Perceptual Requirements for Eye-Tracked Distortion Correction in VR

Supplementary Material

PHILLIP GUAN, Reality Labs Research, Meta
OLIVIER MERCIER, Reality Labs Research, Meta
MICHAEL SHVARTSMAN, Reality Labs Research, Meta
DOUGLAS LANMAN, Reality Labs Research, Meta

1 ADDITIONAL USER STUDY DETAILS

1.1 Psychophysics with Traditional Methods

Traditional psychophysics techniques, such as the method of constant stimuli, collect human judgment from a regular grid in the space of stimuli. As such, they require mk^n observations for m stimulus repetitions in n dimensions with k grid steps. More advanced adaptive staircase methods such as Quest [Watson and Pelli 1983] adaptively select stimuli in one intensity dimension while using a grid in the others. This reduces the requirement to lk^{n-1} for l staircase steps, but is still exponential in the number of stimulus dimensions. In our context, for three spatial dimensions of eye tracking performance, a traditional staircase would fix the x- and y-axis bias over a grid (of e.g. 5×5 points), and for each value of this grid we use an adaptive staircase to determine the acceptability threshold on eye tracking performance. Estimating acceptable eye tracking performance over this coarse grid would conservatively take 2,500 trials ($5 \times 5 \times 100$), assuming 100 staircase trials were sufficient to estimate a threshold function for the z-axis. Adding a fourth dimension, such as eye tracking latency, increases the minimum required samples from 2,500 by another factor of five to 12,500 trials. Even so, such a coarse grid may be insufficient to understand fine-grained trade-offs between errors in different dimensions, but doubling the number of grid points per dimension (still in 3D) would quadruple the number of trials to 10,000 ($10 \times 10 \times 100$). This illustrates why traditional psychophysics are rarely employed for even moderately high-dimensional problems. And these estimates are conservative; in practice, many more than just the minimum trials are required because the bounds around the problem space are unknown *a priori*, so some data collection must take place to determine what the values along the fixed dimensions should be.

1.2 User Study Data

The raw data indicating trials where distortion from inaccurate DDC was correctly (gray dots) or incorrectly (open white circles) identified are shown below. The modeling framework used to analyze our data, AEPsych [Owen et al. 2021], provides the posterior probability of correct detection as a function of stimulus parameters, the median of which is shown using the colormap in the following figures. Solid black lines indicate the estimated 75% correct detection threshold based on the posterior median. Because the thresholds are circular (and therefore not unique in the stimulus dimensions), it is not obvious how to define conventional confidence intervals around them in this setting. To illustrate the model's uncertainty, we instead use dashed contours to represent the estimated 75% detection threshold at the 2.5% and 97.5% percentiles of the model's posteriors (i.e. they are the thresholds at the edges of the model's 95% posterior density interval rather than the 95% posterior density interval of the threshold, which are subtly different quantities). These wide uncertainty bounds are likely overly conservative in the sense that the posterior median tends

to track the true threshold much more accurately than the bounds would indicate in a variety of benchmarking problems [Letham et al. 2022].

1.2.1 Pancake with Text. Six subjects participated in this experiment and their data is presented in Figure 1 by relative sensitivity. The number of trials collected by each observer is shown in the bottom left of the first column and total collection time ranged from two to four hours for each subject. Participant P2 collected approximately 400 pilot trials which are included in the analysis, but removing them does not substantially affect the model fit.

