# D-SAV360: A Dataset of Gaze Scanpaths on 360° Ambisonic Videos

Edurne Bernal-Berdun, Daniel Martin, Sandra Malpica, Pedro J. Perez,
Diego Gutierrez, Belen Masia, and Ana Serrano

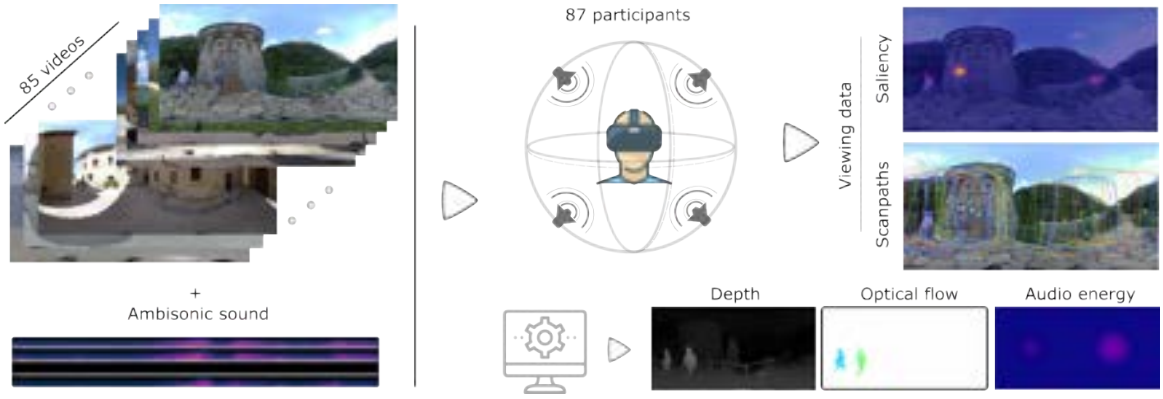


Fig. 1: We present D-SAV360, the most extensive dataset of viewing behavior on 360° ambisonic videos to date. We have collected gaze and head data from 87 different participants viewing 85 dynamic 360° videos with directional ambisonic sound, leading to a total of 4,609 scanpaths, larger than previously available datasets of comparable scope. We have thoroughly analyzed this gathered data, and provide valuable insights about viewing behavior and the importance of factors such as viewing conditions, gender, or the type of content shown. We additionally discuss potential applications for our dataset, including benchmarking of audiovisual saliency models, scanpath prediction, or stitching quality assessment, among others. Our dataset is available at https://graphics.unizar.es/projects/D-SAV360.

**Abstract**—Understanding human visual behavior within virtual reality environments is crucial to fully leverage their potential. While previous research has provided rich visual data from human observers, existing gaze datasets often suffer from the absence of multimodal stimuli. Moreover, no dataset has yet gathered eye gaze trajectories (i.e., scanpaths) for dynamic content with directional ambisonic sound, which is a critical aspect of sound perception by humans. To address this gap, we introduce D-SAV360, a dataset of 4,609 head and eye scanpaths for 360° videos with first-order ambisonics. This dataset enables a more comprehensive study of multimodal interaction on visual behavior in virtual reality environments. We analyze our collected scanpaths from a total of 87 participants viewing 85 different videos and show that various factors such as viewing mode, content type, and gender significantly impact eye movement statistics. We demonstrate the potential of D-SAV360 as a benchmarking resource for state-of-the-art attention prediction models and discuss its possible applications in further research. By providing a comprehensive dataset of eye movement data for dynamic, multimodal virtual environments, our work can facilitate future investigations of visual behavior and attention in virtual reality.

**Index Terms**—Gaze, Saliency, Fixations, Ambisonics, 360° Videos, Dataset

## 1 Introduction

As virtual reality (VR) techniques and applications continue to blossom, creating engaging experiences that exploit their potential becomes increasingly important. To achieve this, understanding and being able to systematically predict human visual behavior plays a fundamental role. For instance, a detailed understanding of visual behavior in VR can enhance storytelling by enabling creators to design more engaging experiences [45, 48], or inform the development of efficient content-aware compression [46] and rendering techniques [40] that take into account visually attractive regions to reduce computational costs. This, in turn, requires the availability of a substantial dataset that contains a wide range of scenarios, with corresponding gaze data collected from a diverse and extensive group of observers.

Sitzmann et al. [46] made one of the earliest attempts to create a comprehensive dataset of gaze data in VR. The authors recorded viewing data from 169 users in twenty-two 360° environments, analyzing it to obtain meaningful insights from which they derived applications such as alignment of cuts, panorama thumbnail generation, video synopsis, or saliency-aware image compression.

However, this dataset has two main limitations. First, all twenty-two scenes used are static, meaning that there is no motion or plot that could affect the observers' attention. This lack of dynamic scenes restricts the generalizability of the conclusions drawn from the data. And second, it is limited to visual-only stimuli, while our perception of the world around us is *multimodal*, involving inputs from multiple senses [30]. In particular, although vision is usually our predominant source of information [4, 51], auditory cues can complement our perception of the world, making it more realistic and believable. In fact, some visual stimuli may appear incomplete and break immersion without a coherent sound source [22].

While follow-up datasets have tackled these limitations by collecting gaze data for 360° videos, they either do not include audio sources [13, 26, 59] or overlook *sound directionality* [58], which is a key aspect. Humans inherently perceive sound through a combination of frequency, amplitude, and direction, which determines the two main binaural cues: the interaural time differences (ITD) between the sound's arrival at our inner ears, and the interaural sound intensity or

• *Edurne Bernal-Berdun, Daniel Martin, Sandra Malpica, Pedro J. Perez, Diego Gutierrez, Belen Masia, and Ana Serrano are with Universidad de Zaragoza - I3A. E-mail: edurnebernal | danims | smalpica | 756642 | diegog | bmasia | anase@unizar.es.*

Table 1: Comparison of D-SAV360 to currently available datasets. D-SAV360 is the first to include head and gaze data in 360° videos with directional sound (first-order ambisonics). It surpasses previous datasets in the amount of captured gaze scanpaths and also provides additional information such as audio energy maps, additional stereoscopic videos, optical flow, and depth estimation. Our dataset maintains common technical characteristics such as resolution and frame rate.

| Dataset | Head Data | Gaze Data | Ambisonic Sound | Nº of Observers | Nº of Scanpaths | Stereoscopic Images | Depth Estimation | Opt. Flow Estimation | Nº of Videos | Duration | Resolution | FPS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 360AVD [42] | ✗ | ✗ | ✓ | - | - | ✗ | ✗ | ✗ | 256 | 10s | 1K to 5K | 24 to 60 |
| Morgado et al. [37] | ✗ | ✗ | ✓ | - | - | ✗ | ✗ | ✗ | 5506 | 10s to 5h | 1K to 5K | 24 to 60 |
| Urban Soundscapes [14] | ✗ | ✗ | ✓ | - | - | ✗ | ✗ | ✗ | 130 | 60s | 4K | 30 |
| ASOD60K [58][1] | ~ | ~ | ~ | 20 | 1,340 | ✗ | ✗ | ✗ | 67 | 29.6s | 4K | 24 to 60 |
| Chao et al. [7] | ✓ | ✗ | ✓ | 15 | 675 | ✗ | ✗ | ✗ | 15 | 25s | 4K | 24 to 60 |
| **D-SAV360 (Ours)** | ✓ | ✓ | ✓ | 87 | 4,609 | ✓ | ✓ | ✓ | 85 | 30s | 4K | 60 |

[1] A recent update of this dataset (now called PAVS10K [57]) includes ambisonic sound; however, the scanpaths were collected with mono sound.

level difference (ILD) caused by the shape of our head and outer ears. This ability to locate the sound source significantly impacts our visual behavior by frequently diverting our attention toward the direction of sound sources [38]. Therefore, gaze data collected without sound directionality does not fully capture the multimodal interactions that drive human visual behavior [34, 53].

In this work, we present D-SAV360, a dataset of 4,609 gaze trajectories (i.e., scanpaths) with eye and head tracking data in 85 diverse 360° videos featuring *dynamic* scenes and directional sound recordings using first-order ambisonics. To the best of our knowledge, this is the first dataset with these characteristics (see Table 1), which we hope helps researchers derive more accurate models of human visual behavior, and develop more engaging virtual and augmented reality applications. In addition, we have also captured gaze data for *stereoscopic* viewing, to assess the impact of binocular disparity on gaze behavior.

Furthermore, we have conducted a detailed analysis of our collected data, leading to notable observations. For instance, we have identified a high inter-observer congruency, indicating the existence of underlying patterns in our data that drive human attention. Additionally, we have detected an equator bias, which had previously been observed in the context of static [46] and dynamic [59] scenes without audio. Our statistical analysis also shows that factors such as stereoscopic vs monoscopic viewing, scene content, and gender affect the tendencies of eye movements.

Lastly, we discuss various potential applications of our dataset, including its use as a benchmark for audiovisual saliency models and as a valuable resource for training novel audiovisual scanpath predictors, among others.

