Modeling the Non-uniform Retinal Perception for Viewport-Dependent Streaming of Immersive Video

Peiyao Guo, Wenjing Su, Xu Zhang, Hao Chen, and Zhan Ma, *Senior Member, IEEE*

***Abstract*—Viewport-dependent streaming (VDS) of immersive video typically devises the attentive viewport (or FoV - Field of View) with high-quality compression but low-quality compressed content outside of it to reduce bandwidth. It, however, assumes uniform compression within the viewport, completely neglecting visual redundancy caused by non-uniform perception in central and peripheral vision areas when consuming the content using a head-mounted display (HMD). Our work models the unequal retinal perception within the instantaneous viewport and explores using it in the VDS system for non-uniform viewport compression to further save the data volume. To this end, we assess the just-noticeable-distortion moment of the rendered viewport frame by carefully adapting image quality-related compression factors like quantization stepsize q and/or spatial resolution s zone-by-zone to explicitly derive the imperceptible quality perception threshold with respect to the eccentric angle. Independent validations show that the visual perception of the immersive images with non-uniform FoV quality guided by our model is indistinguishable from that of images with default uniform FoV quality. Our model can be flexibly integrated with the tiling strategy in popular video codecs to facilitate non-uniform viewport compression in practical VDS systems for significant bandwidth reduction (e.g., about 40% reported in our experiments) at similar visual quality.**

***Index Terms*—Quality perception threshold, non-uniform visual sensation, viewport-dependent streaming**

I. Introduction

Immersive videos have been adopted in applications at a fast pace in the past years, attributing to the advances of affordable 4K/8K 360° cameras used for immersive (or virtual reality - VR) content acquisition, high-speed networks like 5G or Wi-Fi 6 for inter-connection, and commodity head-mounted display (HMD) devices conveniently used for interactive content consumption [1]–[4]. In immersive applications, users usually wear an HMD device (e.g., HTC Vive, Meta Quest, Apple Vision Pro, etc.) to immerse themselves in a virtual space to consume the content through interactive navigation (see Fig. 1(a)).

Background. As seen, at a specific moment, just the content of the current viewport rendered on the HMD display is actually perceived. This is consistent with our natural viewing behavior in the physical world, where we are only capable of seeing the scene just in the front. Such a phenomenon thus motivates extensive explorations on viewport-dependent streaming (VDS) of immersive/omnidirectional video content to greatly save transmission bandwidth [5]–[13]. A practical VDS approach often applies a popular two-tier framework, where it streams high-quality (high-bitrate) representations of the fixated (attentive) viewport at the enhancement tier, and delivers low-quality (low-bitrate) omnidirectional video as the base tier. Such a two-tier VDS approach can avoid the scene freezing (or blackout) when navigating from one viewport to another. To guarantee the uncompromised quality of experience (QoE), notable studies have also been conducted to predict the next viewport for high-quality content prefetching [14]–[16]. Nevertheless, considering an omnidirectional video with 8K spatial resolution at 60 FPS (frame per second), to sustain the service with satisfied QoE, a typical solution in [6] yet demands a stable connection close to and even surpassing a hundred Mbps (megabits per second), which is still impractical for most application circumstances.

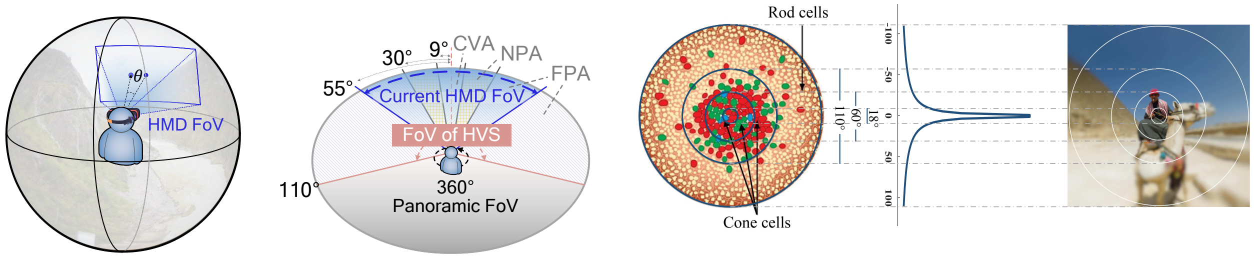
This urges the further reduction of bandwidth consumption in the VDS system to increase the application deployment for more service provisioning. In typical VDS systems, the high-quality (high-bitrate) viewport content instantaneously rendered on the HMD screen contributes a significant portion of total bandwidth consumption, e.g., ≈80% according to our measurements. They usually assume uniform visual sensation within the current field of view (FoV) and apply uniformly compressed (high-quality) tiles across the entire viewport. Actually, only the central vision (or macular) area (CVA) of our retina requires ultra-high resolution and high fidelity [17], [18], while the peripheral area has significantly reduced sensitivity [19], [20]. Such non-uniform visual sensitivity is attributed to the highly non-uniform distribution of cones along with retinal eccentricity [21], as illustrated in Fig. 1(c). Similarly, user perception when wearing the HMD would have a similar non-uniform visual sensation since the central and peripheral vision areas are included in the current FoV, as shown in Fig. 1(b). As a result, leveraging such unequal retinal perception of the human visual system (HVS) within the instantaneous FoV can potentially reduce the data volume for immersive video streaming without sacrificing the perceptual quality.

Method. In our work, we aim to develop a quantitative perception model to guide the non-uniform compression within current FoV for reducing bandwidth requirements in VDS streaming without perceptual degradation, which measures the imperceptible quality perception thresholds over the current viewport.

Given that the unequal retinal perception is mainly attributed to the non-uniform density distribution of cones on the retina [21] concerning retinal eccentricity θ, we rephrase the perception model as the concrete Quality Perception

Peiyao Guo, Wenjing Su, Hao Chen, and Zhan Ma are with the Nanjing University, Nanjing, 210093 China. (e-mails: {peiyao, wenjing_su}@smail.nju.edu.cn, {chenhao1210, mazhan}@nju.edu.cn).

Xu Zhang is with the University of Leeds, U.K. (e-mail: x.zhang15@leeds.ac.uk).



(a) (b) (c)

Fig. 1. (a) A virtual space to consume immersive content is illustrated by wearing a head-mounted display (HMD), where the instantaneous FoV/viewport within HMD is highlighted, and the FoV's or viewport's coverage can be characterized using the retinal eccentricity θ; (b) although the FoV offered by the HMD display, i.e., one-side $[0^\circ, 55^\circ]$, is a sub-area of the biological FoV of the HVS, i.e., $(0^\circ, 110^\circ]$, it still consists of the vision areas that can be specified using θ, e.g., central vision area (CVA) with one-side $\theta \in (0^\circ, 9^\circ]$, near peripheral area (NPA) with one-side $\theta \in (9^\circ, 30^\circ]$ and far peripheral area (FPA) with one-side $\theta \in (30^\circ, 55^\circ]$; (c) the distribution of cones on the human retina is highly non-uniform, contributing to unequal visual sensitivity.

Threshold (QPT) function with respect to the horizontal θ. This function represents the minimum quality level that can be achieved without any perceptual loss, and it is controlled via the quantization stepsize q and spatial resolution s in mainstream codec settings. Specifically, we assess the just-noticeable-distortion (JND) moment of a zonally-compressed viewport content at a set of given θs, which involves adapting the quality of individual frame zones by adjusting q and/or s according to the specific θ (dubbed θ-zone for convenience). The QPTs are derived from the imperceptible compression quality before JND measure at each θ, denoted with relevant compression factors like q-threshold (maximum q), s-threshold (minimum s), or their combination q-s-threshold. In the end, analytical models along with θ like $q(\theta)$, $s(\theta)$ or $q(s, \theta)$ could be formulated in the unified form to measure the separate impact of q or s and the joint impact of q and s perceptually, which is largely different from existing immersive quality assessment works assuming the uniform-quality viewport [7], [22], [23].

To verify our proposed model, we invited hundreds of subjects with normal vision to participate in quality assessment experiments. Experimental results have shown that both the threshold values of q and s can be well modeled using a generalized parametric Gaussian model in terms of the θ separately. We have further discovered that q-threshold can be independent of the s through the joint q-s-threshold exploration. With these models, we are capable of setting non-uniform q and/or s zone-by-zone to compress immersive content rendered on the HMD screen with noticeable bitrate saving but the same overall perceptual quality. Independent validations with subjective assessments are conducted to evaluate the perceived quality of content with the proposed non-uniform quality versus the default uniform quality, where high correlation indexes demonstrate the efficacy of our model in maintaining visual perception with non-uniform immersive quality.

Besides, our model can provide specific quality thresholds for various tiling strategies, which facilitates the flexible unequal quality setup in immersive applications without perceptual degradation. When applying the perception model to viewport-dependent VR streaming using the official platform of Grand Challenge on 360-degree Video-on-demand Streaming[1], quantitative results show a significant bandwidth reduction (over 40% on average) in comparison to the existing VDS solution, promising its potential in practice for navigating the prospective content with ultra-high definition fidelity and high frame rate.

