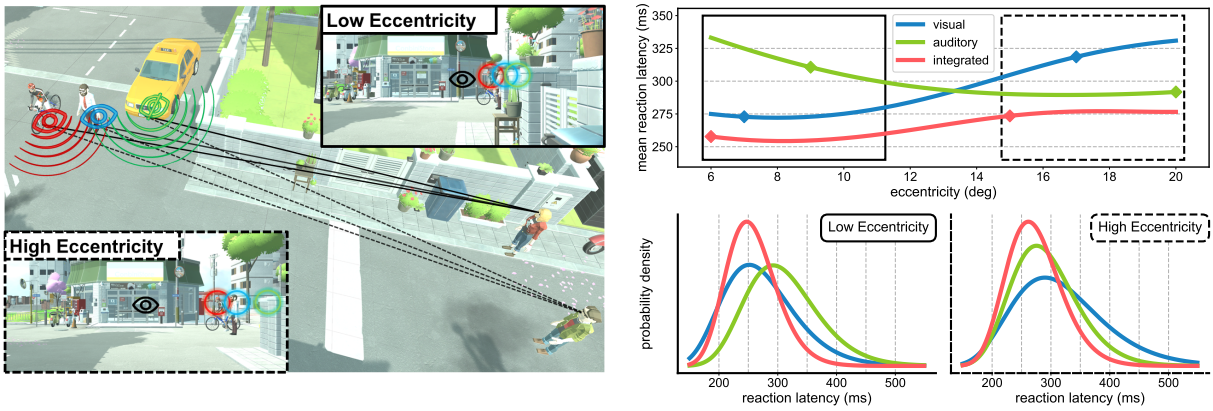
Measuring and Predicting Multisensory Reaction Latency: A Probabilistic Model for Visual-Auditory Integration

Xi Peng, Yunxiang Zhang, Daniel Jiménez-Navarro, Ana Serrano, Karol Myszkowski, and Qi Sun



(a) perceiving targets with different eccentricities and sensory cues (b) probablistic prediction of corresponding reaction latency

Fig. 1: *Predicting the reaction latency under multi-sensory integration and eccentricities. (a) illustrates a virtual environment where two VR individuals perceive three targets: a biker ringing a bell (red, visual and audible), a pedestrian (blue, visual but silent), and a honking car behind the wall (green, audible but invisible). Their reaction time to each target differs based on their standing positions (different eccentricities) and sensory cues. The insets illustrate the first-person views assuming gazing at the center. (b) shows our probabilistic model predicting the mean (marked with diamond dots) and probability distribution of reaction latency for the two individuals. The reaction time for visual/auditory targets exhibits opposite effects with eccentricities.*

Abstract— Virtual/augmented reality (VR/AR) devices offer both immersive imagery and sound. With those wide-field cues, we can simultaneously acquire and process visual and auditory signals to quickly identify objects, make decisions, and take action. While vision often takes precedence in perception, our visual sensitivity degrades in the periphery. In contrast, auditory sensitivity can exhibit an opposite trend due to the elevated interaural time difference. What occurs when these senses are simultaneously integrated, as is common in VR applications such as 360° video watching and immersive gaming?
We present a computational and probabilistic model to predict VR users' reaction latency to visual-auditory multisensory targets. To this aim, we first conducted a psychophysical experiment in VR to measure the reaction latency by tracking the onset of eye movements. Experiments with numerical metrics and user studies with naturalistic scenarios showcase the model's accuracy and generalizability. Lastly, we discuss the potential applications, such as measuring the sufficiency of target appearance duration in immersive video playback, and suggesting the optimal spatial layouts for AR interface design. We will release our source code and model upon acceptance.

Index Terms—Virtual reality, augmented reality, human perception, visual-audio, reaction latency

1 Introduction

In immersive applications, such as VR gaming or 360° video-watching, we synchronously integrate and interpret information from multiple independent sensory sources, including vision and hearing. The speed in integrating the multisensory cues and then taking subsequent action dictates our ability to successfully complete a task in time. Imagine we are playing a VR game. Is it possible we could fail to notice an approaching enemy because our focus shifts too slowly?

Extensive research has delved into how visual content may affect target-shifting behaviors [45, 73]. The discoveries have also catalyzed applications in accelerated rendering for virtual experience [2], gaming for immersive interaction [16], and viewing comfort for 360° stereoscopic videos watching [59]. The impact of auditory cues on human reaction performance is also independently investigated [28, 35, 50], and are found to influence the visual performance once integrated [14, 33, 47]. Based upon those observations, it is imperative to establish a characterized and computational model to guide VR applications, where content may have arbitrary properties and positions throughout the field of view. However, existing psychophysical measurements are primarily performed with singular sensory modality. To our best knowledge, there is no operationalized and algorithmic model that can comprehensively predict the reaction latency, concerning visual-audio stimuli variations and human behavioral noise [84] under various dimensions.

- *Xi Peng is with New York University. E-mail: xp2011@nyu.edu*
- *Yunxiang Zhang is with New York University. E-mail: yunxiang.zhang@nyu.edu*
- *Daniel Jiménez-Navarro is with Max Planck Institute for Informatics. E-mail: djimenez@mpi-inf.mpg.de*
- *Ana Serrano is with Universidad de Zaragoza. E-mail: anase@unizar.es*
- *Karol Myszkowski is with Max Planck Institute for Informatics. E-mail: karol@mpi-inf.mpg.de*
- *Qi Sun is with New York University. E-mail: qisun@nyu.edu*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

To bridge the knowledge gap between statistical observations and practical immersive applications, we first experimentally study human target-shifting latency when presented with synchronized visual-auditory stimuli. The latency is measured with the onset timing of rapid saccadic eye movements. Using both uni- and multi-sensory stimuli in a virtual environment, we investigate various dimensions of perceptual characteristics (visual contrast and auditory volume) and eccentricities across the field of view. Motivated by prior discoveries and our statistical analysis, we develop a computational probabilistic model by regressing its parameters using our acquired behavioral data. Our regression framework is inspired by decision-making modeling via stochastic drift-diffusion [16, 44, 81].

Statistical analysis and user studies with VR evidenced our model's accuracy in different tasks, as well as natural and complex environments. We further showcase the model as a proof-of-concept target perceptibility metric that provides potential guidelines for multisensory (visual-audio) VR interactive applications, AR interface, and video playback, as well as understanding the potential detrimental consequences of audio-canceling devices. To summarize, we make the following main contributions:

- a comprehensive set of VR psychophysical studies for collecting large-scale human reaction latency data during target-reaching, involving a wide range of visual-auditory stimuli characteristics;
- a regressed, analytical model that predicts the probabilistic distribution of human reaction latency to an appeared visual and/or auditory targets in VR;
- the model-derived potential design guidance for reaction-optimized immersive applications, including visual-auditory interface layout and video playback speed.

2 Related Work

2.1 Gaze-Aware Perception in VR/AR

Wide-field immersive displays such as the Vision Pro and Hololens 2 are commonly equipped with low-cost, high-accuracy eye trackers. The run-time gaze information, while combined with the wide-field displays, allows for eccentricity-aware graphics systems. In particular, with computational models depicting the degraded visual acuity in peripheral vision [48, 49], computer graphics systems may be guided to reduce computation workload [41, 42, 61, 82, 86], transmission bandwidth [12, 34, 37], or enhance visual depth perception [40, 77]. However, current gaze-aware graphics systems primarily leverage visual-only perceptual limits, leaving their potential interference with audio unattended.

2.2 Reaction Time Performance

The latency of our reaction to a surrounding event is a core metric of human-centric task performance in broad scenarios such as driving [76], esports [75], and VR/AR [29]. Measuring reaction latency requires precisely timing the onset of identifiable human action. Due to the challenges of reliably detecting cognitive activities, biometric markers have been leveraged to identify the moment of action; examples include brain waves [65] and muscle response [38]. Among these markers, eye movement stands out as a convenient and accessible measurement [68]. In particular, saccade – the ballistic eye movement when humans switch targets – may be detected by gaze velocity patterns [18]. The time interval between the appearance of the target and the onset of saccades is a widely utilized metric for measuring reaction performance [25, 88].

Human behaviors, including reaction latency, exhibit probabilistic patterns due to sensory and control noise [7, 45]. To formulate those uncertainties, drift-diffusion model (DDM) [24] is extensively validated as a computational framework for both uni- [16, 17] or multi- [72] sensory cues (see Section 4.1). With DDM, reaction latency is formulated as the probabilistic distribution of decision-making after accumulated evidence reaches a threshold [16, 24]. To our knowledge, current computational models for reaction performance primarily concentrate on specific or invariable target characteristics. However, it is crucial to systematically functionalize the interplay across multiple axes to facilitate downstream interactive graphics applications in the wild. Therefore, we aim to bridge this gap by developing a unified model that focuses on the distinct characteristics of individual sensory cues and their interference, concerning target eccentricity, visual contrast, and auditory volume.

2.3 Multimodal Perception and Interaction

In certain circumstances, vision may dominate our multisensory perception [62]. However, as eccentricity increases and visual acuity degrades, auditory cues gradually gain more sensitive and could even overwrite visual information [26, 32, 71, 89]. For instance, the ventriloquist effect – audio localization being biased toward visual stimuli – becomes less noticeable in the far periphery [10]. Meanwhile, studies have also found that conjugate audio may enhance performance in visual search tasks [50, 85]. Therefore, while visual and auditory evidence coexists, they jointly influence target acquisition and reaction performance.

Prior studies have presented hypotheses and experiments to understand how humans integrate and react to multisensory inputs. Bayesian inference is commonly used to learn visual-audio perceptual perception, considering cognitive and behavioral noises [5, 8, 83]. Studies have also evidenced that humans might employ statistically optimal strategies when weighing individual sensory cues [15, 20, 22]. When trying to explain reaction time, various hypotheses have been proposed, including temporal window integration [13, 14], normalization [57], or averaging [81]. Given the lack of consensus on the underlying neurological model for temporal integration of multiple sensory inputs, our aim, instead, is to form an executable and holistic model, predicting the probability of reaction latency for diverse targets in natural applications. Our model is regressed from and validated with human reaction time data collected from VR with diverse conditions.

3 Pilot Study: Target-Reaching Latency with Visual-Auditory Stimuli

Setup and participants As shown in Figure 2a, the study was performed using a Varjo Aero VR headset (Table 1) with Sony XM5 noise-canceling headphones (Table 2). We recruited 6 participants (ages 23-29, 3 female) with normal or corrected-to-normal vision and hearing conditions.In this experiment, we included a smaller participant group for larger individual sample sizes to derive a probabilistic model in Section 4. For each participant, an eye-tracking calibration was applied. During the study, participants remained seated and perceived the visual and/or audio stimuli through the headset and headphones. Their eye movement data were recorded at 200 Hz with the gaze tracker provided by the headset. The study was approved by the Institutional Review Board (IRB).