We publicly release our dataset to support future research, which includes 85 dynamic 360° ambisonic videos, 50 of which are also available in stereoscopic format, as well as their corresponding raw scanpaths and saliency maps. The dataset also includes six fish-eye camera recordings from each of our captured videos and additional computed information such as audio energy maps (AEM), optical flow, and depth estimations. Our dataset is available at https://graphics.unizar.es/projects/D-SAV360.

In summary, our main contributions are as follows:

- We have gathered 4,609 scanpaths for 87 participants viewing eighty-five 360° videos with ambisonic audio.
- We have studied the impact of binocular disparity on eye movements by capturing scanpaths for a subset of our dataset consisting of 50 stereoscopic 360° videos with ambisonic sound.
- We have analyzed multiple aspects of our dataset and gaze data covering aspects such as inter-observer congruency, the presence of an equator bias, gender differences, or the impact of scene content on visual behavior.
- We finally discuss further applications of our dataset and showcase its use as a benchmark for evaluating state-of-the-art audiovisual attention prediction models.

## 2 Related Work

### 2.1 360° Video Datasets

In recent years 360° video has gained significant interest, leading to the creation of several datasets for different purposes. For instance, Morgado et al. [36] collected a dataset of 360° videos to generate and align spatialized audio taking into account the visual content [36, 37]. Although this dataset contains a large number of videos, they were batch downloaded from YouTube, leading to inconsistencies in length, resolution, and frame rate, as well as a limited variety of scenes due to a lack of curation. Similarly, Rana et al. [42] created 360AVD, a dataset for learning to generate ambisonics from visual cues featuring short 10-second clips. De Coensel et al. [14] gathered a dataset of immersive audiovisual recordings of cityscapes to evaluate the perceptual influence of noise control and soundscaping measures through auralization. However, despite the usefulness of these datasets, none of them include head and gaze data, which is crucial for gaining insights into how users perceive and process auditory and visual stimuli in immersive environments.

In this direction, some works have gathered datasets of 360° videos with associated eye and head movement data to analyze the exploration behavior of users [13, 26, 55, 59]. However, their videos were played without directional audio or even without sound, which is an important element for immersion and has been shown to affect participants' visual behavior [7]. More recently, Zhang et al. [58] introduced ASOD60K, an audiovisual 360° dataset that also captured gaze and head movements. However, in their studies, their videos were presented only with mono sound, which limits the immersive experience. Further, spatialized sound has been shown to play an important role in guiding viewers' visual attention in 360° content [33, 34]. Closer to our contribution, Chao et al. [7] introduced a dataset that included ambisonics, which they used to compare viewing behavior between muted, mono, and ambisonic sound. While their dataset does provide head data, it does not include gaze data and is limited to fifteen videos and fifteen participants.

To our knowledge, there is currently no comprehensive dataset that includes 360° videos with ambisonic sound together with head and gaze data with a sufficient number of participants or videos (see Table 1 for a summary of existing datasets). To address this gap, we introduce D-SAV360, a large dataset consisting of 85 videos with ambisonic sound, accompanied by head and gaze data from 87 participants. All our stimuli have high resolution (4K) and a high frame rate (60 fps). We also provide additional data, including both monoscopic and stereoscopic images for a subset of the videos, audio energy maps, optical flow estimation, and depth estimation (see Figure 1 for a glimpse).

### 2.2 Analyzing and Predicting Viewing Behavior in VR

**Analysis** Understanding viewing behavior in VR is a key challenge for developing more engaging experiences. Some of the first works towards this goal captured and analyzed viewing data in static 360° images [41, 46], and found very relevant insights, such as the existence of an equator bias when visualizing this content. Other works [43, 45] analyzed common behaviors when viewing dynamic content, and found

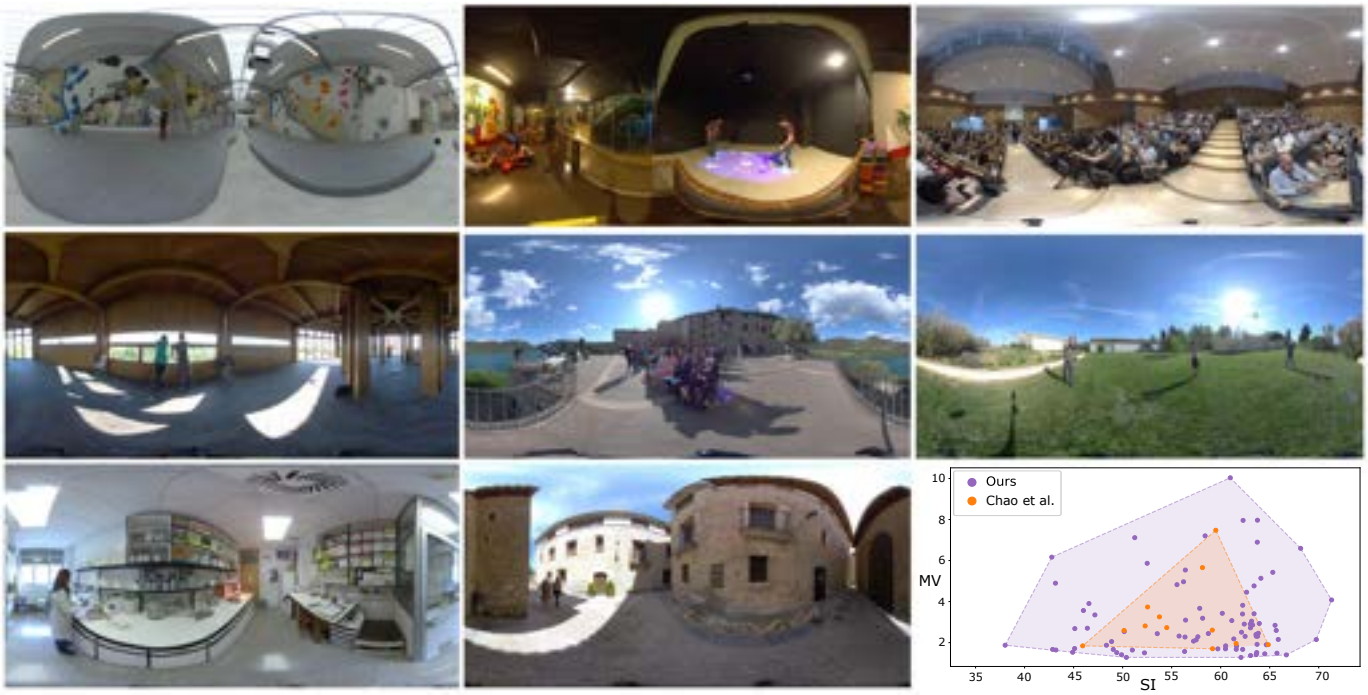


Fig. 2: Sample frames from eight of our captured videos. We have designed our dataset to contain videos with varied scene content and conditions (see Section 3), as it is reflected by the distribution of Spatial Information (SI) and Motion Vector (MV) values in the bottom-right graph. SI measures the level of spatial detail, and MV quantifies the temporal changes (i.e., movement) between frames. The higher number of videos in our dataset allows us to cover a larger MV-SI space than other works of comparable scope [7] (see Table 1 for more details about previous datasets).

that the existence of clear regions of interest affect users' viewing behavior. Other works have focused on analyzing how different visualization conditions affect eye motion. For instance, Ozcinar et al. [40] studied the relationship between the number of fixations and scene complexity. Further, Skaramagkas et al. [47] and Da Silva et al. [11] studied how emotional and cognitive processes affect some of the most common eye-tracking metrics, such as duration and number of fixations, saccadic amplitude, or blink ratio. Leveraging our dataset, we compute these metrics to analyze how stereoscopic viewing, gender, and content type affect eye movement behavior.

**Prediction** Alongside with achieving a deeper understanding of viewing behavior, several works have attempted to model and predict it. First approaches tackled this problem from a static, single-image perspective, resorting to saliency [32, 56] or scanpath [2, 31, 61] prediction. However, as VR environments are often dynamic, these models may not be sufficient for certain applications. To address this, some recent works have focused on attention prediction in 360° videos [3, 9, 12]. Nevertheless, all these models only take visual stimuli as input, and therefore they do not take into account the potential influence of sound in VR environments [30]. Particularly, auditory cues can strongly influence users' attention [28, 33]. More recent works have addressed this problem, taking into account visual and auditory information, both for traditional media [49] and 360° videos [8, 10, 60]. All of these models rely on data-driven deep-learning approaches and their accuracy depends on the availability of large and diverse datasets. Therefore, we expect that our dataset will facilitate research in this area both for benchmarking existing models and training novel ones.

## 3 Dataset Overview

### 3.1 Video Data

Our dataset comprises a total of 85 monoscopic 360° videos with first-order ambisonic sound, consisting of 50 new videos captured by us, which contain varied scenes and scene distributions, and 35 videos curated from the dataset of Morgado et al. [37], which we will refer to as Morgado in the rest of the paper for brevity. Our dataset additionally includes the same 50 videos that we captured in stereoscopic format, which is critical in virtual reality to create a sense of depth and realism for a more immersive experience. The addition of these videos allows us to investigate potential differences in user behavior for stereoscopic videos (Section 5.5). Our captured videos contain a balanced distribution of indoor and outdoor scenes, featuring both simple and complex scenarios that encompass both natural and urban environments. They also include a diverse range of exploratory scenes with multiple visual regions of interest, as well as simple scenes that offer clear regions of focus. Additionally, our ambisonic recordings feature different layouts of auditory regions of interest as well as scenes with background sound. See Figure 2 for some example frames and Section S.1 in the supplementary for representative frames of all our videos.

