A Database for Perceived Quality Assessment of User-Generated VR Videos

Yuming Fang, Senior Member, IEEE, Yiru Yao, Xiangjie Sui, and Kede Ma, Member, IEEE

Abstract—Virtual reality (VR) videos (typically in the form of 360° videos) have gained increasing attention due to the fast development of VR technologies and the remarkable popularization of consumer-grade 360° cameras and displays. Thus it is pivotal to understand how people perceive user-generated VR videos, which may suffer from commingled authentic distortions, often localized in space and time. In this paper, we establish one of the largest 360° video databases, containing 302 user-generated videos with rich content and distortion diversities. We capture viewing behaviors (i.e., scanpaths) of 139 users, and collect their opinion scores of perceived quality under four different viewing conditions (two starting points \(\times\) two exploration times). We provide a thorough statistical analysis of recorded data, resulting in several interesting observations, such as the significant impact of viewing conditions on viewing behaviors and perceived quality. Besides, we explore other usage of our data and analysis, including evaluation of computational models for quality assessment and saliency detection of 360° videos. We have made the dataset and code available at https://github.com/Yao-Yiru/VR-Video-Database.

Index Terms—360° videos, virtual reality, scanpaths, video quality assessment, saliency detection.

1 INTRODUCTION

As virtual reality (VR) acquisition and display systems become widely accessible, people are getting used to capture, edit, and interact with VR content, which is evidenced by the accelerated proliferation of 360° videos uploaded to popular video sharing and social media platforms (e.g., Bilibili and Youtube). A practical issue arising from user-generated 360° videos is that they are often born with complex visual artifacts (i.e., the so-called authentic distortions) due to scene complexity, lens imperfection, sensor limitation, non-professional shooting, and stitching inaccuracy. The acquired videos may subsequently undergo several stages of processing, including compression, editing, transmission, and transcoding, leading to additional video impairments [1]. Thus understanding how people perceive 360° video distortions in virtual environments is central to many VR-enabled video applications.

Different from traditional planar videos, an omnidirectional video, by its name, records/generates the scene of interest by capturing/tracing light from all directions at possibly varying viewpoints through time. This gives rise to a 360° \(\times\) 180° spherical field of view (FoV) at any time instance. With the help of a head-mounted display (HMD), users can freely explore the virtual scene using head and gaze movements as if they were in the real world. Such immersive and interactive viewing experience renders existing quality assessment methods for planar videos [11,17] ineffective in predicting the perceived quality of 360° videos. Although several subjective quality studies [2,10] on omnidirectional videos have been conducted, they may have three limitations. First, most of the resulting databases contain synthetic distortions only, with compression artifacts being the most representative. This is an oversimplification of the real-world situation, where user-generated 360° videos may suffer from commingled authentic distortions, often localized spatiotemporally. The extensively studied compression artifacts may no longer dominate the perceptual quality. Second, the databases assume the viewing conditions such as the starting point and the exploration time to be fixed, which is over-constrained.

\(1\)In this paper, we use the terms “360°”, “omnidirectional”, “spherical”, and “panoramic”, interchangeably.
when viewing virtual scenes with HMDs. If such constraints are relaxed, the visible distortions of a 360° video are probably not perceived for some viewing conditions, and thus it is reasonable to rate the perceptual quality as perfect. Third, given a fixed human labeling budget, the number of unique reference 360° videos in the databases is determined by the number of synthetic distortion types and levels, which is limited to a few dozens (if not fewer). As such, these databases fail to sufficiently represent real-world videos with diverse content, distortion, and motion complexities.

In view of the above limitations of existing 360° video databases, many important problems remain under-explored for understanding the perceived quality of 360° videos. For example, how consistent are human behaviors under the same viewing condition? How are human behaviors affected by viewing conditions? How does perceived quality change with viewing conditions? Can we effectively make quality prediction and saliency detection under different viewing conditions?

In an attempt to answer these questions, we establish a large 360° video database, which contains 502 panoramic video sequences to span diverse video content, including cityscape, landscape, shows, sports, and computer-generated (CG) content. The videos exhibit a wide range of complex authentic distortions, covering the full quality spectrum. To assess the perceived quality of a panoramic video as a function of viewing conditions, we invite subjects to watch the video from different starting points and time durations. By doing so, a total of 40,268 perceptual opinion scores together with the scanpaths (as viewing behaviors) from 139 users are recorded. We then provide an in-depth analysis of our data, investigating the impact of viewing conditions on viewing behaviors and perceived quality. In addition, we leverage our data and analysis to evaluate existing computational models for quality prediction and saliency detection. We show how to adapt planar video quality models to 360° videos, which demonstrate competitive performance.

### 2 Related Work

In this section, we first introduce existing 360° video quality assessment (VQA) databases with mean opinion scores (MOSs). We then review objective VQA models that are adapted to or specifically designed for assessing panoramic content. Last, we summarize 360° saliency detection models.

#### 2.1 Subjective Quality Assessment of Panoramic Videos

Singla et al. [2] constructed one of the first databases to study the impact of H.265 compression and spatial resolution on 360° video quality. The database contains six reference videos and 60 distorted videos at two resolutions and five bitrates. Curcio et al. [3] performed a subjective quality experiment of 360° videos under the tile-based streaming setting [18]. The visual stimuli (24 distorted videos at four quality levels and two resolutions) were carefully selected to probe whether the background tile should be encoded with higher resolution or higher fidelity given the same bitrate budget. Tran et al. [4] established a small database containing 60 mobile distorted videos by five levels of H.264 compression and four resolutions. Duan et al. [5] studied how MPEG-4 compression and spatial resolution affect the perceptual quality of 360° videos. Zhang et al. [6] proposed a large omnidirectional video dataset, including 16 reference and 384 distorted videos, covering H.264, H.265, VP9 compression, and simulated packet loss.

In view of the above limitations of existing 360° video databases, many important problems remain under-explored for understanding the perceived quality of 360° videos. For example, how consistent are human behaviors under the same viewing condition? How are human behaviors affected by viewing conditions? How does perceived quality change with viewing conditions? Can we effectively make quality prediction and saliency detection under different viewing conditions?

