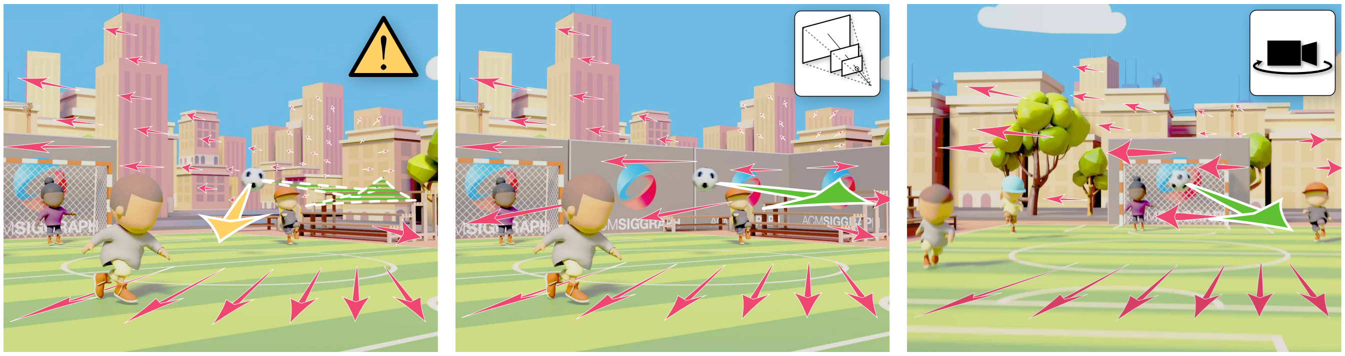
Evaluating Visual Perception of Object Motion in Dynamic Environments

BUDMONDE DUINKHARJAV, New York University, USA
JENNA KANG, Georgia Tech, USA and New York University, USA
GAVIN S. P. MILLER, Adobe Research, USA
CHANG XIAO, Adobe Research, USA
QI SUN, New York University, USA



(a) object motion misestimation

(b) mitigation via change in scene design

(c) mitigation via change in camera angle

Fig. 1. *Predicting and reducing human mis-estimation of object motion.* (a) shows a common animation and gaming scenario where the observer has to estimate the motion of a soccer ball in a 3D scene. Due to perceptual errors in estimating motion vectors in screen-displayed graphics, we may significantly misestimate the direction of the soccer ball (estimated leftward yellow vector vs. actual rightward green vector). For an animated example, see the *supplementary video*. To mitigate motion estimation inaccuracies and enhance task performance and perceptual fidelity, our perceptual model measures the correlation between estimation error, and scene dynamics and content. These predicted measures lead to content designs optimized for minimizing perceptual errors in motion estimation, such as by adjusting scene depth with an added wall (b) or by altering camera angles (c).

Precisely understanding how objects move in 3D is essential for broad scenarios such as video editing, gaming, driving, and athletics. With screen-displayed computer graphics content, users only perceive limited cues to judge the object motion from the on-screen optical flow. Conventionally, visual perception is studied with stationary settings and singular objects. However, in practical applications, we—the observer—also move within complex scenes. Therefore, we must extract object motion from a combined optical flow displayed on screen, which can often lead to mis-estimations due to perceptual ambiguities.

We measure and model observers' perceptual accuracy of object motions in dynamic 3D environments, a universal but under-investigated scenario in computer graphics applications. We design and employ a crowdsourcing-based psychophysical study, quantifying the relationships among patterns of scene dynamics and content, and the resulting perceptual judgments of object motion direction. The acquired psychophysical data underpins a model for generalized conditions. We then demonstrate the model's guidance ability to significantly enhance users' understanding of task object motion in gaming and animation design. With applications in measuring and compensating for object motion errors in video and rendering, we hope the research establishes a new frontier for understanding and mitigating perceptual errors caused by the gap between screen-displayed graphics and the physical world.

CCS Concepts: • **Computing methodologies** → **Perception**; **Motion processing**.

Additional Key Words and Phrases: motion perception, perceptual error, optical flow, camera control

ACM Reference Format:
Budmonde Duinkharjav, Jenna Kang, Gavin S. P. Miller, Chang Xiao, and Qi Sun. 2024. Evaluating Visual Perception of Object Motion in Dynamic Environments. *ACM Trans. Graph.* 43, 6, Article 277 (December 2024), 11 pages. https://doi.org/10.1145/3687912

1 INTRODUCTION

When driving on the road, we must accurately estimate and respond to the motion of various objects in a dynamic environment, including other vehicles and pedestrians. How users perceive object motion is also a universal metric in computer graphics applications, such as guiding camera trajectories in video playback [Kang and Cho 2019], controlling game difficulties [Caroux et al. 2013], compressing videos [Furht et al. 2012], and reducing simulator sickness [Hu et al. 2019; Park et al. 2022]. In these real-world scenarios, both the objects and we ourselves may move within dynamic 3D environments. In such situations, extracting scene-relative object motion solely from the mixed and anisotropic optical flow on the screen can lead to

Authors' addresses: Budmonde Duinkharjav, budmonde@gmail.com, New York University, Brooklyn, USA; Jenna Kang, jennajk@gatech.edu, Georgia Tech, Atlanta, USA and New York University, Brooklyn, USA; Gavin S. P. Miller, gmiller@adobe.com, Adobe Research, San Jose, USA; Chang Xiao, cxiao@adobe.com, Adobe Research, San Jose, USA; Qi Sun, qisun@nyu.edu, New York University, Brooklyn, USA.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in *ACM Transactions on Graphics*, https://doi.org/10.1145/3687912.

misinterpretations due to its ambiguous nature [Dokka et al. 2019]. Therefore, we ask, "How accurately can we perceive moving objects in scenes featuring different motion dynamics?".

Prior studies have observed that perceptual errors can occur when estimating object movements during self-movements [Dokka et al. 2019; Xing and Saunders 2022] and in 3D scenes [Cornilleau-Pérès and Gielen 1996; Van den Berg and Brenner 1994a]. These errors are attributed to how humans exhibit inherent cognitive biases that mislead our motion estimates when crucial perceptual cues are absent (e.g., vestibular cues obtained from whole-body movement) [Xie et al. 2020]. However, to provide design guidance in downstream graphics applications, a quantified understanding of the variability of these errors across different scene dynamics is still missing. Filling this knowledge gap poses a remarkable challenge due to the need of sampling a diverse range of conditions, conducting repeated experiments, and involving a wide population to account for variations in individuals' sensory and perceptual variances [Xing and Saunders 2022].

In this paper, we measure and analyze the errors in our visual perception of screen-displayed object motion, particularly in relation to concurrent global scene movements which result in dynamic environments. To this aim, we present a series of large-scale psychophysical studies comprising over 10,000 trials, which correlate object motion perception and scene dynamics characterized by scene movements and content depths. We employ and validate a crowdsourcing approach to tackle the unique challenges posed by the need for large sample sizes in both population and trial repetitions.

Additionally, we also showcase how the model can guide animation and game design to reduce perceived errors in object motion by viewers. We hope the research will contribute to a new frontier in the computer graphics community, focusing on understanding the visual performance limitations introduced by displays and exploring design strategies to compensate for them. Refer to www.github.com/NYU-ICL/motion-estimation for the anonymized data and model implementations.

Limitation overview. We present this research to raise awareness of how designs of scene content and dynamics can affect human perceptual accuracy of motion in common computer graphics contexts. However, we do not propose automated algorithms that would directly apply these discoveries to optimize user performance.

2 RELATED WORK & TERMINOLOGY

2.1 Image-Space Motion Description and Estimation

Optical flow is often used to depict the spatio-temporal motion of video [Huang et al. 1995; Neumann 1984]. It can be estimated by forward analysis [Beauchemin and Barron 1995; Xu et al. 2011] or learning-based approaches [Hu et al. 2018; Hui et al. 2018]. Typically, global scene motions caused by camera movements induce global optical flows, while isolated object motions result in more localized patterns. Therefore, optical flows are instrumental for detecting object motion in images and 3D space [Talukder and Matthies 2004].

As visualized in Figure 2a, optical flow patterns of rigid translational movements exhibit a stationary on-screen point, known as the *Focus of Expansion* (FOE), from which flow vectors diverge [Jain 1983]. The locations of these FOEs serve as visual cues that humans use to infer the direction of object and scene motions [Jain 1984; Warren Jr and Hannon 1988]. Thus, prior works have examined the relationship between the dynamics of the FOE, which capture the observer-relative scene motions in 3D space, and human perception of motion [Jain 1984; Lappe et al. 1999; Warren Jr and Hannon 1988].