Table 1: Varjo Aero

Display Resolution	2880×2720 pixels per eye
Display Refresh Rate	90Hz
Eye Tracking Frequency	200Hz
Eye Tracking Accuracy	Sub-Degree

Table 2: WH-1000XM5 headphone

Frequency Response	20 Hz - 40000 Hz
Sensitivities	102 dB/mW (unit turned on)

Stimuli and conditions As in Figure 2b, the stimuli were a fixation-assisting indicator [80], a pair of visual targets, and/or a co-located spatial sound; all synthesized in Unity. We characterize the target stimuli with three primary dimensions, eccentricity ($e \in \{5°, 10°, 15° 20°\}$), visual contrast ($c \in \{0, 0.1, 0.5, 4\}$), as well as auditory volume ($v \in \{0, 40\text{ dB}, 52\text{ dB}, 66\text{ dB}\}$) measured by a RISE-PRO decibel meter. Note that $v = 0/c = 0$ indicates visual-only/audio-only stimuli. The sample points for each dimension were selected based on the preliminary study (see Appendix B for illustration). The experiment was conducted with 12 visual-only + 12 audio-only + 36 integrated = 60 conditions.

For $c \neq 0$ conditions, the visual targets were a pair of identical E letters, both toward the left or right. For $c = 0$ (audio-only) condition,

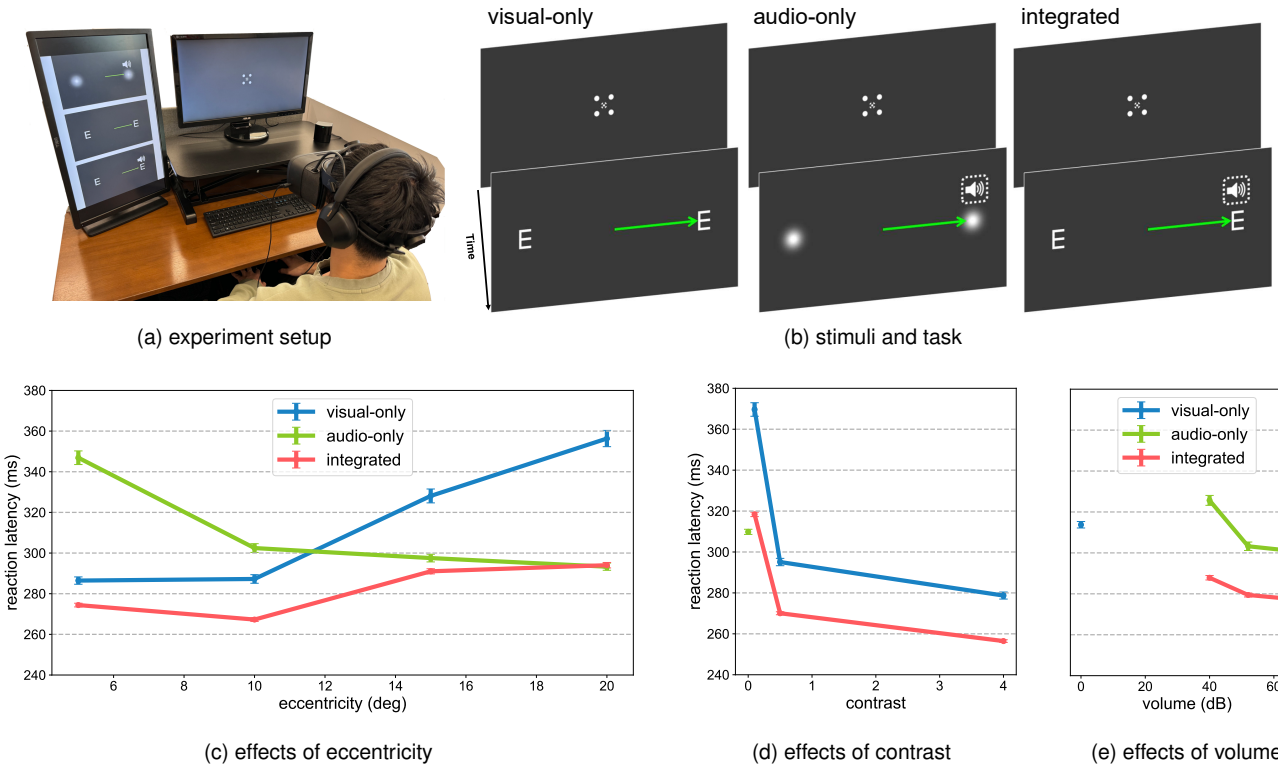


Fig. 2: *Pilot study protocol and results.* (a) shows the hardware and user settings. (b) denotes the stimuli and task, including visual-only (audio volume $v = 0$dB), audio-only (visual contrast $c = 0$), and visual-audio integrated conditions. (c) visualizes the aggregated reaction latency (Y-axis) with regard to eccentricity (X-axis). It compares the three stimuli groups; varied eccentricity effects can be observed with visual-only/audio-only/integrated stimuli. (d)/(e) shows the reaction latency data (Y-axis) in three stimuli groups aggregated with contrast/volume (X-axis). Note that contrast=0/volume=0 indicates audio-only/visual-only conditions. Error bars represent standard error. Please refer to Appendix A for detailed histogram plots.

we substitute the "E"s with a pair of identical Gaussian Blob. Participants had to rely only on their perception of audio stimulus to perform the task without gaining the visual cues. To minimize the reaction time influence from the visual blobs, we intentionally designed the blobs to have high contrast ($c = 4$) across all trials. The audio stimulus was a 20 Hz-20K Hz white noise in all conditions.

Task and procedure The task was a target-reaching using saccade eye movement. Specifically, the procedure was:

1. a fixation indicator appeared at the screen center;
2. after a successful fixation for 0.5 seconds, the indicator disappeared;
3. after a short delay randomly selected between 300ms and 500 ms, visual and/or audio stimuli simultaneously appear;
4. participants analyzed the
   - in visual-only/ integrated, the "E"s' orientations, and saccaded to the correct "E" based on its direction. (If "E"s were facing to the right, the participant saccaded to the "E" on the right side.)
   - in audio-only, audio stimulus's direction, and saccaded to the blob accompanied by the audio.

Please refer to the supplementary video for an animated visualization.

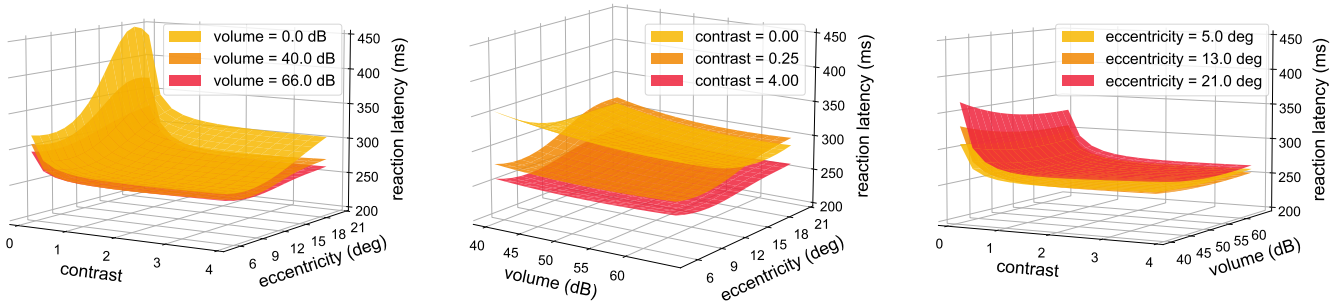
For each participant, the experiment was performed with 6 repetitive sessions, each was partitioned into 3 blocks (visual, audio, and integrated). The order of blocks in different sessions followed the balanced Latin square. Within each block, the trial conditions and the left/right of the targets were fully randomized to prevent performance bias or carry-over effects. The procedure took about 3 hours for each participant, including a short training session and breaks between sessions. With each condition repeated 36 times for each participant, we collected 12960 trials in total.

During the study, we considered the participants' saccading in the wrong direction to be insufficient perception of the visual and/or auditory cues, so we rejected the corresponding trial and let users redo it. Notably, the redo of the rejected trial has not been executed immediately. The program randomly injected the rejected trial into the rest of the undo trials to ensure the users had no prior knowledge of the upcoming trials.

Data processing As in prior literature [19, 39], we measure users' reaction latency as the duration between the onsets of target and the primary saccade. We use the statistical approach from Engbert and Mergenthaler [18] to detect the saccade onset. Studies have observed significant individual variances in reaction latency [30, 67, 84]. Therefore, similar to [16], a cross-subject calibration is performed using each one's mean latency of a neutrally challenging condition ($c = 0.5, v = 52$ dB$, e = 10°$),

3.1 Results

The bottom row of Figure 2 shows the complete reaction latency data in aggregated statistical analysis. As in Figure 2c, with increasing eccentricity, reaction latency elevates in visual-only conditions (from 286.21 ± 1.78 ms at 5° to 355.76 ± 3.96 ms at 20°), but decreases the latency with the audio-only conditions (from 346.71 ± 3.35 ms at 5° to 293.01 ± 1.77 ms at 20°). We compare our measured reaction time in visual-only conditions with prior work [16]. Our measured mean reaction time increases from 275.77 at $c = 0.5, e = 5°$ to 446.44 at $c = 0.1, e = 20°$. This is similar to the data measured under the same conditions in [16], which reported reaction times of 278.29 and 431.68, respectively. Additionally, we leverage a statistical goodness-of-fit Kolmogorov–Smirnov (K-S) test [51] to compare the two distributions. K–S is a non-parametric statistical test for the fitting between two data samples; a significant difference ($p < .05$) by the test indicates that the two samples are drawn from different distributions. The K-S test



(a) contrast-eccentricity (b) volume-eccentricity (c) contrast-volume

Fig. 3: *Visualization of our visual-auditory-integrated DDM model.* Each sub-figure visualizes two axes of the mean value derived from the three-dimensional (c,v,e) latency distribution function $L \coloneqq L(\alpha(c,v), r(c,v,e))$.

$(D = .10, p = 1.00)$ failed to reject the null hypothesis that the datasets from our experiment and [16] are drawn from the same distribution, indicating their alignment.