In addition, we estimate the optical flow for each of the videos using the deep learning model called RAFT [50] and compute the audio energy maps (AEM) with the decoder employed by Morgado et al. [37]. The optical flow provides insights about the motion patterns in the video, while the AEMs represent the spatial distribution of sound, allowing for a better understanding of how the sound interacts with the virtual environment. These additional pieces of information are valuable resources for analyzing and understanding the videos in more detail.

We evaluate the diversity of our dataset using a state-of-the-art method for characterizing 360° content [15]. Specifically, we use spatial information (SI) and motion vectors (MV) as metrics. SI measures the level of spatial detail in an image, while MV quantifies the temporal changes or movement between frames in a video sequence. To compute SI, we project the equirectangular video frames into cubemaps and apply Sobel kernels, following the approach of De Simone et al. [15]. For MV, we estimate motion vectors between consecutive frames using RAFT [50]. See Figure 2 for the results of this evaluation. Although we observe a higher representation of lower MV values due to the static camera setup used during video capture, our dataset covers a wide MV-SI space.

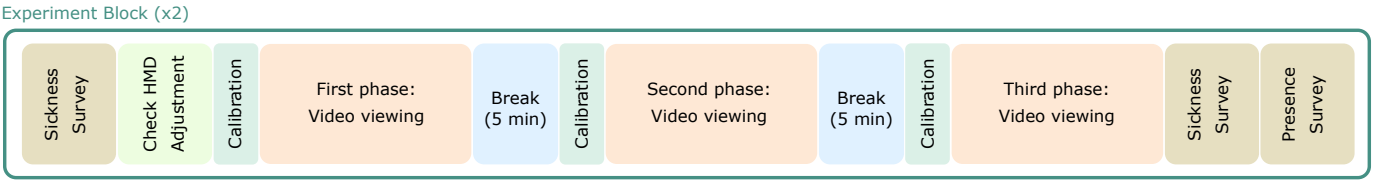


Fig. 3: Our data collection procedure can be summarized as follows: Each session is formed of two blocks - one for stereoscopic videos and the other for monoscopic videos, with a large break of fifteen minutes separating them. The diagram represents the steps of each one of these blocks. At the beginning of each block, participants complete a sickness questionnaire to asses their current state. Following this, they are instructed to adjust their HMD and perform a calibration procedure, after which they begin viewing the videos. After each phase of the block, participants are given a short break of five minutes to prevent discomfort or exhaustion. Once all phases are completed, participants fill out sickness and presence questionnaires (Section 4.2).

### 3.2 Gaze Data

Head and eye tracking data is a crucial component of our dataset as it enables comprehensive analyses of user behavior (Section 5) and facilitates future research in diverse applications in the field (Section 6). Our dataset includes head and eye tracking data for each of the 85 videos, as well as the 50 additional stereoscopic versions. This tracking data captures participants' visual behavior while watching the videos for 30 seconds. In total, we recorded 4,609 head and gaze trajectories, providing a rich source of information for future research. Figure 4 shows some representative videos with their saliency maps extracted from our collected gaze data. The collected tracking data includes head position and orientation, pupil diameter, eye openness, eye gaze vector, and gazed image coordinates, all sampled at 120Hz.

Before collecting gaze data (Section 4), we conducted power analyses to determine the necessary number of visualizations (i.e., scanpaths) per video. For detecting at least medium-sized effects (effect size of $f = 0.25$, power $1 - \beta = 0.80$), we set the minimum number of visualizations per video to 30. To ensure a large enough dataset for studying the difference between monoscopic and stereoscopic viewing, we also set the minimum number of visualizations to 30 per stereoscopic video. The total number of participants in our experiment (Section 4.2) satisfied all of these constraints. Participants explored the videos under a free-viewing condition, and they were not instructed to complete any specific task. We set for each video a fixed, random longitudinal starting position for participants to start exploring. This approach allows to start collecting data already from a point of consensus, which is usually achieved after approximately 5 seconds of free video viewing [13].

## 4 Data Collection

### 4.1 Video Collection

**Video acquisition** We captured 50 stereoscopic videos using a Kandao Obsidian S, equipped with six fish-eye lenses. Each camera records at a resolution of 3000$x$2160 pixels and a frame rate of 50 fps. We used the software provided with the camera[1] to stitch the videos and obtain depth estimations. To capture first-order ambisonic audio, a Zoom H2n microphone was used[2]. The camera was mounted statically on a tripod, ensuring that its height was similar to that of a standing person and that there was no camera movement. After the stitching process, we obtained an equirectangular representation of both the depth and video, with a resolution of 5760$x$2880 pixels per eye panorama and a frame rate of 50 fps. To obtain monoscopic videos from our 50 stereoscopic recordings, we used the right-eye video as the video for both eyes. This approach circumvents the need to extract the central view from the input left and right eyes, which could introduce significant artifacts into the resulting video [44].

In addition to our captured videos, our dataset includes 35 additional monoscopic videos from the collection of Morgado, which contains thousands of videos for monoscopic viewing. However, as the videos were extracted from YouTube in batches for a different purpose (spatial audio generation), they did not undergo a thorough curation and inspection process. Therefore, we carefully selected a diverse set of videos based on technical specifications and content variety. Our selection process focused on videos with a uniform resolution, adequate duration, and semantic diversity. We set the minimum technical requirements for the resolution at 3840$x$1920 pixels, and a minimum frame rate of 30 fps, with a preference for videos at 60 fps.

**Video processing** We standardized our videos before recording gaze data by downsampling them to a resolution of 3840$x$1920 pixels[3] per eye panorama and unifying the frame rates to 60 fps using the motion compensation interpolation method from the FFmpeg tool. Gaze data recording was limited to the most relevant 30 seconds of each video to minimize participant fatigue and maintain attention. We made this decision based on previous studies [13] and datasets gathering gaze data for sequences of similar duration [7, 58]. To evaluate the effect of these post-processing steps on video quality, we conducted a small informal study with five participants and four different videos, which showed that most participants perceived the videos before and after post-processing to have the same quality.

### 4.2 Gaze Data Collection

**Apparatus** Our stimuli were presented on an HTC Vive Pro Eye head-mounted display (HMD) with a horizontal field of view (FoV) of 110 visual degrees and a vertical FoV of 110 visual degrees, a resolution of 1440$x$1600 pixels per eye, and a frame rate of 90 fps. We used three HTC sensors to track participants' position, which was logged at 120Hz. For collecting eye tracking data we used the SRanipal Unity SDK [4] developed for the Tobii eye-tracker integrated into the HTC Vive Pro Eye. This SDK provides automated calibration and captures several gaze parameters at a high frequency of 120Hz, including the eye gaze vector, eye openness, pupil diameter, eye sensor position, and image coordinates for each gaze vector, and constitutes the primary eye-tracking data recorded in our experiments.

Our data collection study was supported by a highly customizable data collection and visualization system that we have developed for this project. To create our capture pipeline, we utilized Unity 2020.3.25f. We outline its key features in Section S.2 of the supplementary. In order to facilitate future user studies, our system is publicly available at `https://graphics.unizar.es/projects/D-SAV360`.

**Participants** A total of 87 participants voluntarily participated in the data collection study, including 41 females and 46 males, with no participants identifying themselves as non-binary, not listed, or preferring not to disclose their gender. The mean average age of the participants was 25.29 years old (STD = 8.77). Only 16% of participants reported using VR in an HMD frequently, while 38% reported never having used an HMD before. All participants were economically compensated and provided written consent for their voluntary participation in the study. They were naïve about the final purpose of the study, and they all reported normal or corrected-to-normal vision and audition. We additionally conducted tests of visual and auditory acuity (see

[1] https://www.kandaovr.com/
[2] https://zoomcorp.com/
[3] The raw videos with full resolution can be available upon request.
[4] https://developer-express.vive.com/resources/vive-sense/eye-and-facial-tracking-sdk/