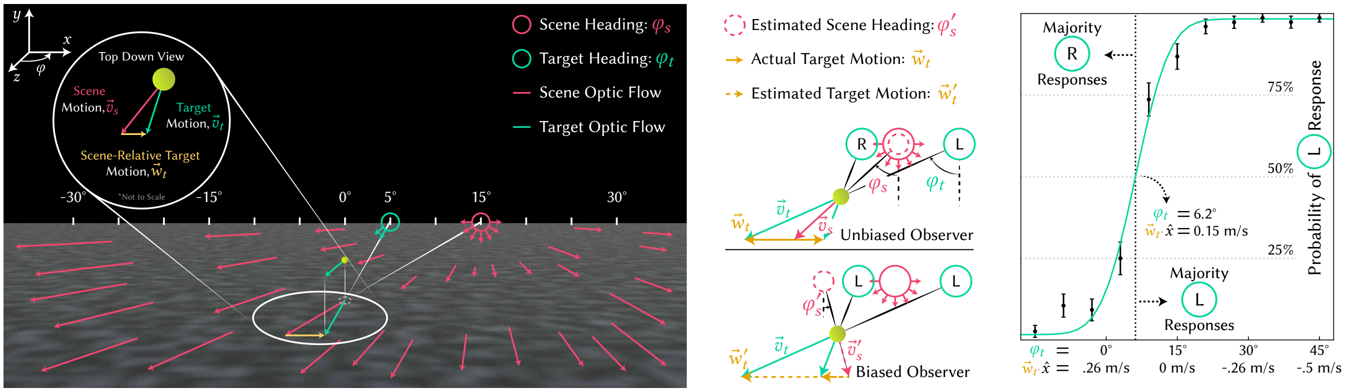
2.2 Human Perception of Scene and Object Motion

In most computer graphics scenarios, users often have to remain physically stationary. To create the sensation of dynamics such as self- (a.k.a., vection [Howard and Howard 1994; Hu et al. 2019]) or scene content movements, we typically synthesize illusions of motion via graphics rendering, resulting in a complex and rich visual percept [Lappe et al. 1999; Nishida et al. 2018]. Examples include playing a racing car video game or watching a roller coaster video. Therefore, when observing moving objects on a computer screen, our primary source of motion cues is the displayed optical flow.

Unlike physical scenarios, such on-screen visual stimuli often lack crucial information that assists us in accurately estimating 3D motion, including vestibular cues [DeAngelis and Angelaki 2012] as well as depth cues via stereopsis [Didyk et al. 2011] and accommodation [Murray 1994]. Specifically, Xie et al. [2020] demonstrate that the accuracy of human motion perception relies on the combination of multiple cues, indicating that we are most adept at motion estimation in real-life. Consequently, recent studies observed human errors in extracting object motion from its surrounding 3D environment under dynamic camera movements [Dokka et al. 2019; Layton and Fajen 2016; Li et al. 2018; Xie et al. 2020; Xing and Saunders 2022]. Several hypotheses have been presented to explain the underlying neurological mechanism for cognitively decomposing the two motions [Beck et al. 2011; Kim et al. 2022; Sasaki et al. 2017]. Although existing research has identified some of the neurological mechanisms, we aim to address the knowledge gap in quantifying the accuracy and precision of human imperfections in extracting object motion. Additionally, we provide functional guidance for content design in downstream graphics applications to enhance task performance. To achieve this, we benchmark and model human object motion estimation based on the movement velocities of the scene and target object, and the depth disparity between them.

2.3 Visual-Acuity-Aware Computer Graphics

Human visual acuity is imperfect, and is affected by various sources of environmental, cognitive, and physiological noise [Deering 1998; Van Beers 2007]. Examples of limitations include reduced resolution in peripheral vision [Watson 2014], diminished color perception [Cohen et al. 2020], and flicker imperception [Tyler 1987]. Researchers have capitalized on these limitations to optimize applications for fast [Kaplanyan et al. 2019; Krajancich et al. 2021], power-efficient [Duinkharjav et al. 2022], and higher-quality [Montalto et al. 2015; Park et al. 2023] imagery, as well as automatically generated visual illusions [Freeman et al. 1991]. Additionally, these perceptual limitations have been exploited to enable optimized shading rates [Denes et al. 2020; Jindal et al. 2021]. However, while most existing methods focus on enhancing system performance within acceptable perceptual thresholds, our work aims to compensate for these perceptual errors to ensure safer and more accurate user task performance.


y
x
z
φ
Top Down View
Scene Motion, $\vec{v}_s$
Target Motion, $\vec{v}_t$
Scene-Relative Target Motion, $\vec{w}_t$
*Not to Scale
-30°
-15°
0°
5°
15°
30°
Scene Heading: φ_s
Target Heading: φ_t
Scene Optic Flow
Target Optic Flow
Estimated Scene Heading: φ_s'
Actual Target Motion: $\vec{w}_t$
Estimated Target Motion: $\vec{w}_t'$
R
L
φ_s
φ_t
$\vec{v}_t$
$\vec{w}_t$
$\vec{v}_s$
Unbiased Observer
φ_s'
$\vec{w}_t'$
$\vec{v}_s'$
Biased Observer
Majority R Responses
Majority L Responses
75%
50%
25%
φ_t = 6.2°
$\vec{w}_t \cdot \hat{x}$ = 0.15 m/s
Probability of L Response
φ_t = 0° 15° 30° 45°
$\vec{w}_t \cdot \hat{x}$ = .26 m/s 0 m/s -.26 m/s -.5 m/s

(a) object movement and self movement

(b) unbiased vs biased observer

(c) left/right threshold measurement

Fig. 2. *Illustration and analysis of biased perception during self-motion.* (a) Accurate reconstruction of the scene-relative target motion $\vec{w}_t$, requires observers to subtract their percept of scene motion $\vec{v}_s$, from the observed on-screen target motion $\vec{v}_t$. The divergence point of optical flow fields due to scene and target motions, a.k.a., FOE, denoted as circles at the horizon. (b) Unbiased "perfect" observers can perfectly estimate the scene heading, φ_s, to determine the direction of scene-relative target motion. Observer L(eft)/R(right) responses are annotated inside the FOE circle for each target motion condition. Biased human observers make judgment errors due to mis-estimation of the scene heading, $\varphi_s' \leq \varphi_s$. Biased estimations denoted as dashed arrows. (c) The psychometric curve visualizes the probability of observers L/R responses for various target motion conditions. The curve indicates that when the target moves through the scene at a speed of 0.15 m/s to the right (equivalent to an observed target heading of $\varphi_t = 6.2°$) observers believe the object to not be moving sideways, on average. Data used for curve fitting is shown as a scatter plot (with SEM error bars).

3 STUDYING OBJECT MOTION PERCEPTION

In a dynamic scenario, a target object moves in the scene ($\vec{w}_t$), which simultaneously appears to be moving to the observer who is also in motion ($\vec{v}_s$), as visualized in Figure 2a. Figure 2b (top) illustrates that an unbiased "perfect" observer can accurately understand $\vec{w}_t$ and $\vec{v}_s$ by analyzing their vector combination, $\vec{v}_t$, as it appears on-screen. Refer to Figure 3 for a reference to all target and scene motion-related symbols used throughout the manuscript. However, this ideal scenario may not reflect reality. As depicted in Figure 2b (bottom), we are imperfect in estimating either motion due to the decomposition ambiguity [Xie et al. 2020; Xing and Saunders 2022]. First, depending on scene dynamics, our perception of scene and target heading often exhibits a "central bias", meaning an *under*-estimation [Xie et al. 2020; Xing and Saunders 2022]. Second, when observers lack visual cues to determine the target distance, the ambiguous optical flow further exacerbates the mis-estimation [Van den Berg and Brenner 1994a]. For example, in Figure 2a, it is ambiguous whether the ball is large and moving at a farther depth or small and moving at a closer depth. Therefore, we study (Section 3.1), quantify, and model (Section 3.2) the perceptual bias scale of target motions under various scene dynamics and content.

3.1 Psychophysical Study

Participants. We recruited subjects for the study through the crowdsourcing platform *Prolific*. A strict screening protocol was enforced to mitigate potential confounds arising from task misinterpretation and attention lapses, ensuring high-quality data (see *Filtering*). As such, we consider the data from $n = 38$ subjects (ages 20 – 56, 21 male) screened from an initial pool of 78. All study protocols were approved by an institutional review board (IRB), and subjects were compensated at a rate of $\$15/h$. Refer to the supplementary video for animated visualizations of all study procedures.