TABLE I
ABBREVIATIONS

Abbr.	Description
HMD	Head-mounted display
VDS	Viewport-Dependent Streaming
FoV	Field of View
θ	Retinal eccentricity or eccentric angle
θ-zone	viewport's regional zone specified by the θ range
q (QP)	quantization stepsize (quantization parameter)
s	spatial resolution
JND	Just Noticeable Distortion
QPT	Quality Perception Threshold
q-threshold	maximum q or minimum s corresponding to QPT
s-threshold	
CVA	central vision area
NPA/FPA	near/far peripheral area
UFQ/NUFQ	uniform/non-uniform FoV quality

Contributions of this work are summarized below:

- To the best of our knowledge, our work is the *first* one to characterize the unequal retinal perception with explicit compression-quality (q/s) perception threshold functions regarding the eccentricity θ when consuming immersive content with the HMD.
- Hundreds of human subjects from various ages, majors, and genders are invited to rate for quality perception thresholds on different scenes. Independent validation assessment reveals almost the imperceptible quality difference between the rendered content within HMD that are respectively compressed using default uniform quality setting and proposed non-uniform quality setting guided by our model.
- Proposed models could be directly used to guide the non-uniform viewport/FoV content compression with flexible tiling strategies to realize data reduction without noticeable perceptual degradation. Quantitative comparisons with the existing VDS solution report over 40%

[1] https://2024.acmmmsys.org/gc/360-vod/

bandwidth reduction on average, promising its potential in practice.

Table I lists frequently used abbreviations in this work.

II. Related Work

This section briefly reviews relevant explorations, including quality assessments of immersive content, peripheral vision, and related studies, as well as viewport-dependent streaming of immersive or VR content.

Quality Assessments. In a panoramic scene, only partial content is rendered within the current FoV at a specific instant, and signal degradation at different positions contributes unequally to the final perception of the content. Many works thus have extended traditional image and video quality metrics like PSNR (Peak signal-to-noise ratio), SSIM (Structural SIMilarity), etc., by introducing pixel inequality to reflect such viewing behavior [24]–[26] to measure the objective quality of immersive content, such as the WS-PSNR (weighted spherical PSNR) taking the positional weights on the spherical surface into the account [24], and the SSIM360 index assuming more stretching area makes less contribution to final quality score [26].

However, as reported in [27], [28], the objective metrics mentioned above can not accurately predict the subjective quality. A collection of learning-based methods [29]–[34] have been devised to predict the perceptual quality of immersive content via extracted scene features and user viewing traces in an end-to-end way. For instance, Kim *et al.* [30] encoded positional and visual information of each image patch to estimate patch-wised weight and quality score, which were then aggregated to derive the overall assessment. Sun *et al.* [31] firstly transformed the sphere image into a projected cubemap for multi-plane feature aggregation and final quality score estimation. Zhou *et al.* [33] further considered the projection conversion differences among various view directions in the overall perceptual assessment via a transformer-based architecture.

Unfortunately, the aforementioned metrics often lack direct connections with compression quality control factors like quantization, for which they can not be easily used to adapt the underlying codec in practical streaming services. As a result, another set of explorations has attempted to characterize the overall perceptual quality as the function of resolution and quantization [22], [23], [35], in which they generally assume uniform visual sensation over the viewport for assessment and model development.

A great review article with comprehensive studies on the perceptual quality of immersive image/video can be found in [2]. Till now, existing works have hardly considered the unequal visual perception within the FoV when viewing the content on the HMD screen for immersive quality assessment.

Peripheral Vision and Foveated Rendering. A notable number of efforts have been devoted to studying the visual acuity or sensitivity in the central vision area and periphery. A review on peripheral vision can be found in [36]. Specifically, in 1998, Duchowski *et al.* [37] conducted experiments to test the perceptibility of spatial degradation in the visual periphery of video frames. The result revealed image resolution, which empirically decreased at the rate of visual acuity, produced an imperceptible impact on subjective quality. Recently, Rai *et al.* have performed serial explorations to understand the perceptual quality in the visual periphery [38]–[40], including the impacts of content features, e.g., texture, color, motion, and flicker. Results have revealed that flicker and color distortions are particularly important in the periphery.

Such non-uniform quality sensation in the periphery inspires subsequent investigations such as the foveated rendering in [41]–[44]. They leverage lower visual sensitivity in the peripheral vision to degrade the luminance signal along with the retinal eccentricity to reduce rendering costs without perceptual loss. Nevertheless, existing works do not end up with a quantitative model capable of measuring the perception at a given eccentric angle to guide the non-uniform compression of the content within current viewport.

Viewport-Dependent Streaming of Immersive/VR Video. In practice, viewport-dependent streaming (VDS) usually transmits immersive videos in two quality scales [5]–[13], where the current viewport or FoV is set with high-quality scale while reduced-quality elsewhere. This is also known as a two-tier scheme, which ensures high-fidelity perception within the viewport and quick response against sudden bandwidth dynamics or viewport re-orientation. Such a two-tier system can be implemented by adapting compression settings in tiles, providing a similar QoE but with significant bandwidth reduction, in the way of adapting spatial resolutions [10], [11], quantization parameters [5], [6], [9], etc. Then, the streaming algorithm would jointly consider current network conditions, buffer status, viewport prediction, etc., to determine which video chunk to transmit at which quality scale. However, existing works usually assume uniform compression in the viewport and choose some typical parameter settings to realize various quality allocations, which could not be optimal for immersive visual perception. This work makes the first attempt to explore and derive explicit models to guide non-uniform compression within the viewport for more bandwidth reduction without visual quality degradation.

III. Modeling the Non-uniform Retinal Perception

First of all, the density of photoreceptors on the retina [21] is non-uniformly distributed as shown in Fig. 1(c), leading to non-uniform visual perception between central and peripheral vision area. Tyler [47] thus proposed a power function to quantify the cone's density[2] (e.g., measured by the number of cones per mm^2) with respect to the retinal eccentricity θ from 0.2° to 20°, i.e.,

$$\rho(\theta) = 50000 \cdot \left(\frac{\theta}{300}\right)^{-\frac{2}{3}}, \quad \theta \in [0.2^\circ, 20^\circ]. \tag{1}$$

When $\theta > 20^\circ$, $\rho(\theta) \approx 4000$ cones/mm^2 for a large area in peripheral retina [49].

On the other hand, it is known that the HVS presents non-uniform perception capacity, which is related to the density of

[2] We assume the application scenario in the daytime where perception is dominated by the cone-mediated vision [48].

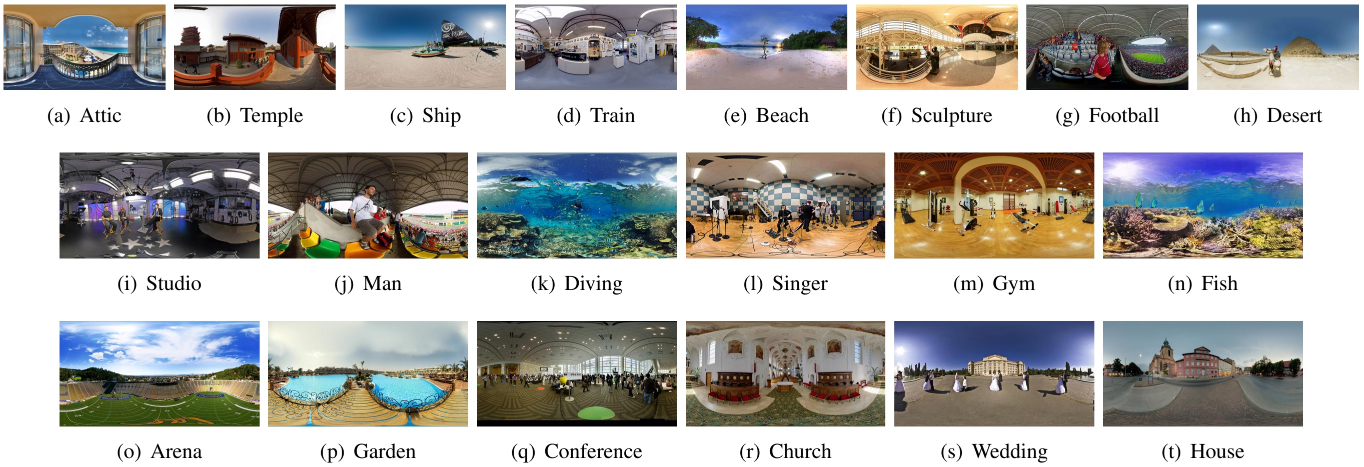


Fig. 2. Immersive images used for subjective assessment and model development (a)-(h) and independent model validation (i)-(t). They are selected from the SUN360 database [45] to cover a sufficiently wide range of content characteristics. Images are rendered with meaningful saliency in the FoV or viewport of the HMD display following prior studies [23], [46].

cones [50]. Specifically, users have higher perception acuity in the central vision area with more cones, in which even a small quality variation of rendered image zones in this area is potentially detectable. On the other hand, users present lower visual acuity in the periphery with the lower density of cones, suggesting that reduced fidelity of the rendered scene in this area may be imperceptible. Such non-uniform visual perception within instantaneous HMD FoV should correspond to the cone density distribution $\rho(\theta)$. As the frame quality variation is often controlled via compression factor q (quantization stepsize) or s (spatial resolution) [51] in practical, the retinal perception measurement is specifically rephrased to model the quality perception threshold (i.e. the maximum q or minimum s) as a function of the eccentricity θ without noticing the immersive image quality degradation. Towards this goal, we have assessed the JND moments of viewport content at a set of given θs, where we gradually adapt the frame compression quality by θ-related zones within the current HMD FoV. The imperceptible compression quality before the JND moment at each θ represents the measured QPT, corresponding to q-threshold, s-threshold, and their combination q-s-threshold in separate q- or s-impact and joint q-s- impact exploration. Lastly, analytical models like $q(\theta)$, $s(\theta)$, and even $q(s,\theta)$ are concretely developed to guide the non-uniform compression for the HMD FoV content without noticeable degradation.