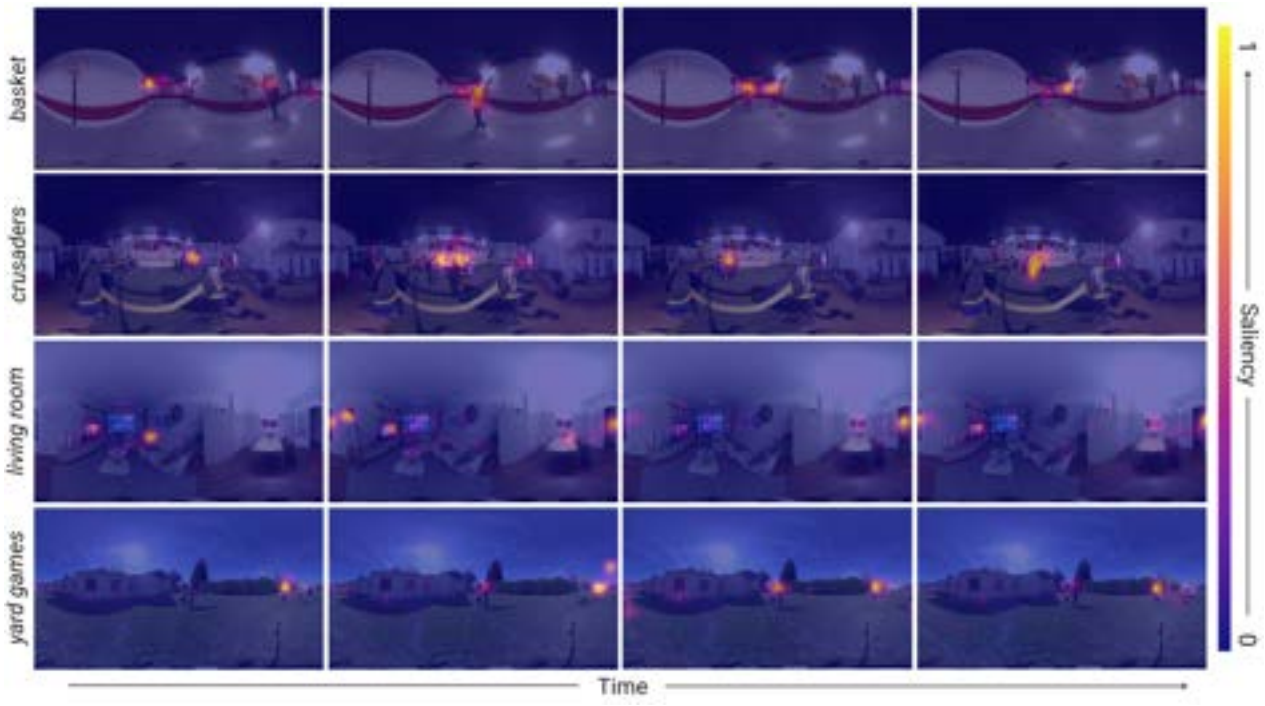


Fig. 4: Example saliency maps for four sample videos, showing the evolution of participants' attention over time (from left to right). It can be seen that attention changes over time, being mostly directed to visual (e.g., warriors in the second row) or auditory (e.g., youngsters in the middle of the frame on the last row) regions of interest. Warmer colors correspond to areas with higher saliency.

Section 4.2 for more details). As a result of these assessments, seven participants out of the initial pool of 94 participants were excluded from the study. The research protocol was approved by the *Comité de Ética de la Investigación de la Comunidad Autónoma de Aragón* (CEICA).

Procedure After the reception, participants were informed about the data collection procedure and signed a written consent for participation. Then, they filled out a pseudo-anonymous demographic questionnaire, and a pre-experiment short version of the sickness questionnaire (SSQ) introduced by Kennedy et al. [21]. These questionnaires are included in Section S.3 in the supplementary. When both questionnaires were completed, we evaluated the visual and auditory acuity of the participants. We established several inclusion criteria for our study. Specifically, we required that participants passed the Snellen chart test, the Ishihara test, and the Titmus stereoscopic test to confirm their visual acuity, color vision, and stereopsis, respectively. Additionally, we administered our stereo sound perception test to confirm participants' ability to perceive stereo sound. More details about these tests can be found in Section S.4 in the supplementary. Participants who did not meet these criteria were not included in our study to avoid confounding effects resulting from different forms of vision or auditory impairment. Investigating the visual behavior of participants with visual or auditory impairments is outside the scope of this project and would be an important step for future studies.

After the aforementioned tests were completed, the experiment started. The experimenter explained to the participants the different phases and tasks of the experiment and helped them correctly adjust the HMD. Participants were instructed to freely explore each of the videos by rotating in place while in a standing position. The experiment was split into two main blocks, one where we showed the participant 25 stereoscopic videos and another one where we showed 30 monoscopic videos. Between blocks, participants had a rest of approximately fifteen minutes. Our procedure ensured that each visualization block contained either stereoscopic or monoscopic videos to help prevent participants from getting fatigued or frustrated from having to switch between the two types of videos. Figure 3 summarizes the different steps performed in each block. Note that videos from both blocks were different, and thus participants never saw the same video more than once. For each participant, we randomly selected which block to start with, and within each block, the video order was also randomized.

Each block was divided into three phases consisting of ten videos each, except for the third phase of the stereoscopic videos block, which consisted of only five videos. Between phases, there was a short break that lasted up to five minutes. Additionally, when the first block was completed, they were asked to take off the HMD and rest for at least fifteen minutes. At the beginning of each block, and before each subset of ten videos, we performed the integrated HTC Vive Pro Eye calibration procedure to ensure correct eye tracking throughout the whole experiment. Following the procedure established by Sitzmann et al. [46], to guarantee that all participants started at the same position for a given video, they had to find a red cube in a black room and lock their gaze on it for the next video to start. In order to maintain participants' engagement throughout the experiment and detect non-compliant participants we included a simple four-alternative forced choice sentinel question in each of the phases, after one randomly chosen video. Participants had to use the controller to select the correct answer. These questions were carefully designed to be very simple and to not interfere with the free viewing condition of our experiment. Refer to Section S.5 in the supplementary for more details and a compilation of these sentinel questions.

When each block was completed, participants filled out a post-experiment sickness questionnaire identical to the pre-experiment one, and a presence questionnaire to evaluate the experience. The whole procedure took one hour and a half on average, and participants were then economically compensated for their participation. After each experiment, we conducted several important sanity checks on our eye tracking data. These included verifying sufficient eyes' openness, valid pupil diameter values, and discarding points with outlier velocities (see Section 5.1). Furthermore, we checked for consecutive measures to identify any instances of eye tracker loss and ensured that the time between eye tracking samples was approximately 8ms (i.e., 120 Hz).

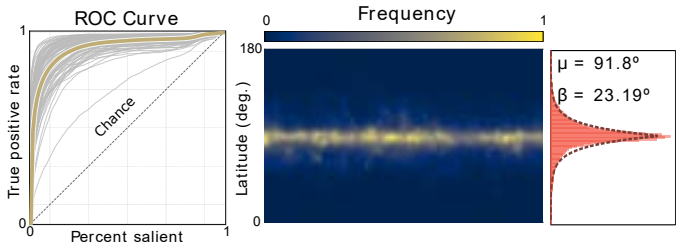


Fig. 5: Left: ROC curve showing inter-observer congruency for each video (gray) and the average for all videos (yellow). The rapid convergence to the maximum suggests a strong agreement between participants. Center: Equator bias present in our data represented as an averaged saliency map obtained from all videos' fixations across time. Warmer colors indicate higher fixation frequencies. Right: Laplacian fit describing the equator bias of the fixation distribution along latitude.

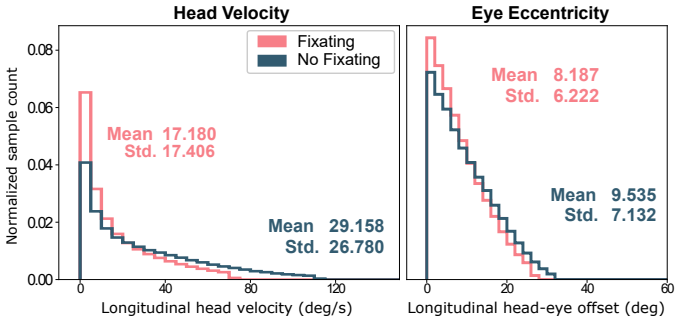


Fig. 6: Distribution of longitudinal head velocity (left) and eye eccentricity (right) while the observer was fixating (pink) and not fixating (blue). Our data reveal a lower tendency towards re-orienting conducts than that found in 360° images [46].

## 5 Analysis

Leveraging our collected gaze data we perform here a detailed analysis of viewing behavior while watching dynamic content. We focus on gaze data since we did not find significant effects of the studied factors in the presence and sickness questionnaires. We first assess inter-observer congruency, which is the basis for further exploration (Section 5.2), as well as the presence of the well-known equator bias in our data (Section 5.3). We then evaluate whether other insights derived for static content—such as the existence of two different viewing modes [46]—hold when participants are watching dynamic content (Section 5.4). The varied nature of our dataset further enables us to contrast monoscopic with stereoscopic viewing (Section 5.5), and to assess the impact of the type of visual content on viewing behavior (Section 5.6). Finally, and while not the main scope of this work, we succinctly look into gender differences in viewing patterns (Section 5.7).