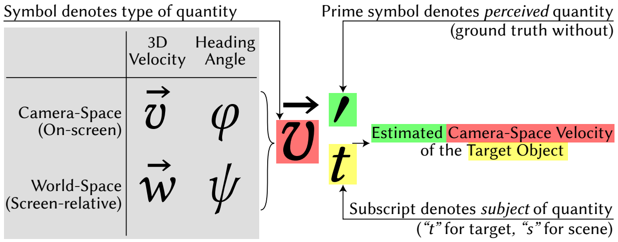
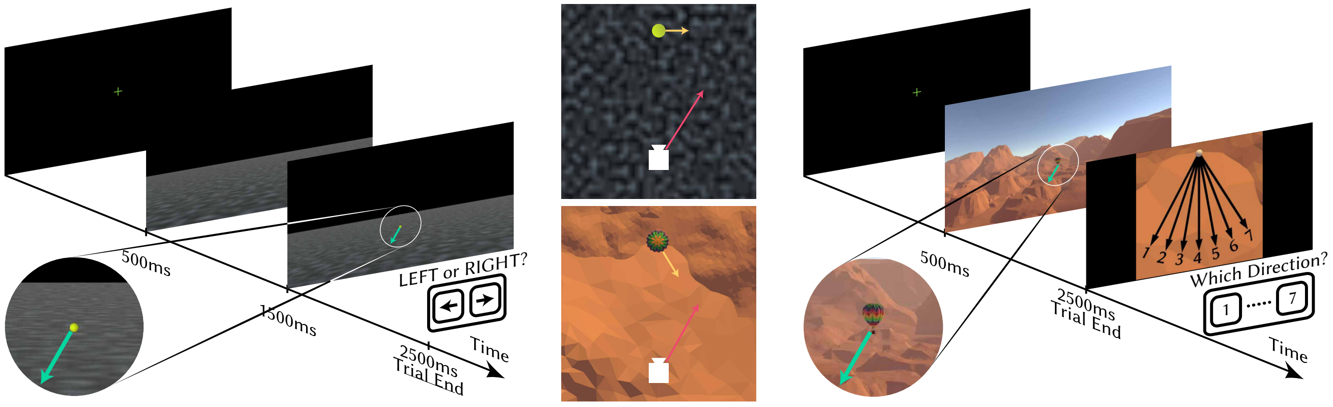

Symbol denotes type of quantity
Prime symbol denotes *perceived* quantity (ground truth without)
3D Velocity
Heading Angle
Camera-Space (On-screen)
World-Space (Screen-relative)
$\vec{v}$
φ
$\vec{w}$
ψ
$\vec{v}$
$'$
t
Estimated Camera-Space Velocity of the Target Object
Subscript denotes *subject* of quantity ("t" for target, "s" for scene)

Fig. 3. *Motion-related variable notation used throughout Sections 3 and 5 and Figs. 2 and 8.*

Stimuli and procedure. The study was conducted via a web-based application on a computer screen. A screen calibration procedure ensured that all subjects viewed the stimuli at approximately 50° *Field of View* (FOV). After calibration, they received a text-based introduction to the stimuli and task.

Subjects initiated each trial by pressing a button. As shown in Figure 4a, they were presented a fixation cross at the screen center for .5 s at the beginning of each trial and instructed to maintain their gaze stationary. After the cross disappeared, a 2 s video (recorded at 60 fps) was shown. Initially, a flat ground surface with Perlin noise texture is visible, conveying forward scene motion with variable speed, v_s, and heading direction, φ_s, to an observer at variable height h_s. The ground texture was chosen to avoid tuning to specific spatial frequency ranges, and instead incorporate a broad spectrum of frequencies, similar to [Xing and Saunders 2022]. After 1 s, a yellow probe (target object) was introduced at a height, h_t, positioned 6 m



(a) pilot study protocol: judge left/right

(b) top-down view

(c) application study protocol: choose direction

Fig. 4. *Study protocols.* (a) In the psychophysical study, a fixation cross is displayed for .5 s at the beginning of each trial. Subsequently, a video plays depicting a scene moving towards the observer at a non-zero heading angle (arrow in (b)). After 1 s, a moving yellow probe (green arrow) is added to the screen. Once the 2 s video finishes, the subject is asked whether the probe was moving left or right. The probe does not have a forward velocity (top of (b)). (c) In the application study, the protocol is near-identical, with three differences. The target object is added at the start of the trial, it has forward velocity (bottom of (b)), and the subject is asked to choose one of seven options to indicate the direction of the object's motion.

in front of the observer at 5° eccentricity below fixation ($h_s - h_t = .52$ m). The target object then moved either left or right relative to the scene at various speeds, w_t, for the rest of the clip (1 s). The object remained visible throughout all trials.

At the end of the video, subjects were prompted to indicate, via button press, whether the probe was moving left or right *relative to the scene*. If they didn't respond after 10 s, the trial expired and prompted a *screening* trial before retrying. No feedback was provided during trials to prevent learning effects.

Prior to the study, subjects participated in an interactive *training* session to familiarize themselves with the task and interfaces. The session comprised eight unique trials of the same protocol. During training, subjects were provided with feedback on their performance after each trial and shown a top-down visualization (see Figure 4b). Subjects were required to respond correctly to all training trials before being allowed to progress. Training conditions were selected to prevent external bias (see *Conditions*).

Metrics. The procedural goal of the study was to determine the threshold heading of the target object, μ, at which subjects perceive the target's scene-relative velocity to be zero: $\vec{w}'_t = 0$ (a.k.a., bias and inaccuracy). During each trial, the subject is presented with targets of different velocities, $\vec{w}_t$, which appear on-screen to be moving along

$$\vec{v}_t = \vec{w}_t + \vec{v}_s, \tag{1}$$

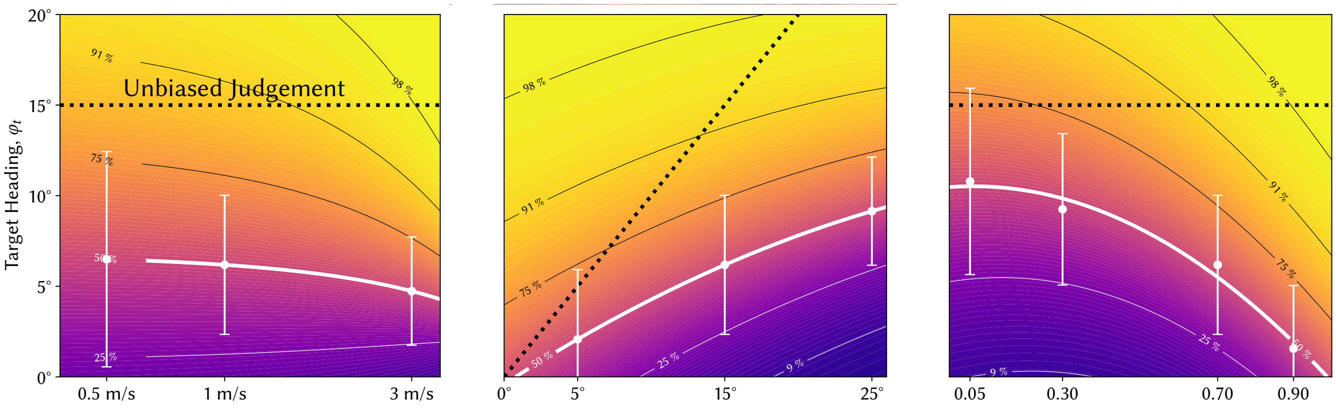
as illustrated in Figure 2a. By aggregating subject responses for different target velocities, $\vec{w}_t$, each corresponding to a different target heading direction, φ_t (see Figure 2b), we fit a psychometric curve, f (see Figure 2c). This allows us to determine the threshold target heading, $\varphi_t = \mu$, at which observers perceive that the target is neither moving left nor right [Wichmann and Hill 2001]:

$$f(\varphi_t; \mu, \sigma, \lambda) = \lambda + (1 - 2\lambda) \times 0.5 \left[1 + \operatorname{erf}\left((\varphi_t - \mu) / \sqrt{2\sigma^2}\right)\right]. \tag{2}$$

Here, σ denotes the slope, and indicates how consistent (or precise) an observer's bias measurement is trial-to-trial, φ_t, and λ denotes the guess/lapse-rate of subjects (i.e., how often subjects make errors irrespective of their observed stimulus). To determine the threshold, μ, and slope, σ, of this psychometric curve, in our study we sampled 11 target heading, φ_t, stimulus levels uniformly sampled between $[-\varphi_s, +3\varphi_s]$.

Conditions. Beyond determining the psychometric parameters of a single condition, we aim to investigate how these parameters vary with scene motion, and depth. To this aim, we anchor our measurements to a reference condition, where $\{v_s = 1 \text{ m/s}, \varphi_s = 15°, h_s = 1.75 \text{ m}\}$, and explore test conditions where only one attribute of the reference changes. These test conditions vary in scene dynamics in speed, $v_s \in \{0.5 \text{ m/s}, 3 \text{ m/s}\}$ and heading, $\varphi_s \in \{5°, 25°\}$, as well as scene content in height, $h_s \in \{.55 \text{ m}, .74 \text{ m}, 5.22 \text{ m}\}$, resulting in a total of 8 study conditions. Note that we vary the observer height h_s to examine the corresponding scene's depth disparity to the target. To provide a more intuitive representation of depth disparity, we henceforth express these conditions via a dimensionless target-scene depth disparity coefficient: $d = h_t/h_s \in \{.05, .3, .9\}$ for each scene height condition, and $d = .7$ for the reference.