In an attempt to answer these questions, we establish a large 360° video database, which contains 502 panoramic video sequences to span diverse video content, including cityscape, landscape, shows, sports, and computer-generated (CG) content. The videos exhibit a wide range of complex authentic distortions, covering the full quality spectrum. To assess the perceived quality of a panoramic video as a function of viewing conditions, we invite subjects to watch the video from different starting points and time durations. By doing so, a total of 40,268 perceptual opinion scores together with the scanpaths (as viewing behaviors) from 139 users are recorded. We then provide an in-depth analysis of our data, investigating the impact of viewing conditions on viewing behaviors and perceived quality. In addition, we leverage our data and analysis to evaluate existing computational models for quality prediction and saliency detection. We show how to adapt planar video quality models to 360° videos, which demonstrate competitive performance.

#### 2.2 Objective Quality Assessment of Panoramic Videos

Existing computational models for evaluating panoramic content are mainly adapted from planar image quality assessment (IQA) and VQA methods to one of three data formats: (projected) 2D plane, spherical surface, and (projected) rectilinear viewports.

Methods in the planar domain [20,22] aim to compensate for the non-uniform sampling caused by the sphere-to-plane projection. If the equiangular projection (ERP) is used, current planar IQA/VQA methods can be augmented by latitude-dependent weighting schemes. Craster parabolic projection can also be used to ensure uniform sampling density [21]. Kim et al. [22] explored an adversarial loss for learning patch-based quality estimators using content and position features. Li et al. [9] trained a convolutional neural network (CNN) for panoramic video quality assessment, making use of HM and EM data.

---

**Table 1: Summary of VR VQA databases.** ERP, RCMP, and TSP stand for the equiangular projection, the reshaped cube map projection, and truncated square pyramid projection, respectively. The numbers in the "#videos" column are in the form of "#reference videos / #distorted videos".

| Database       | Year | Projection | #videos | #subjects | Resolution | Duration (sec) | HM/EM data | Distortion type                  |
|----------------|------|------------|---------|-----------|------------|----------------|-------------|----------------------------------|
| Singla et al.  | 2017 | ERP        | 6 / 60  | 30        | 1,920 × 1,080 to 3,840 × 2,160 | 10            | HM          | H.265 compression, Downsampling  |
| Curcio et al.  | 2017 | ERP        | 3 / 24  | 12        | 3,840 × 1,920          | 21            | HM          | Tile-based H.265 compression     |
| Tran et al.    | 2017 | ERP        | 3 / 60  | 37        | 1,440 × 720 to 3,840 × 1,920 | 30            | N/A         | H.264 compression                |
| VQAD2017      | 2017 | ERP        | 10 / 150| 13        | 4,096 × 2,048          | 15            | N/A         | MPEG-4 compression, Downsampling |
| Zhang et al.   | 2017 | ERP        | 16 / 384| 23        | 4,096 × 2,048          | 10            | N/A         | H.264 compression, VP9 compression, Simulated packet loss |
| Zhang et al.   | 2018 | ERP        | 10 / 50 | 30        | 3,600 × 1,800          | 10            | N/A         | H.265 compression, Downsampling  |
| Lopes et al.   | 2018 | ERP        | 6 / 79  | 37        | 960 × 480 to 7,680 × 3,840 | 10            | N/A         | H.265 compression, Downsampling  |
| VQA-ODV       | 2018 | ERP, RCMP, TSP | 60 / 540| 221       | 3,840 × 1,920 to 7,680 × 3,840 | 10 to 23    | HM + EM     | H.265 compression, Projection    |
| VOD-VQA       | 2018 | ERP        | 18 / 774| 160       | 3,840 × 1,920          | 10            | N/A         | H.264 compression, Downsampling  |
| Proposed      | 2021 | ERP        | - / 502 | 139       | 1,280 × 720 to 5,120 × 2,560 | 15            | HM + EM     | Authentic distortion            |
Methods in the spherical domain (e.g., S-PSNR [23] and S-SSIM [24]) calculate and aggregate local quality estimates over the sphere. Yu et al. [23] incorporated importance weighting derived from the empirical distributions of HM and EM data. Methods in the viewport domain focus on extracting viewports that are likely to be seen for quality computation. Xu et al. [25] used graph convolutional networks to model spatial relations of extracted viewports, which, however, does not necessarily reflect the human viewing process. Li et al. [26] proposed a two-stage approach - viewport proposal and quality assessment using spherical convolution [27]. Recently, Sui et al. [19] suggested to convert a panoramic image to planar videos by sampling, along users’ scanpaths, sequences of rectilinear projections of viewports. By doing so, mature planar IQA/VQA methods can be directly applied. In this work, we will continue along this path to take into account viewing conditions, with the goal of boosting existing quality predictors for panoramic videos.

2.3 Saliency Detection of Panoramic Videos

360° saliency detection aims at predicting objects/regions of interest in VR environments. It identifies visually important information in panoramic videos, and plays a key role in panoramic video streaming and rendering. Saliency detection may also be important to visual quality assessment [9, 23, 26, 29].

Current saliency detection methods for 360° videos can also be divided according to their operating domain: (projected) 2D plane, spherical surface, and (projected) rectilinear viewport. At present, most panoramic saliency detection algorithms build up traditional planar methods, while taking into account some specific properties of the 2D projection [28–30]. Nguyen et al. [28] proposed PanSalNet by fine-tuning a planar image saliency detector on two 360° video databases [31, 32]. Cheng et al. [29] trained a CNN-based saliency detection model for 360° video with a cube padding trick to alleviate projection distortions and boundary discontinuities. Xu et al. [30] predicted HM positions for 360° videos based on deep reinforcement learning. In the spherical domain, Bogdanova et al. [33] extended the classic saliency model [34] by extracting and combining intensity, chroma, and orientation features from the spherical Gaussian pyramid. Zhang et al. [35] proposed the spherical U-Net for saliency detection, where the sliding of the convolution kernel “translates” to kernel rotation on the sphere. In the viewport domain, Lebreton et al. [36] extracted features from viewports based on orientation analysis, which were back-projected to the ERP domain. Concurrently, they [37] extended BMS360 [36] to V-BMS360, enabled by the optical flow-based motion detection. A recent subjective user study on stereoscopic 360° images [38] suggests weak effect of viewing conditions on the gathered saliency maps. However, the results may no longer be valid when watching user-generated 360° videos, which are thoroughly analyzed in the paper.

