.
- **Top-1 Error rates based on the mean of RMSA (E_{mean})**: Using the mean of RMSA as the benchmark, in the 2D dataset, G1 for E_{mean} is $RMSA < 0.5508$, and G2 is $RMSA \geq 0.5508$.

Table 2: Model performances on MTS-2D dataset. We report means (standard deviations in parentheses) across different seeds.

Model	$E_{clust}(G1)$	$E_{clust}(G2)$	$E_{mean}(G1)$	$E_{mean}(G2)$	$E@1$	$E@2$
Border-based	0.7844(0.0068)	0.8318(0.0080)	0.7694(0.0109)	0.8307(0.0050)	0.8003(0.0042)	-
Distance-based	0.2223(0.0089)	0.2285(0.0088)	0.2021(0.0098)	0.2463(0.0046)	0.2243(0.0039)	0.0657(0.0034)
Expert(s-f)	0.1367(0.0035)	0.1681(0.0153)	0.1239(0.0029)	0.1705(0.0140)	0.1473(0.0072)	0.0400(0.0021)
Expert(s-h)	0.1189(0.0053)	0.1637(0.0152)	0.1109(0.0041)	0.1569(0.0121)	0.1341(0.0077)	0.0359(0.0018)
Expert(w-h)	0.1156(0.0073)	0.1558(0.0158)	0.1061(0.0044)	0.1521(0.0132)	0.1292(0.0087)	0.0349(0.0015)
MAGNeT (1-Shot)	0.1158(0.0058)	0.1584(0.0152)	0.1069(0.0044)	0.1533(0.0130)	0.1303(0.0085)	0.0357(0.0031)
MAGNeT (3-Shot)	0.1110(0.0027)	0.1554(0.0132)	0.1025(0.0018)	0.1494(0.0106)	0.1261(0.0049)	**0.0341**(0.0019)
MAGNeT (5-Shot)	0.1122(0.0055)	0.1535(0.0141)	0.1012(0.0028)	0.1509(0.0130)	0.1262(0.0074)	0.0353(0.0016)
MAGNeT (10-Shot)	**0.1093**(0.0047)	**0.1524**(0.0128)	**0.1002**(0.0028)	**0.1474**(0.0114)	**0.1239**(0.0064)	0.0351(0.0030)

In the 3D dataset, G1 for E_{mean} is $RMSA < 0.4671$, and G2 is $RMSA \geq 0.4671$.

6.3 Comparative Results

To evaluate the effectiveness of MAGNeT, we conducted comparative experiments against several approaches. 1) The Border-based method determines whether a target is successfully selected based on whether the user's touch point falls within the radius of the target. This method serves as a standard baseline for moving target selection tasks. 2) The Distance-based method identifies the user's intended target as the one with the smallest Euclidean distance to the touch point. 3) The Expert models refer to expert models, the specific definitions of which have been detailed in earlier sections. We compare MAGNeT with few-shot training samples with aforementioned methods, and following the previous works [9, 27], we refer 1-shot setting as each participant contributes only one selection sample under each combination of target size (W) and speed (V). And similar to the 2-shot, 3-shot, 5-shot, and 10-shot settings. Based on the performance comparison results presented in Table 2 and 3, we summarize the key findings as follows:

