Q-STAR+: A Perceptual Quality Model for Viewport-Based Immersive Video Streaming

Yu Meng, Zhan Ma, Xun Cao and Qiu Shen

Abstract—Users typically wear the head mounted display (HMD) to navigate and enjoy the immersive video content. In practice, a viewport-based immersive video streaming strategy is intuitively preferred for constrained network bandwidth, but it still lacks of a theoretical framework to resolve this fundamental bandwidth-quality optimization problem. Therefore, in this work, we have developed an analytical model, so called “Q-STAR+”, to connect the perceptual quality of a compressed viewport video (viewed via wearing the HMD) with respect to its Spatial (frame size), Temporal (frame rate), and Amplitude (quantization) Resolutions (STAR). There are overall four parameters that can be easily extracted from the content itself. The complete model correlates well with the mean opinion scores (MOS) with a Pearson Correlation Coefficient (PCC) of 0.95, and a Spearman’s Rank Correlation Coefficient (SRCC) of 0.95, according to an independent validation test, yielding the state-of-the-art performance in comparison to those famous objective quality metrics (e.g., MS-SSIM [1], VIFP [2], FSIM [3], and VSI [4]). Together with the rate model, we could make the perceptual optimized viewport video streaming of immersive content analytical tractable, so as to maximize the visual quality subjectively under any given network bandwidth constraint.

Index Terms—Perceptual quality, content adaptive, viewport video, immersive content, rate-quality optimization

I. INTRODUCTION

Recently, we have witnessed the exponential growth of the video related applications, thanks to the advances in the areas of high speed networking [5], video processing (e.g., capturing [6], [7], compression [8] and display [9]), etc. Particularly, panoramic and immersive videos are becoming popular, because of the recent introduction of a variety of powerful virtual reality (VR) devices, such as the HTC Vive, Oculus Rift, Samsung Gear VR, etc.

Similar to the conventional videos that were delivered to diverse users over the heterogeneous networks, it is desired to have the immersive video applications accessed remotely. However, networks often exhibit dynamic behaviors, such as channel fading, congestion, etc, resulting in unpredictable bandwidth fluctuation. To tackle the network dynamics, a variety of standards and transport protocols are developed, such as HTTP Live Streaming (HLS) [10], Dynamic Adaptive Streaming over the HTTP [11], etc, where multiple copies at various quality levels (i.e., at different combinations of spatial, temporal and amplitude resolutions) for a same video are cached for adaptation according to the estimated network bandwidth. The most straightforward way to facilitate the immersive video content streaming is directly using the existing DASH or HLS backed mechanism. However, a significant amount of bandwidth is demanded to deliver the immersive video entirely. In practice, motivated by the characteristics of our human visual system (HVS), a viewport\(^1\) adaptive streaming scheme is often preferred [12], [13] with a noticeable bandwidth reduction. Intuitively, a DASH or HLS compliant viewport video streaming adaptation would be of great help to combat the network dynamics. Recent omnidirectional media format (OMAF) [14] also suggests the viewport-based immersive content streaming with underlying encapsulation using DASH. But how to ensure the optimal perceptual quality, when performing the bandwidth constrained viewport streaming to select the best STAR combination, still remains unclear.

Aforementioned VR devices offer the revolutionary stunning and immersive viewing experience, especially given a very wide field of view (FoV), and free navigation inside a virtualized space, shown in Fig. 1(a). The sensation of a vividly changed reality can be dramatic compared to traditional viewing of videos on fixed display screens having very limited FoVs in Fig. 1(b).

Users would often focus their attention on a few salient areas close to the equator [15], [16] that can be covered by a specific viewport. Therefore, we presume that the quality of an entire immersive content is highly correlated with the quality of its salient viewport. A topic of interest in this context is predicting the perceived quality of immersive videos at current viewport or FoV.

Furthermore, we have emphasized our studies on the impacts of different Spatial resolution (SR, or frame size \(s\)), Temporal resolution (TR, or frame rate \(t\)) and Amplitude Resolution (Quantization parameter QP or equivalent quantization

\(^1\)We use “viewport” and “Field of View (FoV)” interchangeably, unless specified otherwise.
A variety of STAR combinations for current viewport video generally constitutes a repository of DASH/HLS compliant content segments for adaptation.