3 PSYCHOPHYSICAL EXPERIMENT

In this section, we summarize our effort towards recording a database that contains human perceptual data - MOSs and scanpaths for users watching 360° videos under different viewing conditions.
VR shooting: These moving-camera videos typically receive a considerable number includes beautiful natural scenes, such as waterfalls, moun-

3.1 Data Gathering
Visual Stimuli Our database contains 502 unique user-generated 360° video sequences, with frame rate ranging from 20 to 60 frames per second (fps) and resolution ranging from 1,280 × 720 to 5,120 × 2,560 pixels. All videos are downloaded from the Internet, carrying Creative Commons licenses. Each video is cropped to a duration of about 15 seconds and stored in the ERP format without further compression. In our dataset, we mainly include 360° videos shot by a (nearly) static camera. This is because videos captured by a moving camera have a higher probability to cause dizziness [39], reducing the reliability of the collected MOSs. We only select 32 videos with strong camera motion, and guarantee the visual comfort by a posteriori questionnaire [40]. These moving-camera videos typically receive a considerable number of likes on video sharing platforms, and bring users a stronger sense of immersion and interaction.

The videos are selected to span a diversity of scenes that are ideal for VR shooting: Cityscape, Landscape, Shows, Sports, CG, and Others (see Fig. 3). Cityscape contains different places of interest around the world like the Roman Colosseum and other famous historical sites. Landscape includes beautiful natural scenes, such as waterfalls, mountains, volcanoes, etc. Shows represent different forms of entertainment, including band performance, living theatre, and street improvisation. Sports gather various sporting events, e.g., car racing, skiing, and riding. CG is a collection of rendered videos by mature CG techniques. Finally, the Others category is reserved for scenes that do not belong to the previous five classes. As suggested by Winkler [41], we quantify the content diversity of the proposed database using two low-level statistics: spatial information (SI) and temporal information (TI), with higher values indicating more complexities. Fig. 4 shows the 2D scatter plot together with 1D histograms, from which we see that the selected stimuli provide fairly wide coverage in the SI-TI space.

Different from existing 360° video databases (see Table 1), we primarily focus on authentic distortions, manifesting themselves as complex mixtures of multiple visual artifacts that arise during 360° video creation [42]. Fig. 5 shows the entire 360° video processing pipeline, and we see that the creation of 360° videos consists of two steps: optical acquisition with a multi-camera rig and stitching of multiple planar videos with limited and overlapping FoVs. Visual distortions from the optical acquisition are often the consequences of the combination of scene complexity, lens imperfection, sensor limitation, and non-professional shooting, which include under/over-exposure, out-of-focus and motion blurring, sensor noise, annoying shaky motion, flickering [43], jerking [44] and floating [45]. Stitching distortions are mainly due to the limitation of the stitching algorithm itself and the negative influence of visual distortions from the previous acquisition step (e.g., stitching images of different luminance levels tends to create artificial boundaries, as shown in Fig. 5(d)). Visually, stitching distortions are abrupt lumiance/structure change, object with missing parts, ghosting, and motion discontinuity localized in space and time. Of particular interest is the artificial converging points visible at two poles (see Fig. 5(h)). These authentic distortions inevitably affect the whole video processing pipeline, and ultimately be perceived by end users.

Viewing Conditions We use an HTC Eye Pro to display 360° videos, which provides an FoV of 110° and a binocular resolution of 2,880 × 1,600 pixels. Subjects are asked to seat on a swivel chair using the HMD to watch videos. EM and HM data can be collected by a built-in Tobii Pro eye-tracking system with a sampling rate of 2±3 fps. Video playback is supported by a high-performance server with an AMD Ryzen 9 3950X 16-Core CPU, 128 GB RAM, and NVIDIA GeForce RTX 2080 Ti GPU. The graphical user interface is customized using the Unity Game Engine.

An important consideration in our psychophysical experiment is that we vary two viewing conditions: the starting point and the exploration time. We intentionally choose Starting Point I to give users a poor initial viewing experience. This includes viewport that contain localized distortions or intensive spatiotemporal information. Another example is the initial viewport from the side when there is strong camera motion guidance. On the contrary, Starting Point II is selected to encourage a good initial viewing experience, and is at least 120° (in longitude) apart from Starting Point I. An example is shown in Fig. 6, where the viewport extracted from Starting Point I contains visible over-exposure and color cast distortions, while the viewport extracted from Starting Point II is of high quality. We also set two exploration times: one spanning the entire duration (i.e., about 15 seconds) and the other set to the half of the former (i.e., 7 seconds). More exploration time allows more viewpoints to be extracted and viewed. As shown in Fig. 7, the “F BridgeOpening2” video records a cruise ship sailing out of a dark cave. Viewing from Starting Point I in the first seven seconds, the user may be exposed to distortions like under-exposure and over-exposure, which injures her/his viewing experience. In the subsequent eight seconds, the cruise ship has sailed out of the cave, and the viewer may see high-quality content without artifacts, which improves the viewing experience. The starting points and exploration times are combined in pairs, giving rise to a total of four viewing conditions.

Subjective Methodology The single stimulus continuous quality evaluation method described in the ITU-R BT 500.13 recommendation [44] is employed for our psychophysical study. Subjects need to rate the perceived quality of a 360° video on a continuous scale of [1, 5], labeled by five quality levels (“bad”, “poor”, “fair”, “good”, and “excellent”). To reliably collect MOSs, the experimental procedure consists of three phases: pre-training, training, and testing, as shown in Fig. 8.

In the pre-training phase, basic non-sensitive user information such as age, gender, and whether to wear glasses are recorded. Subjects are familiarized with the experiment procedure and the rating guideline. We find it relatively time-consuming to teach subjects to use the hand controller for rating. Therefore, one of the authors is responsible for recording the opinion scores, read out by the subjects.