- **Border-based methods yield high error rates in both 2D and 3D datasets, indicating that accurately selecting targets is non-trivial, which highlights the critical importance of incorporating intent understanding to assist in target selection.** As shown in Tables 2 and 3, the Border-based method suffers from substantial error across both datasets, with E_{mean} values as high as 0.7694 (2D) and 0.8357 (3D). These high error rates underscore the difficulty of directly selecting moving targets—especially under complex selection conditions.
- **Distance-based methods achieve partial improvement but struggle in complex interaction scenarios.** Compared with Border-based methods, Distance-based approaches significantly reduce selection error (e.g., $E_{mean}(G1)$ drops from 0.7694 to 0.2021 in 2D). However, their performance remains inferior to Experts, especially in scenarios with environmental noise and user variability. Given the complex dynamics of real-world settings, Euclidean distance proves insufficient to capture user intent, particularly when movement-induced noise and ambiguity are present—highlighting the need for more expressive models.
- **Expert models offer further gains but show limited generalization across contexts.** The Expert models, trained on specific conditions such as walking, sitting, or vibration, perform better than baseline methods. However, the performance across Experts varies (e.g., Expert(s-h) vs. Expert(s-f)), revealing sensitivity to the context in which they were trained. In particular, their generalization to new or unseen contexts is constrained, as no adaptive mechanism is available to adjust their weights based on the current interaction scenario.
- **MAGNeT is effective in low-data regimes and generalizes well across motion conditions.** Across both MTS-2D and MTS-3D datasets, MAGNeT exhibits remarkable learning capabilities. Its performance remains consistent across different random seeds (G1 and G2), with low variance in error metrics. Importantly, this consistency holds under both sparse data (e.g., 1 or 2 samples per condition) and varying interaction dynamics, confirming that its self-calibrating mixture of experts and context encoding mechanisms generalize well across interaction complexities.
- **MAGNeT achieves significant increments on the MTS-3D dataset, validating the effectiveness of adaption.** MAGNeT demonstrates superior performance across metrics in the 3D setting. For instance, the 2-Shot variant achieves an exceptionally low $E_{mean}(G1)$ of 0.0027, outperforming all baselines by a wide margin. Given the challenge of selecting 1 target from 5 under motion, such performance illustrates MAGNeT's strength in combining contexts with real-time expert adaptation, enabling robust intent prediction even in three-dimensional environments.

6.4 Ablation of Experts Selection

To investigate the impact of expert models collected under diverse scenarios when utilized as initialization values on model performance, we conducted ablation experiments on a 2D dataset with a 10-shot training set. As shown in Table 4, key observations from the results are summarized as follows:

- **Limited impact of single-expert absence, but significant degradation when all experts are missing.** The absence of any individual expert model exhibits a marginal effect on overall performance, demonstrating the model's adaptive learning capability. However, removing all expert models (w/o all) leads to increase in error rate, showing that expert models serve as effective prior knowledge for initialization.
- **Scenario-specific sensitivity in expert removal.** Eliminating the model collected during walking (w/o w-h) results in incremental error rate escalation in the high RMSA scenario (G2) compared to removing the seated-condition model (w/o s-h).

Table 3: Model performances on MTS-3D dataset. We report means (standard deviations in parentheses) across different seeds.

Model	$E_{clust}(G1)$	$E_{clust}(G2)$	$E_{mean}(G1)$	$E_{mean}(G2)$	$E@1$	$E@2$
Border-based	0.8417(0.0146)	0.8519(0.0108)	0.8357(0.0170)	0.8548(0.0164)	0.8458(0.0097)	-
Distance-based	0.4235(0.0288)	0.3757(0.0397)	0.4389(0.0297)	0.3721(0.0387)	0.4047(0.0294)	0.2214(0.0219)
Expert(3D)	0.3121(0.0261)	0.2184(0.0350)	0.3265(0.0225)	0.2273(0.0317)	0.2759(0.0168)	0.1179(0.0152)
MAGNeT (1-Shot)	0.0101(0.0071)	0.0166(0.0093)	0.0076(0.0073)	0.0173(0.0109)	0.0125(0.0074)	0.0006(0.0013)
MAGNeT (2-Shot)	**0.0042**(0.0062)	**0.0048**(0.0039)	**0.0027**(0.0054)	**0.0061**(0.0054)	**0.0044**(0.0047)	**0.0000**(0.0000)

Table 4: Ablation results.