Towards this goal, we have performed the subjective quality assessments on viewport videos extracted from the common immersive sequences chosen by IEEE 1857.9 Virtual Reality Unit (VRU) [17] and Joint Video Expert Team (JVET) [18]. People might argue that why not directly assess the entire immersive video without constraining the subjective rating on the particular salient viewports. One reason was that the user would require a certain duration to stabilize their focus, e.g., \(\approx 10\) seconds as observed in [19]. This would introduce more rating noise for the subjective assessment. On the other hand, after focusing their fixation to the salient viewports, users typically stay for a noticeable period (e.g., at least \(>5\) seconds) as measured in datasets provided by [19], [20] to consume the content, without motion (e.g., head, eye, body, etc) induced viewport adaptation. Thus, it is reasonable and practical to prepare the viewport videos for users, to alleviate or eliminate unexpected rating noise, when wearing the HMD. These viewport videos are selected to represent the salient regions of immersive content and cover a variety of spatial and temporal activities to ensure the model generalization. Each viewport video is sampled at different STAR combinations to collect the corresponding subjective ratings (e.g., mean opinion score - MOS).

Then an analytical model, so called Q-STAR+, is developed as a product of separable exponential functions of STAR variables, i.e., frame size \(s\), frame rate \(t\) and quantization stepsize \(q\). Model parameters are derived from the underlying content within current viewport. This model development is motivated by our previous Q-STAR [21], [22], sharing the similar mathematical forms, but with different model parameters due to the dramatic viewing experience when wearing the HMD compared with the conventional rendering on a flat display screen.

Extensive experiments show a high degree of correlation between the scores predicted using the proposed objective model and those collected by the subjective assessments, in terms of Pearson correlation coefficient (PCC), Spearman correlation coefficient (SRCC), and relative root mean squared error (rRMSE). The performance evaluations are conducted on a different viewport video set than those samples used to derive the model. Our proposed Q-STAR+ provides the state-of-the-art performance on perceptual quality prediction with noticeable margin, in comparison to those famous objective quality metrics [1]–[4], [22].

The contributions of this work are highlighted as follows:

- Through extensive subjective assessments (i.e., \(618^3\) test samples of 14 distinct viewport videos covering four spatial resolutions, three temporal resolutions and four quantization-induced amplitude resolutions), we have demonstrated that proposed Q-STAR+ correlates the perceptual quality accurately, when viewing the viewport videos by wearing a HMD. (see Section IV)

- Model parameters can be predicted via content features extracted at, either viewport frame level or tile level. By carefully examining the structure and material of the HMD, we have also provided reasonable explanations on the parameter differences between the original Q-STAR [22] and Q-STAR+, evidencing the generalization of our proposed model to various videos (i.e., Q-STAR for conventional video, and Q-STAR+ for immersive video). (see Section V)

- We are making our database of viewport videos (and corresponding immersive content) publicly available as a resource for the video quality and virtual reality communities. We envision that these samples (together with the MOSs) will prove to be useful for future subjective and objective quality assessment research studies. (see Section III)

- By combining developed Q-STAR+ model, and rate model in [50], upon the instantaneous sustain network, we could ensure the optimal STAR selection analytically tractable to maximize the perceptual quality, for practical viewport-based immersive video streaming (see Section VI).

The remainder of this work proceeds as follows: Section II performs the literature reviews on the quality assessment and model development of the immersive content, followed by the subjective quality assessment configuration and data post-processing in Section III. Analytical model derivation is presented in Section IV, together with the model parameter prediction and performance comparison. In the meantime, we have presented plausible inferences to justify the model generalization from the original Q-STAR to the Q-STAR+ in this work. A perceptual optimized viewport streaming application is given in Section VI, and finally, Section VII has drawn the conclusion of this work and discussed the future exploration avenues.

\(^2\) Quantization removes the high frequency components of the input signal, resulting in the signal amplitude distortion.

\(^3\) \(384\) PVSs were used for model development, and another \(234\) PVSs were generated for independent validations.
II. RELATED WORK

There are many attempts to study the subjective quality of immersive images and videos by far [15], [23]–[27]. For example, Zhou et. al investigated the spatial resolution impact on the perceptual quality of the immersive images rendered in a head mounted display (HMD) [24], which is then extended to discuss the joint impacts of spatial resolution and quantization [27]. Entire immersive image is used for quality assessment and model development in [24], [27]. In practice, our HVS only perceives the content scene in the front, i.e., current viewport or FoV momentarily. Therefore, the quality model developed for an entire immersive image [24], [27] may not be applicable for the corresponding viewport-based image or video, by including unnecessary information outside of current FoV.

Guo et. al [25] have studied the image quality at current viewport, but with the focus on the peripheral vision impact. This is because the FoV range has been expanded significantly when viewing the content with HMD, including not only the central vision area within single-side 9° eccentrically, but also the peripheral vision areas outside4. However, due to the unequal density distribution of the photoreceptors, our HVS exhibits very distinct quality sensitivities across various vision areas. This work will help to allocate different quality scales (as well as bit rates) even within current FoV for bandwidth reduction, but without subjective quality degradation. In the meantime, Xie et. al [26] have attempted to model the impacts of quality variation when performing the viewport adaption that is often incurred naturally in immersive environment. This work can be devised to maximize the visual quality for the viewport adaptive immersive video streaming under the network bandwidth constraint.