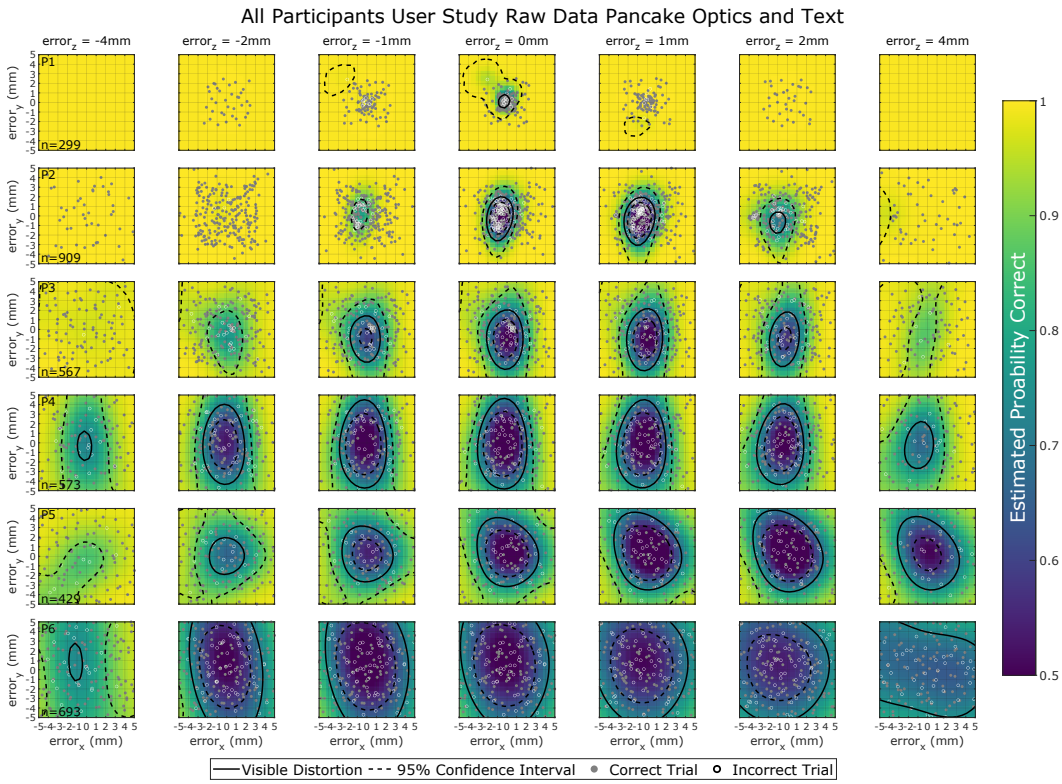

Fig. 1. *Pancake Eye Tracking Requirements for Text:* Eye tracking requirements to eliminate perceptible distortions using dynamic distortion correction for the pancake viewing optics while viewing text. Each row represents a different observer and each column represents a fixed value of eye relief error where negative values place the eye closer to the display and positive values place the eye farther away. Each plot shows acceptable eye tracking bias for different observer in x-, y-, and z-dimensions of the eyebox that can support a perceptually-distortion-free viewing experience. The participant ID is shown in the top left and the number of trials collected by each participant is shown in the bottom left of each subplot in the first column.

1.2.2 Multiple Optics and Scenes. The data for P2 and P3 are shown in Figure 2 for the optical designs and scenes shown in Figure 8.