Lastly, in the training session, to avoid introducing external bias to subjects' judgment, the trials were deliberately designed as (1) significantly different from trials in the study, and (2) sufficiently easy for classification, yet difficult enough to mitigate potential misinterpretation of the task. So, we selected four trials with $\varphi_s = 40°$, and $\varphi_t \in \{\pm 30°, \pm 40°\}$. The trial with $\varphi_s = 40°$ and $\varphi_t = 30°$ satisfied the requirement (2) above and thus was reused as a *screening* trial to identify subjects who misinterpreted the task even after the training.



(a) scene speed, v_s

(b) scene heading, φ_s

(c) target-scene depth disparity, d

Fig. 5. *Psychophysical Study Results.* Psychometric curves along (a) scene speed, (b) scene heading, and (c) target-scene depth ratio are fitted from the study data, and interpolated via polynomial regression. Yellow colors indicate majority left responses in the left/right study protocol described in Section 3.1. Each curve's threshold is denoted as a scatter with error-bars indicating the *Just Noticeable Difference* (JND) offset, or stimulus levels at 25/75% response probability. Contour lines represent JND step-sizes. "Perfect" unbiased observer's thresholds, as depicted in Figure 2b, are visualized as comparison via dotted black lines. Refer to supplementary video for user study conditions which correspond to various points across the heatmaps.

The *screening* trial was repeated 24 times throughout the study, Each trial was mirrored to ensure left/right balance, resulting in a total of $(11 \times 8 + 24) \times 2 = 224$ main trials (median completion in 21 min).

Filtering. To ensure high-quality data from crowd subjects, we employed a two-layer statistical screening. First, we screened inattentive subjects who only made random guesses. An informal pre-pilot study suggested that subjects almost always gave correct responses when $\varphi_t = 3\varphi_s$ as these were easy-to-answer trials. We leveraged this observation and required an accuracy of $\geq 90\%$, or a guess rate of $\lambda < 10\%$, to pass this screen (random guess accuracy is 50%). Second, we screened for subjects who misinterpreted the task and indicated object motion directions relative to the *observer*. To this end, we required an accuracy of $\geq 50\%$ on *screening* trials (where observer-relative accuracy is 0%). Refer to Supplement A for study results reported without screening trial-based filtering.

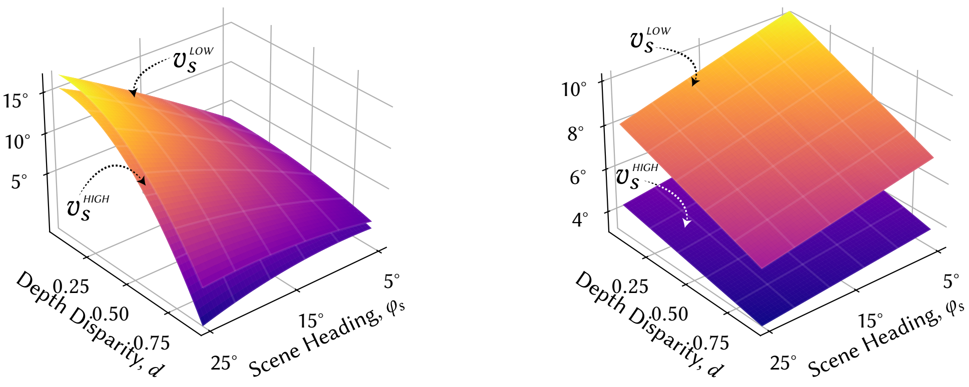
Results. From the initial 78 subjects, we removed 4 (5%) from the attentiveness screen and 36 (46%) from task understanding screen, within a normal range for such crowdsourcing studies [Brühlmann et al. 2020]. In total, 6, 688 trial results were used for further analysis. Prior to combining the left and right heading conditions, we conducted a one-way *Analysis of Variance* (ANOVA) which showed that the direction of heading did not have a significant effect on the subject-aggregated responses ($F_{1,174} = .1, p = .75$).

As described in *Metrics*, we statistically summarized study responses by fitting psychometric curves, extracting the low-dimensional parameters of the threshold, μ, and slope, σ, for each condition separately (with a fixed $\lambda = 1.6\%$ across all conditions found via the attentiveness screen guess rate). Curve parameters for each series of conditions that varied along a single attribute were interpolated via polynomial regression (quadratic for μ, and linear for σ). The results are visualized in Figure 5. See Supplement B for individual curve parameters and polynomial term coefficients.

Discussion. The statistical analysis demonstrates that we can safely aggregate heading directions in a left-right agnostic manner. The central bias persists across all studied conditions, as evidenced by the measured thresholds below the "unbiased judgment" line in Figure 5. This suggests that objects moving to the *right* at a heading angle between the 50% threshold and the unbiased judgment line will be perceived as moving to the *left* by most observers.

We observe other notable trends from the visualization. **From Figure 5a**, we observe a steady increase in both bias and consistency. That is, at higher scene speeds, judgments across subjects become more consistent, yet inaccurate. **From Figure 5b**, the threshold for the scene heading model intersects at zero degrees, indicating that for forward headings, our perception of lateral motion directions becomes accurate due to the lack of asymmetric optical flow cues. Comparing the unbiased judgment line with the threshold fit suggests that the scale of motion estimation bias is roughly proportional to the scene heading, φ_s. **From Figure 5c**, our perceptual errors increase with the depth disparity between the target and the surrounding scene (i.e., $\uparrow d$). Intuitively, this reveals that if the scene content is too far (e.g.,, the sky), it no longer appears to move nor offer cues to target motion. Conversely, if the scene overlaps with the target (i.e., $d \longrightarrow 0$), we still observe a significant bias.

Our 2D-monitor-based study results notably reveal stronger bias compared to prior literature with similar stimuli but in VR (12° when $\varphi_s = 15°$ [Xie et al. 2020; Xing and Saunders 2022]). This aligns with previous findings of stereo cues on motion perception [Burlingham and Heeger 2020; Van den Berg and Brenner 1994a,b]. The stronger bias observed in 2D displays underscores the crucial need to thoroughly measure, predict, and compensate for human errors in the prevailing computer graphics medium today. This also motivates the future development of 3D displays. In the following section, we utilize our study data to establish a perceptual model predicting human errors in target and scene heading judgment.



(a) threshold, μ

(b) slope, σ

Fig. 6. *Full model parameters.* The combined model parameters are visualized as 2D surface slices at two different scene speeds, $v_s^{HIGH} = 3$ m/s and $v_s^{LOW} = 0.5$ m/s. The threshold, μ indicates the critical heading of observed targets, φ_t, at which observers, on average indicate that the target is moving neither left nor right toward the observer. The slope, σ indicates the confusability between different target headings (i.e., higher σ indicates that the ability to discriminate two target headings are poorer). As reported in Section 3.1, increasing the scene movement speed increases the perceptual bias (meaning lower threshold) for observers, while decreasing the confusability between targets moving along different heading directions.

3.2 Modeling Target Motion Errors

Model Extrapolation. In Section 3.1, we conducted three separate polynomial fits to distinct subsets of the study data, each sharing only the reference condition of $\{v_s = 1 \text{ m/s}, \varphi_s = 15°, d = .7\}$. By factoring out the parameters of the reference from the fitted models, we express each model as $\mu(v_s) = \mu_r k_v(v_s)$, $\mu(\varphi_s) = \mu_r k_\varphi(\varphi_s)$, and, $\mu(d) = \mu_r k_d(d)$, where μ_r represents the psychometric threshold of the reference; $k_{v/\varphi/d}$ denote the three individually fitted polynomial models with μ_r factored out. That is, these models show how the threshold changes due to a change in condition from the reference, meaning, $k_v(v_s = 1 \text{ m/s}) = k_\varphi(\varphi_s = 15°) = k_d(d = 0.7) = 1$. To integrate these individual models into a unified holistic one, we employ a first-order approximation and assume the absence of cross-condition effects. Then, we express the overarching model as:

$$\mu(v_s, \varphi_s, d) = \mu_r k_v(v_s) k_\varphi(\varphi_s) k_d(d). \tag{3}$$

This formulation ensures that the trends of each model are extended across a broader spectrum of conditions without compromising the predictive accuracy of the existing conditions. We acknowledge that closer analysis of cross-condition effects could reveal more intricate trends in motion perception errors and is an interesting direction of study, but in the scope of this work, we aimed to determine only the first-order effect, and explore the interesting applications that such a model can enable.

In Figure 6, we present a visualization of the predicted psychometric parameters of the combined model. The extended model features combinations of prominent features discussed in Section 3.1 such as the decrease in estimation errors as the target-scene depth disparity, d, decreases, and the proportional errors with heading direction, φ_s.