Besides these analytical modeling developments, Xu et. al [15] have presented a subjective visual quality assessment (VQA) method to collect the differential mean opinion score (DMOS) and vectorized DMOS as the perceptual quality index of a panoramic video. In addition, user behaviors have been also explored when viewing the panoramic videos with the HMD, including the salient region (heat map) distribution, viewing direction consistency across different subjects, etc. These observations could generally help to design an appropriate VQA methodology (e.g., test material screening) and subsequent data post-processing (e.g., region-wise weighted evaluation). Similar studies can be also found in [16], [28].

It is still worth to mention that several objective quality metrics have been proposed to reflect the viewing difference in HMD, compared with traditional flat display, including [29], [30] and other alternatives. These metrics implement the distortion weights according to the projection strategy, but the principle behind is still the norm error between corresponding pixels.

Another exploration avenue discusses the perceptual impacts of various STAR combinations [21], [22], [31]–[34], but almost all of them focused on the conventional video rendered on a flat display having very limited FoV, rather the immersive video often rendered in a head mounted display.

III. SUBJECTIVE QUALITY ASSESSMENT

This section details the rating procedure, including test sequence pool selection, subjective opinion score collection of viewport video when viewing immersive content with HMD, and data post-processing for MOS derivation.

A. Test Sequence Pool

We select the immersive test videos by enforcing the following criteria. First, these native immersive videos are used in well-known standards team, e.g., VRU and JVET, for specific standardization purposes (such as immersive video projection or compression tool evaluations). These videos are already screened once by the experts in aforementioned group to ensure the content generalization. Second, because a particular FoV or viewport of an immersive video is rendered on a HMD display, we further would like to ensure that the viewport area covers salient region with sufficient and meaningful spatio-temporal complexity distribution.

4For example, mainstream VR devices, such as HTC Vive or Oculus Rift, offer single-side 50° horizontal range.

Fig. 3. Illustration of spatial and temporal information indices of the viewport videos extracted from corresponding test videos. Videos annotated with same marker belong to the same viewing sequence category (VSC).

Fig. 4. Subjective test platform using the HTC Vive system to deliver an immersive image experience (a) PC, (b) two tracking stations, (c) human subject, (d) two hand controllers, and (e) a headset. This original image is from http://vrplay.com.au/
Nine immersive videos, i.e., “AerialCity”, “BearAttack”, “BroadWay”, “Canolafied”, “Harbor”, “Highway”, “Jamsession”, “PoleVault”, and “KiteFlite”, are illustrated in Fig. 2. The first eight sequences are used for our Q-STAR+ model development while the last one (“KiteFlite”) for the assessment training. The native spatial and temporal resolution of each test sequence is sampled at 4K (3840×1920) and 30 Hz, respectively. For those sequences originally produced at 8K (8192×4096) or even higher resolution, are downscalled to the 4K using the default bicubic filter in FFmpeg (https://www.ffmpeg.org).

These videos are rendered and displayed using the HTC Vive HMD for user interaction inside a panoramic virtual environment. Expect a few seconds in the very beginning, subjects typically stabilize their focus quickly to the salient region [15], [16]. To reduce the possible rating noise induced by unexpected viewport movement, we have extracted the specific viewport video of each immersive content by asking a few subjects to identify the interested area manually. These cropped salient areas are consistent across different users and close to the equator. Similar observations are also presented in [15]. Corresponding viewport windows are highlighted for test videos shown in Fig. 2, associated with wide spatial (SI) and temporal information (TI) indices annotated in Fig. 3 to cover a variety of contents.

Given that HTC Vive HMD display covers 110° horizontal FoV, it actually corresponds to a 1280×960 viewport when rendering the native 4K content. To study the impacts on subjective quality with respect to various STAR combinations, we have prepared the test clips by sampling the viewport video extracted from the original immersive content to four spatial resolutions (i.e., s = 1280×960, 960×720, 640×480, and 320×240), three temporal resolutions t at 30 Hz, 15 Hz and 7.5 Hz, and four amplitude resolutions encoded via H.264/AVC [35] compliant four distinct quantization parameters (QP) (i.e., QP 22, 28, 36, and 44). Note that test clips with resolution less than 1280\times 960 are upscaled and rendered on HMD display, while the playback rate is consistent with the actual frame rate to ensure the same 10 second rating duration. Overall, for each viewport video, we will produce 48 processed video sequence (PVS) for subjective assessment, covering a variety of quality scales.