Metric	MAGNeT	w/o s-f	w/o s-h	w/o w-h	w/o all
$E_{clust}(G1)$	0.1093	0.1108	0.1101	0.1093	0.1122
$E_{clust}(G2)$	0.1524	0.1531	0.1513	0.1535	0.1573
$E_{mean}(G1)$	0.1002	0.1007	0.1007	0.1002	0.1038
$E_{mean}(G2)$	0.1474	0.1494	0.1471	0.1481	0.1509
$E@1$	0.1239	0.1252	0.1241	0.1243	0.1275
$E@2$	0.0351	0.0334	0.0351	0.0348	0.0363

This indicates cross-scenario correlations and shared knowledge representation across different environmental contexts.

6.5 Analysis of Adaptive Weight Learning

To investigate the effect of weight learning on context awareness, we analyzed test results from MAGNeT, trained with 10-shot learning on MTS-2D, focusing on two users in two distinct poses on the test set. As shown in the figure 6, our observations are as follows:

- **Weight adaptation demonstrates well perception of user-specific information.** We observed that for different users in Pose 1 (tablet in hand), the touch points (marked with gray 'X') were generally closer to the actual targets (orange) compared to Pose 2. Specifically, cases (a), (e), and (g) show user selections closer to the target than cases (d), (f), and (h). This is reflected in the fusion weights, where the system assigns higher weights to the expert model representing a seated, handheld gesture.
- **MAGNeT effectively encodes environmental information with stability.** In relatively stable conditions, such as cases (a), (b), (e), and (f), the system consistently assigns higher weights to the second expert model. However, in situations with significant movement or shaking, the weights are adjusted accordingly.
- **Model Adjusts Weights for Complex Environments or High User Uncertainty.** For complex environments or user gestures with high uncertainty, the model dynamically adjusts weights to fuse different expert models. As shown in cases (c) and (h), where acceleration and vibration data exhibit significant fluctuations, the system assigns higher weights to the third expert model, which represents a walking, handheld gesture.

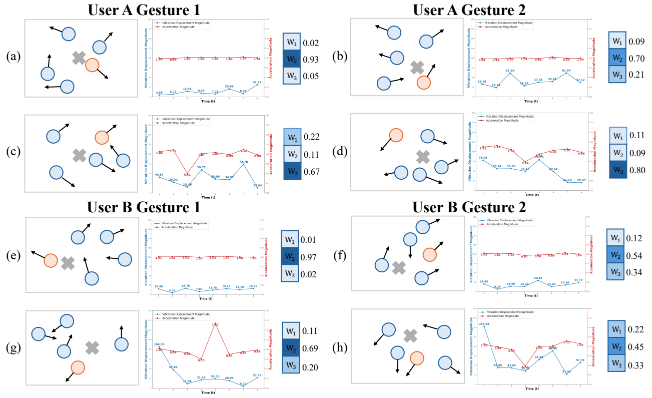

Figure 6: Cases showing how weights are adaptively learned. For each sub-figure, we show a zoom-in scene of moving targets while user touching the screen (left), the trend of accuracy and vibration changing 3 seconds before touching (middle), the fusion weights for experts (right).

7 Conclusion

This paper introduces MAGNeT, a novel approach designed to address the challenge of understanding user intent when selecting moving targets. MAGNeT leverages multi-modal information for adaptive learning, significantly reducing error rates by integrating expert models with only a small number of samples. Its effectiveness has been validated through experiments on both 2D and 3D datasets.