### 5.1 Data Pre-Processing

Prior to the analysis, we classify data into fixations and saccades, and compute saliency information for our videos. Classification of eye-tracking data is done using a Velocity-Threshold Identification (I-VT) algorithm. For this algorithm, fixations are defined as gaze points with velocities below 30% of the maximum velocity (computed for each video and participant) and a minimum duration of 100ms. Prior to the classification, we remove outliers by discarding gaze points with velocities in the top 2% [45], and filter velocities with a running average of two samples [46]. We derive saliency maps from the fixations by tallying the number of fixations at each pixel and then convolving the resulting data with a Gaussian with a standard deviation of 5° of visual angle. This convolution yields continuous saliency maps, which provide a representation of the regions in the video that are most salient to the observers. The longitudinal standard deviation of the Gaussian kernel is scaled proportionally to the latitude of the panorama to account for the distortions present in the equirectangular projection [46]. Since our videos have a frame rate of 60 fps, each frame lasts for 16.67 ms. Given that fixations typically last for at least 100 ms, to compute saliency maps we chose to group fixations performed over an 8-frame interval, which corresponds to 133.36 ms. This approach ensures that each saliency map includes a sufficient number of fixations, while providing a more informative visualization of the data.

### 5.2 Inter-Observer Congruency

The goal of this section is to assess whether viewing behavior is similar between participants watching the same video. To achieve this, we follow common practice and employ the receiver operating characteristic (ROC) curve [23, 31, 46]. The ROC curve is computed by measuring the percentage of fixations from each participant that fall within the top n% most salient regions of the ground-truth saliency map, which is obtained with the data of all other participants, excluding the respective participant. ROC curves are computed for each participant for one-second intervals, and then averaged across participants for each of our videos. Figure 5 (left) shows the resulting ROC curves for all our videos and the average across videos. The rapid convergence of these curves towards the maximum value of 1 indicates a high level of agreement between participants. This is consistent with findings from previous studies on traditional displays [20] and static 360° images [46].

### 5.3 Equator Bias

Previous research has consistently shown an equator bias in visual attention in traditional images [20, 39], 360° panoramas [46], and 360° videos [55, 59]. To determine whether this equator bias is present in our data, we average the saliency maps across all videos, and exclude the first five seconds of fixations from each video since participants started at a fixed equatorial position. As depicted in Figure 5 (center), our results show a higher density of fixations along the equator, consistent with prior findings. Following Sitzmann et al., this equator bias can be further quantified by fitting a Laplacian distribution to the longitudinal component of the overall fixations (Figure 5, right). Interestingly, the fit to our data yields as parameters $\mu = 90°$ and $\beta = 27°$, which are very close to those observed by Sitzmann et al. in static panoramas.

### 5.4 Relation between Head and Gaze Statistics

Sitzmann et al. [46] identify two modes of behavior of observers while freely viewing 360° static images, and they term these modes *attention* and *re-orientation*. Participants show lower head velocities and eye eccentricities[5] when they are fixating, revealing an attention mode, than when they are not (re-orientation mode). We investigate whether these two modes are also observable when viewing 360° *videos*.

The results are shown in Figure 6, depicting head velocities and eye eccentricities when observers are fixating (pink histograms) and when they are not (blue histograms). While we do observe a certain difference between these two modes, it is much smaller than the one described by previous work for the case of static images. Static images yielded differences of 30 deg/s between mean head velocities, and of 4 deg in mean eye eccentricities [46], whereas in our case these differences are around 10 deg/s and 1 deg, respectively. Further, distributions of head velocity and eye eccentricity in videos, both when fixating and not fixating, are closer to the ones found in images when observers were fixating (so-called *attention* mode).

This discrepancy between images and videos could be due to a lack of wide re-orientation movements in video viewing. Video content typically exhibits smooth transitions between regions of interest, thereby discouraging frequent broad viewport changes. However, when presented with images, observers engage in more exploratory behaviors due to the static nature of the content and the consequent lack of a temporal thread. We therefore do not observe both *attention* and *re-orientation* modes in 360° video viewing, but rather a lower tendency towards re-orientation.

[5] Offset, in visual degrees, between the directions of the head and the eye.

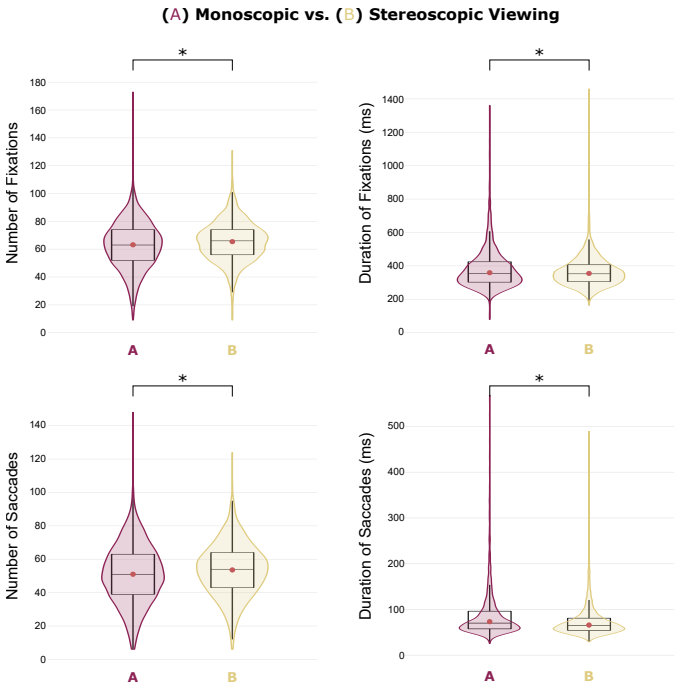


Fig. 7: Violin plots showing the distribution of the metrics with significant differences (Yuen-Welch test) when comparing monoscopic (wine) and stereoscopic (sand) viewing. Our results show that stereoscopic viewing leads to a significantly higher number of fixations and saccades with shorter durations. Trimmed means are denoted by red circles, and (*) indicates a significant difference of small effect size.

### 5.5 Monoscopic vs. Stereoscopic Viewing

Previous datasets have mainly focused on monoscopic content (see Table 1). However, since stereoscopic viewing involves additional factors like vergence and accommodation, it may influence the patterns of fixation and saccade movements. We therefore explore here the influence of stereo viewing mode (monoscopic vs. stereoscopic) on gaze behavior in 360° videos.

Statistical analysis We follow the same procedure in Sections 5.5, 5.6 and 5.7. As dependent variables, we computed for each video and participant the number of fixations, average fixation duration, number of saccades, average saccade duration, and average saccade amplitude. For hypothesis testing we used the Yuen-Welch test for trimmed means for dependent samples [54] that is robust to non-normality and heterogeneity, and supports dependent samples like ours. Then, we used the Algina-Keselman-Penfield (AKP) robust standardized difference for calculating the effect sizes. We report the results of the analyses indicating the Yue-Welch test parameter and statistic, the significance (p-value), and the AKP effect size estimate with 95% confidence intervals. We set the threshold for significance at $p = 0.01$.

Results and discussion Figure 7 displays the most relevant results for our discussion, while the remaining results can be found in Section S.6 in the supplementary. Our analysis revealed significant differences of small effect size in the number ($t_{Yuen}(985) = 6.48$, $p < 0.001$, $CI_{AKP} = [0.14, 0.23]$) and duration ($t_{Yuen}(985) = 5.11$, $p < 0.001$, $CI_{AKP} = [0.11, 0.20]$) of fixations, as well as in the number ($t_{Yuen}(985) = 6.11$, $p < 0.001$, $CI_{AKP} = [0.11, 0.22]$) and duration ($t_{Yuen}(985) = 6.72$, $p < 0.001$, $CI_{AKP} = [0.12, 0.24]$) of saccades. No significant difference was found in saccade amplitude.

These results show that stereoscopic viewing leads to a significantly higher number of fixations and saccades that were shorter in duration. This is consistent with previous studies on stereoscopic images [19] and movies [17] visualized in traditional displays. One possible explanation for this behavior is that the additional depth information provided by the stereoscopic content may prompt the viewer to shift their gaze more frequently to different depth planes. While these differences were statistically significant, the small effect size indicates that their practical impact may not be substantial. Nonetheless, these findings provide a preliminary analysis of the potential differences between stereoscopic and monoscopic viewing in 360° videos and underscore the need for further research to investigate the underlying mechanisms of visual processing during stereoscopic viewing.

Additionally, we investigate the effect of stereoscopic visualization on overall saliency. We compute the widely-used linear correlation coefficient (CC) metric [23] to compare the saliency maps obtained for the same videos under monoscopic and stereoscopic viewing. This metric ranges from -1 (perfectly inversely correlated) to 1 (perfectly correlated). Across all videos, the average CC score is 0.77, indicating that regions that attract viewers' attention in both stereoscopic and monoscopic videos are consistent. These results have important implications for visual attention prediction and suggest that the choice of viewing condition may not strongly impact visual attention.

### 5.6 Exploratory vs. Focused Content

Existing studies have shown that the presence of regions of interest (ROIs) significantly affects observers' viewing behavior during visualization of 360° narrative content [29, 45]. Building upon these findings, we set out to analyze whether the visual content of the video also has an influence in gaze behavior, even in free exploration viewing of videos that lack a strong narrative story (as in the case of our dataset). To assess this, we classified our videos into two categories based on the presence of ROIs: *exploratory* and *focused* videos. Exploratory videos are those that lack clear visual or auditory ROIs, whereas focused videos feature distinct ROIs, either auditory, visual, or both (see Section S.7 in the supplementary for details).