In the training phase, we select six video sequences that are not included in the proposed dataset. For the first four videos, we let subjects freely view the virtual scenes, and point out the distortions they encounter during exploration. We find that this level of instruction is necessary to familiarize subjects with distortions that are likely to occur in the testing phase. We try not to over-instruct the subjects, for example avoiding providing a reference MOS for each of the four videos. For the remaining two videos, we ask the subjects to give quality scores with no instructions. A discussion on how the subjects arrive at such ratings is held to make sure they understand the evaluation process. No feedback is provided on their scores. More importantly, if the subjects feel any consistently with the background.

Fig. 4: 360° video processing pipeline, from optical acquisition to content consumption via an HMD. The optical acquisition and stitching are two main steps for 360° video creation, where authentic distortions arise.
discomfort during this phase, the experiment is interrupted immediately. These are not invited to conduct the subsequent experiments.

In the testing phase, we divide the 502 videos into eight sessions to reduce fatigue and discomfort caused by possible long-time viewing. To further minimize such effects, the subjects can take a break at any time during this phase. Each session contains about 60 videos with a 5-second mid-gray screen in between. We gather human data from 139 subjects (75 females and 64 males with ages between 17 and 26). All participants report normal or corrected-to-normal color vision. The subjects are divided into two groups, according to two different sets of starting points. Each subject takes part in at least two sessions. Each video is viewed only once by one subject to ensure that user data is collected without prior knowledge of the scene.

3.2 Data Processing

Opinion Scores After obtaining the raw human scores, we detect and remove outliers using the method in [46]. Specifically, we first determine whether the subjective scores given to a 360° video are normally distributed by calculating the kurtosis coefficient:

$$\kappa = \frac{\mu_4}{(\sigma)^4},$$

where $\mu_4$ is the fourth central moment and $\sigma$ is the standard deviation. If they are normally distributed (i.e., $\kappa \in [2, 4]$), an outlier is detected if the score is out of range $[\mu - 2\sigma, \mu + 2\sigma]$. If they are not well fitted by Gaussian, we extend the valid range to $[\mu - \sqrt{20}\sigma, \mu + \sqrt{20}\sigma]$. We compute the MOS by

$$q_j = \frac{1}{M} \sum_{i=1}^{M} q_{ij},$$

where $q_{ij}$ is the opinion score of the $i$-th observer for the $j$-th video sequence. In our study, each of the 502 videos in each viewing conditions receives at least 20 ratings. Fig. 9 shows the MOSs with the corresponding 95% confidence intervals.
4 UNDERSTANDING VIEWING BEHAVIORS IN VR

With the recorded data, we gather insights and investigate a number of questions about the behaviors of users when watching user-generated VR videos. Here we focus on analyzing one particular type of viewing behaviors - the scanpath - because it is the most relevant to the perceived quality of a 360° video by the corresponding user.

4.1 Viewing Behavior Metrics

To compare multiple scanpaths, we adopt two wide-used metrics: temporal correlation [48] and similarity ring metric (SRM) [49]. We also consider comparing the saliency maps (i.e., the heatmaps) as spatial aggregations of scanpaths for further analysis.

**Temporal correlation** uses PLCC to calculate the correlations between the longitude values and the latitude values of two scanpaths $s^{(i)} = [\phi^{(i)}, \theta^{(i)}]$ and $s^{(j)} = [\phi^{(j)}, \theta^{(j)}]$, followed by simple averaging:

$$TC(s^{(i)}, s^{(j)}) = \frac{1}{2} \left( PLCC\left(\phi^{(i)}, \phi^{(j)}\right) + PLCC\left(\theta^{(i)}, \theta^{(j)}\right) \right),$$

where $\phi^{(i)}$ and $\theta^{(i)}$ represent the longitudes and latitudes of the $i$-th scanpath, respectively. The mean temporal correlation over $M$ subjects exploring the video is calculated by

$$mTC = \frac{2 \sum_{s=1}^{M-1} \sum_{t=1}^{M} TC(s^{(i)}, s^{(j)})}{M(M-1)}.$$  

**Similarity ring metric (SRM)** measures whether different subjects have been watching the same video parts at the same time. It is less likely that all scanpaths completely overlap, but it is reasonable to determine if they fall within a certain range, i.e., passing through the same ring. As suggested by [49], we focus on the longitude of the scanpath, and set the radius and the center of the ring to be $r = FoV/2$ and the mode of longitude values from $M$ scanpaths at the same time instance:

$$c_t = \text{mode}\left(\phi^{(1)}_t, \phi^{(2)}_t, \ldots, \phi^{(M)}_t\right).$$

A longitude value out of the ring means that the corresponding subject does not watch the same content with respect to other subjects at the corresponding time instance. The instantaneous similarity at the $t$-th time instance for the $i$-th scanpath is then defined as

$$IS^{(i)}_t = \begin{cases} 1, & \text{if } \phi^{(i)}_t \in [c_t - FoV/2, c_t + FoV/2]; \\ 0, & \text{otherwise}, \end{cases}$$

based on which we compute the SRM by averaging across all scanpaths and over all time instances:

$$SRM = \frac{100}{MT} \sum_{t=1}^{T} \sum_{i=1}^{M} IS^{(i)}_t.$$  

SRM is scaled to lie within $[0, 100]$, with a larger value indicating higher consistency.

**Heatmap** reflects the salient areas that users pay attention to, and can be considered as a spatial aggregation of scanpaths. To generate dynamic heatmaps for omnidirectional videos, we apply the density-based spatial clustering (DBSCAN) algorithm [50] to scanpaths of all subjects, and define fixations as the cluster centroids that span at least 200 ms [51], during which the gaze direction remains roughly unchanged. Noisy fixation points will be automatically filtered out. For each second of the video clip, we compute a fixation map by DBSCAN. The saliency of each location in the fixation map is determined by the total spherical (i.e., great-circle) distances from the location to all fixation points [30], normalized by the computed maximum distance. To compare the similarity of two heatmaps, we follow [38] and use PLCC as the quantitative measure.