However, this study has several limitations that open avenues for future research:

- The current approach still relies on pre-built user profiles for perceptual understanding. Future work could explore incorporating real-time sensors, such as cameras, to dynamically generate user profiles and extract user features during cold-start scenarios. This would enhance MAGNeT's ability to adapt to new users without prior data.
- The current dataset is relatively small, and due to experimental constraints, our vehicular experiments were limited to a single loop-shaped road. This restricts the variety of environmental factors considered. Future research will focus on expanding the dataset to include new road types and diverse environments, such as maritime scenarios (e.g., ship cabins), to further validate MAGNeT's robustness and generalizability.
- The current experimental setup did not account for varying numbers of targets. While the principles of Bayesian pointing and MAGNeT's proposed method are theoretically transferable to scenarios with different target quantities, the generalization capability remains unexplored. Future work will investigate MAGNeT's performance and adaptability when the number of moving targets changes.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 62332017, 62172397, 62277035), the Youth Innovation and Technology Support Program of Shandong Provincial Higher Education Institutions under Grant [number 2022KJN028].

References

[1] NAIS An and S SI. 1997. Mechanical Vibration and Shock-Evaluation of Human Exposure to Whole-Body Vibration-Part 1: General Requirements. (1997).

[2] Bing Cao, Yiming Sun, Pengfei Zhu, and Qinghua Hu. 2023. Multi-modal gated mixture of local-to-global experts for dynamic image fusion. In *Proceedings of the IEEE/CVF international conference on computer vision*. 23555–23564.

[3] Sonia Dodd, Jeff Lancaster, Andrew Miranda, Steve Grothe, Bob DeMers, and Bill Rogers. 2014. Touch Screens on the Flight Deck: The Impact of Touch Target Size, Spacing, Touch Technology and Turbulence on Pilot Performance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58, 1 (2014), 6–10. arXiv:https://doi.org/10.1177/1541931214581002 doi:10.1177/1541931214581002

[4] Huajian Fang and Timo Gerkmann. 2023. Uncertainty estimation in deep speech enhancement using complex Gaussian mixture models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[5] Wensheng Gan, Zhenyao Ning, Zhenlian Qi, and Philip S Yu. 2025. Mixture of Experts (MoE): A Big Data Perspective. *arXiv preprint arXiv:2501.16352* (2025).

[6] Samuel J Gershman and David M Blei. 2012. A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology* 56, 1 (2012), 1–12.

[7] Khalad Hasan, Tovi Grossman, and Pourang Irani. 2011. Comet and target ghost: techniques for selecting moving targets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 839–848. doi:10.1145/1978942.1979065

[8] Jin Huang, John J. Dudley, Stephen Uzor, Dong Wu, Per Ola Kristensson, and Feng Tian. 2022. Understanding user performance of acquiring targets with motion-in-depth in virtual reality. *International Journal of Human-Computer Studies* 163 (2022), 102817. doi:10.1016/j.ijhcs.2022.102817

[9] Jin Huang and Byungjoo Lee. 2019. Modeling error rates in spatiotemporal moving target selection. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[10] Jin Huang, Feng Tian, Xiangmin Fan, Huawei Tu, Hao Zhang, Xiaolan Peng, and Hongan Wang. 2020. Modeling the Endpoint Uncertainty in Crossing-Based Moving Target Selection. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376336

[11] Jin Huang, Feng Tian, Xiangmin Fan, Xiaolong (Luke) Zhang, and Shumin Zhai. 2018. Understanding the Uncertainty in 1D Unidirectional Moving Target Selection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3173574.3173811

[12] Jin Huang, Feng Tian, Nianlong Li, and Xiangmin Fan. 2019. Modeling the Uncertainty in 2D Moving Target Selection. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology* (New Orleans, LA, USA) *(UIST '19)*. Association for Computing Machinery, New York, NY, USA, 1031–1043. doi:10.1145/3332165.3347880

[13] Michael Victor Ilich. 2009. *Moving target selection in interactive video.* Ph. D. Dissertation. University of British Columbia.

[14] Heon-Jeong Kim and Bernard J Martin. 2013. Biodynamic characteristics of upper limb reaching movements of the seated human under whole-body vibration. *Journal of applied biomechanics* 29, 1 (2013), 12–22.

[15] Jiamin Li, Qiang Su, Yitao Yang, Yimin Jiang, Cong Wang, and Hong Xu. 2023. Adaptive gating in mixture-of-experts based language models. *arXiv preprint arXiv:2310.07188* (2023).