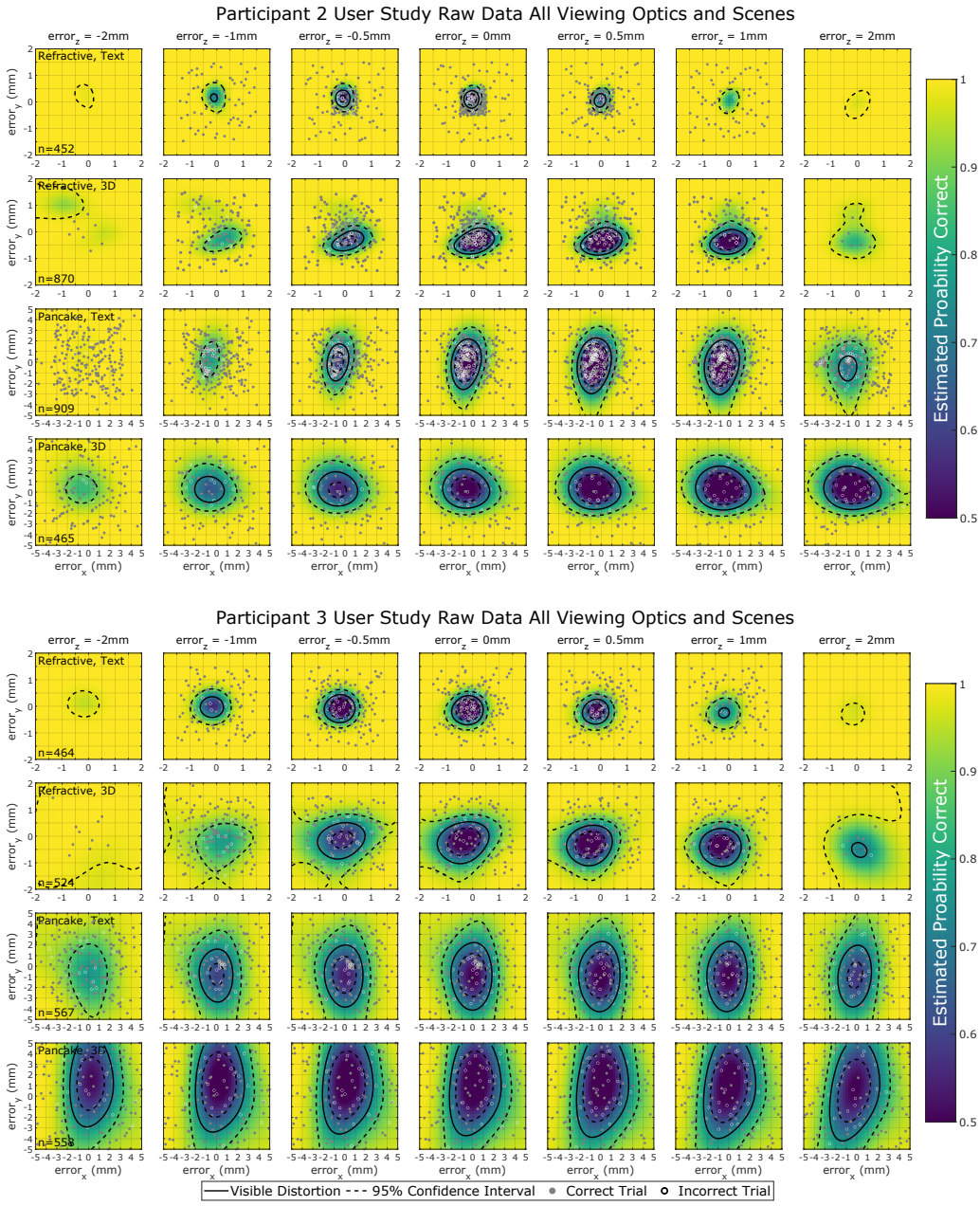

Fig. 2. *P2 and P3 Data for Two Viewing Optics and Two Scenes:* The viewing optic and scene are indicated in the top left and the number of trials are indicated in the bottom left of each subplot in the first column. Each column represents a fixed value for eye relief error where negative values place the eye closer to the display. (Top) Participant 2 Data. (Bottom) Participant 3 data.

1.2.3 Eye Tracking Latency. While specific strategies for adaptive sampling for multidimensional psychophysics are an area of active work, in an informal sense they should all target points close to hypothesized threshold locations. In observing the data collected using the adaptive EAVC method (Figure 3), we see a clear contrast to the remaining figures: there are very few trials evaluated at high tracking errors. This is because the model can rapidly determine that the threshold is at lower errors. Nonetheless, using a flexible model lets us take advantage of less efficiently-sampled data as well, and we can combine the 3D dataset collected non-adaptively (seen by itself in Figure 1) with the adaptive 4D collection, to give a combine model given in Figure 4 and reduce overall posterior uncertainty. In Figure 4 the data is merged with P2's data from Figure 1.

Fig. 3. *P2 Adaptive Data Only*: 442 trials collected using an adaptive sampling method.