A. Subjective Assessment

1) Test Preparation: Since adapting quantization and spatial resolution mainly affects the spatial quality of immersive content, immersive images are used for subjective assessment and model derivation. Subsequent applications could apply these derived models uniformly across video frames.

Eight immersive images from the SUN360 database [45] are chosen and uniformly downscaled to the spatial resolution of 4096×2160 as the testing images, as shown in Fig. 2(a)-2(h). Another twelve images shown in Fig. 2(i)-2(t) are used for validation. They represent typical scenarios of immersive video applications with the spatial information (SI) indexes [52] covering a wide range of content characteristics. Besides, each image contains meaningful saliency to fill up the user's FoV when rendered on the HMD screen for consumption [23], [46].

The HTC Vive system [53] with its associated HMD is set up to perform subjective quality assessments, which provides the binocular 110° FoV at 2160×1200 spatial resolution refreshed at 90Hz (or frame per second, FPS). The same methodology applies to other VR systems as well.

To ensure the derived model could describe the visual perception of compressed content in daily media stream, each image is compressed with multiple quality levels, by different combinations of the s or/and q via mainstream encoders (e.g., x264). Three independent tests are performed to study the separate and joint impact of q and s. Specifically,

- for evaluating the independent impact of q, we enforce the image at its native resolution, but apply ten different qs via equivalent quantization parameters (QPs) increasing uniformly from 22 to 49;
- to study the impact of s, we adapt eight distinct resolution levels (from 4K to 240p) for each raw image;
- for the joint impact of q and s, we still use ten different QPs, but with only four resolution levels for each QP. It is to reduce test cases for each subject as the subjects feel dizzy after a long rating process. Normally, each subject's test duration should be less than 30 minutes [22].

To characterize the unequal perception concerning the eccentricity θ, the 110°-wide FoV of the HMD display viewport is divided into three zones following the retinal structures [17], [20], [54], i.e., CVA with one-side eccentricity $\theta \in (0^\circ, 9^\circ]$, NPA with $\theta \in (9^\circ, 30^\circ]$ and the rest $\theta \in (30^\circ, 55^\circ]$ for FPA (see Fig. 1(b)). We implement an interactive UI to gradually adapt the test material's quality by zones for assessing the perception. Although applying fine-grained θ with more zone partitions would make the analytical model more accurate, the increased experimental complexity makes assessments hard to perform. The current three-zone setup already costs about 30 minutes per subject. More zones need longer rating duration, causing subjects to feel fatigued and give noisy data.

2) Test Procedure: As users commonly fixate on the salient viewport for a reasonable duration without noticeable movement when navigating the immersive content within HMD [46], the subjects are kindly asked to stay steady by fixating on the FoV center without head and body movement for measuring non-uniform retinal perception conveniently within current FoV. A tiny green cross is overlaid to assist subjects in fixing their gazes quickly, as also suggested in [44].

In general, we show image pairs sequentially to determine the Quality Perception Thresholds by zones and the corresponding compression parameters (e.g., q-threshold, s-threshold, etc.) under the guideline of the double stimulus [55] and JND criteria. Each sample in a pair is displayed for about 3 seconds, with 1 second in between and another 1-second pause to record the subjective JND opinion. There is a 1-minute interval for subjects to rest between two distinct scenes.

For each rating pair, one sample is the *anchor* image, and the other is the *testing* sample. Specifically, the initial anchor image is presented at its native spatial resolution $s_{\max}$ and uniformly compressed with $q = q_{\min}$ (i.e., $\mathrm{QP}_{\min} = 22$).

- First, we increase q or reduce s step-wisely to degrade the compression quality of the testing sample until the subject perceives distortion - this is referred to as the JND moment. We retrieve the recorded q_c or s_c for the quality just before the JND moment to infer the quality perception threshold of the CVA zone, referred to as q-threshold or s-threshold. In other words, with $q \leq q_c$ or $s \geq s_c$, we will not sense any perceptual difference between the anchor image and the testing sample.
- Afterwards, we replace the anchor image with an image that is compressed using q_c or s_c uniformly in the entire image, then we fix the content quality at the CVA of each test sample using the q_c or s_c and degrade the quality in both the NPA and FPA until the subject notices the distortion. We record the corresponding thresholds for NPA, e.g., q_{np} or s_{np}, respectively.
- Finally, we replace the anchor image by one that is compressed using q_c or s_c in the CVA, and by q_{np} or s_{np} in the NPA and FPA, and for the test sample, we fix the quality in the CVA and NPA with the q_c/s_c and q_{np}/s_{np}, respectively, and continue to degrade the quality in the FPA separately until the subject can detect the difference perceptually. The corresponding threshold, e.g., q_{fp} or s_{fp} are marked.

To ensure the model generalization, we have tried our best to make the assessments reliable for subsequent modeling. First, the subjects could retract their decisions three times at most if they do not have sufficient confidence, by adjusting to the preceding level for re-evaluation. Additionally, we perform three independent assessments with different subjects for collecting the q-threshold, s-threshold, and q-s-threshold separately. Results show a consistent trend, revealing the generalization of the proposed methodology.

With measured QPTs across CVA, NPA, and FPA regions specified by θ, analytical perception models are formulated as $q(\theta)$, $s(\theta)$, or $q(s, \theta)$ for subsequent applications.

3) Test Participants: We invite 175 students, including 101 males and 74 females, from different majors in Nanjing University for subjective assessments. All viewers have normal vision (or after correction) and color perception after performing the respective Snellen and Ishihara tests [55]. About 90% of viewers are naïve with video processing, subjective assessment, or virtual reality. They volunteer to anonymously participate in these tests after being clearly explained clear explanations of the experimental purpose and procedures.

TABLE II
PARAMETERS FOR VISUAL QUALITY THRESHOLDS MODEL $\hat{q}(\theta)$ OR $\hat{s}(\theta)$.

	a	b	c	d
$\hat{q}(\theta)$	2.2	0.08	1.38	0.05
$\hat{s}(\theta)$	2.2	0.033	$c(\mathbf{x})$	0.06

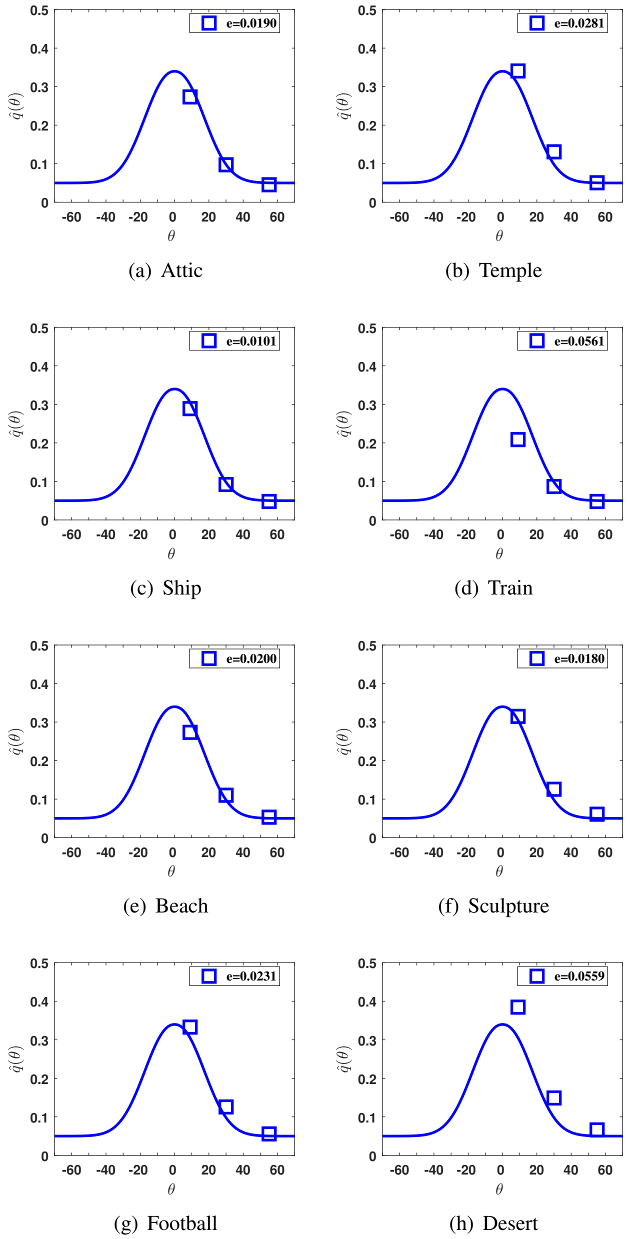
4) Data Post-Processing: It takes about 26 minutes for each subject to view a test sequence, including various quality scales for all images. Approximately 40-50 viewers assess each test sequence. We collect and screen all raw data to remove outliers (to reduce the rating noise). Specifically, we first generate the probability distribution of the q-thresholds for each image with all ratings, and then calculate the mean (μ) and standard deviation (σ). For the j-th image rated by the i-th subject on q-threshold of the CVA, if $|q_{j,i,c} - \mu_{q_{j,c}}| > 2 \times \sigma_{q_{j,c}}$, we would exclude this number. Here, $\mu_{q_{j,c}}$ and $\sigma_{q_{j,c}}$ are the mean and standard deviation of all the measured q-thresholds in the CVA for i-th image from all subjects. Similar procedures are conducted for screening s-threshold or q-s-threshold data. All ratings of a subject will be removed if its individual rating is excluded twice or more. After data screening, each test sample has about 35 valid threshold ratings for each vision area and the means of these valid data are considered as the final q-, s- or q-s-thresholds of this image.