A one-way repeated measures ANOVA shows a significant main effect of eccentricity on both the visual-only condition ($F_{3,2443} = 133.19, p < .0001$) and audio-only condition ($F_{3,2468} = 109.17, p < .0001$). Pearson coefficients validate the significant positive and negative correlations for the visual-only ($r(2445) = .36, p < .0001$) and auditory-only ($r(2470) = -.29, p < .0001$) conditions, respectively. We also observe their intersecting effect with eccentricity: visual-only stimuli trigger faster reactions with low eccentricities (<=10°) than audio-only, but it is becoming surpassed as the eccentricity further increases. In the integrated condition, the reaction latency is also influenced by the eccentricity ($F_{3,7368} = 133.41, p < .0001$), but with less standard deviation (281.43 ± 11.1) compared to visual-only (314.20 ± 29.3) and audio-only (309.84 ± 21.5) conditions.

Figures 2d and 2e show that higher contrast and volume generally lower reaction latency, although the volume effect is weaker. Pearson coefficients indicate negative correlations between contrast and latency in visual-only ($r(2445) = -.36, p < .0001$) and integrated ($r(7370) = -.39, p < .0001$) conditions, also between volume and latency in auditory-only ($r(2470) = -.16, p < .0001$) and the integrated ($r(7370) = -.08, p < .0001$)conditions.

3.2 Discussion

The statistical analysis and visualization led us to several findings. First, we found that reactions are faster to audio-only cues at larger eccentricities, *contrary to* visual-only conditions where peripheral reactions are slower. Second, each sense is influenced by its specific attributes, such as visual contrast and auditory volume. Third, we observe that visual/auditory-only cues receive quicker reactions and even reach the "generally faster" integrated cue in smaller/larger eccentricity, explaining that audio-visual integration may be optimized for the more efficient and reliable sensory cue [20]. In an extreme case (e.g., $e = 20°$ and $c = 0.1$), reaction latency slightly increases in the integrated condition compared to auditory-only, indicating that singular sense can outperform integrated when one sensory cue is prominent and the other is greatly degraded. Based on the comparison between our measured data and the existing visual-only reaction time model [16], our results show statistical alignment with prior work and, therefore, remain credible. Those conclusions motivate us to develop a probabilistic computational model in Section 4.

4 A Multisensory Model for Novel Conditions

We develop a computational model using a widely applied computational framework, the drift-diffusion model (DDM, Section 4.1). In Section 4.2, we present a visual-auditory DDM-based formulation to depict the probability of users' reaction time given target contrast c (0 for audio-only), volume v (0 for visual-only), and eccentricity e. Finally, in Section 4.3, we fit the model with our data collection.

4.1 Drift-Diffusion Framework for Reaction Latency

Drift-diffusion model (DDM) [24, 63, 69] has been evidenced as an effective framework to model human decision-making over time in neuroscience, psychology, and economics [6, 43, 44, 69], especially the accuracy and latency of human action performance [16, 52, 70]. DDM simulates a stochastic diffusion process to accumulate evidence with a non-zero drift while we examine the content. The cumulative level at time t is modeled as A_t whose trajectory is a Brownian motion with mean drift rate $r > 0$:

$$A(t;r) = rt + W(t), t \geq 0, \tag{1}$$

where $W(t)$ is a noise term depicting human behavioral uncertainty.

Once A reaches a threshold $\alpha > 0$, a decision or action is initiated with a reset. Therefore, the corresponding reaction latency $L(\alpha;r)$ can be formulated as:

$$L(\alpha,r) = \inf_{t>0}\{A(t;r) \geq \alpha\}. \tag{2}$$

Instancing $W(t)$ as a Gaussian noise of standard deviation t derives:

$$L(\alpha,r) \sim \overline{W}\left(\frac{\alpha}{r}, \alpha^2\right), \tag{3}$$

where $\overline{W}$ is an inverse Gaussian distribution [23]. Next, we establish a DDM-based model that predicts users' reaction latency given the stimuli's visual-auditory characteristics.

4.2 Integrated Reaction Time Model

We discuss and formulate how our investigated multisensory stimuli characteristics influence and determine these parameters in a DDM.

Threshold α Intuitively, the evidence threshold α indicates how much evidence needs to be gathered, i.e., "how hard" a task is. In fact, prior literature has also suggested that it is primarily determined by the nature of a decision-making task [60, 64]. Interestingly in a multisensory setting, the presence of each individual modality fundamentally changes the task nature. Therefore, we model the task based on which one(s) of the visual and auditory cues are present. That is

$$\alpha \coloneqq \alpha(D(c), D(v)) \tag{4}$$

, where $D(x) \coloneqq \begin{cases} 0, x = 0 \\ 1, x > 0 \end{cases}$ is a Dirichlet function. Note that α is not correlated to the target eccentricity e, which does not alter the task nature. The sensory-dependent α plays a core role in integrating both uni- and multi-sensory tasks.

Drift rate r The difficulty of processing sensory inputs is usually modeled to determine the drift rate [16, 60]. Our pilot study analysis in Section 3.2 also shows the significance of the stimuli characteristics, including visual contrast c, audio volume v, and eccentricity e. Motivated by these observations, we model visual-only, auditory-only, and visual-auditory-integrated evidence accumulation rates as $r \coloneqq r(c,v,e)$.

Statistics \ Test Set	$\mathbf{P}_1$	$\mathbf{P}_2$	$\mathbf{P}_3$	$\mathbf{P}_4$	$\mathbf{P}_5$	$\mathbf{P}_6$
K-S test D	.06	.10	.10	.13	.10	.06
K-S test p	1.00	.99	.99	.95	.99	1.00
Wasserstein Dist	4.14	6.46	10.75	13.26	20.08	4.40

Statistics \ Test Set	$\mathbf{R}_1$	$\mathbf{R}_2$	$\mathbf{R}_3$	$\mathbf{R}_4$	$\mathbf{R}_5$	$\mathbf{R}_6$
K-S test D	.06	.06	.10	.10	.06	.10
K-S test p	1.00	1.00	.99	.99	1.00	.99
Wasserstein Dist	9.75	6.09	8.37	8.21	6.58	8.52

Table 3: *Evaluation of generalization to unseen users.* D and p denote the distance metric and its corresponding p-value for the K-S test, respectively. The null hypothesis cannot be rejected in the K-S test due to $p \gg .05$. With the mean reaction latency across all pilot study trials being 293ms, a Wasserstein distance(WD) of 8.88 indicates a 3.03% mean deviation.

Integration The computational model shall obtain the explicit representation of $\alpha(c,v)$ and $r(c,v,e)$ as presented above. There has been extensive neurological literature and hypothesis on the process of multimodal integration, such as time-windowing [13], or averaging [81]. While weighing individual modalities, performance optimality (minimal variances) have been considered [20,46], while in visual-audio tasks, both optimal and sub-optimal patterns might be observed [3,4,81]. However, instead of seeking a neurological explanation, we aim to develop a characterized and executable algorithm to assist downstream graphics applications. Therefore, we simplify the integrated model as a pure computational regression from our acquired data, such as both $\alpha(\cdot)$ and $r(\cdot)$ are formulated as a neural network with the learning process detailed below.

4.3 Fitting Model Parameters: r and α

Using the psychophysical data collected in Section 3, we compute and fit the DDM parameters $\{\alpha, r\}$ by noting their relationships with the mean and variance of the latency distributions L, which are modeled as inverse Gaussians:

$$\mathrm{E}[L] = \frac{\alpha}{r}, \ \mathrm{Var}[L] = \frac{\alpha}{r^3}. \tag{5}$$

Therefore, for each presented sensory identified by $\{\hat{D}_c, \hat{D}_v\}$ and the corresponding reaction latency data $\{\hat{L}\}$, we can approximate the evidence threshold α as:

$$\alpha(\hat{D}_c, \hat{D}_v) = \sqrt{\mathrm{E}[\hat{L}]^3 / \mathrm{Var}[\hat{L}]}. \tag{6}$$

Similarly, we can approximate the drift rates by sampling the subset $\{\hat{L}\}$ from our entire dataset as only those trials from condition $\hat{c}, \hat{v}, \hat{e}$:

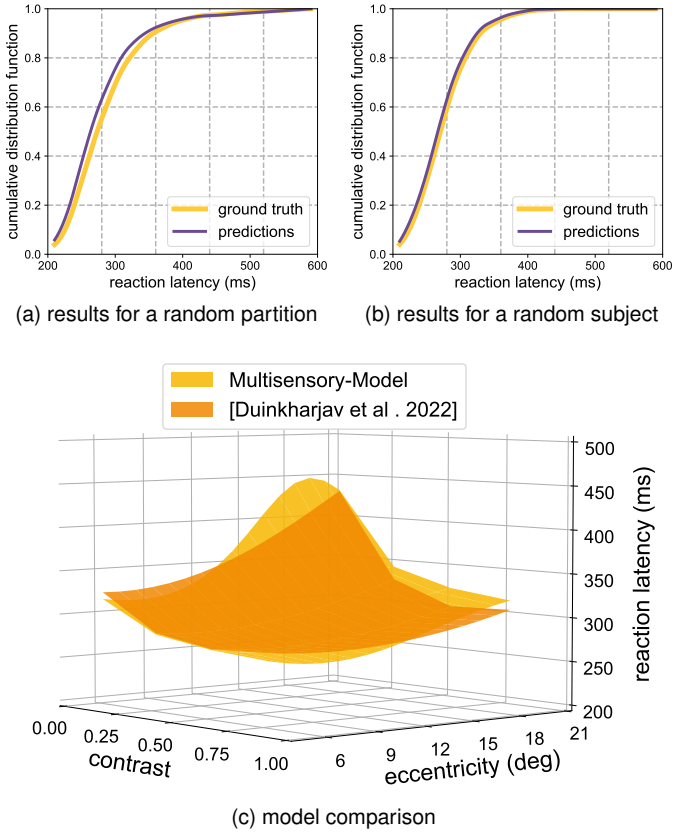
$$r(\hat{c}, \hat{v}, \hat{e}) = \sqrt{\mathrm{E}[\hat{L}] / \mathrm{Var}[\hat{L}]}. \tag{7}$$

However, there is a remarkable distinction in learning α and r: All α can be directly approximated from our data given which sensory inputs are presented, However, r analytically depends on the exact values of each sensory cue. To ensure the local smoothness, we formulate the underlying $r(c,v,e)$ as a Radial Basis Function Neural Network (RBFNN) and optimize the parameters therein using sample-computed target values. The final model is visualized in Figure 3. The drift rate component $r(\cdot,\cdot,\cdot)$ of our visual-auditory-integrated DDM, which is a function of the contrast c, volume v, and eccentricity e, was modeled using Radial Basis Function Neural Network. Specifically, we modeled our function as

$$r(c,v,e) = \sum_{i=1}^{N} \lambda_i \rho \left(\left\| \begin{bmatrix} 0 & \text{if } c = 0 \\ \log_{10}(c) & \text{if } c \neq 0 \\ v & \\ e & \end{bmatrix} - b_i \right\|, \sigma_i \right) \tag{8}$$

where ρ is the Gaussian Basis function and b_i/σ_i is the radial basis center/ Gaussian deviation for i_{th} neuron. Since we observed a fast drop down of the reaction latency in response to the contrast from 0.1 to 0.5, we performed a log function to the contrast value to stabilize the nonlinear relationship. We chose the $N = 25$, L_2 loss and Adam optimizer to train the RBFNN for 5k epochs. The learning rate started at .01 for neural network training, and a $10\times$ learning rate decay was performed at epoch 2.5k. Since our dataset of $\{(c,v,e), \hat{r}(c,v,e)\}$ pairs is fairly small, we performed full-batch training instead of mini-batch training, i.e. the whole dataset was used for each parameter update. For visual-only/ auditory-only conditions, we set $v = 0/c = 0$ as the input to our model. The training was conducted using a single Nvidia RTX 4090 GPU and took less than 1 minute to complete.