Results and discussion The Yuen-Welch test revealed significant differences of medium size for the number ($t_{Yuen}(350) = 10.59$, $p < 0.001$, $CI_{AKP} = [0.32, 0.57]$) and duration ($t_{Yuen}(350) = 13.06$, $p < 0.001$, $CI_{AKP} = [0.49, 0.67]$) of fixations, and the number ($t_{Yuen}(350) = 9.57$, $p < 0.001$, $CI_{AKP} = [0.31, 0.47]$) and amplitude ($t_{Yuen}(350) = 8.58$, $p < 0.001$, $CI_{AKP} = [0.28, 0.44]$) of saccades. No significant difference was found for the duration of saccades. These effects can be seen in Figure 8 (please refer to Section S.6 in the supplementary for the complete results).

Our analysis shows that videos lacking clear ROIs result in more frequent and shorter fixations, also with more saccades and with greater amplitudes. These findings are in agreement with previous works in narrative content [29, 45], and suggest that participants exhibit a more exploratory behavior when there is no clear focal point, frequently shifting their gaze across multiple regions of the video.

### 5.7 Gender Differences

Since our dataset includes a balanced number of male and female participants[6], we explore potential differences in gaze behavior between genders.

Results and discussion We found significant differences between genders of small effect size in the number of fixations ($t_{Yuen}(1215) = 7.48$, $p < 0.001$, $CI_{AKP} = [0.15, 0.24]$), and the number ($t_{Yuen}(1215) = 8.73$, $p < 0.001$, $CI_{AKP} = [0.17, 0.28]$), duration ($t_{Yuen}(1215) = 10.66$, $p < 0.001$, $CI_{AKP} = [0.19, 0.28]$) and amplitude ($t_{Yuen}(1215) = 6.73$, $p < 0.001$, $CI_{AKP} = [0.11, 0.21]$) of saccades. We did not find significant differences in the duration of fixations. Figure 9 illustrates these effects (see Section S.6 in the supplementary for additional results).

Specifically, female participants exhibit a lower number of fixations and saccades. Also, their saccades present a smaller amplitude and longer duration compared to male participants. These findings align with prior research on gender differences in conventional (2D) image exploration [27, 35]. Our insights highlight the importance of considering potential gender differences in datasets and study designs.

[6] Our questionnaire follows gender-inclusive language in research guidelines and provides options for non-binary, unlisted, or undisclosed gender identities. However, all participants identified as male or female.

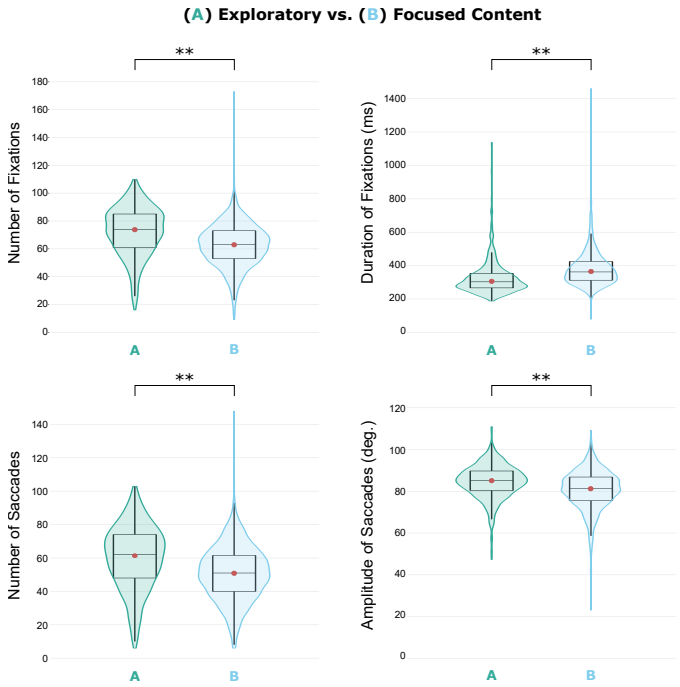


Fig. 8: Violin plots showing the distribution of the metrics with significant differences (Yuen-Welch test) when comparing exploratory (teal) and focused (cyan) content. Videos with exploratory content (i.e., lacking clear ROIs) produce more frequent and shorter fixations and more saccades with greater amplitudes. Trimmed means are denoted by red circles, and (**) indicates a significant difference of medium effect size.

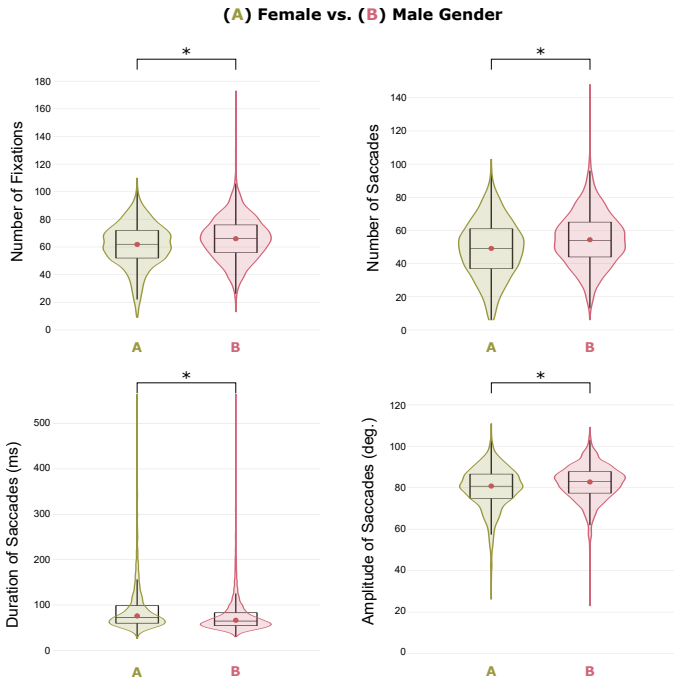


Fig. 9: Violin plots showing the distribution of the metrics with significant differences (Yuen-Welch test) when comparing female (olive) and male (pink) genders. Female participants show a lower number of fixations and saccades. Also, their saccades present smaller amplitudes and last longer compared to male participants. Trimmed means are denoted by red circles, and (*) indicates a significant difference of small effect size.

## 6 Applications of our Dataset

D-SAV360 is a versatile and valuable resource that can benefit a wide range of research areas. We outline in this section some potential applications for our dataset.

Benchmarking of audiovisual saliency models. Saliency prediction is a well-established area of research in computer vision and graphics. Over the years, numerous computational models have been developed to predict where people look in images and videos using various approaches, including the emerging field of 360° content. However, comparing the performance of different models is difficult as they are typically evaluated on particular subsets of their data. To address this issue, researchers have utilized existing datasets such as the widely-used MIT saliency benchmark for traditional images [5] and the dataset introduced by Sitzmann et al. [46] specifically designed for 360° images. These datasets have played a critical role in evaluating the efficacy of saliency models and benchmarking their performance against each other. Similarly, our dataset can serve as a resource for benchmarking predictive saliency models of audiovisual attention in 360° videos. We showcase this application scenario by evaluating the performance of recent state-of-the-art audiovisual 360° saliency predictors on D-SAV360, specifically those proposed by Cokelek et al. [10] and Chao et al. [8]. We chose to implement Cokelek et al.'s method on top of SST-Sal [3], as it requires a video saliency predictor as a base architecture. Please refer to Section S.8 in the supplementary for implementation details. We show in Table 2 the performance of these two methods using three widely used metrics [16]. As we show in Table 1, our dataset is the first to provide gaze data for 360° videos with ambisonic audio. Therefore, a direct comparison of these models' performance with previous datasets is challenging in this particular application, as the trained models would be inherently different due to the different training data these other datasets provide.

Scanpath prediction. The prediction of scanpaths (complete paths of observer fixations over time), similar to saliency, is gaining attention for its potential applications in fields such as image compression and foveated rendering [1, 6, 31]. With our dataset's rich gaze tracking data, novel data-driven scanpath prediction models could be developed. Furthermore, the availability of gaze tracking data from multiple viewers watching the same 360° video provides the opportunity to evaluate the effectiveness of these algorithms in predicting scanpaths across different individuals, thereby facilitating the development of more robust and generalizable models.

Stitching quality assessment. Our six-fisheye recordings and stitched videos provide an opportunity for researchers to explore and create new stitching algorithms leveraging 360° stitching quality assessment metrics [25, 52]. Our dataset includes a diverse range of videos with varying light conditions, indoor-outdoor scenes, and temporal and spatial complexity, presenting challenging video conditions for testing the performance of novel stitching algorithms.

Motion parallax. Captured 360° video fails to deliver experiences with motion parallax due to the lack of wide parallax information resulting from the fixed viewpoint used during the recording. This static viewpoint can potentially induce motion sickness or lack of immersion, as the observers tend to move their heads slightly while viewing the content through HMD. To tackle this limitation, state-of-the-art techniques, such as the one proposed by Serrano et al. [44], leverage depth estimations to generate a dynamic viewpoint that simulates motion parallax. Our dataset can serve as a valuable resource for evaluating and validating the effectiveness of such methods, allowing researchers to improve and enhance the immersive experience for viewers.