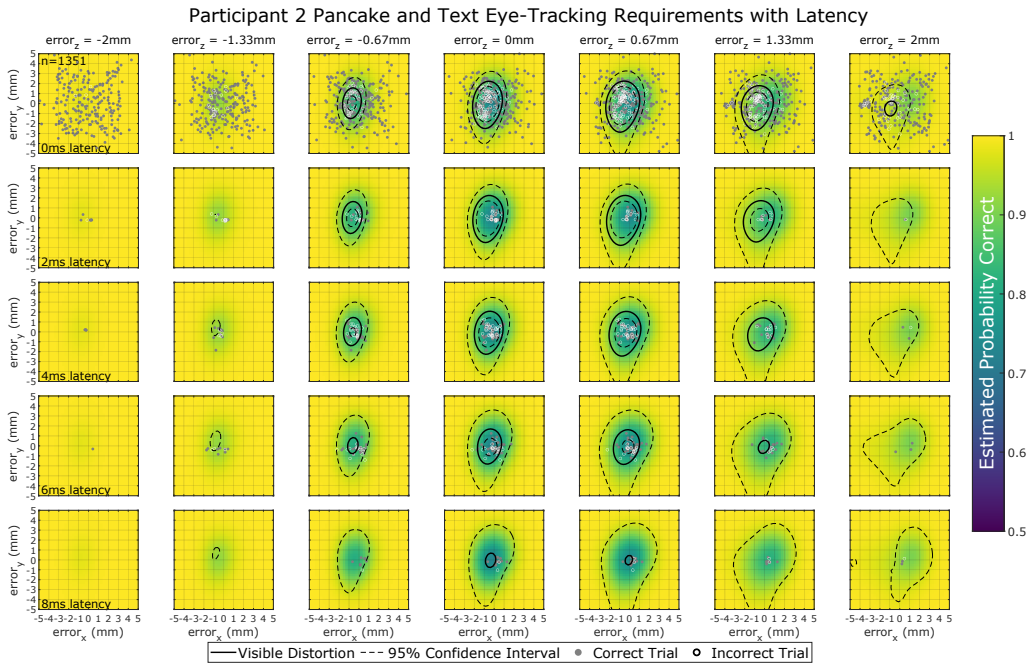

Fig. 4. *P2 Adaptive Data Combined with Existing Data:* Trials from Figure 3 merged with 3D data from Figure 1.

1.3 User Study Protocol

We elected to employ a two-interval forced choice (2IFC) protocol shown in Figure 5 to identify the thresholds in our study instead of a faster, but more subjective yes/no task to reduce variability across subjects. Data was collected in timed sessions (approximately 4 hours total for each participant in the main experiment) rather than running the experiment for a fixed number of trials. For conditions in Sections 1.2.1 and 1.2.2 of this document, the limits of the synthesized eye tracking errors were adjusted for each subject prior to the start of each new session (three to five sessions depending on participant, each session between 45-90 minutes), based on their performance from previous sessions, to more efficiently sample the parameter space for each participant.

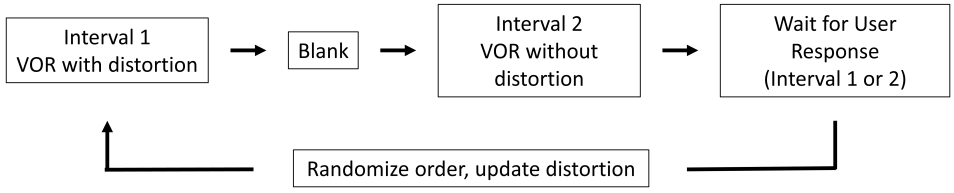

Fig. 5. *Two-interval forced choice procedure:* In the 2IFC task participants compare pupil swim during VOR to a distortion-free reference in every trial compared to a yes/no procedure where subjects compare potential distortion during VOR to a mental model of a distortion-free reference. 2IFC reduces variance across subjects because every trial is compared against the same reference, but increase data collection time.

2 VR HEADSET SIMULATOR

2.1 Hardware

Even with the advent of high-dimensional psychophysical data collection methods employed in our user study, the parameter space for quantifying optical distortions is too large to study without enforcing consistent eye movements across trials. To this end we designed our VR headset simulator to facilitate horizontal VOR eye movements for the reasons outlined in Section 3.2 of the main text. The simulator has a chin rest and can hold a bite bar to stabilize a user's head. The chin rest is mounted to a rotation bearing (CUI AMT10) and a linear encoder (EM2-2000-L) is used to read the rotation angle from a two inch, 25,000 count per revolution encoding disk (US Digital DISK2). A second linear encoder is used to measure linear translation along a sliding rail using a 2,000 line per inch transparent linear encoder strip (US Digital LIN). The sides of the chin rest are fabricated from solid pieces of aluminum to resist deformation from the forces imposed by supporting a user's rotating head and body while they turn their head. Generic motorcycle handlebar buttons on each of the handles collect user responses and input during user studies. Left eye images are rendered with a white square in the bottom left corner of the OLED TV and a Mouser OPL536A photodiode is used to synchronize the lenses from Optoma ZD302 shutter glasses which were placed into a custom housing mounted to the chin rest. The fast response time of the OLED is necessary to avoid crosstalk between the left and right eyes. The television itself is mounted on a distance adjustable track allow studies targeting specific angular resolution, field of view, or focal distance.