5 Evaluation



(a) results for a random partition (b) results for a random subject

(c) model comparison

Fig. 4: *Cross Validation and model comparison with prior work.* (a) shows the results for one held-out partition $\mathbf{R}_1$ with random sampling. (b) shows the testing results on $\mathbf{P}_1$ with the model trained using the other 5 participants' data. (c) shows the model prediction of our model in visual-only target compared with the prior literature [16].

We present a quantitative assessment of our model's prediction accuracy and generalization ability (Section 5.1). In addition, we conduct user studies to compare the predictions with user behaviors in VR driving and gaming scenarios (Section 5.2).

5.1 Model Accuracy and Generalizability

Dataset setup We split the reaction time data from Section 3 into non-overlapping training-testing partitions for cyclic cross-validation. We first randomly sample $1/6$ of the entire dataset to compose the evaluation set to measure the model's prediction accuracy and repeated six times ($\mathbf{R}_{1,2,\ldots,6}$). We also split the data by participants ($\mathbf{P}_{1,2,\ldots,6}$) to evaluate the model's generalizability to new users. Specifically, each of the 6 testing datasets contains all participant trials, with the remaining reserved to re-train our model.

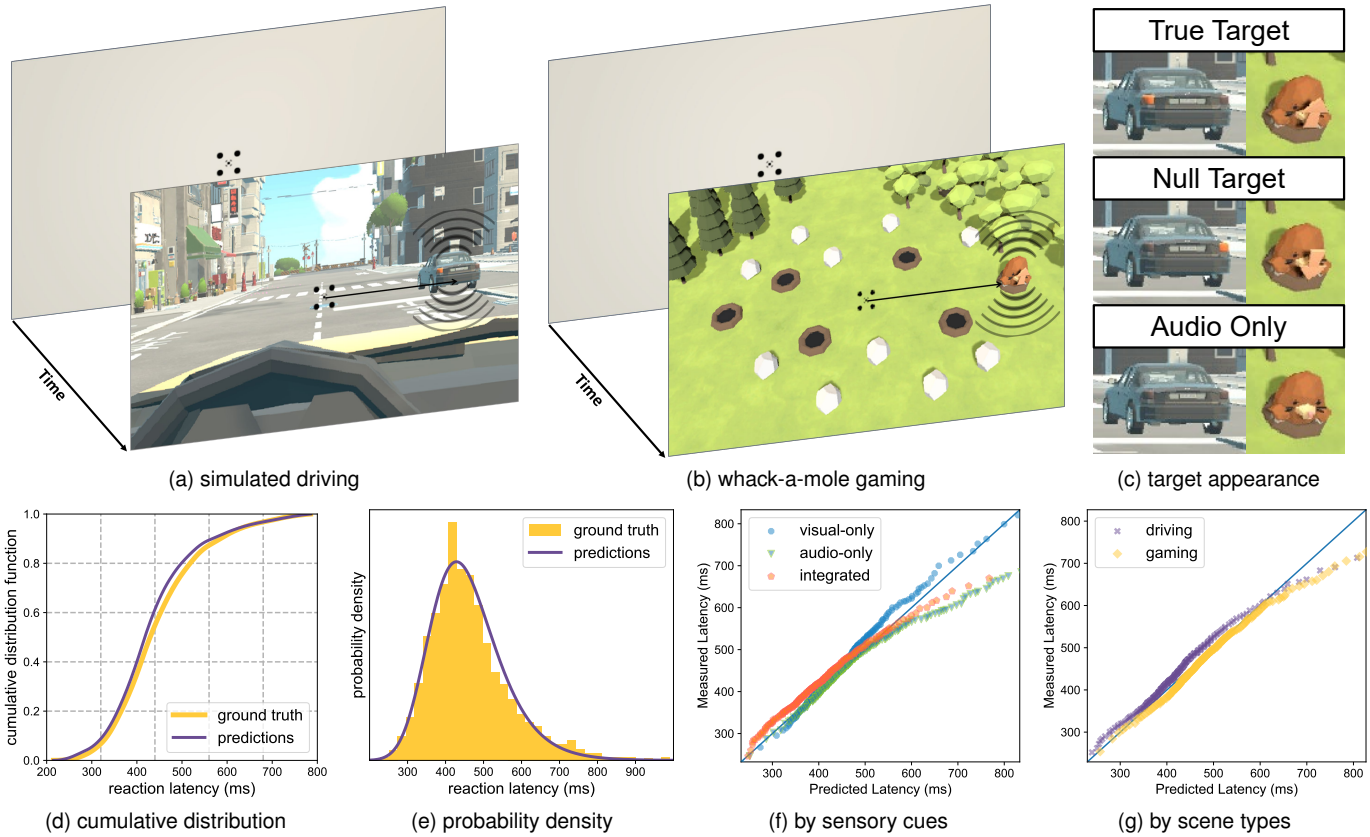


Fig. 5: *User study results.* (a)/(b) shows the scene and task. (c) indicates the stimuli appearance for true/null/audio-only target corresponding to (a) and (b). (d)/(e) shows the cumulative probability/probability density of aggregated data vs. model predictions. (f)/(g) visualizes the model prediction vs. user data split with different sensory cues/scene types in Q-Q plot.

Results and discussion When comparing between a testing dataset and model predictions, we adopt two metrics, the K-S test and Wasserstein distance (WD) [36]. WD provides a continuous-domain metric on the distinction between the two samples.

Figure 4 visualizes the cumulative distribution function (CDF) of measured data and model predictions for both random (**R**) and participant (**P**)-based partitions. Table 3 shows all numerical analysis results.

In summary, all K-S tests on the 12 partitions show $p \geq .95$, failing to reject the null hypothesis that the model prediction and the test data are drawn from the same distribution. The WD for random/participant-based partitions is $7.92 \pm 1.23 / 9.85 \pm 5.63$. Given the mean latency of 293 ms from the entire dataset, the model prediction may be approximated as exhibiting $< 2.70\% / 3.36\%$ deviation. These results validate the distributional alignment between test data and our model's predictions. Specifically, the analysis with random partitions confirms our model's prediction accuracy. Moreover, the analysis with participant-based partitions demonstrates the model's generalizability to new users in practical applications.

Comparison with prior literature As in Figure 4c, we further compare our model with prior open-source literature from Duinkharjav et al. [16]. We test with a median $f = 1.5$ spatial frequency (of their range $\{.5, 1, 2, 4\}$) as their frequency input , maintaining consistent contrast and eccentricity to evaluate and compare the reaction time predictions. Since the previous model only predicts the visual-only condition as our subset, we set $v = 0$ for our model. Our model aligns with the [16] across various contrasts and eccentricities ($D = .10, p = 1.00$) from a K-S test.

5.2 Study: Predicting Target-Reaching Latency in Multi-modal Naturalistic Scenarios

From the pilot study (Section 3.2) and our model (Section 4), we observed how the visual-audio interplay significantly influences the reaction time. This study evaluates our model's applicability in varied target-identification tasks, as well as realistic and representative scenarios – driving and gaming.

Participants and tasks Twelve adults (ages 21-46, 3 females) with normal or corrected-to-normal vision and hearing participated in the study. The appearance orders of the three conditions and the two scenes were randomized and counter-balanced across participants. Rather than examining a pair of identical targets in Section 3, we aim to validate the model with a different go/no-go target-reaching task [27, 56]. At the beginning of each trial, participants were shown a grey background with a fixation indicator at the center of the display. After a successful fixation, the scene with the task object appeared (a car or a mole). Then, after a randomized $.75 - 1$ second delay, an additional task-relevant stimulus (turning light/arrow with or without a spatial sound) appeared on the object to indicate true or null targets (see "scenes and setup" below for details). Participants were directed to make a saccade towards the true target or to maintain the gaze fixed if perceiving a null target. Similar to other go/no-go paradigms, we included null trials to ensure that participants were perceptually analyzing the targets instead of blindly responding to any changes. Please refer to the supplementary video for animated visualization.

Scenes and setup As shown in Figures 5a to 5c, we designed virtual environments to simulate two realistic scenarios. In the *driving scene*, the true target was the car with an overtaking turn signal (left/right-side car turning right/left light on) and/or with a spatially co-located car horn sound (60 dB); the null target was the car with the

opposite turn signal and/or a spatially misaligned sound. In the *whack-a-mole gaming scene*, the true target was the mole with an up arrow and/or a spatially co-located rustled sound (55 dB); the null target was the mole with a down arrow and/or a spatial misaligned rustled sound. Visual contrasts were approximated as the averaged Weber contrast from a three-layer Laplacian pyramid. Auditory volume was physically measured with the same device as Section 3. The hardware remained the same as Section 3.

Conditions For each scene, we also experimented with three sensory modalities (visual-only, audio-only, and integrated). In the audio-only condition, all targets were designed as visually unidentifiable shown in Figure 5c. For example, the mole didn't hold any arrow in its hand, and the car didn't turn on its turn signal. Therefore, the true target could only be identified by whether the auditory direction is on the same side as its visual appearance. Aiming to evaluate our model's generalizability in unseen scenarios, we intentionally selected different eccentricity levels from the pilot study (Section 3). Within each condition, we experimented with three horizontal eccentricities not presented in the pilot study (7°, 14° and 21°). We also introduced vertical eccentricities from 1° to 4° to enable validation of the generalization capability. Each participant performed 16 repeats $\times$ 3 eccentricities $\times$ 2 scenes $\times$ 3 sensory modalities, resulting 288×12 participants $= 3456$ trials. They also did $3456 \times 50\% = 1728$ null trials. Overall, we collected $3456 + 1728 = 5184$ trials in total.