B. Analytical Models

1) Separate Impact of Quantization and Spatial Resolution: We normalize the q and s for a unified presentation to model the q- and s-thresholds, i.e., $\hat{q} = q_{\min}/q$, and $\hat{s} = s/s_{\max}$ (note that s indicates the total number of pixels). When q = 64, the corresponding $\hat{q}$ is 0.125; while s = 2048 × 1080, $\hat{s} = 0.25$.

The discrete points in Fig. 3 are measured q-thresholds corresponding to the CVA, NPA, and FPA, respectively. Here, we use eccentric angles at 9°, 30°, and 55° which are typical boundaries across neighbor regions to represent their corresponding vision areas. We call them "border θ" for simplicity. Note that we measure the JND region-wisely, revealing the lowest quality that can be just sensible in this specific region (e.g., CVA, NPA, and FPA). In the meantime, the visual quality acuity/sensitivity gradually degrades eccentrically following the distribution of cone cells on the retina [17]–[19]. Thus, we apply the border θ of each region to reflect its just imperceptible quality threshold since the farthest border θ corresponds to the lowest quality in this specific region accordingly. Similarly, Fig. 4 shows measured s-thresholds in s-impact test.

We aim to derive an analytical model for these quality perception thresholds as a continuous function of the eccentric angle θ based on the measured thresholds at the critical points. By examining the trend of how quality threshold changes with θ, it is found that the generalized Gaussian function could fit



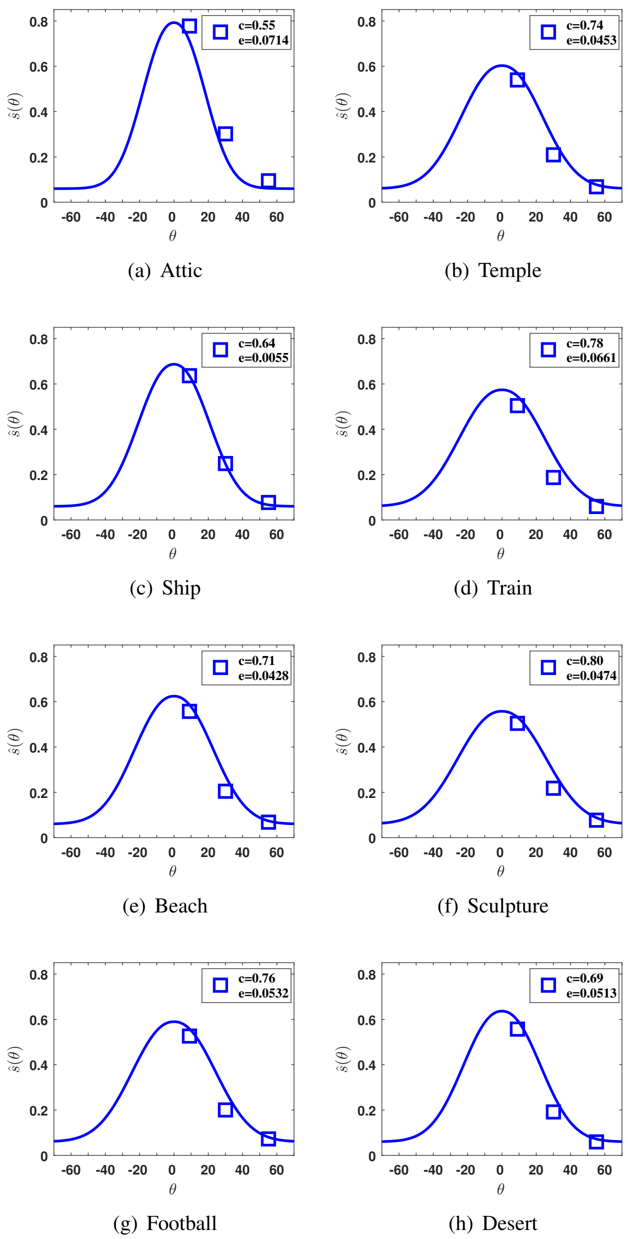
(a) Attic (b) Temple

(c) Ship (d) Train

(e) Beach (f) Sculpture

(g) Football (h) Desert

Fig. 3. Measured quantization thresholds and fitted threshold model versus eccentric angle, i.e., $\hat{q}(\theta)$. e represents the root mean square error (RMSE). Parameters are fixed for all image content.



(a) Attic (b) Temple

(c) Ship (d) Train

(e) Beach (f) Sculpture

(g) Football (h) Desert

Fig. 4. Measured spatial resolution thresholds and fitted threshold model versus eccentric angle, i.e., $\hat{s}(\theta)$. e represents the root mean square error (RMSE). Parameters except c are fixed, while c is content dependent.

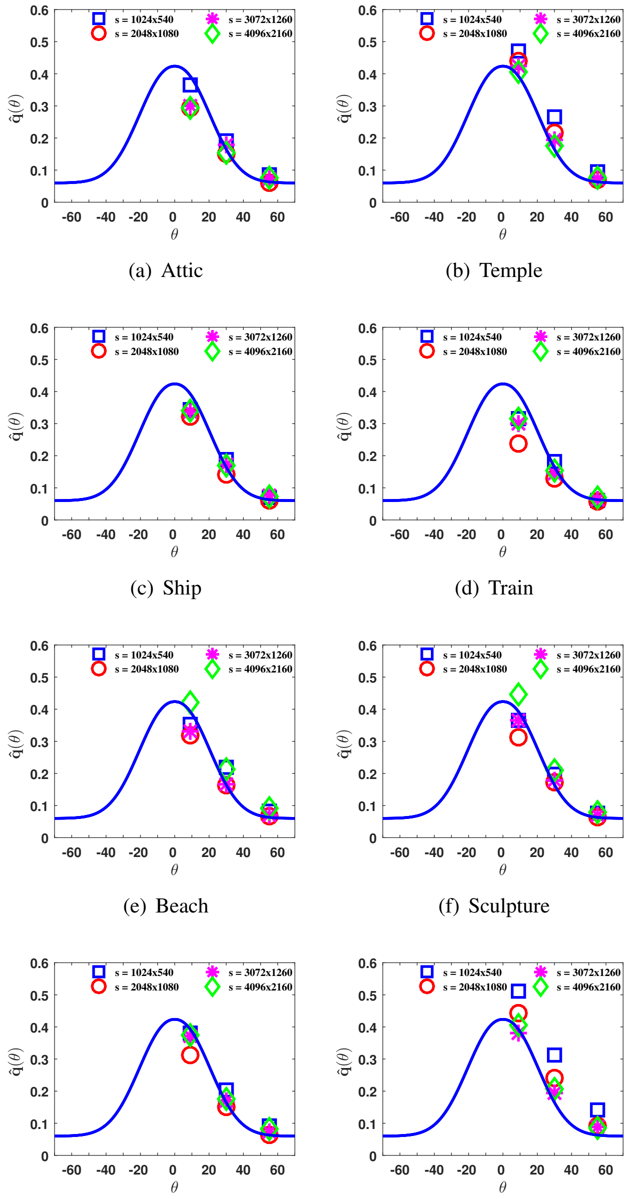
the measured points well, i.e.,

$$\hat{q}(\theta), \text{or} \quad \hat{s}(\theta) = \frac{1}{c\sqrt{2\pi}} \times e^{-\frac{|(b\cdot\theta)^a|}{2c^2}} + d, \tag{2}$$

where a, b, c, d are model parameters derived by fitting the measured thresholds in Eq. (2) for the average error minimization. This model also presents a similar trend with the density distribution of cones $\rho(\theta)$.

Table II shows fitted parameters for respective $\hat{q}(\theta)$ and $\hat{s}(\theta)$. Parameters differ between q and s due to the visual distortion variations caused by quantization and down-sampling.