Predicting Scene-Relative Target Heading. Thus far, our psychophysical study, and analysis have concentrated on measuring motion judgment errors under the simple condition where the scene-relative target's motion, $\vec{w}_t$, was constrained along a single axis leftward or rightward (illustrated by dashed yellow vectors in Figure 2). But how do these results generalize to conditions where target objects can move in various directions? In order for our model to be applicable for any practical scenarios, it is imperative to establish a framework for extending our perceptual model to accommodate target motions beyond simple lateral movements.

As shown in Figure 2b and supported by the relation in Equation (1), the poor estimation of the two motions—the scene motion ($\vec{v}_s$) and scene-relative target motion ($\vec{w}_t$)—are dependent on each other. This relationship is expressed as $\vec{w}_t = \vec{v}_t - \vec{v}_s$, where $\vec{v}_t$ represents the target's observer-relative velocity. Hence, an observer's misjudgment of scene-relative target movement corresponds to an opposite misjudgment of scene movement:

$$\vec{w}_t' = \vec{v}_t - \vec{v}_s'. \tag{4}$$

In our study, the psychophysical thresholds indicate the critical value $\vec{v}_t$, with a corresponding heading of $\varphi_t = \mu(v_s, \varphi_s, d)$, at which $\vec{w}_t' = 0$. By incorporating these results into Equation (4), we conclude that our model yields the perceived heading of scene motion, which our study has shown to deviate from the actual heading:

$$\varphi_s' = \mu(v_s, \varphi_s, d). \tag{5}$$

Ultimately, by combining Equations (4) and (5), we derive an expression for estimating the perceived scene-relative target motion:

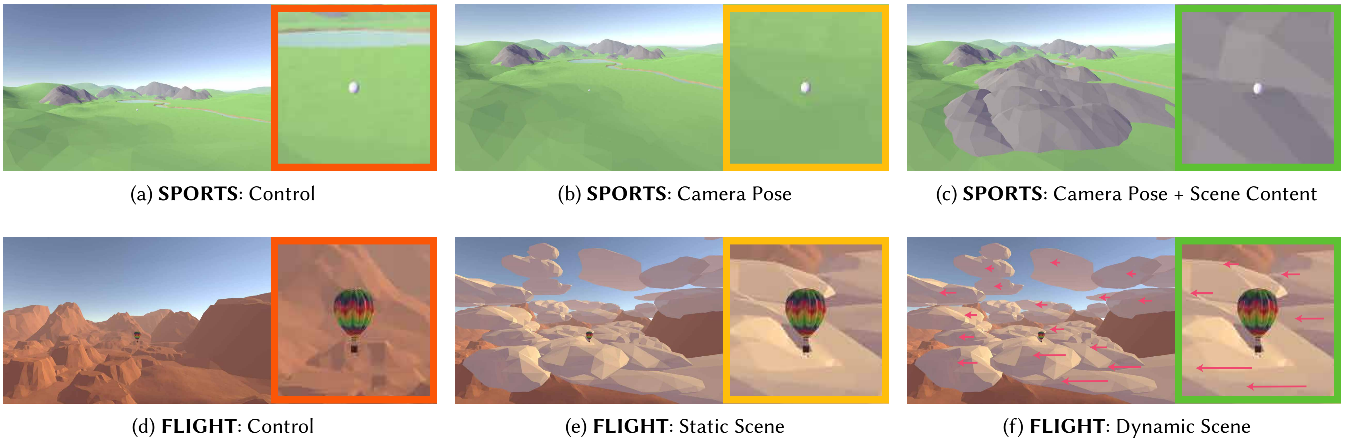
$$\vec{w}_t' = \vec{v}_t - \vec{v}_s' = (\vec{w}_t + \vec{v}_s) - \vec{v}_s' = \vec{w}_t + \vec{v}_s - (R_\mu \hat{z}) v_s \tag{6}$$

where $R_\mu \hat{z}$ represents the *forward* unit vector (see Figure 2a) laterally rotated by $\mu(v_s, \varphi_s, d)$. We visualize this vector sum in Figure 8a.

4 MODEL VALIDATION

4.1 Measuring Model Robustness

To ensure model robustness, we conduct a numerical validation by fitting the model to half of the experimental data, and measure its goodness-of-fit to the other half of the data unseen by the fitted


(a) **SPORTS**: Control

(b) **SPORTS**: Camera Pose

(c) **SPORTS**: Camera Pose + Scene Content

(d) **FLIGHT**: Control

(e) **FLIGHT**: Static Scene

(f) **FLIGHT**: Dynamic Scene

Fig. 7. *Application case study protocols and scenes.* (a)/(d) shows the original animations of the target and camera simultaneously moving in a 3D scene. Both the model prediction and our study results indicate that the animation design induces significant perceptual errors in users' perceptual error of target motion. To reduce such errors, our model enables predictive suggestions for design optimizations, such as adjusting camera poses (b), as well as adding static (c)/(e) and dynamic (f) background geometries.

model. Specifically, each of the $n = 38$ subjects' data is randomly partitioned into either a model fitting or evaluation group. We then assess the model's prediction accuracy compared to the observed data using the R^2 coefficient for each study condition. Due to the arbitrary nature of the subject partitioning operation, we repeated this procedure $N = 20$ times, and observed that the lowest score recorded was .61, while the mean score across all conditions and repeats to be .95, compared to the full model's self-fitting score of .98, indicating acceptable fits [Ozili 2023].

4.2 Generalizability Over Population

We validate whether the psychometric curves fitted from the sample population in Section 3.1 can generalize to unseen subjects. To this aim, we conducted a smaller-scale user study featuring only the *reference* condition from our main study in Section 3.1 on a new subject group ($n = 23$, ages $22 - 52$, 11 males). This study replicated the study protocol, stimuli, and crowdsourcing-based recruitment methods of Section 3.1.

Conditions. Our goal in this study was to investigate the variability of motion judgment errors across different subjects and to use the results to validate our main study in Section 3.1. To keep the study duration and cost feasible, we only studied the reference condition from the main study (i.e., $\{v_s = 1 \text{ m/s}, \varphi_s = 15°, d = .7\}$) and increased the number of repetitions for each trial (10 repeats) to sufficiently fit corresponding psychometric curves for individual subjects. Step sizes between target heading levels, φ_t, were decreased to 4.2° to ensure higher precision measurements. Overall, the study consisted of 80 *measurement* trials, 20 *filler* trials featuring random conditions to prevent categorical judgments [Xing and Saunders 2022], and 48 *screening* trials (see Section 3.1 for details) for a total of 148 trials completed in 15 min by the median subject.

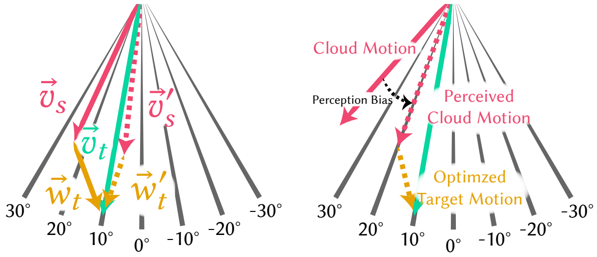
Results and discussion. We fit individual psychometric curves to each of the subjects' aggregated study responses, and observed a mean threshold, $\mu_{\text{avg}} = 4.6° \pm 1.1°$ *Standard Error Mean* (SEM) and mean slope, $\sigma_{\text{avg}} = 6.2° \pm 1.4°$ SEM for the condition identical to the reference of our main study. A single sample t-test indicates that the mean threshold and slope from the main study $\mu = 6.2°$ and $\sigma = 5.7°$ is not significantly different from the distribution of thresholds and slopes in the evaluation study, $t(22) = -1.4, p = .18$ and $t(22) = .35, p = .73$, respectively.

The statistical analysis demonstrates that the psychometric threshold found for the reference condition in our main study lies within acceptable limits of thresholds of out-of-population individuals. While the approach for establishing representative psychometric curve parameters utilized in this evaluation study are more robust due to the larger volume of samples we collect per-subject, we note that conducting a main study of similar scale in terms of different conditions studied becomes unfeasible in practice due to prohibitively high study durations and costs.

5 APPLICATION CASE STUDY: ANIMATION DESIGN GUIDANCE

Scene dynamics, including camera and object motion control [Hsu et al. 2013], as well as scene content, such as depth [Kellnhofer et al. 2013], are crucial factors in animation design [Jiang et al. 2021; Lino and Christie 2015], video editing [Kang and Cho 2019], and game development [Caroux et al. 2013]. Traditionally, the design of these factors has been implicitly driven by aesthetics or storytelling.

We investigate observers' perceptual errors in the target dynamics with two 3D animations. Subsequently, we propose model-guided design alterations, including optimizing camera pose, adjusting the placement of scene objects, and introducing subtle motions to them, to mitigate the predicted perceptual errors. We evaluate the effectiveness of these scene design improvements by conducting multiple-choice user studies.