Results and discussion In comparison with our model prediction and evaluation data, we provide Figure 5d and Figure 5e visualize the aggregated cumulative distribution and probability density of ground truth and predictions. In addition, Figure 5f and Figure 5g visualize the quantile-quantile (Q-Q) plot across different sensory modalities and scenes, which the samples approximately lying on the diagonal line indicate the distributions between the model and data are compared as similar. We also statistically validate our model with user data across modalities and scenes using goodness-of-fit (K-S) and distribution distance metrics (WD) as detailed in Table 4. K-S test failed to reject the null hypothesis with($D = .06$, $p = 1.00$) and WD$= 21.41$ for the entire dataset.

Statistics \ Test Set	All	Visual	Auditory	Integrated	Driving	Gaming
K-S D	.06	.10	.10	.10	.13	.06
K-S p	1.00	1.00	1.00	1.00	.96	1.00
WD	21.41	15.94	41.86	23.69	22.85	15.01

Table 4: *Evaluation of generalization to natural scenes.* The K-S test analysis and Wasserstein Distance across sensory cues/ scene types.

The analysis demonstrates our model's applicability in naturalistic scenes, maintaining consistent predictions with variations in scene types or novel target characteristics.

6 Application Case Studies

6.1 Multisensory Content Perceptibility Assessment

In scenarios like VR games (e.g., Beat Saber), immersive simulators (e.g., VR driving simulation), or content creation (e.g., commercial video), observers may overlook a quickly appearing target (such as a gaming block, a virtual pedestrian, or a promotional product). Our model can approximate the likelihood of users reacting to visual and/or auditory events in time, assessing the perceptibility of multisensory content. The assessment can thereby offer guidelines for the design of VR interactive applications and video creation.

Perceptibility approximation Our model predicts the probabilistic distribution of users' reaction latency for each target. Then, We approximate the likelihood of missing a target as the accumulated probability of the reaction time exceeding the target appearance duration. That is, for each target defined with $\{e, c, v\}$ and appearance duration T, the probability of missing a target is:

$$P_{miss}(e,c,v,T) \coloneqq P(L(\alpha(c,v), r(e,c,v)) > T)$$
$$= 1 - \Phi\left(\sqrt{\frac{\alpha^2}{T}}\left(\frac{rT}{\alpha} - 1\right)\right) - e^{2\alpha r}\Phi\left(-\sqrt{\frac{\alpha^2}{T}}\left(\frac{rT}{\alpha} + 1\right)\right), \quad (9)$$

where Φ is the cumulative distribution function of Gaussian.

Dataset Here, we demonstrate the proof-of-concept with a public video dataset for virtual reality applications. We experiment with a visual-audio omnidirectional video dataset from [9]. It contains 17 omnidirectional footage (2K-4K resolution @ 24-60 FPS) with first-order B-format ambisonic surround sound @ 4,000 Hz.

Experiment and results We rendered an omnidirectional video in VR ($120° \times 90°$). We extracted the targets via the instance segmentation of MMDetection [11] and estimated the audio sources' directions from the ambisonics (see Figure 6a). We visualized the probability of missing targets in different viewpoints and target appearance time shown as Figure 6c.

The visualization indicates that across visual/auditory/integrated conditions, varying viewpoints and time of appearance (240ms to 330ms) influence target-miss likelihood, with respective shifts of 17.3%/21.0%/8.3% and 42.9%/44.8%/62.0%.

6.2 Reaction-Optimized Layout Guidance for Multisensory Immersive Interface

VR/AR interfaces can be designed with not only visual but also auditory cues [1, 31]. So far, we have little quantitative guidance on how such multisensory cues should be spatially designed so that users react faster to the individual elements.

As shown in Figure 7a, we leverage our model to suggest whether audio or visual (if not both) cue is more beneficial in terms of triggering faster user reaction at given eccentricities. Specifically, the model compares the mean user reaction time to visual-only/audio-only cues given the contrast/volume and determines which sensory modality has a faster reaction within explicit eccentricity ranges. Their isosurface is visualized in Figure 7b. That is, for eccentricities (Z-axis) above/below the surface, auditory/visual interfaces are more beneficial for faster reaction time. Two example 2D slices from the 3D isosurface are shown in Figure 7c. The audio-only cue triggers the faster reaction in a larger area (audio-driven) while $c = 0.05$, but in a smaller area (visual-driven) while $v = 45$dB.

6.3 Suggesting Video Playback Speed

For faster content consumption, today's video streaming platforms commonly support accelerating playback speed (e.g., $0.25\times$ to $2\times$ on YouTube). However, we still lack a comprehensive guideline that captures the experiential consequences of these adjustments. For instance, if a video is accelerated, is there a risk of crucial events being missed before users can react to them? In fact, studies have identified drawbacks of accelerated video play on learning effectiveness [74]. Here, by leveraging the perceptibility assessment from Section 6.1, we can predict users' reaction time to a certain target during video watching. The higher playback speed reduces the target appearance time and makes it more likely to be overlooked. Therefore, our model can suggest maximal playback speed so that the video is not over-accelerated and that observers do not miss crucial events.

Dataset preprocessing We reused the dataset in Section 6.1, standardized the footage, and annotated the directions of visual and auditory sources. Since the intended targets may be arbitrary depending on individuals, we simplify by identifying the (visual/auditory/integrated) target as shown below. For consistency in both spatial and temporal resolutions across all frames, we normalized all footage by downsampling to 320×240 @ 10 Hz.

- Identifying visual targets: without loss of generality, we assume the potential visual target as human characters. Specifically, in each plane frame, we detect the two characters with the highest confidence scores by MMDetection [11]. The visual eccentricities are directly

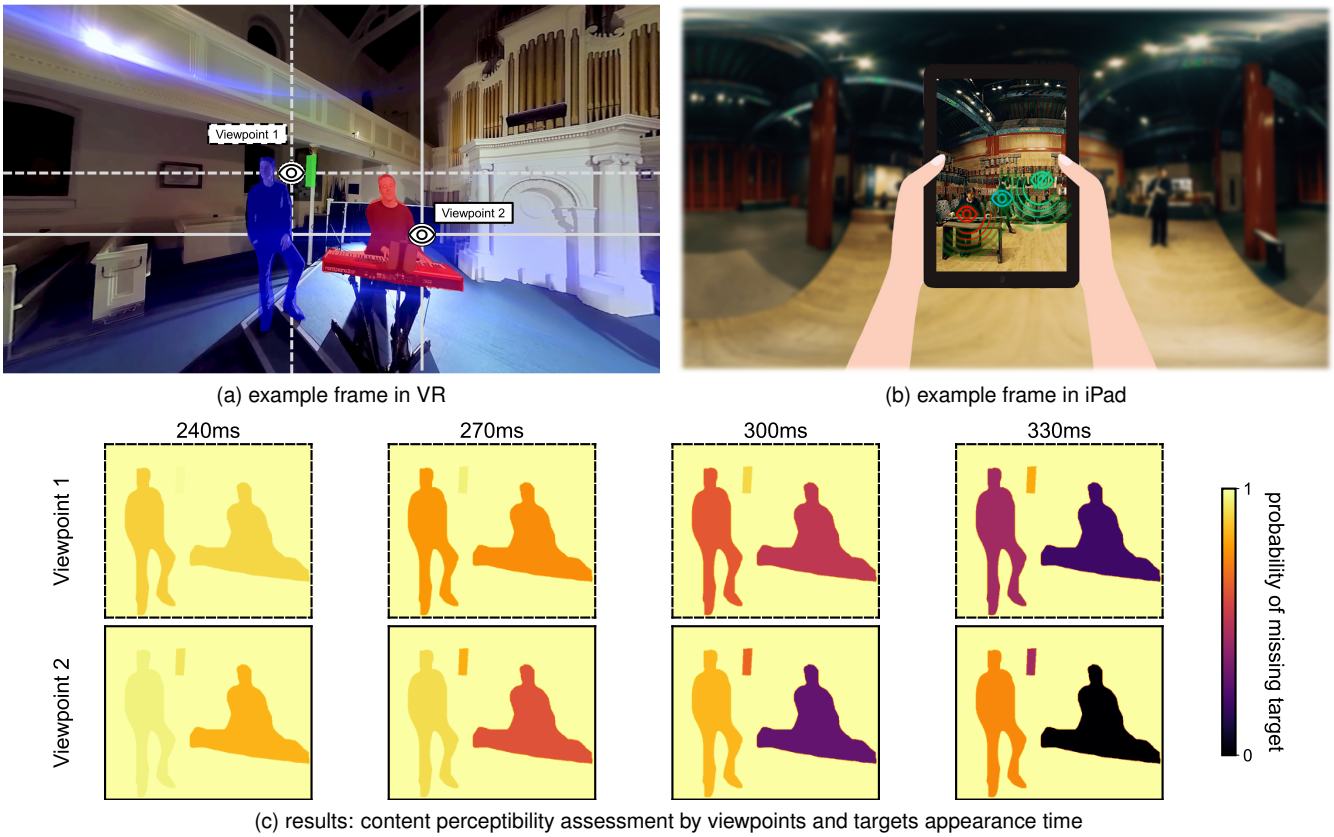


Fig. 6: *Application: analysis of video target-missing probability with a VR video dataset.* (a) shows the example frame rendered in VR with different viewpoints. The blue/green/red masks indicate the visual-only/audio-only/visual-auditory integrated targets. (b)shows the example of 40° FoV tablet-rendered frames (foreground) from the panorama video data (background). The blue/green/red icon indicates a visual-only character/an audio-only organ/a visual-auditory integrated piano player character. (c) visualizes the target-missing probabilities under various target appearance times and viewpoints in the colormap.

obtained from the detection. Their contrasts were approximated similarly to Section 5.2.