- Parameter a reflects the decay speed of the visual sensitivity with the increasing θ. Parameter b differs for the impacts of q and s on the visual sensitivity. Parameter c is generally content-dependent. But for $\hat{q}$, we can still use a fixed c for all images due to negligible accuracy loss but significant model complexity reduction. Parameter d indicates the quality perception threshold when θ goes to the max value (i.e., where the number of cones goes to zero). For q, d = 0.05 corresponds to QP = 48. In fact, the subjective tests show that subjects cannot distinguish between QP = 51 and 48. For s, d = 0.06 corresponds to s = 840×630 as it is impractical to have s = 0 for rendering. Thus, we set the parameter d with the least model prediction error.
- Parameter c for model $\hat{s}(\theta)$ is content-dependent. We have further explored how to predict parameter c from content



(a) Attic

(b) Temple

(c) Ship

(d) Train

(e) Beach

(f) Sculpture

(g) Football

(h) Desert

Fig. 5. Normalized $\hat{q}(\theta)$ at different spatial resolutions. Discrete points are measured data; while the curve is fitted model.

features. Intuitively, image quality is mainly determined by its spatial complexity, color distribution, and local orientation. Through careful examination, it is found that c could be predicted by the linear combination of $\rho_{c_{\mathtt{SI}}}$, ρ_{μ_I} and $\rho_{\mu_{\gamma_v}}$, i.e.,

$$c = -0.002 \cdot \rho_{c_{\mathtt{SI}}} + 0.4342 \cdot \rho_{\mu_I} + 3.9029 \cdot \rho_{\mu_{\gamma_v}} + 0.2557, \quad (3)$$

where $\rho_{c_{\mathtt{SI}}}$ is the SI of the image's partial content located in the CVA of current FoV (This is because we constrain the saliency region in the central vision), i.e.,

$$\rho_{c_{\mathtt{SI}}} = \frac{1}{N} \sum_{\{x|x \in \mathtt{CVA}\}} \sqrt{\mathtt{Sobel}_h^2(x) + \mathtt{Sobel}_v^2(x)}, \quad (4)$$

with $\mathtt{Sobel}_h^2(x)$ and $\mathtt{Sobel}_v^2(x)$ for horizontal and vertical Sobel computation, ρ_{μ_I} is the averaged intensity of the original image within current FoV in HSI color space, and $\rho_{\mu_{\gamma_v}}$ refers to the mean amplitude of vertical orientation which is calculated using a 3×3 Gabor filter.

TABLE III
PARAMETERS FOR $\hat{q}(\theta)$ AT DIFFERENT SPATIAL RESOLUTION. FITTING ERROR e REPRESENTS RMSE.

	s	a	b	c	d	e
$\hat{q}(s,\theta)$	4096×2160	2.2	0.05	1.2	0.05	0.0330
	3072×1260	2.2	0.05	1.3	0.05	0.02365
	2048×1080	2.4	0.06	1.2	0.06	0.04789
	1024×540	2.4	0.05	1.1	0.08	0.04733
$\hat{q}_s(\theta)$		2.2	0.055	1.1	0.06	0.04567

2) Joint Impacts of Quantization and Spatial Resolution: This section investigates the joint impacts of the quantization and spatial resolution on the perceptual quality with respect to the eccentricity θ. It is hard to conduct tests with exhaustively joint (q,s) settings for thorough exploration. Motivated by the previous work [23], we have performed the test where the q-threshold is studied at different spatial resolutions for joint impact discussion. To reduce the overall rating duration, we use a few typical spatial resolutions but still allow ten distinct quantization levels to cover a variety of quality scales.

We plot the normalized $\hat{q}(\theta)$ at different spatial resolution s in Fig. 5. It is found that discrete $\hat{q}(s,\theta)$ measurements are almost overlapped for different spatial resolutions. This implies that a single analytical model may be sufficient to explain the q-threshold at different s, though individually fitted models at different s can describe the perceptual sensation more precisely. Nevertheless, we directly fit the discrete $\hat{q}(\theta)$s using Eq. (2) as $\hat{q}_s(\theta)$ which is listed in Table III, first assuming the independent parameters at different spatial resolution, and then enforcing the same parameters for all spatial resolutions, via the least squared error criteria. As seen, the optimal s-dependent parameters do not differ significantly at different spatial resolutions and are quite close to the s-independent parameters. Since the prediction errors of the above two proposals for different s show homoscedasticity but don't meet the normality assumption of a one-way ANOVA, we use the Kruskal-Wallis test [56] to compare the effects of these two parameter sets, and the result ($P = 0.6099$) demonstrates that there is no difference at a 5% significance level. Thus, to minimize the number of model parameters, we propose to apply fixed parameters for the following discussion.

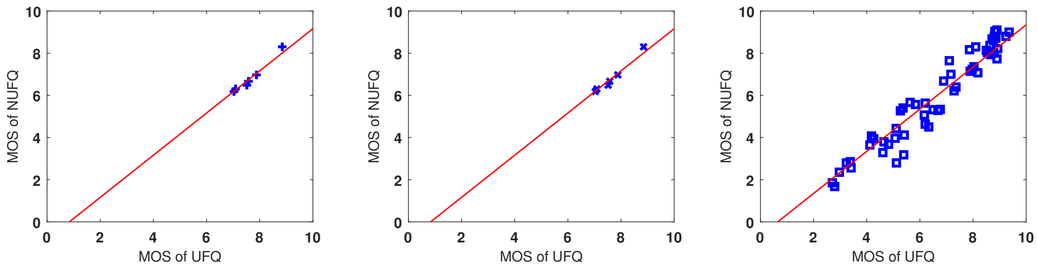
IV. Independent Model Validation

This section details the validation of our proposed analytical models for maintaining visual perception with non-uniform immersive content quality.

A. Validation of $q(\theta)$ and $s(\theta)$

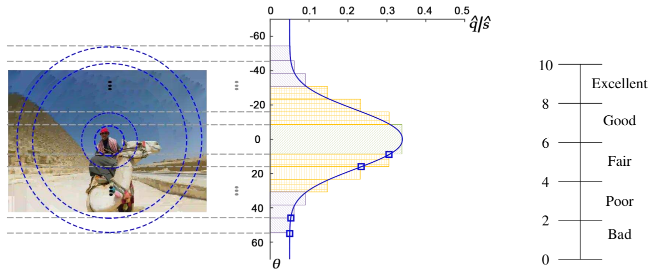
We invite another set of individual subjects to participate in the independent validation assessments with extra six scenes, as illustrated in Fig. 2(i)-2(n).

Differing from the model development in Section III, we propose to measure the mean opinion scores (MOS) directly



(a) $q(\theta)$ (b) $s(\theta)$ (c) $q_s(\theta)$

Fig. 6. Illustration of measured MOS on average for the images with uniform FoV quality (UFQ) versus corresponding ones with non-uniform FoV quality (NUFQ) guided by the separate model $q(\theta)$ or $s(\theta)$ and the joint model $q_s(\theta)$.



(a) (b)

Fig. 7. (a) An image with non-uniform compression using various q in central and peripheral areas via model (2); (b) The rating scale of the subjective assessment.

TABLE IV
STAIRCASE REPRESENTATION FOR $\hat{q}(\theta)$ AND $\hat{s}(\theta)$ OF THE SCENE *Studio*.

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
θ	0°-9°	9°-16°	16°-23°	23°-30°	30°-38°	38°-46°	46°-55°	> 55°
$\hat{q}$	0.3399	0.3052	0.2345	0.1562	0.0978	0.0640	0.0529	0.0503
$\hat{s}$	0.7192	0.6599	0.5325	0.3744	0.2348	0.1306	0.0820	0.0644

TABLE V
STAIRCASE REPRESENTATION FOR $\hat{q}(\theta)$ AND CORRESPONDING QP(θ) AT DIFFERENT SPATIAL RESOLUTION s.

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
θ	0°-9°	9°-16°	16°-23°	23°-30°	30°-38°	38°-46°	46°-55°	> 55°
$\hat{q}$	0.4236	0.3930	0.3262	0.2418	0.1649	0.1049	0.0751	0.0632
QP	29	30	32	34	38	42	44	46

[*] for $\theta \in [0°, 55°)$, s is a constant, e.g., 4096×2160, 3584×1890, 3072×1260, 2048×1080, 1524×810,1024×540.

of each image pair for model validation, of which one is uniformly compressed using a fixed $q_{\min}$ or $s_{\max}$; and the other is compressed with non-uniform quality using $q(\theta)$ or $s(\theta)$, where the image is partitioned into θ-related regional zones accordingly, as shown in Fig. 7(a)[3]. Associated q or s (or corresponding $\hat{q}$ or $\hat{s}$) for each θ-related zones are derived through the developed threshold model in Eq. (2) and parameter settings in Table II.

For example, the predicted quality thresholds for the scene *Studio* are given in Table IV. For other scenes, the distribution of q-threshold remains constant while that of s-threshold depends on the image's content. Except for those fixed parameters, content features are extracted from the images to derive the corresponding c via Eq. (3) explicitly.