(a) baseline prediction

(b) design optimization

Fig. 8. *Predicting and compensating target motion estimation in animation design.* (a) Similar to the illustration in Figure 2a, an observer may erroneously perceive the target motion $\vec{w}_t$ as $\vec{w}_t'$ by judging from $\vec{v}_t$ on screen. As shown in Figure 7, we leverage our model to alter the scene designs in various ways to reduce the error. (b) We take the "Dynamic Scene" condition in **FLIGHT** (Figure 7f) as example. The model-guided cloud motion alters observers' perception so that $\vec{w}_t'$ becomes closer to $\vec{w}_t$ (as evidenced in Figure 9).

Participants and procedure. We conducted two user studies via crowdsourcing and recruited $n = 22$ subjects (ages $20 - 64$, 10 male) for each. Unlike the two-alternative forced choice (left vs. right judgment) tasks in Section 3.1, subjects in this study directly indicated perceived scene-relative directions of target motion. As shown in Figures 4c and 8, they chose from one of seven options, each representing a scene-relative target heading of $\psi_t \in \{\pm 30°, \pm 20°, \pm 10°, 0°\}$. After viewing a 2 s video featuring a moving target within a moving scene, subjects referred a top-down view presented at the end of each trial and pressed a button to indicate their choice.

Stimuli. Two realistic scenes, along with corresponding target objects, were used to simulate common gaming and simulation animations: (1) sports gaming with golf (**SPORTS**), and (2) flight simulation (**FLIGHT**), as shown in Figure 7. In both scenes, as depicted in Figure 4b, the scene moves towards the observer at a heading of $\varphi_s = 25°$ with a speed of $v_s = 1$ m/s and .5 m/s for **SPORTS** and **FLIGHT**, respectively (the scene and target sizes were re-scaled to align with the scaling of our model).

Each scene features a target object: a golf ball, and a hot-air balloon. At the start of each trial, the target object appears at a random location within 10° from the fixation point, and a distance of $12 - 14$ m and $6 - 7$ m from the observer for each scene. The target moves towards the observer along the 3rd trajectory in Figure 8 at a heading of $\varphi_t = 10°$ and a speed of $v_t = 2.8 \times v_s$. The observer-relative motion of the target is equivalent to a scene-relative motion along the 6th trajectory in Figure 8, or $\psi_t = -20°$.

Each subject completed 10 repetitions of these trials as well as 5 more *filler* trials with random target object headings to prevent categorical responses. We provided mirrored motions for each trial to ensure left-right balance for a total of 30 trials per study condition. Similar to our psychophysical study in Section 3.1, subjects also completed a pre-study *training* session with a straight-ahead heading $\varphi_s = 0°$, and targets moving along 1st, or 7th trajectory (i.e., $\psi_t \in \pm 30°$). The median completion time was 15 min.

Conditions. For each scene, we prepared two content re-design "treatment" conditions without changing the original camera motion trajectory, when compared to the control conditions shown in Figures 7a and 7d. As evidenced in Figure 5c, decreasing target-scene depth disparity, d, reduces perceptual errors. Thus, to address this issue, in **SPORTS**, the first re-design elevates the camera height, and lowers the viewing angle for a more "birds-eye" view (Figure 7b). As a more aggressive re-design, we also added scene elements behind the target golf ball to further decrease depth disparity (Figure 7c). Across these three scenes, the average scene-target depth disparities were $d = .1/.6/.7$, respectively. Using our model and target heading prediction framework of Section 3.2, we determined $\psi_t' = 16°/10°/-5°$ for the three conditions respectively.

Similarly, for **FLIGHT**, we first added static cloud objects into the scene to decrease the depth disparity from $d = .8$ to $d = .4$ as shown in Figure 7e. For the second treatment, we took a different approach by attempting to simulate a different scene heading by adding a horizontal drift velocity, $v = .25$ m/s, to the clouds relative to the rest of the scene to reinforce the lateral direction of optical flow and induce a higher perceived scene heading angle of $\varphi_s = 37°$ (see Figure 8b). In effect, our model predicts that the perceived scene-relative target heading for the target hot-air balloon was $\psi_t' = 22°/-1°/-12°$, respectively.

Analysis and results. For both studies, we summarize the mean response of each subject and each condition by aggregating across the 20 recorded trials. Figure 9 compares the acquired distributions of target headings ψ_t' with the model-prediction. Across subjects, in **SPORTS**, the measured mean and SEM target headings were $\psi_t' = 9.1° \pm .91°$, $4.8° \pm .60°$ and $-5.5° \pm 1.2°$ for the control, camera pose and additional scene content conditions respectively, while in **FLIGHT**, the measurements were $\psi_t' = 6.5° \pm .71°$, $-1.8° \pm$

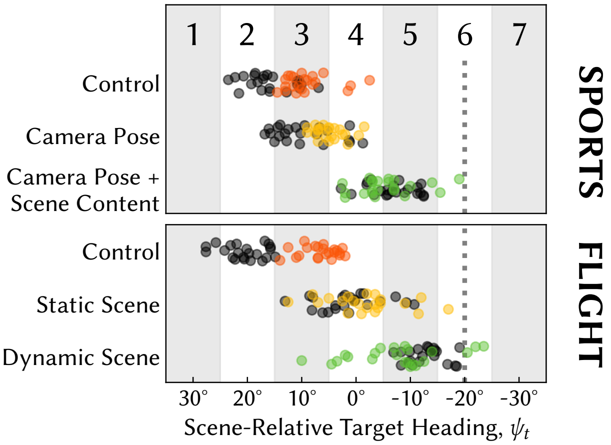


Fig. 9. *Results of the application case study.* The x-axis shows the scene-relative target heading angles corresponding to individual options (1-7) provided in the study. The red and yellow/green points represent the distribution of per-subject aggregated mean response data in control and our model-suggested re-designed animations, respectively. The black points represent the corresponding response distribution simulated from our model prediction. The points (ψ_t') are vertically jittered for plot visibility.

1.5° and $-7.5° \pm 1.8°$ for the control, static scene and dynamic scene conditions, respectively. Across all conditions, the ground-truth scene-relative target heading was $\psi_t = -20°$. A repeated measured ANOVA shows that the conditions within each study had a significant effect on the mean responses for both **SPORTS** ($F_{2,42} = 94.0, p < .01$) and **FLIGHT** ($F_{2,42} = 65.6, p < .01$) scenes.

Discussion. As shown by the ANOVA results, the model-guided content re-design significantly improved the accuracy of target heading judgments for the subjects. Our model was able to predict the overall trend of heading judgment errors, although the exact numerical predictions were slightly inaccurate. We attribute this performance regression to the introduction of higher-order cognitive cues in the more realistic stimuli and discuss its implications further in Section 6. Nevertheless, our model is still capable of providing a first-order approximation of the relationship between observer-relative scene and target velocities ($\vec{v}_s$ and $\vec{v}_t$) and the scene-relative target velocity ($\vec{w}_t$). In real-world applications, we can leverage these predictions to provide guidance and feedback on the overall estimation difficulty, and anticipated motion judgment errors users are likely to make when observing dynamic imagery.

6 LIMITATIONS AND FUTURE WORK

Additional cues. Beyond image space, stereo [Burlingham and Heeger 2020] and vestibular [DeAngelis and Angelaki 2012] cues from emerging 3D displays may also alter motion perception, together with semantic and cognitive influences, including human body pose [Blake and Shiffrar 2007], visual path information [Li et al. 2009], and object shadows [Kersten et al. 1997]. Meanwhile, many of these phenomena rely on higher-order cognitive cues beyond low-level visual operators. For example, understanding the relationship between the motion of objects and the shadows they cast requires spatial reasoning and is a non-intrinsic, learned skill in humans [Van de Walle et al. 1998]. In this work, we chose to first establish a baseline for human perception at an abstraction level where all high-level cues were absent, and the only source of information was the optical flow derived from motion within a 3D environment. After confirming significant perceptual errors under these abstract baseline conditions, we then constructed a more realistic synthetic scene in Section 5 to determine whether any of the baseline estimation errors persist and to assess if our model can still mitigate these errors within the scope of our chosen parameterization, despite the introduction of high-level factors. We believe that these experiments successfully demonstrate the effective application of optimizing animation design pipelines as a first-order measurement and mitigation of human perceptual errors.

Cross-conditions. In Section 3, we characterize the scene dynamics with self movement (direction and speed) and content depths (with regard to the object). Exploring additional combinations of scene and object dynamics, such as rotations and vertical movements, leads to a prohibitively large number of trials. This poses challenges due to participants' limited attentive capacity for maintaining data accuracy, as well as the associated financial costs or running long studies. Therefore, this research focuses on separately measuring the effects from individual dimensions. To study the cross-conditions while maintaining feasibility, we plan to first analyze a primary effect via a pilot study similar to [In 2017], and extend the work towards a dimension-reduced study.