4.2 Does the Viewing Condition Affect Viewing Behaviors?

To assess whether viewing behaviors are affected by the viewing conditions, we calculate $mTC$ in Eq. (4) and $SRM$ in Eq. (7) under different viewing conditions, as listed in Table 2. We also employ analysis of variance (ANOVA) to see whether such differences in viewing behavior consistency as measured by $mTC$ and $SRM$ are statistically significant, as listed in Table 3. From the tables, we find that human viewing
Table 2: Viewing behavior consistency in terms of mTC and SRM (and the associated standard error) under different viewing conditions.

| Factor  | mTC | SRM  |
|---------|-----|------|
|         | 7-second | 15-second | 7-second | 15-second |
| Starting Point I | 0.394 (±0.046) | 0.289 (±0.038) | 72.997 (±7.221) | 63.721 (±5.371) |
| Starting Point II | 0.395 (±0.041) | 0.286 (±0.034) | 74.702 (±7.870) | 65.128 (±5.603) |

Table 3: p-values in the ANOVA test for mTC and SRM. A p-value below the threshold of 0.05 represents that the corresponding factor has a significant impact on mTC and SRM, i.e., viewing behavior consistency.

| Factor             | mTC | SRM |
|--------------------|-----|-----|
| Starting point      | 0.615 (±0) | 0.415 (±0) |
| Exploration time    | 0 (±0) | 0 (±0) |

Table 4: Perceived quality analysis in terms of MOS (and the associated standard error) under different viewing conditions.

| Video attribute             | Average MOS |
|-----------------------------|-------------|
| Low-resolution              | 2.176 (±0.024) |
| High-resolution             | 3.081 (±0.021) |
| No camera motion            | 2.856 (±0.021) |
| Camera motion               | 2.779 (±0.037) |

5. UNDERSTANDING PERCEIVED QUALITY IN VR

Understanding how people perceive visual distortions in VR is challenging due to the differences in viewing conditions between planar 2D and immersive 360° videos. In this section, we analyze the effects of VR viewing conditions as well as video attributes on the perceived quality of omnidirectional videos.

5.1 Does the Viewing Condition Affect Perceived Quality?

Previous studies [2–10] assume that viewing conditions have a negligible impact on the perceived quality of 360° videos, which is reasonable if synthetic artifacts (e.g., video compression) with global distortion appearances are considered. This is because for any head orientation at any time instance, the extracted viewport has a high probability of containing the same main artifacts. However, this is not the case when it comes to user-generated VR videos, where we are dealing with authentic distortions, localized in space and time. Whether and when to encounter such spatiotemporal local distortions may have a different influence on the perceived quality. To test the hypothesis, we average the MOSs in the proposed dataset for different viewing conditions in Table 4. Several useful findings are worth mentioning. First, compared to 15-second exploration, seven seconds mean fewer viewports of the scene to be observed, highlighting the importance of the starting point to the perceived video quality. Second, if a longer exploration time is allowed, the viewports close to the end are more likely to affect the overall viewing experience due to the recency effect [12]. This explains that for the 7-second exploration, the subjects tend to give low-quality scores when viewing from Starting Point I, where distortions appear in the initial viewports (see Sec. 3.1 for the definitions of the two types of starting points). If they are allowed more time, the subjects would consciously re-orient their heads to avoid viewing distorted viewports, and look for those with better quality, leading to an improved viewing experience. On the contrary, from Starting Point II where the distortions may not be noticed initially, the subjects are less likely to observe distortions with less allowable time, explaining the higher average MOS of the 7-second exploration.

5.2 Does the Video Attribute Affect Perceived Quality?

We consider two video attributes: camera motion and spatial resolution. The camera motion is often generated by body-mounted cameras. We select 32 videos with complex camera motion (e.g., camera mounted on the roller coaster or held by a surfer). From Table 5, we find, as expected, that videos with camera motion generally obtain lower MOSs irrespective of viewing conditions. This validates that using camera motion as strong visual guidance may be accompanied by annoying shaky motion, impairing the user viewing experience. We also show in Table 5 that high-resolution videos receive a higher average MOS than low-resolution ones (for all viewing conditions) despite being downsampled in the HMD. It remains to be seen that such downsampling has a positive impact on the perceived quality by “concealing” certain types of distortions (e.g., compression artifacts and high-frequency noise).

5.3 Significant Impact Analysis on Perceived Quality

To test whether the four factors: the starting point, the exploration time, the spatial resolution, and the camera motion have a statistically significant effect on perceived quality, we apply the multi-factorial ANOVA to the MOS values between factors. The results are listed in Table 6, from which we confirm that spatial resolution is a significant individual factor. The effect of the camera motion alone is not statistically significant due in part to the limited inclusion of such videos to avoid visual discomfort. Two viewing conditions (i.e., the starting point and the exploration time) together play a compound decisive role in the perceived quality of user-generated VR videos. It turns out that the spatial resolution and camera motion also have an interplay effect.

6. EVALUATING VQA MODELS FOR VR VIDEOS

In this section, we first enable existing planar IQA/VQA methods to assess the perceived quality of 360° videos, taking advantage of human
We then evaluate several representative quality models, including one that is specifically designed for omnidirectional content, on the proposed dataset. Note that full-reference IQA/VQA models such as the mean squared error (MSE) and the structural similarity (SSIM) index \(SSIM\) are not applicable here because no original 360° videos of pristine quality are available as reference.

### 6.1 Model Selection

The implementations of existing IQA/VQA methods are fairly unambiguous for planar images and videos. But there are many ways one can adapt these methods to 360° videos. The primary issues are: Resolution. To span the entire 360°×180° FoV with high fidelity, the spatial resolution of an omnidirectional video is extraordinarily high. To speed up computational prediction, it is necessary to consider spatial downsampling as one of the pre-processing steps. But, how should we determine the downsampling factor?

**Frame rate.** Similar to spatial resolution, the frame rate of an omnidirectional video is suggested to be as high as possible to enhance the sense of presence and to reduce motion sickness. Should we also perform temporal downsampling for computational reasons?