- Identifying auditory targets: The original ambisonic audio in the dataset is at a high frequency of 48,000 Hz, where each audio sample encodes a primary sound source [87]. To robustly determine the direction of the audio targets, we uniformly downsampled the audio to 150 samples per video frame and estimated the direction by averaging their horizontal eccentricities. [78]. We approximate the audio source's volume by mapping the ambisonic W-channel to the scale of 44 dB - 63 dB.
- Identifying visual-auditory integrated cues: we define a pair of visual and auditory targets as integrated if their spherical angular distance is less than 12°, otherwise, as visual-only or audio-only.
- Approximating target appearance duration: regardless of the video content, eye rotations change our retinal image and thus target visual eccentricities. Therefore, we approximate the duration of target appearance (defined by unchanging {e,c,v}) as an averaged fixation duration of approximately 400 ms [66]. While performing content-aware studies with eye-tracked datasets such as [73] is an interesting future avenue, it falls beyond our main focus in this research, as discussed in Section 7. Nonetheless, our model and the analysis framework are still applicable to target-adaptive durations.

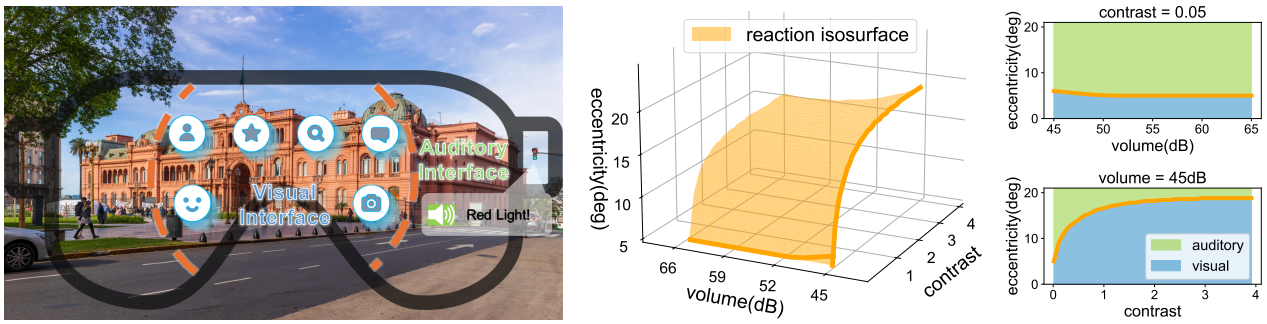
Experiment and results To simulate viewing with an iPad Pro tablet (22 cm × 28 cm) in portrait mode (3:4 aspect ratio), we uniformly re-framed 9 plane videos (40° × 53°) along the equator (0° latitude) of each panoramic video (as illustrated in Figure 6b) and obtained 1648/931/389 frames with visual-only/audio-only/integrated targets. For fair cross-comparisons, we randomly sampled 350 frames for each sensory condition. As a proof-of-concept analysis, we approximate the user's gaze at a fixed position of display before each target-reaching. Therefore, We assume each user's "current" eye gaze is at the display center with a 30cm viewing distance. Figure 8a show the overall statistics of the dataset. The predicted target-missing probabilities (P_{miss}) across the dataset is visualized in Figure 8b. Without loss of generality, we analyze with 40% chance of missing as an "acceptable" threshold, i.e., $P_{miss} \leq 0.4$. As in Figure 8b, we observe that $1.35 \pm .06\times/1.28 \pm .04\times/1.46 \pm .04\times$ to be the maximal playback speed for visual-only/audio-only/integrated targets.

6.4 Consequence of Audio Cancelling

With the growing production, many users wear noise-canceling headphones outdoors, even when crossing a street. Eliminating the audio from critical events such as approaching cars/trains has caused life-loss tragedies [54]. Here, we leverage the model to approximate the elevated likelihood of this risk. According to the "decision latency threshold" to avoid collisions (0.41 second [55]), our model shows that, in large eccentricity ($e = 20°$), the probabilities of missing a honking car (a visual-audio integrated target) may increase by 1.30% to 16.52% (assuming $c = 0.5$ and $v = 65$dB) if the audio cue is fully muted.

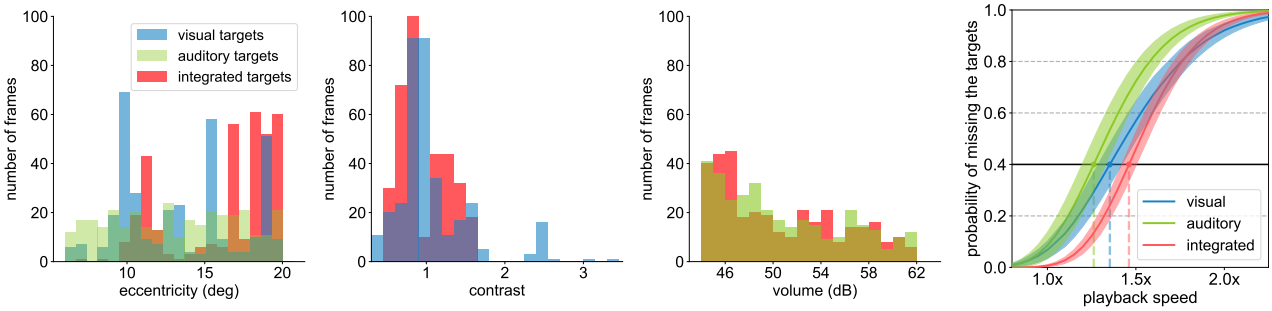
7 Limitations and Future Work

Stimuli characteristics Our focus is on predicting the effects of uni- and multi-sensory targets by varying their eccentricity, contrast, and volume, factors known to influence reaction performance [16, 21, 45]. Building a full-spectrum perceptual metric may require other factors, such as object motion and size, or visual/auditory frequencies These additions could substantially expand the number of dimensions for establishing a statistical model. A future direction is to



(a) example visual-auditory cues layout in AR interface (b) reaction isosurface (c) marked subset of (b)

Fig. 7: *Guiding visual-auditory AR interface using our model.* (a) shows an example placement layout of AR interface; designers may have options to place audio or visual cues at different eccentricities of the device. Our model provides quantitative design guidance and optimization in terms of triggering faster user reaction to individual cues. (b) shows our model-derived isosurface on which the visual-only and audio-only cues trigger identical reaction time. The X-/Y-/Z-axis indicates audio volume/visual contrast/eccentricity. Visual-only stimuli trigger faster/slower user reaction than the audio-only with the eccentricity below/above the surface. (c) visualizes two example 2D slices marked in (b). The blue/green area illustrates the cue (visual/auditory) triggering the faster reaction. We selected a visual-driven subset ($v = 45$dB) and an audio-driven subset ($c = 0.05$).



(a) data summary (b) results: playback speed

Fig. 8: *Application: suggesting playback speed.* (a) visualizes the statistical distribution of the content characteristics in our collected video dataset. Specifically, (b) shows the mean and standard deviation of target-missing probabilities (Y-axis), by varying playback speed (sampled at 30cm viewing distance), respectively.

study the cross-condition effects toward dimensionality reduction with principal component analysis, multidimensional scaling, or geodesic transformation [79] in the feature space.

Metrics For accuracy and consistency, we measure the reaction time by saccade onset, similar to [16]. However, studies also found discrepancies between saccade-based and limb-based latency, e.g., key-pressing [19]. As an exciting future direction, we plan to extend the research by exploring reaction markers such as controller tracking or more precise biometrics, including electromyography (EMG) [53].

Inter-target interference While performing the experiment and modeling, we ensured the prior knowledge of a single target throughout the field of view. The users were given the defined target before their reaction, while the non-targets were assumed as non-influential backgrounds. However, users may process arbitrary and/or multiple targets in natural tasks due to the selective attention The multi-target interference may further influence reaction patterns [58]. It may be particularly outstanding for auditory stimuli that are not spatially separable. Adapting the model with predicted saliency [73] may enable more precise prediction in the wild, especially for film-watching scenarios as we leveraged in Section 6.

Run-time adaptive video playback In the application of suggesting video playback speed (Section 6.3), the suggestion is adapted to both video content (visual-audio targets) and users' real-time gaze information, which we assumed as display center as a proof-of-concept analysis. To achieve pre-recorded and universal suggestion without eye-tracking, we plan to apply approximation approaches such as saliency-based gaze estimation [73].

8 Conclusion

We measured and predicted multisensory reaction latency. This is accomplished via a psychophysical study in VR and eye movement detection. The acquired data further derive a probabilistic model that propagates to new conditions. Beyond validating its accuracy and generalizability, we demonstrate the model's real-world application in predicting a viewer's likelihood of missing the target in video, VR/AR interface optimal applicable range for each sensory modality, and noise-cancelling-induced risk. The discoveries provide quantitative evidence for user-aware content creation and consumption, VR/AR and interactive applications, and noise-canceling consequences. We hope the research will develop new attention in the community concerning the ubiquitous visual-auditory interplay and how it influences content consumers' and VR/AR users' behaviors. For instance, how do we determine whether users are aware of critical events in VR? Also, how could the VR/AR designer place the interface for better interactive cueing? Also, how should video-sharing platforms recommend proper/maximal playback speed to preserve the perceived content?

Acknowledgments

This work has been partially supported by grant PID2022-141539NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU, an academic gift from Meta, and the National Science Foundation (NSF) grants #2225861 and #2232817.

References

[1] R. Altosaar, A. Tindale, and J. Doyle. Physically colliding with music: Full-body interactions with an audio-only virtual reality interface.