Motion degrees of freedom. We study perceptual errors for horizontal motion patterns along transverse (horizontal) planes—the more common human motion [Hummel et al. 2016]. However, both object and scene motions together form a complex 12 degrees of freedom (DoF) problem (6 DoF each for the self and the object) across all planes, including the coronal and sagittal. In such case, a rotating observer or object will elicit a moving FOE [Danz et al. 2020]. Therefore, introducing a temporal movement factor to the FOE, a.k.a., its *locus*, could be a key to modeling arbitrary motions [Rangarajan and Shah 1992]. Additionally, camera motion analysis using a large-scale egocentric motion dataset (e.g., Ego4D [Grauman et al. 2022]) could establish a coordinate system tailored for the most prevalent human motion patterns.

Perceptual attention and confidence. In highly complex scenarios, various objects may move in different directions. The confounding optical flow may further compromise observers' perception in understanding the motion [Warren et al. 1988; Warren Jr and Hannon 1988]. Moreover, because of humans' selective attention, the movement of multiple objects can also interfere with the visual sensitivity towards a specific target [Min and Corso 2019]. Our current model assumes full attention to a single target. In the future, we plan to explore the influence from optical flow entropy toward a more content-aware probabilistic model.

7 CONCLUSION

In this study, we investigate a universal yet under-investigated factor in computer graphics: human perception of how displayed objects move in dynamic environments. Our data reveals a consistent and systematic bias in how object motions are interpreted, influenced by scene dynamics and content depth. Additionally, we demonstrated how the findings provide quantifiable guidance for animation and game design, helping to reduce users' perceptual errors. We hope this work will inspire future research in the community, towards developing predictive models that compensate for perceptual limitations and enhance human performance with computer graphics.

ACKNOWLEDGMENTS

This project is partially supported by the National Science Foundation grant #2232817/#2225861, the DARPA Intrinsic Cognitive Security program, and a generous gift from Adobe. Part of the work was conducted during an internship at Adobe Research. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

Steven S. Beauchemin and John L. Barron. 1995. The computation of optical flow. *ACM computing surveys (CSUR)* 27, 3 (1995), 433–466.

Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. 2011. Marginalization in neural circuits with divisive normalization. *Journal of Neuroscience* 31, 43 (2011), 15310–15319.

Randolph Blake and Maggie Shiffrar. 2007. Perception of human motion. *Annu. Rev. Psychol.* 58 (2007), 47–73.

Florian Brühlmann, Serge Petralito, Lena F Aeschbach, and Klaus Opwis. 2020. The quality of data collected online: An investigation of careless responding in a crowdsourced sample. *Methods in Psychology* 2 (2020), 100022.

Charlie S Burlingham and David J Heeger. 2020. Heading perception depends on time-varying evolution of optic flow. *Proceedings of the National Academy of Sciences* 117, 52 (2020), 33161–33169.

Loïc Caroux, Ludovic Le Bigot, and Nicolas Vibert. 2013. Impact of the motion and visual complexity of the background on players' performance in video game-like displays. *Ergonomics* 56, 12 (2013), 1863–1876.

Michael A Cohen, Thomas L Botch, and Caroline E Robertson. 2020. The limits of color awareness during active, real-world vision. *Proceedings of the National Academy of Sciences* 117, 24 (2020), 13821–13827.

V Cornilleau-Pérès and CCAM Gielen. 1996. Interactions between self-motion and depth perception in the processing of optic flow. *Trends in Neurosciences* 19, 5 (1996), 196–202.

Adam D Danz, Dora E Angelaki, and Gregory C DeAngelis. 2020. The effects of depth cues and vestibular translation signals on the rotation tolerance of heading tuning in macaque area MSTd. *Eneuro* 7, 6 (2020).

Gregory C DeAngelis and Dora E Angelaki. 2012. Visual–Vestibular Integration for Self-Motion Perception. *The Neural Bases of Multisensory Processes* (2012).

Michael F Deering. 1998. The limits of human vision. In *2nd international immersive projection technology workshop*, Vol. 2. 1.

Gyorgy Denes, Akshay Jindal, Aliaksei Mikhailiuk, and Rafał K Mantiuk. 2020. A perceptual model of motion quality for rendering with adaptive refresh-rate and resolution. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 133–1.

Piotr Didyk, Tobias Ritschel, Elmar Eisemann, Karol Myszkowski, and Hans-Peter Seidel. 2011. A perceptual model for disparity. *ACM Transactions on Graphics (TOG)* 30, 4 (2011), 1–10.

Kalpana Dokka, Hyeshin Park, Michael Jansen, Gregory C DeAngelis, and Dora E Angelaki. 2019. Causal inference accounts for heading perception in the presence of object motion. *Proceedings of the National Academy of Sciences* 116, 18 (2019), 9060–9065.

Budmonde Duinkharjav, Kenneth Chen, Abhishek Tyagi, Jiayi He, Yuhao Zhu, and Qi Sun. 2022. Color-perception-guided display power reduction for virtual reality. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.

William T Freeman, Edward H Adelson, and David J Heeger. 1991. Motion without movement. *ACM Siggraph Computer Graphics* 25, 4 (1991), 27–30.

Borko Furht, Joshua Greenberg, and Raymond Westwater. 2012. *Motion estimation algorithms for video compression.* Vol. 379. Springer Science & Business Media.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 18995–19012.

Ian P Howard and Antonie Howard. 1994. Vection: the contributions of absolute and relative visual motion. *Perception* 23, 7 (1994), 745–751.

Wei-Hsien Hsu, Yubo Zhang, and Kwan-Liu Ma. 2013. A multi-criteria approach to camera motion design for volume data animation. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2792–2801.

Ping Hu, Qi Sun, Piotr Didyk, Li-Yi Wei, and Arie E Kaufman. 2019. Reducing simulator sickness with perceptual camera control. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–12.

Ping Hu, Gang Wang, and Yap-Peng Tan. 2018. Recurrent spatial pyramid CNN for optical flow estimation. *IEEE Transactions on Multimedia* 20, 10 (2018), 2814–2823.

Yan Huang, Kannappan Palaniappan, Xinhua Zhuang, and Joseph E. Cavanaugh. 1995. Optic flow field segmentation and motion estimation using a robust genetic partitioning algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17, 12 (1995), 1177–1190.

Tak-Wai Hui, Xiaoou Tang, and Chen Change Loy. 2018. Liteflownet: A lightweight convolutional neural network for optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 8981–8989.

Nadine Hummel, Luigi F Cuturi, Paul R MacNeilage, and Virginia L Flanagin. 2016. The effect of supine body position on human heading perception. *Journal of vision* 16, 3 (2016), 19–19.

Junyong In. 2017. Introduction of a pilot study. *Korean journal of anesthesiology* 70, 6 (2017), 601–605.

Ramesh Jain. 1983. Direct computation of the focus of expansion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1 (1983), 58–64.

Ramesh Jain. 1984. Complex logarithmic mapping and the focus of expansion. *ACM SIGGRAPH Computer Graphics* 18, 1 (1984), 24–24.

Hongda Jiang, Marc Christie, Xi Wang, Libin Liu, Bin Wang, and Baoquan Chen. 2021. Camera keyframing with style and control. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–13.

Akshay Jindal, Krzysztof Wolski, Karol Myszkowski, and Rafał K Mantiuk. 2021. Perceptual model for adaptive local shading and refresh rate. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–18.

Kyoungkook Kang and Sunghyun Cho. 2019. Interactive and automatic navigation for 360 video playback. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–11.

Anton S Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. 2019. DeepFovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–13.

Petr Kellnhofer, Tobias Ritschel, Karol Myszkowski, and Hans-Peter Seidel. 2013. Optimizing disparity for motion in depth. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 143–152.

Daniel Kersten, Pascal Mamassian, and David C Knill. 1997. Moving cast shadows induce apparent motion in depth. *Perception* 26, 2 (1997), 171–192.

HyungGoo R Kim, Dora E Angelaki, and Gregory C DeAngelis. 2022. A neural mechanism for detecting object motion during self-motion. *Elife* 11 (2022), e74971.

Brooke Krajancich, Petr Kellnhofer, and Gordon Wetzstein. 2021. A perceptual model for eccentricity-dependent spatio-temporal flicker fusion and its applications to foveated graphics. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–11.

Markus Lappe, Frank Bremmer, and Albert V van den Berg. 1999. Perception of self-motion from visual flow. *Trends in cognitive sciences* 3, 9 (1999), 329–336.

Oliver W Layton and Brett R Fajen. 2016. The temporal dynamics of heading perception in the presence of moving objects. *Journal of neurophysiology* 115, 1 (2016), 286–300.

Li Li, Jing Chen, and Xiaozhe Peng. 2009. Influence of visual path information on human heading perception during rotation. *Journal of Vision* 9, 3 (2009), 29–29.

Li Li, Long Ni, Markus Lappe, Diederick C Niehorster, and Qi Sun. 2018. No special treatment of independent object motion for heading perception. *Journal of Vision* 18, 4 (2018), 19–19.