**Projection.** Omnidirectional videos are typically stored in ERP format, which introduces severe geometric distortions at high latitudes. Thus the direct application of existing planar IQA/VQA models may achieve suboptimal quality prediction performance. Among many map projection methods in cartography, which one should be adopted in tandem with current quality models?

**Temporal Pooling.** When applying IQA models for VQA, temporal pooling is a necessary step to aggregate frame-level quality estimates into an overall quality score. Among popular temporal pooling strategies [59], which one should be implemented?

We use bicubic interpolation to downsample the 360° videos, and find that the quality prediction performance degrades gracefully with the increasing downsampling factor. To strike a balance between speed and accuracy on the proposed dataset, we choose a downsampling factor of two. Due to the fact that our database does not include high frame rate videos (e.g., > 60 fps), we choose not to downsample in the temporal dimension. Moreover, to reproduce the user viewing process, we transform an omnidirectional video into a set of planar videos by sampling, along the users’ scanpaths, sequences of rectilinear projections of viewpoints [19]. The pixel value in the viewport can be easily retrieved by first projecting its positions onto the unit sphere and then onto the ERP plane. In [59], temporal pooling has demonstrated effectiveness in boosting IQA models for VQA. As pursuing the best temporal pooling strategy is not our focus, we decide to use simple average pooling to give prominence to the adopted IQA models.

In this paper, we select eight representative blind IQA/VQA models.

- **BRISQUE** [54], the Blind/Referenceless Image Spatial Quality Evaluator, is a blind IQA model that extracts nature scene statistics (NSS) in the spatial domain to quantify the “unnaturalness” of the test image.
- **NIQE** [55], the Natural Image Quality Evaluator, is a completely blind IQA model without training on MOSs. It measures the deviation of the test image from statistical regularities observed in natural undistorted images.
- **UNIQUE** [56], the Unified No-reference Image Quality and Uncertainty Evaluator, is a DNN-based model trained on multiple image quality datasets simultaneously. It can assess both synthetic and realistic distortions.
- **VSFA** [14], the Video Semantic Feature Aggregation, is a blind VQA model exploiting the content-dependency and temporal memory effects. The content-aware features are extracted by a pre-trained DNN model for object recognition, and the temporal memory is modeled by gated recurrent units [60].
- **TLVQM** [15], the Two-Level Video Quality Model, is a blind VQA model based on two sets of features of different complexities. Support vector regression (SVR) is implemented for final quality prediction.
- **PVQ** [16], the Patch Video Quality, is a blind VQA model that uses PaQ2PiQ [61] and ResNet3D [62] to compute 2D and 3D video features, respectively. The pooled features are fed into an InceptionTime [63] network for quality estimation.
- **MLSP-VQA** [17], the Multi-Level Spatially Pooled deep features for VQA, is a blind model based on the multi-level features from the pre-trained Inception network.
- **MC360IQA** [57], the Multi-Channel CNN for blind 360° IQA, uses six viewports that cover the panoramic scene as input. Six parallel hyper-ResNet34 networks [64] are used to extract features, which are fed into an image quality regressor.

The implementations of all competing models are obtained from the original authors. We apply planar IQA/VQA models directly in ERP format as baselines. To distinguish the two types of methods, we add a “V-” in front to name the viewport-based methods that make use of viewing behaviors. Given the scanpath of the \(i\)-th user, the viewport-based methods are able to sample a planar video, and produce a quality estimate, denoted by \(\hat{q}^{(i)}\). The final score is computed by averaging the estimated scores across subjects:

\[
\hat{q} = \frac{\sum_{i=1}^{M} \hat{q}^{(i)}}{M}.
\]

### 6.2 Performance Comparison

In addition to PLCC, we use a second evaluation metric to quantify the quality prediction performance that has been widely adopted in the IQA/VQA literature: Spearman rank-order correlation coefficient (SRCC). A higher SRCC value indicates better prediction monotonicity.

To compensate for the nonlinearity between model predictions and MOSs, we fit a four-parameter logistic function before computing SRCC. A higher SRCC value indicates better prediction monotonicity.

\[
q = (\beta_1 - \beta_2) \frac{1}{1 + e^{-\frac{q - \beta_0}{\beta_3}}} + \beta_2,
\]

where \(\{\beta_i\}_{i=1}^4\) are the parameters to be fitted.
Table 7: Performance comparison of different IQA/VQA methods on the proposed dataset under different viewing conditions. The best results are highlighted in bold.

| Model         | Starting Point I | Starting Point II | Overall  |
|---------------|------------------|-------------------|----------|
|               | 7s    | 15s   | 7s      | 15s    | 7s  | 15s |
| BRSQUE        | 0.176  | 0.240  | 0.279  | 0.289  | 0.236 |
| NIQE          | 0.368  | 0.461  | 0.471  | 0.483  | 0.424 |
| UNIQUE        | 0.260  | 0.270  | 0.213  | 0.230  | 0.229 |
| VSFA          | 0.259  | 0.251  | 0.316  | 0.305  | 0.266 |
| TVQM          | 0.583  | 0.668  | 0.690  | 0.709  | 0.778 |
| PVQ           | 0.453  | 0.562  | 0.591  | 0.572  | 0.513 |
| MLSP-VQA      | 0.377  | 0.388  | 0.363  | 0.414  | 0.396 |
| MC360IQA      | 0.510  | 0.602  | 0.622  | 0.601  | 0.569 |
| V-BRISQUE     | 0.260  | 0.358  | 0.342  | 0.388  | 0.317 |
| V-NIQE        | 0.520  | 0.637  | 0.613  | 0.628  | 0.567 |
| V-UNIQUE      | 0.504  | 0.623  | 0.591  | 0.579  | 0.551 |
| V-VSFA        | 0.443  | 0.555  | 0.407  | 0.540  | 0.457 |
| V-TLVQM       | 0.842  | 0.899  | 0.895  | 0.903  | 0.824 |
| V-MLSP-VQA    | 0.911  | 0.933  | 0.935  | 0.945  | 0.891 |

Table 8: Performance comparison of different panoramic saliency detection models under different viewing conditions.