In *Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction*, pp. 553–557, 2019. 7

[2] E. Arabadzhiyska, O. T. Tursun, K. Myszkowski, H.-P. Seidel, and P. Didyk. Saccade landing position prediction for gaze-contingent rendering. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 1

[3] D. H. Arnold, K. Petrie, C. Murray, and A. Johnston. Suboptimal human multisensory cue combination. *Scientific Reports*, 9(1):1–11, 2019. 5

[4] P. W. Battaglia, R. A. Jacobs, and R. N. Aslin. Bayesian integration of visual and auditory signals for spatial localization. *Josa a*, 20(7):1391–1397, 2003. 5

[5] U. R. Beierholm, S. R. Quartz, and L. Shams. Bayesian priors are encoded independently from likelihoods in human multisensory perception. *Journal of vision*, 9(5):23–23, 2009. 2

[6] S. Bitzer, H. Park, F. Blankenburg, and S. J. Kiebel. Perceptual decision making: drift-diffusion model is equivalent to a bayesian model. *Frontiers in human neuroscience*, 8:102, 2014. 4

[7] A. Bobu, D. R. Scobee, J. F. Fisac, S. S. Sastry, and A. D. Dragan. Less is more: Rethinking probabilistic models of human behavior. In *Proceedings of the 2020 acm/ieee international conference on human-robot interaction*, pp. 429–437, 2020. 2

[8] D. Burr and D. Alais. Combining visual and auditory information. *Progress in brain research*, 155:243–258, 2006. 2

[9] F.-Y. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic. Audio-visual perception of omnidirectional video for virtual reality applications. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, 2020. doi: 10.1109/ICMEW46912.2020.9105956 7

[10] G. Charbonneau, M. Véronneau, C. Boudrias-Fournier, F. Lepore, and O. Collignon. The ventriloquist in periphery: impact of eccentricity-related reliability on audio-visual localization. *Journal of Vision*, 13(12):20–20, 2013. 2

[11] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, et al. Mmdetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 7

[12] S. Chen, B. Duinkharjav, X. Sun, L.-Y. Wei, S. Petrangeli, J. Echevarria, C. Silva, and Q. Sun. Instant reality: Gaze-contingent perceptual optimization for 3d virtual reality streaming. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2157–2167, 2022. 2

[13] H. Colonius and A. Diederich. Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *Journal of cognitive neuroscience*, 16(6):1000–1009, 2004. 2, 5

[14] H. Colonius and A. Diederich. The optimal time window of visual-auditory integration: a reaction time analysis. *Frontiers in integrative neuroscience*, p. 11, 2010. 1, 2

[15] J. Drugowitsch, G. C. DeAngelis, E. M. Klier, D. E. Angelaki, and A. Pouget. Optimal multisensory decision-making in a reaction-time task. *Elife*, 3:e03005, 2014. 2

[16] B. Duinkharjav, P. Chakravarthula, R. Brown, A. Patney, and Q. Sun. Image features influence reaction time: a learned probabilistic perceptual model for saccade latency. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022. 1, 2, 3, 4, 5, 6, 8, 9

[17] B. Duinkharjav, B. Liang, A. Patney, R. Brown, and Q. Sun. The shortest route is not always the fastest: Probability-modeled stereoscopic eye movement completion time in vr. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023. 2

[18] R. Engbert and K. Mergenthaler. Microsaccades are triggered by low retinal image slip. *Proceedings of the National Academy of Sciences*, 103(18):7192–7197, 2006. 2, 3

[19] E. J. Engelken, K. W. Stevens, and J. D. Enderle. Relationships between manual reaction time and saccade latency in response to visual and auditory stimuli. Technical report, SCHOOL OF AEROSPACE MEDICINE BROOKS AFB TX, 1991. 3, 9

[20] M. O. Ernst and M. S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, 2002. 2, 4, 5

[21] J. Farrell. The effect of increasing music volume on reaction time. *The Journal of Science and Medicine*, 2021. 8

[22] C. R. Fetsch, A. Pouget, G. C. DeAngelis, and D. E. Angelaki. Neural correlates of reliability-based cue weighting during multisensory integration. *Nature neuroscience*, 15(1):146–154, 2012. 2

[23] J. L. Folks and R. S. Chhikara. The inverse gaussian distribution and its statistical application—a review. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):263–275, 1978. 4

[24] D. Fudenberg, W. Newey, P. Strack, and T. Strzalecki. Testing the drift-diffusion model. *Proceedings of the National Academy of Sciences*, 117(52):33141–33148, 2020. 2, 4

[25] K. Fujiwara, K. Kunita, and H. Watanabe. Sports exercise effect on shortening of saccadic reaction time associated with neck extensor muscle activity. *International journal of sports medicine*, 27(10):792–797, 2006. 2

[26] D. N. Gabriel, D. P. Munoz, and S. E. Boehnke. The eccentricity effect for auditory saccadic reaction times is independent of target frequency. *Hearing Research*, 262(1-2):19–25, 2010. 2

[27] P. Gomez, R. Ratcliff, and M. Perea. A model of the go/no-go task. *Journal of Experimental Psychology: General*, 136(3):389, 2007. 6

[28] L. W. Gregg and W. Brogden. The relation between reaction time and the duration of the auditory stimulus. *Journal of Comparative and Physiological Psychology*, 43(5):389, 1950. 1

[29] R. Gruen, E. Ofek, A. Steed, R. Gal, M. Sinclair, and M. Gonzalez-Franco. Measuring system visual latency through cognitive latency on video see-through ar devices. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 791–799. IEEE, 2020. 2

[30] F. M. Henry. Stimulus complexity, movement complexity, age, and sex in relation to reaction latency and speed in limb movements. *Research Quarterly. American Association for Health, Physical Education and Recreation*, 32(3):353–366, 1961. 3

[31] L. Hespanhol, O. Bown, J. Cao, and M. Tomitsch. Evaluating the effectiveness of audio-visual cues in immersive user interfaces. In *Proceedings of the 25th Australian computer-human interaction conference: augmentation, application, innovation, collaboration*, pp. 569–572, 2013. 7

[32] S. Hidaka, Y. Manaka, W. Teramoto, Y. Sugita, R. Miyauchi, J. Gyoba, Y. Suzuki, and Y. Iwaya. Alternation of sound location induces visual motion perception of a static object. *PLoS One*, 4(12):e8188, 2009. 2

[33] A. Hirway, Y. Qiao, and N. Murray. Spatial audio in 360° videos: does it influence visual attention? In *Proceedings of the 13th ACM Multimedia Systems Conference*, pp. 39–51, 2022. 1

[34] L. Hsiao, B. Krajancich, P. Levis, G. Wetzstein, and K. Winstein. Towards retina-quality vr video streaming: 15ms could save you 80% of your bandwidth. *ACM SIGCOMM Computer Communication Review*, 52(1):10–19, 2022. 2

[35] A. Jain, R. Bansal, A. Kumar, and K. Singh. A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students. *International Journal of Applied and Basic Medical Research*, 5(2):124, 2015. 1

[36] L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management science*, 6(4):366–422, 1960. 6

[37] A. S. Kaplanyan, A. Sochenov, T. Leimkühler, M. Okunev, T. Goodall, and G. Rufo. Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. 2

[38] S. Kasahara, J. Nishida, and P. Lopes. Preemptive action: Accelerating human reaction using electrical muscle stimulation without compromising agency. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–15, 2019. 2

[39] R. J. Kosinski. A literature review on reaction time. *Clemson University*, 10(1):337–344, 2008. 3

[40] B. Krajancich, P. Kellnhofer, and G. Wetzstein. Optimizing depth perception in virtual and augmented reality through gaze-contingent stereo rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–10, 2020. 2

[41] B. Krajancich, P. Kellnhofer, and G. Wetzstein. A perceptual model for eccentricity-dependent spatio-temporal flicker fusion and its applications to foveated graphics. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021. 2

[42] B. Krajancich, P. Kellnhofer, and G. Wetzstein. Towards attention-aware rendering for virtual and augmented reality. *arXiv preprint arXiv:2302.01368*, 2023. 2

[43] I. Krajbich, D. Lu, C. Camerer, and A. Rangel. The attentional drift-diffusion model extends to simple purchasing decisions. *Frontiers in psychology*, 3:193, 2012. 4

[44] I. Krajbich and A. Rangel. Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, 108(33):13852–13857, 2011. 2, 4

[45] M. Lisi, J. A. Solomon, and M. J. Morgan. Gain control of saccadic eye movements is probabilistic. *Proceedings of the National Academy of Sciences*, 116(32):16137–16142, 2019. 1, 2, 8

[46] W. J. Ma and A. Pouget. Linking neurons to behavior in multisensory perception: A computational review. *Brain research*, 1242:4–12, 2008. 5

[47] S. Malpica, A. Serrano, J. Guerrero-Viu, D. Martin, E. Bernal, D. Gutierrez, and B. Masia. Auditory stimuli degrade visual performance in virtual reality. In *ACM SIGGRAPH 2022 Posters*, pp. 1–2. 2022. 1

[48] R. K. Mantiuk, M. Ashraf, and A. Chapiro. stelacsf: a unified model of contrast sensitivity as the function of spatio-temporal frequency, eccentricity, luminance and area. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 2

[49] R. K. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021. 2

[50] D. Martin, S. Malpica, D. Gutierrez, B. Masia, and A. Serrano. Multimodality in vr: A survey. *ACM Computing Surveys (CSUR)*, 54(10s):1–36, 2022. 1, 2

[51] F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951. 3

[52] M. Milosavljevic, J. Malmaud, A. Huth, C. Koch, and A. Rangel. The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment and Decision making*, 5(6):437–449, 2010. 4

[53] E. A. Y. Murakami. Reaction time and emg measurement applied to human control modeling. *Measurement*, 43(5):675–683, 2010. 9

[54] Newsweek. Teen killed by train while wearing noise-canceling headphones. https://www.newsweek.com/teen-killed-train-wearing-noise-canceling-headphones-1803738, 2023. Accessed: 2023-01-21. 8

[55] B. Nie, Q. Li, S. Gan, B. Xing, Y. Huang, and S. E. Li. Safety envelope of pedestrians upon motor vehicle conflicts identified via active avoidance behaviour. *Scientific reports*, 11(1):3996, 2021. 8

[56] B. A. Nosek and M. R. Banaji. The go/no-go association task. *Social cognition*, 19(6):625–666, 2001. 6

[57] T. Ohshiro, D. E. Angelaki, and G. C. DeAngelis. A normalization model of multisensory integration. *Nature neuroscience*, 14(6):775–782, 2011. 2

[58] E. Ophir, C. Nass, and A. D. Wagner. Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences*, 106(37):15583–15587, 2009. 9

[59] N. Padmanaban, T. Ruban, V. Sitzmann, A. M. Norcia, and G. Wetzstein. Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos. *IEEE transactions on visualization and computer graphics*, 24(4):1594–1603, 2018. 1

[60] E. M. Palmer, T. S. Horowitz, A. Torralba, and J. M. Wolfe. What are the shapes of response time distributions in visual search? *Journal of experimental psychology: human perception and performance*, 37(1):58, 2011. 4

[61] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn. Towards foveated rendering for gaze-tracked virtual reality. *ACM Transactions on Graphics (TOG)*, 35(6):1–12, 2016. 2

[62] M. I. Posner, M. J. Nissen, and R. M. Klein. Visual dominance: an information-processing account of its origins and significance. *Psychological review*, 83(2):157, 1976. 2

[63] R. Ratcliff and G. McKoon. The diffusion decision model: theory and data for two-choice decision tasks. *Neural computation*, 20(4):873–922, 2008. 4

[64] B. A. Reddi, K. N. Asrress, and R. H. Carpenter. Accuracy, information, and response time in a saccadic decision task. *Journal of neurophysiology*, 90(5):3538–3546, 2003. 4