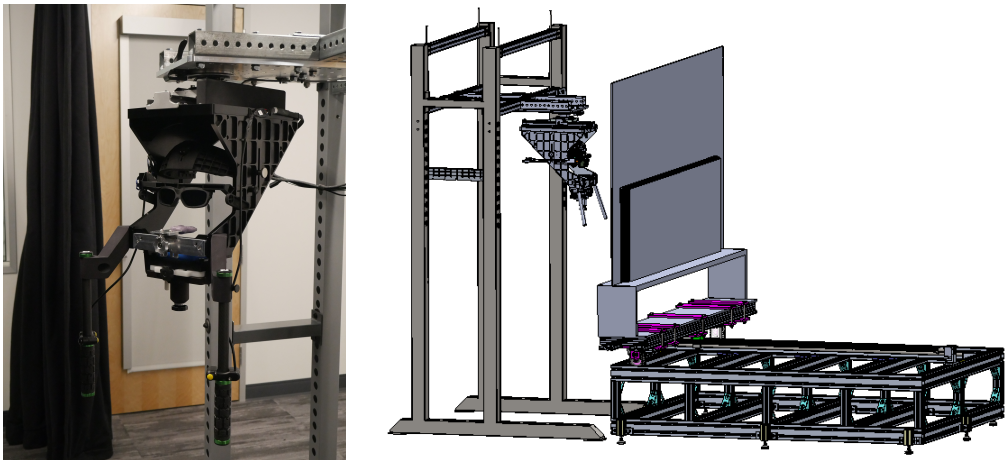

Fig. 6. *VR Simulator Hardware:* (Left) Close up of the rotating chin rest. (Right) CAD of the chinrest shown with the display on its 2m translation stand.)

2.2 Distortion Simulation Pipeline

In order to compute the correct images to display on the television to accurately recreate in-headset viewing in the perceptual testbed, the head and entrance pupil positions from the testbed hardware are combined with light field portals (LFPs) in a standard rendering engine pipeline. There are multiple frames of reference to be considered in this pipeline, as depicted in Figure 7: The rendering engine frame, where the user's entrance pupils can be placed anywhere in the virtual scene; the headset frame, where the user's entrance pupils look through the simulated viewing optics at a small display; and the real world frame, where the user's entrance pupils are placed in front of

the shutter glasses, looking at a large display in the perceptual testbed. Rendering cameras are placed at the entrance pupil locations in all three frames of reference, and these cameras are used as projectors, as well as cameras, to compute the various images of the pipeline, as explained in the remainder of this section.

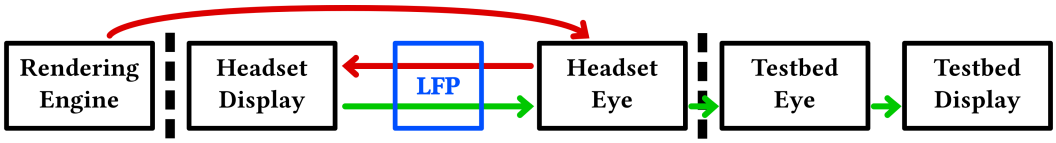

Fig. 7. Rendering pipeline for the perceptual testbed detailed in Section 2.2, which uses three separate frames of references (separated by dashed lines). Rendering is done in two main passes: the headset pass (red arrows) and testbed pass (green arrows). In the headset pass, the engine images are copied to the headset pupil cameras, and projected through LFPs to the headset displays. Then, in the testbed pass, these headset display images are viewed from a potentially different headset pupil position through the LFP, copied to the testbed pupil cameras, and projected to the testbed display which is viewed by a user in the testbed.

The distortion simulation process can be separated into two main passes, each repeated for both eyes. First, the *headset pass* (Figure 7 red arrows) computes the image that would be shown on the headset display. A camera first captures the images in the virtual world. This is where the virtual content lives, and these images are ultimately what the user should see if all distortions were perfectly compensated. The virtual cameras are set according to the interpupillary distance of the user, and they are set to parallel gaze (i.e. gaze distance is set to infinity). The virtual world images are transferred as-is to the headset cameras, and these images are then projected from the headset camera, distorted through the headset LFP, and splatted onto the headset display. This process computes images for the headset display so that a camera located at the headset entrance pupil location would see undistorted render engine images.

Second, the *testbed pass* (Figure 7 green arrows) computes the images to show on the testbed display. In this pass, cameras located at the headset pupils first capture the headset display through the LFP. Note that, since the purpose of this study is to investigate the effect of eye tracker imprecision on dynamic distortion correction (DDC), the headset pupil camera locations are, in general, different during the two passes: In the headset pass, the headset cameras are located at the reported location of the user's entrance pupils, whereas in the testbed pass, the headset cameras are located where the user's entrance pupils actually are. This distinction is key to study the effects of eye tracking errors on DDC. The testbed pass continues by transferring the headset camera images to the testbed cameras. These images are projected as a mesh and splatted onto the testbed display. Note that there are no physical optics in the testbed system, so this final projection happens through air and no LFP is required.