Christophe Lino and Marc Christie. 2015. Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics (TOG)* 34, 4 (2015), 1–12.

Kyle Min and Jason J Corso. 2019. Tased-net: Temporally-aggregating spatial encoder-decoder network for video saliency detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 2394–2403.

Carlos Montalto, Ignacio Garcia-Dorado, Daniel Aliaga, Manuel M Oliveira, and Feng Meng. 2015. A total variation approach for customizing imagery to improve visual acuity. *ACM transactions on graphics (TOG)* 34, 3 (2015), 1–16.

John Murray. 1994. Some perspectives on visual depth perception. *ACM SIGGRAPH Computer Graphics* 28, 2 (1994), 155–157.

Bernd Neumann. 1984. Optical flow. *ACM SIGGRAPH Computer Graphics* 18, 1 (1984), 17–19.

Shin'ya Nishida, Takahiro Kawabe, Masataka Sawayama, and Taiki Fukiage. 2018. Motion perception: From detection to interpretation. *Annual review of vision science* 4 (2018), 501–523.

Peterson K Ozili. 2023. The acceptable R-square in empirical modelling for social science research. In *Social research methodology and publishing results: A guide to non-native English speakers.* IGI global, 134–143.

Mu Hyeon Park, Kwan Yun, and Gerard J Kim. 2022. Focused Area of Movement as an Effective Rest Frame for Reducing VR Sickness. In *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct).* IEEE, 712–713.

Seung Ho Park, Young Su Moon, and Nam Ik Cho. 2023. Perception-Oriented Single Image Super-Resolution using Optimal Objective Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 1725–1735.

Krishnan Rangarajan and Mubarak Shah. 1992. Interpretation of motion trajectories using focus of expansion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, 12 (1992), 1205–1210.

Ryo Sasaki, Dora E Angelaki, and Gregory C DeAngelis. 2017. Dissociation of self-motion and object motion by linear population decoding that approximates marginalization. *Journal of Neuroscience* 37, 46 (2017), 11204–11219.

Ashit Talukder and Larry Matthies. 2004. Real-time detection of moving objects from moving vehicles using dense stereo and optical flow. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, Vol. 4. IEEE, 3718–3725.

Christopher W Tyler. 1987. Analysis of visual modulation sensitivity. III. Meridional variations in peripheral flicker sensitivity. *JOSA A* 4, 8 (1987), 1612–1619.

Robert J Van Beers. 2007. The sources of variability in saccadic eye movements. *Journal of Neuroscience* 27, 33 (2007), 8757–8770.

Gretchen A Van de Walle, Jayne S Rubenstein, and Elizabeth S Spelke. 1998. Infant sensitivity to shadow motions. *Cognitive Development* 13, 4 (1998), 387–419.

AV Van den Berg and Eli Brenner. 1994a. Humans combine the optic flow with static depth cues for robust perception of heading. *Vision research* 34, 16 (1994), 2153–2167.

AV Van den Berg and Eli Brenner. 1994b. Why two eyes are better than one for judgements of heading. *Nature* 371, 6499 (1994), 700–702.

William H Warren, Michael W Morris, and Michael Kalish. 1988. Perception of translational heading from optical flow. *Journal of Experimental Psychology: Human Perception and Performance* 14, 4 (1988), 646.

William H Warren Jr and Daniel J Hannon. 1988. Direction of self-motion is perceived from optical flow. *Nature* 336, 6195 (1988), 162–163.

Andrew B Watson. 2014. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of vision* 14, 7 (2014), 15–15.

Felix A Wichmann and N Jeremy Hill. 2001. The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & psychophysics* 63, 8 (2001), 1293–1313.

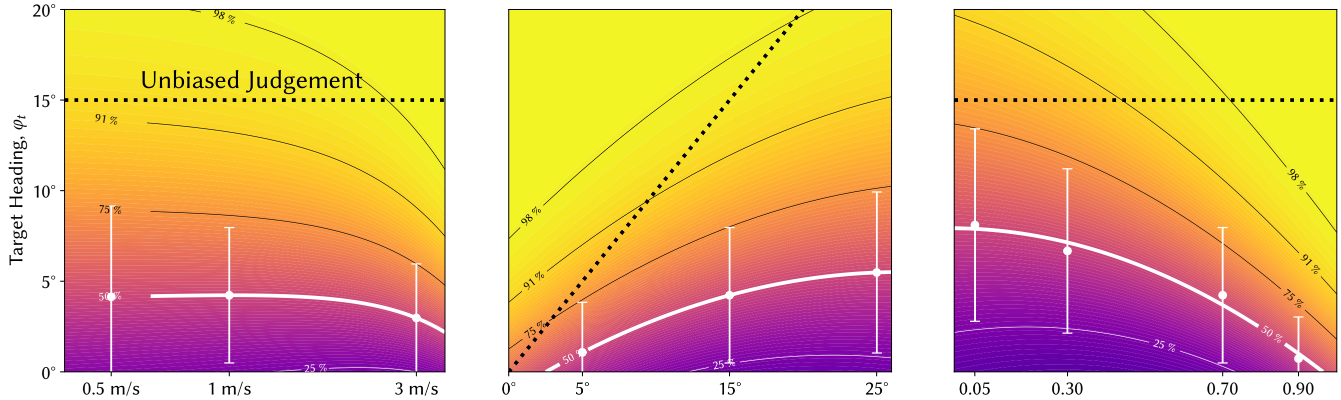
Mingyang Xie, Diederick C Niehorster, Markus Lappe, and Li Li. 2020. Roles of visual and non-visual information in the perception of scene-relative object motion during walking. *Journal of Vision* 20, 10 (2020), 15–15.

Xing Xing and Jeffrey A Saunders. 2022. Perception of object motion during self-motion: Correlated biases in judgments of heading direction and object motion. *Journal of Vision* 22, 11 (2022), 8–8.

Li Xu, Jiaya Jia, and Yasuyuki Matsushita. 2011. Motion detail preserving optical flow estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 9 (2011), 1744–1757.

A UNFILTERED PSYCHOPHYSICAL DATA ANALYSIS

In this work, we rejected a significant number of subjects via our task understanding filter, as described in Section 3.1, to ensure high quality data acquired from crowdsourced study participants. Here, we present the psychometric curve fitting results for the unfiltered data to serve as a comparison to the results included in the main manuscript. In Figure 10, we replicated Figure 5 to serve as a direct comparison between the filtered and unfiltered data. The psychometric threshold for the reference condition was $\mu_r = 4.2°$ when compared to $\mu_r = 6.2°$ as reported for the unfiltered data. The *Discussion* about the trends and patterns of the psychophysical study results in Section 3.1 are largely unchanged for the unfiltered data, albeit with a much stronger bias effect.



(a) scene speed, v_s

(b) scene heading, φ_s

(c) target-scene depth disparity, d

Fig. 10. *Unfiltered Study Data Analysis.* Results of processing the data without applying the task understanding filter are visualized for comparison with Figure 5. See the caption for Figure 5 for details on the visualization designs.

B PSYCHOMETRIC AND POLYNOMIAL FITTING

Below, we list the parameters for all the psychometric curves fitted using the data collected from our psychophysical study of Section 3.1:

Table 1. Psychometric parameters for different scene speeds, headings, and depth ratios.

Attribute	**Value**	**Threshold, μ (°)**	**Slope, σ (°)**
v_s (m/s)	0.5	6.5	8.8
	1	6.2	5.7
	3	4.7	4.4
φ_s (degrees)	5	2.1	5.7
	15	6.2	5.7
	25	9.1	4.4
d	0.05	10.8	7.6
	0.3	9.2	6.2
	0.7	6.2	5.7
	0.9	1.6	5.2

These psychometric parameters were then regressed to fit polynomial curves with fitted parameters $\mu_r = 6.2°$ and $\sigma_r = 5.7°$.:

$$\begin{aligned}
\mu(v_s, \varphi_s = 15^\circ, d = .30) &= \mu_r \times (.931 + .077v_s + .006v_s^2),\\
\mu(v_s = 1\text{ m/s}, \varphi_s, d = .30) &= \mu_r \times (.045 + .054\varphi_s + .001\varphi_s^2),\\
\mu(v_s = 1\text{ m/s}, \varphi_s = 15^\circ, d) &= \mu_r \times (.531 - .171d + 1.390d^2),\\
\sigma(v_s, \varphi_s = 15^\circ, d = .30) &= \sigma_r \times (1.486 - .302v_s),\\
\sigma(v_s = 1\text{ m/s}, \varphi_s, d = .30) &= \sigma_r \times (1.093 - .011\varphi_s),\ \text{and,}\\
\sigma(v_s = 1\text{ m/s}, \varphi_s = 15^\circ, d) &= \sigma_r \times (1.308 - .459d).
\end{aligned} \tag{7}$$