[16] Zhi Li, Maozheng Zhao, Dibyendu Das, HANG ZHAO, Yan Ma, Wanyu Liu, Michel Beaudouin-Lafon, Fusheng Wang, IV Ramakrishnan, and Xiaojun Bi. 2022. Select or Suggest? Reinforcement Learning-based Method for High-Accuracy Target Selection on Touchscreens. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 494, 15 pages. doi:10.1145/3491102.3517472

[17] Yiqin Lu, Chun Yu, and Yuanchun Shi. 2020. Investigating bubble mechanism for ray-casting to improve 3D target acquisition in virtual reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 35–43.

[18] Jennifer Mankoff, Scott E. Hudson, and Gregory D. Abowd. 2000. Providing integrated toolkit-level support for ambiguity in recognition-based interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands) *(CHI '00)*. Association for Computing Machinery, New York, NY, USA, 368–375. doi:10.1145/332040.332459

[19] Neil J Mansfield and Setsuo Maeda. 2005. Effect of backrest and torso twist on the apparent mass of the seated body exposed to vertical vibration. *Industrial Health* 43, 3 (2005), 413–420.

[20] Geoffrey McLachlan. 2000. Finite mixture models. *A wiley-interscience publication* (2000).

[21] Adrian L Jessup Schneider and TC Nicholas Graham. 2023. Supporting aim assistance algorithms through a rapidly trainable, personalized model of players' spatial and temporal aiming ability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–17.

[22] Reza Shadmehr, Maurice A Smith, and John W Krakauer. 2010. Error correction, sensory prediction, and adaptation in motor control. *Annual review of neuroscience* 33, 1 (2010), 89–108.

[23] Lode Vanacken, Tovi Grossman, and Karin Coninx. 2007. Exploring the Effects of Environment Density and Target Visibility on Object Selection in 3D Virtual Environments. In *2007 IEEE Symposium on 3D User Interfaces*. doi:10.1109/3DUI.2007.340783

[24] Jing Yi and Zhenzhong Chen. 2024. Variational Mixture of Stochastic Experts Auto-encoder for Multi-modal Recommendation. *IEEE Transactions on Multimedia* (2024).

[25] Hao Zhang, Jin Huang, Huawei Tu, and Feng Tian. 2023. Shape-Adaptive Ternary-Gaussian Model: Modeling Pointing Uncertainty for Moving Targets of Arbitrary Shapes. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[26] Ziyue Zhang, Jin Huang, and Feng Tian. 2020. Modeling the Uncertainty in Pointing of Moving Targets with Arbitrary Shapes. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. doi:10.1145/3334480.3382875

[27] yawen zheng, Jin Huang, Juan Liu, Chenglei Yang, and Feng Tian. 2021. A Scenario Adaptive Model for Predicting Error Rates in Moving Target Selection on Smartphones. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction* (Toulouse &amp; Virtual, France) *(MobileHCI '21)*. Association for Computing Machinery, New York, NY, USA, Article 10, 15 pages. doi:10.1145/3447526.3472049

[28] Yawen Zheng, Jin Huang, Hao Zhang, Yulong Bian, Juan Liu, Chenglei Yang, Feng Tian, and Xiangxu Meng. 2025. 3D Ternary-Gaussian model: Modeling pointing uncertainty of 3D moving target selection in virtual reality. *International Journal of Human-Computer Studies* 198 (2025), 103454. doi:10.1016/j.ijhcs.2025.103454

[29] Suwen Zhu, Yoonsang Kim, Jingjie Zheng, Jennifer Yi Luo, Ryan Qin, Liuping Wang, Xiangmin Fan, Feng Tian, and Xiaojun Bi. 2020. Using Bayes' Theorem for Command Input: Principle, Models, and Applications. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376771