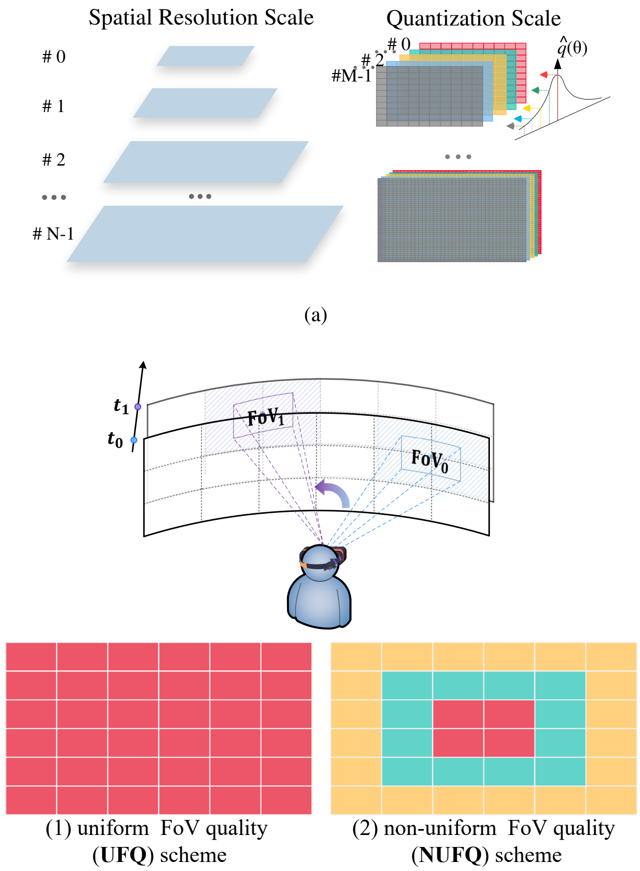
To let the participants familiarize themselves with the quality scales from the worst (MOS = 0) to the best (MOS = 10) as shown in Fig. 7(b), we prepare the training samples (*Attic* and *Dessert* in Fig. 2) for assessment pre-training. During the test procedure, we mix the image pairs from all test images and place them randomly to collect MOSs. The subject is asked to give a score ranging from 0 to 10 for each displayed sample sequentially. Each image sample repeats three times, totaling six repetitions of the same content: three for the sample with uniform FoV quality (UFQ) and another three for the copy with non-uniform FoV quality (NUFQ). Intuitively, the scores for each test sample should be very close to a specific subject. We enforce the repetition to avoid random noise.

For each image sample, all raw scores from all subjects are collected and then screened as discussed in Section III. The averaged value is referred to as its MOS. We then plot the MOSs for the samples with uniform quality versus the MOSs for the samples with non-uniform quality of the same image content, in Fig. 6(a)-6(b) for the respective impact of q and s. We further evaluate the PCC (Pearson correlation coefficient) and SRCC (Spearman's rank correlation coefficient) between averaged scores for each scene image with UFQ and corresponding image with NUFQ in both $q(\theta)$ and $s(\theta)$ validation. The PCC and SRCC are higher than 0.93, suggesting that the MOSs of the NUFQ image are highly correlated with the MOSs of the UFQ image, indicating the high efficiency of our proposed individual $\hat{q}(\theta)$ and $\hat{s}(\theta)$ to model the non-uniform retinal perception quantitatively. However, compared with the score of the UFQ image, there is a slight decrease on that of the NUFQ image, which indicates a reduction of 0.8384 and 0.6970 points respectively for q-impact and s-impact in Fig. 6(a) -6(b). However, users can't distinguish the quality disparity between the UFQ image and the corresponding NUFQ image as the relative deviations of the ratings are less than 1.

B. Validation of $q_s(\theta)$

This section extends the validation to the scenarios with non-uniform FoV quality, considering joint impacts of quantization and spatial resolution. Quality variations of twelve test images in Fig. 2(i)-2(t) are assessed by more than fifty

[3] Note that the same methodology can be easily extended to various tiling strategies with corresponding granularity.



(a)

(1) uniform FoV quality (**UFQ**) scheme

(2) non-uniform FoV quality (**NUFQ**) scheme

(b)

Fig. 8. Illustration of model-driven 360-degree video streaming: (a) multi-scale tile structure considering the variations from both spatial resolution and quantization, guided by Eq. (2); (b) a user navigating the immersive content from t_0 to t_1. The blue lines denote the FoV at t_0 and the purple lines indicate the FoV at t_1. For the content tiles delivered within HMD FoV at each timestamp ($\mathrm{FoV}_{0/1/\ldots}$), the typical VDS solution delivers tiles within the FoV at uniformly high quality (denoted as the **UFQ** scheme), while our proposed **NUFQ** scheme assigns diverse quality scales within the instantaneous FoV.

subjects. We prepare 6 resolution variations for each scene and compress different regional zones with non-uniform $q_s(\theta)$ following the model setting in Table III. Detailed compression parameters are listed in Table V. Fig. 6(c) reveals that similar rating trends of the uniform and non-uniform quality copies of the same content, with very high PCC at 0.9645 and SRCC at 0.9526. Meantime, the score of the NUFQ image is 0.65 points lower than that of the corresponding UFQ image on average which only marginally affects users' perception of the image's quality.

V. Model-Driven 360-degree Video Streaming

We further utilize the proposed non-uniform retinal perception model to optimize 360-degree video streaming for improved efficiency. Unlike typical viewport-dependent streaming approaches for immersive videos that only leverage the unequal content quality inside and outside the user's FoV, our perception models suggest allocating *non-uniform FoV quality* (NUFQ), a.k.a., non-uniform compression of the viewport content, during the streaming to further reduce the bandwidth without perceptual degradation. As shown in Fig. 8, we follow Eq. (2) to process the videos into multiple quality scales (i.e., via a variety of q and s) and then determine the quality scale of each tile within the FoV according to its corresponding retinal eccentricity. For the tiles outside the FoV, we assign the lowest quality following the basic practice used in VDS systems.

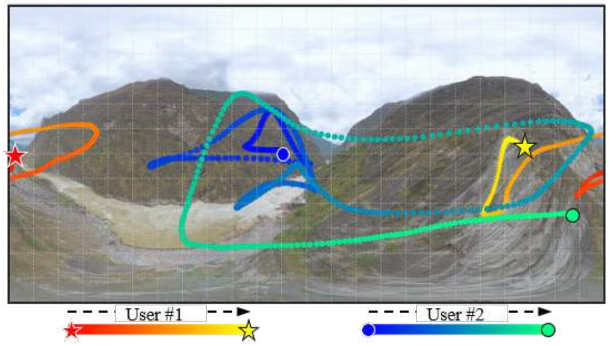


Fig. 9. Illustration of user motion trajectories in the scene "Valley". "User #1" represents slow head movement while "User #2" shows the drastic head movement.

In the following, the NUFQ scheme guided by our model is evaluated on an open platform E3PO[4] in comparison with the existing VDS scheme that applies the uniform compression within the FoV. E3PO, as the official valuation platform of "Grand Challenge on 360-degree Video-on-demand Streaming"[5] in ACM MMsys24, could implement different 360° video streaming approaches for performance comparison, using the same video content and same motion trajectory.

A. Experimental Setup

The pipeline of the VR streaming evaluation in the E3PO platform consists of three stages. Firstly, using the *video preprocessor* in E3PO, the videos were temporally segmented into 1-second chunks, each of which was further spatially segmented into 24×12 tiles. Each tile was independently encoded at different quality scales. Secondly, we employed linear regression and exponential smooth to predict potential user viewports every 10ms for the *streaming simulator* of E3PO. It then sent the detailed streaming actions that include when and which video chunk/tile was transmitted along with the desired quality scale. Here, the user's FoV of the display viewport was configured with a coverage of 90° × 90°. Lastly, we analyzed the whole streaming process from two perspectives. One is the network bandwidth consumed during the streaming. Lower bandwidth consumption demonstrates the ease of deploying VR streaming services. Another is the FoV uncoverage rate, which refers to the ratio or percentage of the FoV that is not covered or predicted accurately. A lower FoV uncoverage rate indicates a closer alignment between the predicted FoV and the actual FoV, resulting in a more immersive and realistic viewing experience.

We selected three 360-degree videos ("valley", "coaster", and "forest") from the grand challenge as the testing materials.

[4] https://github.com/bytedance/E3PO.git

[5] https://2024.acmmmsys.org/gc/360-vod/

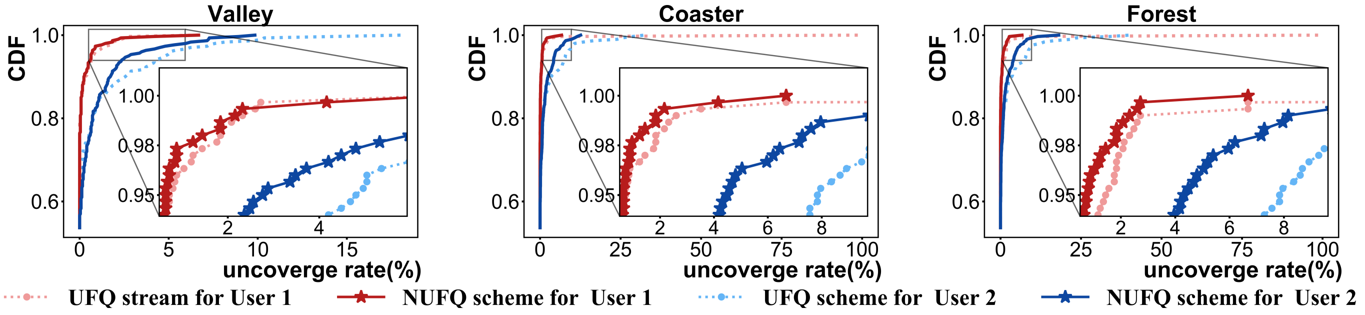


Fig. 10. Illustration of the FoV uncoverage rates in the form of cumulative distribution function (CDF) for both the UFQ and NUFQ schemes using the two user trajectories in various scenes. When the CDF value approaches 1 earlier, it indicates a closer alignment between the predicted FoV and the user's actual FoV.