Binaural ambisonics decoders Our first-order ambisonic recording dataset can be a valuable resource for researchers looking to evaluate and improve binaural decoding techniques. While ambisonics are versatile in reproducing playback on various speaker arrays and easy to manipulate in post-production, decoding ambisonics binaurally through headphones can lead to a reduction in audio quality [24]. Our dataset offers a complex soundscape of real-world audio with overlapping sources, providing a challenging environment for researchers to test and compare different binaural decoding methods.

Table 2: Performance of two state-of-the-art models [8, 10] for 360° audiovisual saliency prediction. Evaluation is carried out using three commonly used metrics: linear correlation coefficient (CC), similarity metric (SIM), and Kullback-Leibler divergence (KLDiv). We show results when benchmarking the original models with our dataset. The mean value of the mean calculated across video frames is presented along with the corresponding mean standard deviation in parentheses.

| Models | CC ↑ | SIM ↑ | KLDiv ↓ |
|---|---|---|---|
| AVS360 [8] | 0.245 (0.098) | 0.248 (0.061) | 10.800 (2.128) |
| Cokelek et al. [10] + SST-Sal [3] | 0.370 (0.121) | 0.313 (0.076) | 9.438 (2.306) |

## 7 Conclusion

We have presented D-SAV360, a comprehensive dataset that includes head and gaze tracking data from 87 participants observing 85 different 360° ambisonic videos in VR, including both stereoscopic and monoscopic videos. Our dataset contains 4,609 distinct scanpaths, making it a valuable resource for studying and modeling human visual behavior in immersive environments. In addition to gaze data, our dataset includes other important features such as audio energy maps, depth, and optical flow estimations.

Leveraging our dataset, we have performed an extensive analysis that has yielded important insights with direct implications for future research. We have confirmed previous findings on equatorial bias and head-gaze coordination statistics observed in static 360° images [46]. Moreover, our findings suggest that stereoscopic viewing results in more fixations and shorter saccades, which is consistent with previous studies on stereoscopic images and movies viewed on traditional displays [17, 19]. We have also observed that content lacking clear regions of interest leads to more explorative behavior with shorter and more frequent fixations, as previously reported for VR cinematic content [29,45]. Our analysis highlights the importance of gender-inclusive research and diverse participant pools. Specifically, we found that females tend to exhibit a lower number of fixations and saccades, which are smaller in amplitude and longer in duration, compared to males when observing 360° videos. Finally, we see great potential for our dataset to be used for a variety of applications such as audiovisual attention modeling and benchmarking, and evaluation of different techniques such as stitching algorithms, motion parallax reproduction, and binaural ambisonic decoding. Our comprehensive dataset and our data collection and visualization system will be publicly available to the research community, and we hope that it will inspire further research and advancements in this field.

Limitations and future work. While our dataset and study offer valuable resources and insights into studying human visual behavior in VR environments, there are still potential areas for future research and improvement. One limitation of our dataset is that it only includes first-order ambisonics, which may not fully capture the complexity of auditory perception in VR. Nevertheless, first-order ambisonics remain widely adopted due to the current high cost of microphones with higher-order array configurations and the challenges associated with storing and transmitting the necessary amount of data for higher orders [18]. Future research could explore more advanced audio techniques to investigate the relationship between auditory perception and visual behavior in more detail.

Additionally, while we focused on gaze behavior in our study, it would be valuable to investigate other modalities such as body posture or physiological responses. For example, combining eye tracking with electroencephalography (EEG) or electromyography (EMG) could provide a more comprehensive understanding of the relationship between visual behavior and physiological responses in VR. While our videos were designed to avoid eliciting intense emotional responses, studying how the emotional valence of videos can influence both gaze behavior and physiological responses is an intriguing line of future work.

Although 360° videos are widely used in numerous applications, they lack important cues such as motion parallax and 6-degrees-of-freedom interaction. These cues are expected to have an impact on the visual behavior of observers [44]. Future research could focus on assessing gaze behavior within synthetically generated content that offers immersive and interactive experiences.

To improve the generalizability of our results, future studies may benefit from recruiting participants from a wider range of ages, backgrounds, and demographics. Although our study achieved a balanced distribution of male and female participants and we used gender-inclusive language in our recruitment process, there is still room for further diversity. Moreover, while we accomplish a varied range of participants ages (18 to 64 years old) there is a skew toward younger populations, thus a wider range of ages could be explored. Building on these foundations, future research can strive for even greater diversity in participant recruitment to explore how cultural, demographic, and age factors may impact visual behavior in immersive environments.

Looking ahead, we see potential for our data collection and visualization system to be leveraged in exploring new scenarios such as interactive scenes or social VR, investigating the influence of tasks, or studying the presence of motion parallax. By making our methodology and data collection system publicly available, we hope to encourage and support future research in this exciting field.

## Acknowledgments

We extend our gratitude to the members of the Graphics and Imaging Lab for their support and collaboration in the video recordings, especially to Maria Plaza for her valuable assistance during the capture process. We would also like to thank the anonymous reviewers for their insightful comments and the participants in the experiment. Our work has received funding from the European Union's Horizon 2020 research and innovation programme (ERC project CHAMELEON, Grant No 682080, and Marie Sklodowska-Curie project PRIME, Grant No 956585). This project was also funded by the Spanish Agencia Estatal de Investigación (projects PID2019-105004GB-I00 and PID2022-141539NB-I00). Additionally, Sandra Malpica, Daniel Martin, and Edurne Bernal-Berdun were supported by a Gobierno de Aragon predoctoral grant (2018-2022, 2020–2024, and 2021–2025, respectively).

## References

[1] E. Arabadzhiyska, O. T. Tursun, K. Myszkowski, H.-P. Seidel, and P. Didyk. Saccade landing position prediction for gaze-contingent rendering. *ACM Trans. on Graphics*, 36(4), 2017. 8

[2] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proc. International Conference on Computer Vision (ICCV) Workshops*, pp. 2331–2338, 2017. 3

[3] E. Bernal-Berdun, D. Martin, D. Gutierrez, and B. Masia. SST-Sal: A spherical spatio-temporal approach for saliency prediction in 360º videos. *Computers & Graphics*, 2022. 3, 8, 9

[4] E. Burns, S. Razzaque, A. T. Panter, M. C. Whitton, M. R. McCallus, and F. P. Brooks. The hand is slower than the eye: A quantitative exploration of visual dominance over proprioception. In *IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR)*, pp. 3–10, 2005. 1

[5] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. Mit saliency benchmark. http://saliency.mit.edu/. 8

[6] F.-Y. Chao, C. Ozcinar, and A. Smolic. Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need. In *IEEE 23nd International Workshop on Multimedia Signal Processing*, pp. 1–6, 2021. 8

[7] 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 *International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6, 2020. 2, 3, 4

[8] F.-Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic. Towards audio-visual saliency prediction for omnidirectional video with spatial audio. In *International Conference on Visual Communications and Image Processing (VCIP)*, pp. 355–358. IEEE, 2020. 3, 8, 9

[9] H.-T. Cheng, C.-H. Chao, J.-D. Dong, H.-K. Wen, T.-L. Liu, and M. Sun. Cube padding for weakly-supervised saliency prediction in 360 videos. In

*Proc. Computer Vision and Pattern Recognition (CVPR)*, pp. 1420–1429, 2018. 3

[10] M. Cokelek, N. Imamoglu, C. Ozcinar, E. Erdem, and A. Erdem. Leveraging frequency based salient spatial sound localization to improve 360° video saliency prediction. In *International Conference on Machine Vision and Applications (MVA)*, pp. 1–5, 2021. 3, 8, 9

[11] A. C. da Silva, C. A. Sierra-Franco, G. F. M. Silva-Calpa, F. Carvalho, and A. B. Raposo. Eye-tracking data analysis for visual exploration assessment and decision making interpretation in virtual reality environments. In *Symposium on Virtual and Augmented Reality (SVR)*, pp. 39–46, 2020. 3

[12] Y. Dahou, M. Tliba, K. McGuinness, and N. O'Connor. ATSal: An attention based architecture for saliency prediction in 360 videos. *Lecture Notes in Computer Science*, pp. 305–320, 2020. 3

[13] E. J. David, J. Gutiérrez, A. Coutrot, M. P. Da Silva, and P. L. Callet. A dataset of head and eye movements for 360° videos. In *Proc. of ACM Multimedia Systems Conference*, pp. 432–437, 2018. 1, 2, 4

[14] B. De Coensel, K. Sun, and D. Botteldooren. Urban soundscapes of the world: Selection and reproduction of urban acoustic environments with soundscape in mind. In *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, vol. 255, pp. 5407–5413, 2017. 2

[15] F. De Simone, J. Gutiérrez, and P. Le Callet. Complexity measurement and characterization of 360-degree content. *Electronic Imaging*, 31:1–7, 2019. 3

[16] J. Gutiérrez, E. David, Y. Rai, and P. L. Callet. Toolbox and dataset for the development of saliency and scanpath models for omnidirectional 360° still images. *Signal Processing: Image Communication*, 69:35–42, nov 2018. 8