2.3 VR Simulator Rendering Pipeline

The scene graph representing the user and our VR simulator are shown in Figure 8 along with the scenes which are available to download as FBX files. The origin of the system is directly below the rotary encoder, in the same horizontal plane as the user's cyclopean eye. We model both eyes as a 24mm sphere, and place the corneas 9.1cm from the origin by aligning the user's corneas to a sighting reticle built into the chin rest (Figure 6). For 3D content we use a bite bar to maintain this eye position for more accurate world-locked rendering. We assume the entrance pupil of the eyes are 7.8mm from the center of rotation, so the render cameras are placed 4.2mm behind the pupil.

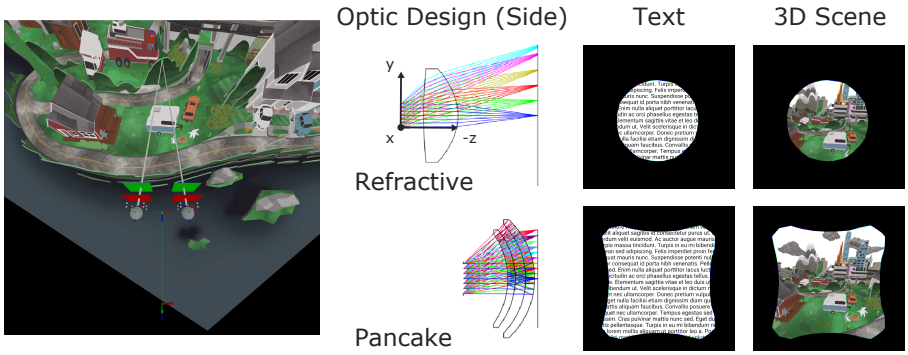

Fig. 8. *Viewing Optics and Scenes* ()Left) Scene graph representing the user and LFPs. (Middle) Illustrations of the refractive and pancake optics used in the user study. (Right) Example images of the scenes used in the user study which are available as downloadable FBX files.

In our experiments the center of the TV is placed 60cm in front of the origin (-60cm using our sign convention), and both scenes are rendered so that the user's fixation point is set to the center of the display. Within the rendering engine the text scene is rendered as a coplanar quad on the display surface and the 3D scene is placed so that the top of a bush is located at the center of the display by using a 0.025x scale, (43.3°, 14.4°, 11.9°) Euler Angle rotation, and (0.84m, -3.4m, -1.145m) translation. With these parameters set, the rendering engine runs according with the following logic:

ALGORITHM 1: VR Simulator Logic

```
set gaze to (0, 0, -60cm);
while rendering stimulus do
    update current rotation angle from rotary encoder;
    update current linear offset from linear encoder;
    update eye rotation to maintain fixation at gaze point;
    if simulating eye tracking error for distortion correction then
        add error to eye position used for distortion correction;
    end
    if simulating eye tracking error for rendering then
        add error to eye position used for rendering;
    end
    render game engine image;
    simulate distortion with LFP;
    project distorted image to TV;
end
```

3 DISTORTION METRIC

The binocular distortion metric is not intended to serve as a definitive model for distortion perception, however, it does leverage LFPs and the rendering pipeline in our simulator to model distortion correction errors while accounting for binocular impacts of optical distortion, scene structure, and user behavior which we believe are important components to develop more advanced metrics. The basic logic for the metric is shown below:

ALGORITHM 2: Binocular Distortion metric

```
foreach pixel do
    cast ray into 3D scene from cyclopean eye;
    project 3D point into left and right eye images without distortion;
    compute disparity for undistorted stereoimage pair;
    project 3D point into left and right eye images with distortion;
    determine disparity for distorted stereoimage pair;
    compute difference between the two disparities;
end
return average difference over all pixels
```

REFERENCES

Benjamin Letham, Phillip Guan, Chase Tymms, Eytan Bakshy, and Michael Shvartsman. 2022. Look-Ahead Acquisition Functions for Bernoulli Level Set Estimation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS)*. https://doi.org/10.48550/ARXIV.2203.09751

Lucy Owen, Jonathan Browder, Benjamin Letham, Gideon Stocek, Chase Tymms, and Michael Shvartsman. 2021. Adaptive Nonparametric Psychophysics. arXiv:2104.09549 [stat.ME]

Andrew B. Watson and Denis G. Pelli. 1983. QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics* 33, 2 (1983), 113–120. https://doi.org/10.3758/BF03202828