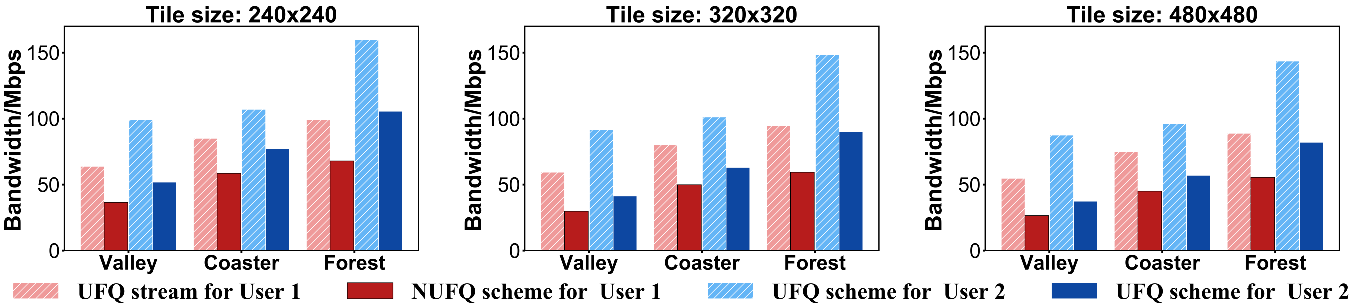


Fig. 11. Illustration of the bandwidth consumption during streaming with different tile sizes. As the tile size decreases, the compression efficiency decreases, resulting in higher bandwidth consumption. However, the proposed NUFQ scheme still achieves significant bandwidth reduction even when using different tile sizes.

Each test video is 10 seconds long and has a resolution of 7680×3840 with 30fps. Additionally, we selected two typical user trajectories, one denoting slow head movement and the other representing drastic head movement, as shown in Fig. 9.

We used the typical VDS method as a baseline, which delivers the tiles within the user's FoV at a high quality and other tiles at a low quality. Since this scheme assigns the same high quality within the FoV, we refer to it as *uniform FoV quality* (UFQ) scheme. In contrast, our NUFQ scheme assigned the tiles within the FoV at diverse quality scales based on the non-uniform retinal perception model. The quality of each tile is determined by its corresponding retinal eccentricity according to Eq. (2), as shown in Fig. 8(a). In the implementation of the UFQ scheme, we encoded the videos using typical quantization parameters (QP) of 22 for high quality and 44 for low quality. For the sake of fairness, we only varied the QP for our proposed NUFQ scheme. Specifically, the QP values were calculated according to Eq. (2). These QP values gradually increased from the center tile of the FoV to the peripheral tiles.

B. Experimental Results

Table VI illustrates the required bandwidth for streaming each video using different user motion trajectories in both the UFQ and NUFQ schemes. As demonstrated in Table VI, streaming with the NUFQ scheme results in an average bandwidth savings of over 40%, effectively reducing the bandwidth requirements for accessing VR streaming services.

TABLE VI
AVERAGED BANDWIDTH REQUIREMENT (MBPS) IN THE VDS METHODS USING THE UNIFORM FOV QUALITY (UFQ)/NON-UNIFORM FOV QUALITY (NUFQ) SCHEMES.

Scene	User	UFQ	**NUFQ**	Scalage
valley	# 1	59.360	**29.846**	49.72%
	# 2	91.469	**41.330**	54.82%
coaster	# 1	80.04	**49.846**	37.72%
	# 2	101.206	**63.017**	37.73%
forest	# 1	94.530	**59.366**	37.20%
	# 2	148.450	**90.080**	39.32%
Average		95.843	**55.581**	42.01%

Particularly, the NUFQ scheme exhibits more significant advantages in bandwidth reduction when the user motion is drastic. By reducing the bandwidth requirements, the NUFQ scheme exhibits a superior capacity to adapt to diverse network conditions, lowering the costs and complexity associated with deploying VR streaming services.

Fig. 10 depicts the cumulative distribution function (CDF) of the FoV uncoverage rate during the streaming processes. Compared to the UFQ scheme, our proposed NUFQ scheme exhibits a lower FoV uncoverage rate. This is because the reduced delivery time in the NUFQ scheme eliminates the need for long-range predictions of user movement, resulting in more accurate user motion prediction. However, in cases where the user motion is drastic, a long tail effect becomes evident in the CDF curve. This means that there is a probability of significant deviation between the predicted FoV and the user's actual FoV. As shown in Fig. 10, the maximum FoV uncoverage rate of the NUFQ scheme is substantially lower than that of the UFQ scheme.

In the above experiment, we observed that using three quality tiles at a resolution of 320×320 within the FoV resulted in an average bandwidth savings of over 40%. To further validate the bandwidth reduction potential of our model's guidance, we explored various tiling strategies for VR videos, such as 16×8 patches at a resolution of 480×480 and 32×16 patches at a resolution of 240×240. As shown in Fig. 11, our perception model successfully achieved significant bandwidth reduction across various tile sizes. However, it is worth noting that smaller tiles with more fine-grained quality scales may suffer from a decrease in compression efficiency, which can limit the overall bandwidth reduction achieved through unequal-quality tile allocation. It deserves more exploration to realize the trade-off between compression efficiency and bandwidth reduction in our future work.

In general, the NUFQ scheme guided by our perception model not only significantly reduces the bandwidth requirements of VR streaming, but also effectively enhances the alignment between the predicted FoV and the user's actual FoV. It shows remarkable robustness against network dynamics [57] and user interactive behaviors.

VI. Conclusion

This paper characterized the non-uniform retinal perception concerning the eccentricity θ, where, at each eccentric angle θ, we measured the JND to determine the quality perception threshold in the form of corresponding compression-related quantization q and/or resolution s, e.g., q-threshold, s-threshold, or q-s-threshold. Then the closed-form theoretical models like $q(\theta)$, $s(\theta)$, and $q_s(\theta)$ are derived to quantitatively offer various compression factors in respective θ-indexed image zones for viewport rendering. As seen, these models enabled non-uniform compression inside the viewport or FoV without impairing the perceptual sensation when wearing the HMD to consume the immersive content, which greatly differs from the existing works. Such a model-driven non-uniform FoV/viewport compression could be easily implemented on typical viewport-dependent streaming methods of immersive or virtual reality content. As demonstrated in a field test, our method with model-driven non-uniform viewport compression provides an average of 40% bandwidth savings without degrading the perpetual quality, promising the practical potential of the proposed models in applications.

Acknowledgment

The authors would like to thank all volunteers for their contribution in subjective assessments and thank all reviewers for their constructive comments.

References

[1] C.-L. Fan, W.-C. Lo, Y.-T. Pai, and C.-H. Hsu, "A survey on 360° video streaming: Acquisition, transmission, and display," ACM Comput. Surv., vol. 52, no. 4, aug 2019. [Online]. Available: https://doi.org/10.1145/3329119 1

[2] M. Xu, C. Li, S. Zhang, and P. Le Callet, "State-of-the-art in 360 video/image processing: Perception, assessment and compression," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 1, pp. 5–26, 2020. 1, 3

[3] A. A. Simiscuka, D. A. Ghadge, and G.-M. Muntean, "Omniscent: An omnidirectional olfaction-enhanced virtual reality 360° video delivery solution for increasing viewer quality of experience," IEEE Transactions on Broadcasting, vol. 69, no. 4, pp. 941–950, 2023. 1

[4] N. Gao, J. Zhou, G. Wan, X. Hua, T. Bi, and T. Jiang, "Low-latency vr video processing-transmitting system based on edge computing," IEEE Transactions on Broadcasting, 2024. 1

[5] F. Duanmu, E. Kurdoglu, S. A. Hosseini, Y. Liu, and Y. Wang, "Prioritized buffer control in two-tier 360 video streaming," in VR/AR Network@SIGCOMM, 2017. 1, 3

[6] L. Sun, F. Duanmu, Y. Liu, Y. Wang, Y. Ye, H. Shi, and D. Dai, "A two-tier system for on-demand streaming of 360 degree video over dynamic networks," IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 9, no. 1, pp. 43–57, 2019. 1, 3

[7] S. Xie, Y. Xu, Y. Li, Q. Shen, Z. Ma, and W. Zhang, "Perceptually optimized quality adaptation of viewport-dependent omnidirectional video streaming," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 1, pp. 146–160, 2019. 1, 2, 3

[8] Z. Jiang, X. Zhang, Y. Xu, Z. Ma, J. Sun, and Y. Zhang, "Reinforcement learning based rate adaptation for 360-degree video streaming," IEEE Transactions on Broadcasting, vol. 67, no. 2, pp. 409–423, 2020. 1, 3

[9] E. Ghabashneh, C. Bothra, R. Govindan, A. Ortega, and S. Rao, "Dragonfly: Higher perceptual quality for continuous 360 video playback," in Proceedings of the ACM SIGCOMM 2023 Conference, 2023, pp. 516–532. 1, 3

[10] L. Hsiao, B. Krajancich, P. Levis, G. Wetzstein, and K. Winstein, "Towards retina-quality vr video streaming: 15ms could save you 80% of your bandwidth," ACM SIGCOMM Computer Communication Review, vol. 52, no. 1, pp. 10–19, 2022. 1, 3