[17] J. Häkkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman. What do people look at when they watch stereoscopic movies? In *Stereoscopic Displays and Applications XXI*, vol. 7524, pp. 129–138, 2010. 7, 9

[18] E. Hellerud, A. Solvang, and U. P. Svensson. Spatial redundancy in higher order ambisonics and its use for lowdelay lossless compression. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 269–272, 2009. 9

[19] L. Jansen, S. Onat, and P. König. Influence of disparity on fixation and saccades in free viewing of natural scenes. *Journal of Vision*, 9(1), 2009. 7, 9

[20] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *Proc. International Conference on Computer Vision (ICCV)*, pp. 2106–2113, 2009. 6

[21] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The International Journal of Aviation Psychology*, 3(3):203–220, 1993. 5

[22] H. Kim and I. K. Lee. Studying the effects of congruence of auditory and visual stimuli on virtual reality experiences. *IEEE Trans. on Visualization and Computer Graphics*, 28(5):2080–2090, 2022. 1

[23] O. Le Meur and T. Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, 2013. 6, 7

[24] H. Lee, M. Frank, and F. Zotter. Spatial and timbral fidelities of binaural ambisonics decoders for main microphone array recordings. *Journal of the Audio Engineering Society*, 2019. 8

[25] J. Li, K. Yu, Y. Zhao, Y. Zhang, and L. Xu. Cross-reference stitching quality assessment for 360° omnidirectional images. In *Proc. of ACM International Conference on Multimedia*, p. 2360–2368. New York, NY, USA, 2019. 8

[26] W.-C. Lo, C.-L. Fan, J. Lee, C.-Y. Huang, K.-T. Chen, and C.-H. Hsu. 360 video viewing dataset in head-mounted virtual reality. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pp. 211–216, 2017. 1, 2

[27] R. Ma, Y. Luo, and K. Furuya. Gender differences and optimizing women's experiences: An exploratory study of visual behavior while viewing urban park landscapes in tokyo, japan. *Sustainability*, 15(5), 2023. 7

[28] S. Malpica, A. Serrano, D. Gutierrez, and B. Masia. Auditory stimuli degrade visual performance in virtual reality. *Scientific Reports (Nature Publishing Group)*, 2020. 3

[29] C. Marañes, D. Gutierrez, and A. Serrano. Exploring the impact of 360° movie cuts in users' attention. In *IEEE Conference on Virtual Reality and 3D User Interfaces (IEEE VR)*, 2020. 7, 9

[30] 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, 3

[31] D. Martin, A. Serrano, A. W. Bergman, G. Wetzstein, and B. Masia. ScanGAN360: A Generative Model of Realistic Scanpaths for 360° Images. *IEEE Trans. on Visualization and Computer Graphics*, 28(5):2003–2013, 2022. 3, 6, 8

[32] D. Martin, A. Serrano, and B. Masia. Panoramic convolutions for 360° single-image saliency prediction. In *CVPR Workshop on Computer Vision for Augmented and Virtual Reality*, 2020. 3

[33] B. Masia, J. Camon, D. Gutierrez, and A. Serrano. Influence of directional sound cues on users exploration across 360° movie cuts. *IEEE Computer Graphics and Applications*, 2021. 2, 3

[34] X. Min, G. Zhai, J. Zhou, X.-P. Zhang, X. Yang, and X. Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Trans. on Image Processing*, 29:3805–3819, 2020. 2

[35] A. Miyahira, K. Morita, H. Yamaguchi, K. Nonaka, and H. Maeda. Gender differences of exploratory eye movements: A life span study. *Life Sciences*, 68(5):569–577, 2000. 7

[36] P. Morgado, Y. Li, and N. Vasconcelos. Learning representations from audio-visual spatial alignment. In *Advances in Neural Information Processing Systems*, 2020. 2

[37] P. Morgado, N. Nvasconcelos, T. Langlois, and O. Wang. Self-supervised generation of spatial audio for 360°video. In *Advances in Neural Information Processing Systems*, vol. 31, 2018. 2, 3

[38] T. Noesselt, D. Bergmann, M. Hake, H.-J. Heinze, and R. Fendrich. Sound increases the saliency of visual events. *Brain Research*, 1220:157–163, 2008. 2

[39] A. Nuthmann and J. M. Henderson. Object-based attentional selection in scene viewing. *Journal of Vision*, 10(8):20–20, 2010. 6

[40] C. Ozcinar and A. Smolic. Visual attention in omnidirectional video for virtual reality applications. In *International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6, 2018. 1, 3

[41] Y. Rai, J. Gutiérrez, and P. Le Callet. A dataset of head and eye movements for 360 degree images. In *Proc. of ACM Multimedia Systems Conference*, pp. 205–210, 2017. 2

[42] A. Rana, C. Ozcinar, and A. Smolic. Towards generating ambisonics using audio-visual cue for virtual reality. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2012–2016, 2019. 2

[43] S. Rossi, C. Ozcinar, A. Smolic, and L. Toni. Do users behave similarly in vr? investigation of the user influence on the system design. *ACM Trans. on Multimedia Computing, Communications, and Applications (TOMM)*, 16(2):1–26, 2020. 2

[44] A. Serrano, I. Kim, Z. Chen, S. DiVerdi, D. Gutierrez, A. Hertzmann, and B. Masia. Motion parallax for 360° RGBD video. *IEEE Trans. on Visualization and Computer Graphics*, 25(5):1817–1827, 2019. 4, 8, 9

[45] A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia. Movie editing and cognitive event segmentation in virtual reality video. *ACM Trans. on Graphics*, 36(4), 2017. 1, 2, 6, 7, 9

[46] 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 Trans. on Visualization and Computer Graphics*, 24(4):1633–1642, 2018. 1, 2, 5, 6, 8, 9

[47] V. Skaramagkas, G. Giannakakis, E. Ktistakis, D. Manousos, I. Karatzanis, N. Tachos, E. E. Tripoliti, K. Marias, D. I. Fotiadis, and M. Tsiknakis. Review of eye tracking metrics involved in emotional and cognitive processes. *IEEE Reviews in Biomedical Engineering*, 2021. 3

[48] Y.-C. Su, D. Jayaraman, and K. Grauman. Pano2vid: Automatic cinematography for watching 360 videos. In *Asian Conference on Computer Vision*, pp. 154–171, 2016. 1

[49] H. R. Tavakoli, A. Borji, E. Rahtu, and J. Kannala. Dave: A deep audio-visual embedding for dynamic saliency prediction. *ArXiV (Preprint)*, 2019. 3

[50] Z. Teed and J. Deng. RAFT: recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 3

[51] N.-O. R. A. Thomas Politzer, O.D. Former NORA President. Vision Is Our Dominant Sense. `https://www.brainline.org/article/vision-our-dominant-sense`, 2008. 1

[52] C. Tian, X. Chai, G. Chen, F. Shao, Q. Jiang, X. Meng, L. Xu, and Y.-S. Ho. VSOIQE: A novel viewport-based stitched 360° omnidirectional image quality evaluator. *IEEE Trans. on Circuits and Systems for Video Technology*, 32(10):6557–6572, 2022. 8

[53] R. Warp, M. Zhu, I. Kiprijanovska, J. Wiesler, S. Stafford, and I. Mavridou. Moved by sound: How head-tracked spatial audio affects autonomic emotional state and immersion-driven auditory orienting response in VR environments. *Journal of the Audio Engineering Society*, may 2022. 2

[54] R. Wilcox. In *Introduction to Robust Estimation and Hypothesis Testing (Third Edition)*, Statistical Modeling and Decision Science, pp. 291–377. 2012. 7

[55] Y. Xu, Y. Dong, J. Wu, Z. Sun, Z. Shi, J. Yu, and S. Gao. Gaze prediction in dynamic 360° immersive videos. In *Proc. Computer Vision and Pattern Recognition (CVPR)*, pp. 5333–5342, 2018. 2, 6

[56] R. Zhang, C. Chen, J. Zhang, J. Peng, and A. M. T. Alzbier. 360-degree visual saliency detection based on fast-mapped convolution and adaptive equator-bias perception. *The Visual Computer*, pp. 1–18, 2022. 3

[57] Y. Zhang, F.-Y. Chao, W. Hamidouche, and O. Deforges. PAV-SOD: A new task towards panoramic audiovisual saliency detection. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(3), feb 2023. 2

[58] Y. Zhang, F.-Y. Chao, and L. Zhang. ASOD60K: An audio-induced salient object detection dataset for panoramic videos. *ArXiV (Preprint)*, 2021. 1, 2, 4

[59] Z. Zhang, Y. Xu, J. Yu, and S. Gao. Saliency detection in 360° videos. In *Proc. European Conference on Computer Vision (ECCV)*, September 2018. 1, 2, 6

[60] D. Zhu, X. Shao, Q. Zhou, X. Min, G. Zhai, and X. Yang. A novel lightweight audio-visual saliency model for videos. *ACM Trans. on Multimedia Computing, Communications and Applications*, 2022. 3

[61] Y. Zhu, G. Zhai, and X. Min. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69:15–25, 2018. 3