[65] W. Ritter, R. Simson, H. G. Vaughan Jr, and D. Friedman. A brain event related to the making of a sensory discrimination. *Science*, 203(4387):1358–1361, 1979. 2

[66] J. A. Roberts, G. Wallis, and M. Breakspear. Fixational eye movements during viewing of dynamic natural scenes. *Frontiers in psychology*, 4:797, 2013. 8

[67] F. Schmiedek, K. Oberauer, O. Wilhelm, H.-M. Süß, and W. W. Wittmann. Individual differences in components of reaction time distributions and their relations to working memory and intelligence. *Journal of experimental psychology: General*, 136(3):414, 2007. 3

[68] J. A. Seideman, T. R. Stanford, and E. Salinas. Saccade metrics reflect decision-making dynamics during urgent choices. *Nature communications*, 9(1):2907, 2018. 2

[69] M. N. Shadlen and R. Kiani. Decision making as a window on cognition. *Neuron*, 80(3):791–806, 2013. 4

[70] N. Shahar, T. U. Hauser, M. Moutoussis, R. Moran, M. Keramati, N. Consortium, and R. J. Dolan. Improving the reliability of model-based decision-making estimates in the two-stage decision task with reaction-times and drift-diffusion modeling. *PLoS computational biology*, 15(2):e1006803, 2019. 4

[71] J. Shelton and G. P. Kumar. Comparison between auditory and visual simple reaction times. *Neuroscience and medicine*, 1(01):30–32, 2010. 2

[72] M. Shinn, N. H. Lam, and J. D. Murray. A flexible framework for simulating and fitting generalized drift-diffusion models. *ELife*, 9:e56938, 2020. 2

[73] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018. 1, 8, 9

[74] K. Song, A. Chakraborty, M. Dawson, A. Dugan, B. Adkins, and C. Doty. Does the podcast video playback speed affect comprehension for novel curriculum delivery? a randomized trial. *Western Journal of Emergency Medicine*, 19(1):101, 2018. 7

[75] J. Spjut, A. Madhusudan, B. Watson, B. Boudaoud, and J. Kim. The esports frontier: Rendering for competitive games. *arXiv preprint arXiv:2208.11774*, 2022. 2

[76] H. Summala. Brake reaction times and driver behavior analysis. *Transportation Human Factors*, 2(3):217–226, 2000. 2

[77] Q. Sun, F.-C. Huang, J. Kim, L.-Y. Wei, D. Luebke, and A. Kaufman. Perceptually-guided foveation for light field displays. *ACM Transactions on Graphics (TOG)*, 36(6):1–13, 2017. 2

[78] V. Tabry, R. J. Zatorre, and P. Voss. The influence of vision on sound localization abilities in both the horizontal and vertical planes. *Frontiers in psychology*, 4:932, 2013. 8

[79] J. B. Tenenbaum, V. d. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000. 9

[80] L. Thaler, A. C. Schütz, M. A. Goodale, and K. R. Gegenfurtner. What is the best fixation target? the effect of target shape on stability of fixational eye movements. *Vision research*, 76:31–42, 2013. 2

[81] B. M. Turner, J. Gao, S. Koenig, D. Palfy, and J. L. McClelland. The dynamics of multimodal integration: The averaging diffusion model. *Psychonomic bulletin & review*, 24:1819–1843, 2017. 2, 5

[82] C. Tursun and P. Didyk. Perceptual visibility model for temporal contrast changes in periphery. *ACM Transactions on Graphics*, 42(2):1–16, 2022. 2

[83] M. Ursino, A. Crisafulli, G. Di Pellegrino, E. Magosso, and C. Cuppini. Development of a bayesian estimator for audio-visual integration: a neurocomputational study. *Frontiers in computational neuroscience*, 11:89, 2017. 2

[84] R. J. Van Beers. The sources of variability in saccadic eye movements. *Journal of Neuroscience*, 27(33):8757–8770, 2007. 1, 3

[85] E. Van der Burg, C. N. Olivers, A. W. Bronkhorst, and J. Theeuwes. Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 34(5):1053, 2008. 2

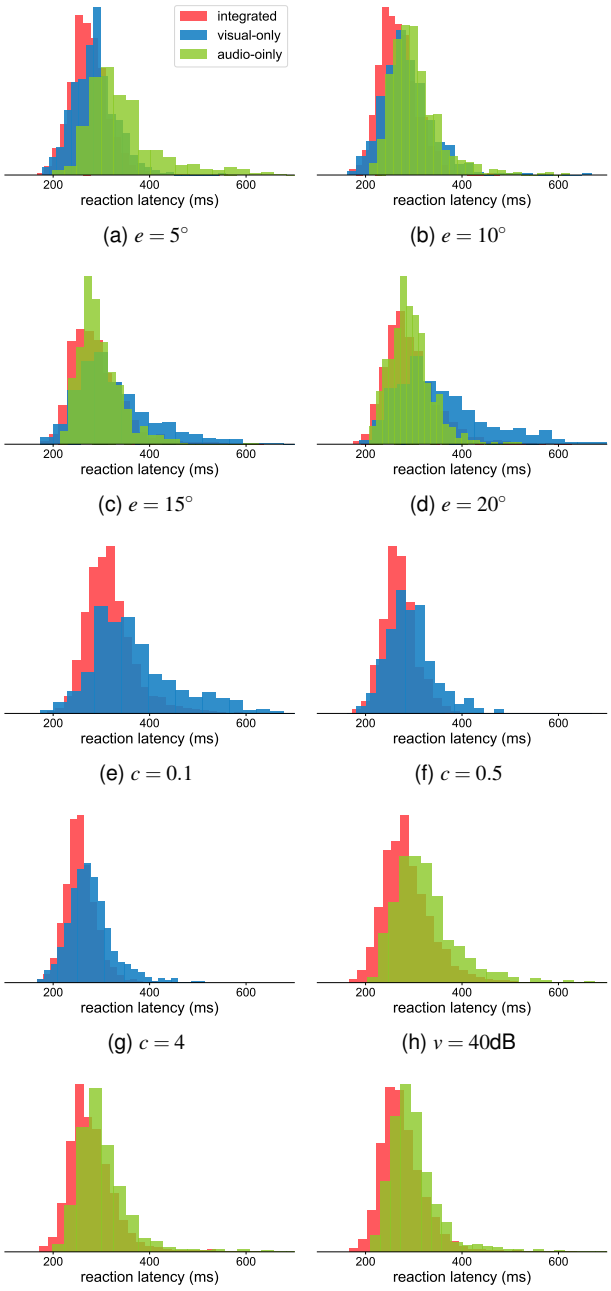
[86] D. R. Walton, R. K. Dos Anjos, S. Friston, D. Swapp, K. Akşit, A. Steed, and T. Ritschel. Beyond blur: Real-time ventral metamers for foveated rendering. *ACM Transactions on Graphics*, 40(4):1–14, 2021. 2

[87] J. Wierzbicki, P. Małecki, and J. Wiciak. Localization of the sound source with the use of the first-order ambisconic microphone. *Acta Physica Polonica A*, 123(6):1114–1117, 2013. 8

[88] Y. Wu, K. Kihara, Y. Takeda, T. Sato, M. Akamatsu, S. Kitazaki, K. Nakagawa, K. Yamada, H. Oka, and S. Kameyama. Eye movements predict driver reaction time to takeover request in automated driving: A real-vehicle study. *Transportation research part F: traffic psychology and behaviour*, 81:355–363, 2021. 2

[89] L. Yao and C. Peck. Saccadic eye movements to visual and auditory targets. *Experimental brain research*, 115:25–34, 1997. 2

A Pilot Study Data Collection in Histogram



(a) $e = 5°$ (b) $e = 10°$

(c) $e = 15°$ (d) $e = 20°$

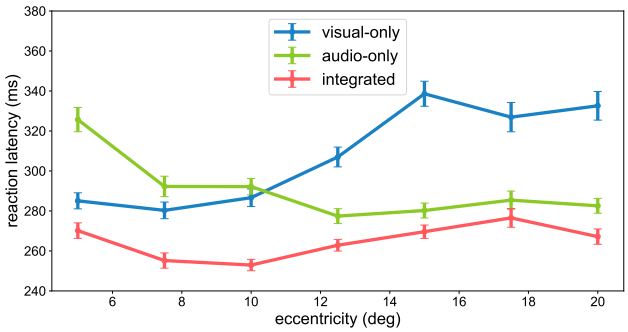
(e) $c = 0.1$ (f) $c = 0.5$

(g) $c = 4$ (h) $v = 40$dB

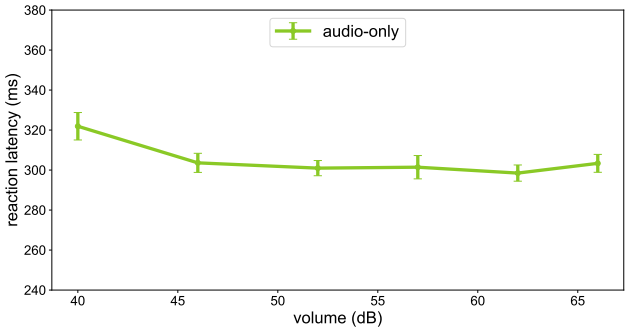
(i) $v = 52$dB (j) $v = 66$dB

Fig. 9: *Pilot study results in histograms.* Histograms show the detailed distribution of aggregated user behaviour data at each eccentricity, contrast, and volume.

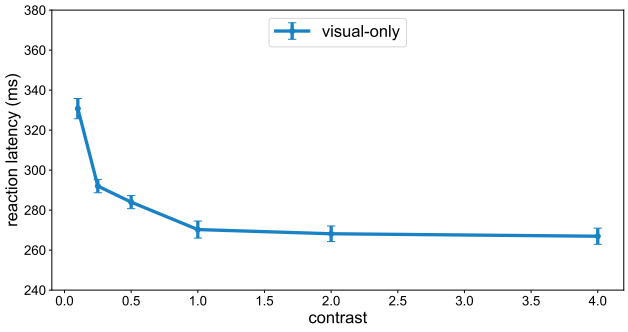
B Preliminary Study



(a) effects of eccentricity



(b) effects of volume



(c) effects of contrast

Fig. 10: *Preliminary study results.* We recruited four users (ages 23-29, 1 female) with normal or corrected-to-normal vision and hearing. The preliminary study contains the same experimental setup and protocol with Section 3 but dense sample points with eccentricity, volume, and contrast. This experiment was to help us design our pilot study by selecting the maximum/ minimum range and neutral challenging sample points for stimuli characteristics. We guaranteed a balanced number of sample points across three dimensions to prevent user fatigue.