[11] J. Chen, Z. Luo, Z. Wang, M. Hu, and D. Wu, "Live360: Viewport-aware transmission optimization in live 360-degree video streaming," IEEE Transactions on Broadcasting, vol. 69, no. 1, pp. 85–96, 2023. 1, 3

[12] Y. Xie, Y. Zhang, and T. Lin, "Deep curriculum reinforcement learning for adaptive 360° video streaming with two-stage training," IEEE Transactions on Broadcasting, 2023. 1, 3

[13] Z. Li, Y. Wang, Y. Liu, J. Li, and P. Zhu, "Just360: Optimizing 360-degree video streaming systems with joint utility," IEEE Transactions on Broadcasting, 2024. 1, 3

[14] 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, no. 11, pp. 2693–2708, 2018. 1

[15] Z. Ye, Q. Li, X. Ma, D. Zhao, Y. Jiang, L. Ma, B. Yi, and G.-M. Muntean, "Vrct: A viewport reconstruction-based 360° video caching solution for tile-adaptive streaming," IEEE Transactions on Broadcasting, vol. 69, no. 3, pp. 691–703, 2023. 1

[16] Y. Wang, J. Li, Z. Li, S. Shang, and Y. Liu, "Synergistic temporal-spatial user-aware viewport prediction for optimal adaptive 360-degree video streaming," IEEE Transactions on Broadcasting, 2024. 1

[17] A. Gupta, S. Mazumdar, and S. Choudhry, "Practical approach to ophthalmoscopic retinal diagnosis," 2010. 1, 4, 5

[18] S. L. Polyak, The retina. Univ. Chicago Press, 1941. 1, 5

[19] J. Besharse and D. Bok, The retina and its disorders. Academic Press, 2011. 1, 5

[20] Wikipedia, "Peripheral vision — wikipedia, the free encyclopedia," 2019. [Online]. Available: https://en.wikipedia.org/wiki/Peripheral_vision 1, 4

[21] C. A. Curcio, K. R. Sloan, R. E. Kalina, and A. E. Hendrickson, "Human photoreceptor topography," Journal of comparative neurology, vol. 292, no. 4, pp. 497–523, 1990. 1, 3

[22] M. Huang, Q. Shen, Z. Ma, A. C. Bovik, P. Gupta, R. Zhou, and X. Cao, "Modeling the perceptual quality of immersive images rendered on head mounted displays: Resolution and compression," IEEE Transactions on Image Processing, vol. 27, no. 12, pp. 6039–6050, 2018. 2, 3, 4

[23] 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, 2021. 2, 3, 4, 7

[24] A. L. Y. Sun and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," IEEE signal processing letters, vol. 24, no. 9, pp. 1408–1412, 2017. 3

[25] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, "Spherical structural similarity index for objective omnidirectional video quality assessment," in 2018 IEEE international conference on multimedia and expo (ICME). IEEE, 2018, pp. 1–6. 3

[26] Facebook. (2018) Quality assessment of 360 video view sessions. [Online]. Available: https://code.fb.com/video-engineering/quality-assessment-of-360-video-view-sessions/ 3

[27] E. Upenik, M. Rerabek, and T. Ebrahimi, "On the performance of objective metrics for omnidirectional visual content," in 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX). IEEE, 2017, pp. 1–6. 3

[28] Z. Jiang, Y. Xu, J. Sun, J.-N. Hwang, Y. Zhang, and S. C. Appleby, "Tile-based panoramic video quality assessment," IEEE Transactions on Broadcasting, vol. 68, no. 2, pp. 530–544, 2022. 3

[29] H.-T. Lim, H. G. Kim, and Y. M. Ra, "Vr iqa net: Deep virtual reality image quality assessment using adversarial learning," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 6737–6741. 3

[30] H. G. Kim, H.-T. 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. 3

[31] W. Sun, X. Min, G. Zhai, K. Gu, H. Duan, and S. Ma, "Mc360iqa: A multi-channel cnn for blind 360-degree image quality assessment," IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 1, pp. 64–77, 2020. 3

[32] X. Zhou, Y. Zhang, N. Li, X. Wang, Y. Zhou, and Y.-S. Ho, "Projection invariant feature and visual saliency-based stereoscopic omnidirectional image quality assessment," IEEE Transactions on Broadcasting, vol. 67, no. 2, pp. 512–523, 2021. 3

[33] M. Zhou, L. Chen, X. Wei, X. Liao, Q. Mao, H. Wang, H. Pu, J. Luo, T. Xiang, and B. Fang, "Perception-oriented u-shaped transformer network for 360-degree no-reference image quality assessment," IEEE Transactions on Broadcasting, 2023. 3

[34] T. Wu, S. Shi, H. Cai, M. Cao, J. Xiao, Y. Zheng, and Y. Yang, "Assessor360: Multi-sequence network for blind omnidirectional image quality assessment," Advances in Neural Information Processing Systems, vol. 36, 2024. 3

[35] R. Zhou, M. Huang, S. Tan, L. Zhang, D. Chen, J. Wu, T. Yue, X. Cao, and Z. Ma, "Modeling the impact of spatial resolutions on perceptual quality of immersive image/video," in Proc. of the IEEE IC3D, Dec 2016. 3

[36] H. Strasburger, I. Rentschler, and M. Juettner, "Peripheral vision and pattern recognition: A review," Journal of Vision, vol. 11, no. 13, pp. 1–82, May 2011. 3

[37] A. T. Duchowski and B. H. McCormick, "Gaze-contingent video resolution degradation," in Human Vision and Electronic Imaging, 1998. 3

[38] Y. Rai, A. Aldahdooh, S. Ling, M. Barkowsky, and P. L. Callet, "Effect of content features on short-term video quality in the visual periphery," 2016 IEEE 18th International Workshop on MMSP, pp. 1–6, 2016. 3

[39] Y. Rai, M. Barkowsky, and P. L. Callet, "Role of spatio-temporal distortions in the visual periphery in disrupting natural attention deployment," Electronic Imaging, vol. 2016, no. 16, pp. 1–6, 2016. 3

[40] Y. Rai, M. Barkowsky, and P. Le Callet, "Does h.265 based peri and para-foveal quality flicker disrupt natural viewing patterns?" in 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), Sep. 2015, pp. 133–136. 3

[41] A. Patney, M. Salvi, J. Kim, A. Kaplanyan, C. Wyman, N. Benty, D. Luebke, and A. Lefohn, "Towards foveated rendering for gaze-tracked virtual reality," ACM Transactions on Graphics (TOG), vol. 35, no. 6, p. 179, 2016. 3

[42] D. Hoffman, Z. Meraz, and E. Turner, "Limits of peripheral acuity and implications for vr system design," Journal of the Society for Information Display, vol. 26, no. 8, pp. 483–495, 2018. 3

[43] O. T. Tursun, E. Arabadzhiyska-Koleva, M. Wernikowski, R. Mantiuk, H. Seidel, K. Myszkowski, and P. Didyk, "Luminance-contrast-aware foveated rendering," ACM Trans. Graph., vol. 38, pp. 98:1–98:14, 2019. 3

[44] N. Deng, Z. He, J. Ye, B. Duinkharjav, P. Chakravarthula, X. Yang, and Q. Sun, "Fov-nerf: Foveated neural radiance fields for virtual reality," IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 11, pp. 3854–3864, 2022. 3, 5

[45] J. Xiao, K. A. Ehinger, A. Oliva, and A. Torralba, "Recognizing scene viewpoint using panoramic place representation," in Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 2695–2702. 4

[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 TVCG, vol. 24, no. 4, pp. 1633–1642, April 2018. 4, 5

[47] C. W. Tyler, "Analysis of human receptor density," in Basic and clinical applications of vision science. Springer, 1997, pp. 63–71. 3

[48] D. Purves, G. Augustine, D. Fitzpatrick, and et al., Neuroscience, 2nd Ed. Sunderland (MA): Sinauer Associates, 2001, ch. Anatomical Distribution of Rods and Cones. 3

[49] C. W. Tyler, "Analysis of visual modulation sensitivity. iii. meridional variations in peripheral flicker sensitivity." Journal of the Optical Society of America. A, Optics and image science, vol. 4 8, pp. 1612–9, 1987. 3

[50] D. McBurney and V. Collings, Introduction to sensation/perception. Prentice-Hall, 1977. 4

[51] Y. Xue, Y.-F. Ou, Z. Ma, and Y. Wang, "Perceptual video quality assessment on a mobile platform considering both spatial resolution and quantization artifacts," in Proc. of PacketVideo, 2010. 4

[52] H. Yu and S. Winkler, "Image complexity and spatial information," in Quality of Multimedia Experience (QoMEX), 2013 Fifth International Workshop on. IEEE, 2013, pp. 12–17. 4

[53] HTC Vive. [Online]. Available: http://www.vive.com/us/ 4

[54] D. O. Harrington, Ed., The Visual Fields: A Textbook and Atlas of Clinical Perimetry. The C. V. Mosby Company, 1981. 4

[55] Rec. ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures," 2002. 5

[56] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," Journal of the American statistical Association, vol. 47, no. 260, pp. 583–621, 1952. 7

[57] S. Lederer, C. Mueller, and C. Timmerer, "Dynamic adaptive streaming over http dataset," in Proceedings of the ACM Multimedia Systems Conference, Feb 2012. 11