| Model         | Starting Point I | Starting Point II | Overall  |
|---------------|------------------|-------------------|----------|
|               | 7s    | 15s   | 7s      | 15s    | 7s  | 15s |
| Chance        | 0.523  | 0.535  | 0.529  | 0.502  | 0.500  | 0.501 |
| Equator       | 0.784  | 0.769  | 0.777  | 0.525  | 0.501  | 0.513 |
| Constant      | 0.815  | 0.811  | 0.813  | 0.507  | 0.494  | 0.501 |
| PanoSalNet    | 0.665  | 0.680  | 0.673  | 0.523  | 0.541  | 0.532 |
| CP360        | 0.685  | 0.694  | 0.690  | 0.587  | 0.595  | 0.591 |
| DHB          | 0.552  | 0.559  | 0.556  | 0.607  | 0.593  | 0.600 |
| ATSal        | 0.711  | 0.711  | 0.711  | 0.536  | 0.540  | 0.538 |
| Human         | 0.800  | 0.795  | 0.798  | 0.754  | 0.800  | 0.777 |

7 EVALUATING SALIENCY DETECTION MODELS FOR VR VIDEOS

In this section, we explore additional use of the proposed dataset for evaluating panoramic saliency detection models.

7.1 Model Selection

We choose three baselines and four state-of-the-art saliency detectors for omnidirectional videos.

- **Chance model** assigns a uniformly distributed value from \([0, 1]\) to each pixel in the heatmap (see Fig. 12(a)).
- **Equator model** consists of a symmetric Gaussian around the equator with a variance to cover 20% of the equator in the \(\theta\)-direction and a degenerate Gaussian with infinite variance in the \(\phi\)-direction (see Fig. 12(b)).
- **Constant model** is the average heatmap across the whole dataset (see Fig. 12(c)).
- **PanoSalNet** employs transfer learning to adapt an existing saliency model [65] for ERP-based saliency detection. A prior filter that encodes inductive viewing biases (e.g., center and equator biases) is used to refine the prediction.
- **CP360** is a cubemap-based weakly-supervised model with a cube padding trick to reduce projection distortions and image border discontinuities.
- **DHB** applies \(M\) workflows in parallel to predict the HM positions of \(M\) subjects for a 360° video. We set \(M = 20\), which is the number of subjects in our database. For each video frame, a heatmap can be obtained by Gaussian blurring the estimated HM positions.
We have put together the first user-generated VR video dataset that explicitly models the viewing conditions performs the best among all models in terms of s-AUC, NSS, and PLCC (see also Fig. 14). Moreover, equator and constant models confirm the effectiveness of the equator bias, and are even top-2 performers under AUC-Judd. Nevertheless, there is significant room for improvements as evidenced by a large performance gap between computational models and humans.

7.2 Performance Comparison

We use four metrics to evaluate the saliency detection performance, including the Judd variant of the area under curve (AUC-Judd) [66], shuffle-AUC (s-AUC) [67], normalized scanpath saliency (NSS) [68], and PLCC [69]. It is noteworthy that in order to avoid the over-sampling problem in ERP format when calculating the metrics, we uniformly sample 1,000 points on the sphere, whose saliency values can be retrieved from the corresponding heatmap of size 180 × 360, as suggested in [70].

We show the overall performance for different starting points in Table 8. To make the results more comparable and interpretable, for each metric, we compute the human consistency as a realistic upper bound for model performance [71]. Specifically, we first compare the fixations of two groups of M observers, where M varies from 1 to 10 (i.e., half of the total 20 observers). We then fit the 10 performance scores to a power function (i.e., $aM^b + c$), and predict the human performance as that of two groups of infinite observers (which is equal to c, for $b < 0$). We also take a closer look at the performance changes over time for different starting points in Fig. 13. We find that DHP [30] that explicitly models the viewing conditions performs the best among all models in terms of s-AUC, NSS, and PLCC (see also Fig. 14). Moreover, equator and constant models confirm the effectiveness of the equator bias, and are even top-2 performers under AUC-Judd. Nevertheless, there is significant room for improvements as evidenced by a large performance gap between computational models and humans.

8 Conclusion and Discussion

We have put together the first user-generated VR video dataset that includes MOSs and viewing behavioral data. This dataset encompasses 139 users viewing 360° videos from four different conditions, resulting in a total of 40,268 opinion scores and scanpaths. We conducted a statistical analysis of various effects on viewing behaviors and perceived quality in VR, identifying viewing conditions to be crucial. We lastly minimize the adverse physiological reactions by means of questionnaire. It would be interesting to design clever psychophysical experiments to disentangle visual discomfort and visual quality in a quantitative way. Second, when collecting viewing behaviors, we expose users to a VR scene only once to eliminate the prior knowledge about the scene configuration. How do viewing behaviors change with multiple exposures to the same scene and how these changes affect the perceived quality are worth exploring. Third, currently, objective quality models tailored to user-generated 360° videos are largely lacking. Our study suggests that a key step in the model development is the incorporation of viewing conditions that faithfully reflect how humans explore VR videos. Fourth, when evaluating existing objective IQA/VQA models for 360° videos, we directly make use of human scanpaths, which have to be estimated in practice. However, scanpath prediction for 360° videos is still in its infancy, especially for long-term prediction (e.g., ≥10 seconds). It is thus desirable to develop better scanpath prediction models that deliver accurate short-term and long-term results, and meanwhile capture the diversity of human scanpaths. Lastly but not least, the current work considers 360° videos as the sole visual stimuli, it is of interest to investigate audio-visual perception of user-generated 360° videos [51] and its consequences on perceived quality.

REFERENCES

[1] R. G. D. A. Azevedo, N. Birkbeck, F. De Simone, I. Janatra, B. Adsumilli, and P. Frossard, “Visual distortions in 360° videos,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 8, pp. 2524–2537, 2020.

[2] A. Singla, S. Fremeery, W. Robitza, P. Lebreton, and A. Raake, “Comparison of subjective quality evaluation for HEVC encoded omnidirectional videos at different bit-rates for UHD and FHD resolution,” in *ACM Multimedia Workshops*, 2017, pp. 51–519.

[3] I. D. Curcio, H. Toukomaa, and D. Naik, “Bandwidth reduction of omnidirectional viewport-dependent video streaming via subjective quality assessment,” in *International Workshop on Multimedia Alternate Realities*, 2017, pp. 9–14.

[4] H. T. Tran, N. P. Ngoc, C. M. Bui, M. H. Pham, and T. C. Thang, “An evaluation of quality metrics for 360° videos,” in *Ninth International Conference on Ubiquitous and Future Networks*, 2017, pp. 7–11.

[5] H. Duan, G. Zhai, X. Yang, D. Li, and W. Zhu, “IVQAD 2017: An immersive video quality assessment database,” in *International Conference on Systems, Signals and Image Processing*, 2017, pp. 1–5.

[6] B. Zhang, J. Zhao, S. Yang, Y. Zhang, J. Wang, and Z. Fei, “Subjective and objective quality assessment of panoramic videos in virtual reality environments,” in *IEEE International Conference on Multimedia Expo Workshops*, 2017, pp. 163–168.

[7] Y. Zhang, Y. Wang, F. Liu, Z. Liu, Y. Li, D. Yang, and Z. Chen, “Subjective panoramic video quality assessment database for coding applications,” *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 461–473, 2018.

• ATSal [55] is a two-stream CNN model, where the ERP-based stream is dedicated to extracting global attention statistics, while the cubemap-based stream aims at learning local saliency features.

The implementation of the four panoramic saliency detectors are again obtained from the original authors, and tested with the default settings.

Fig. 13: Saliency detection performance changes over time for different starting points.
Fig. 14: Heatmaps produced by different saliency detection models. From left to right show the results of the 1-st, 4-th, 7-th, 10-th, and 15-th second, respectively.

[8] F. Lopes, J. Ascenso, A. Rodrigues, and M. P. Queluz, “Subjective and objective quality assessment of omnidirectional video,” in Applications of Digital Image Processing, 2018, pp. 249–265.

[9] C. Li, M. Xu, X. Du, and Z. Wang, “Bridge the gap between VQA and human behavior on omnidirectional video: A large-scale dataset and a deep learning model,” in ACM Multimedia, 2018, pp. 932–940.

[10] Y. Meng and Z. Ma, “Viewport-based omnidirectional video quality assessment: Database, modeling and inference,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 32, no. 1, pp. 120–134, 2022.

[11] Z. Wang, L. Lu, and A. C. Bovik, “Video quality assessment based on structural distortion measurement,” Signal Processing: Image Communication, vol. 19, no. 2, pp. 121–132, 2004.

[12] Y. Wang, T. Jiang, S. Ma, and W. Gao, “Novel spatio-temporal structural information based video quality metric,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 22, no. 7, pp. 989–998, 2012.

[13] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, “C3DVQA: Full-reference video quality assessment with 3D convolutional neural network,” in IEEE International Conference on Acoustics, Speech and Signal Processing, 2020, pp. 4447–4451.

[14] D. Li, T. Jiang, and M. Jiang, “Quality assessment of in-the-wild videos,” in ACM Multimedia, 2019, pp. 2351–2359.

[15] J. Korhonen, “Two-level approach for no-reference consumer video quality assessment,” IEEE Transactions on Image Processing, vol. 28, no. 12, pp. 5923–5938, 2019.

[16] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, “Patch-VQ: ‘Patching Up’ the video quality problem,” in IEEE Conference on Computer Vision and Pattern Recognition, 2021, pp. 14019–14029.

[17] F. Götz-Hahn, V. Hosu, H. Lin, and D. Saupe, “Konvid-150k: A dataset for no-reference video quality assessment of videos in-the-wild,” IEEE Access, vol. 9, pp. 72139–72160, 2021.

[18] M. Graf, C. Timmerer, and C. Mueller, “Towards bandwidth efficient adaptive streaming of omnidirectional video over HTTP: Design, implementation, and evaluation,” in ACM on Multimedia Systems Conference, 2017, pp. 261–271.

[19] X. Sui, K. Ma, Y. Yao, and Y. Fang, “Perceptual quality assessment of omnidirectional images as moving camera videos,” IEEE Transactions on Visualization and Computer Graphics, pp. 1–11, 2021.

[20] Y. Sun, A. Lu, and L. Yu, “Weighted-to-spherically-uniform quality evaluation for omnidirectional video,” IEEE Signal Processing Letters, vol. 24, no. 9, pp. 1408–1412, 2017.

[21] V. Zakharchenko, K. F. Choi, and J. H. Park, “Quality metric for spherical panoramic video,” in Optics and Photonics for Information Processing, 2016, pp. 57–65.

[22] H. G. Kim, H. Lim, and Y. M. Ro, “Deep virtual reality image quality assessment with human perception guider for omnidirectional image,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 4, pp. 917–928, 2019.

[23] M. Yu, H. Lakshman, and B. Girod, “A framework to evaluate omnidirectional video coding schemes,” in IEEE International Symposium on Mixed and Augmented Reality, 2015, pp. 31–36.

[24] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, “Spherical structural similarity index for objective omnidirectional video quality assessment,” in IEEE International Conference on Multimedia and Expo, 2018, pp. 1–6.

[25] J. Xu, W. Zhou, and Z. Chen, “Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 5, pp. 1724–1737, 2021.

[26] C. Li, M. Xu, L. Jiang, S. Zhang, and X. Tao, “Viewport proposal CNN for 360-degree video quality assessment,” in IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10169–10178.

[27] T. Cohen, M. Geiger, J. Kohler, and M. Welling, “Spherical CNNs,” in International Conference on Learning Representations, 2018.

[28] A. Nguyen, Z. Yan, and K. Nahrstedt, “Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction,” in ACM Multimedia, 2018, pp. 1190–1198.

[29] 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 IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1420–1429.

[30] M. Xu, Y. Song, J. Wang, M. Qiao, L. Huo, and Z. Wang, “Predicting head movement in panoramic video: A deep reinforcement learning approach,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41,