B. Test Platform and Rating Procedure

We choose to use the HTC Vive system (https://www.vive.com.cn/) as our subjective assessment platform on which the subjects can view the immersive content via the Vive cinema player [36]. Figure 4 exemplifies the configuration of experimental system. The same setup is also used for our previous work [27] for immersive image quality assessment. This system consists of three components, including (a) a personal computer (PC) for high-performance rendering via a dedicated graphics card, (b) a pair of tracking stations to locate and track user interactions of (c) a subject wearing (d) the HMD. The PC was configured with an i7-6700K CPU, an NVIDIA GTX 1070 GPU, 32 GigaBytes of RAM and a 1 TeraByte hard drive. The subject interacts with the PC via wired connections from the HMD.

The HTC Vive HMD, which is a representative mainstream product, has binocular resolution 2160\times 1200\(^5\) with a 110° effective FoV. Rendered viewport videos are streamed from the PC to the HMD through HDMI connections. In reality, users may also use hand controller to interact. However, to reduce the possibility of additional rating noise, we have asked the subjects to either stand or sit steady and rate the viewport video without letting them navigate using either hand controller or head movement.

The subjective assessment protocol that we followed in the experiments was the ACR (Absolute Category Rating) Single Stimulus method. Various parameters of each experiment are illustrated in Fig. 5. Subject participated in two sessions (Fig. 5(c)). The first was a training session, where the subjects viewed samples of viewport videos of diverse, representative quality levels (24 PVSs with 4 frame sizes, 3 frame rates and two QPs\(^6\)). During the training session, the subjects were exposed to video samples of a wide range of quality levels. This allowed the participants to quickly familiarize the rating protocol and to obtain a sense of the range of viewport video qualities. In the second session, 48 PVSs belonging to the same viewport content were randomly placed and shown to each subject on which they gave their opinion scores verbally, while an assistant recorded them during a three seconds interval between consecutive PVS. There is a twenty seconds rest period between two different test videos (Fig. 5(b)). Note that we only allow two test videos (i.e., 96 PVSs), in addition to a training video (i.e., 24 PVSs), for each subject, resulting in overall 26 minutes period approximately to avoid the dizziness and unconformable feeling induced rating noise. The same observation is also reported in [27] that the duration of subjective assessment with VR HMD is better to kept within 30 minutes.

Therefore, eight test videos are categorized into four rating sequences shown in Fig. 3, where the videos belonging to the same class are annotated using the identical marker. Such categorization is enforced with the specific purposes to ensure

---

\(^5\)HTC Vive system uses an AMOLED display screen at 1080×1200 (e.g., width×height).

\(^6\)We use the maximum and minimum QP to cover the entire quality range.
TABLE I

<table>
<thead>
<tr>
<th>Metrics</th>
<th>PCC</th>
<th>SRCC</th>
<th>rRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.58</td>
<td>0.58</td>
<td>22.50%</td>
</tr>
<tr>
<td>MS-SSIM [1]</td>
<td>0.69</td>
<td>0.79</td>
<td>16.10%</td>
</tr>
<tr>
<td>VIFP [2]</td>
<td>0.81</td>
<td>0.82</td>
<td>14.20%</td>
</tr>
<tr>
<td>FSIM [3]</td>
<td>0.60</td>
<td>0.73</td>
<td>17.72%</td>
</tr>
<tr>
<td>VSI [7]</td>
<td>0.62</td>
<td>0.74</td>
<td>17.31%</td>
</tr>
</tbody>
</table>

| Q-STAR+ v1 | 0.71 | 0.72 | 20.38% |
| Q-STAR+ v2 | 0.99 | 0.99 | 4.30% |
| Q-STAR+ v3 | 0.97 | 0.97 | 6.09% |

Q-STAR+ v1: features are extracted at frame level.
Q-STAR+ v2: features are extracted at tile level.

that two videos in the same class exhibit very different content characteristics, and present the ranges of SI and TI as wider as possible.

We have invited 80 people to participate the assessments. The human subjects were naive students from Nanjing University, from widely diverse academic majors, of ages 18 to 30. Of these, 80% had no prior VR experience. All of the viewers were tested and found to have normal visual (after correction) and color perception. The numbers of male and female subjects were 38 and 42, respectively. Male and female participants are evenly distributed for tests. Roughly, each PVS is rated by 20 different subjects.

The rating scores ranged from 1 (Unacceptable) to 10 (Imperceptible), using the Likert marks shown in Fig. 5(a). Other rating ranges, such as 0-5, 0-100, etc, would yield similar outcomes without noticeable statistical difference, as reported in [37].

C. Data Post-processing

The raw ratings are first converted to Z-scores [38], [39] based on the mean and standard deviation of all the scores of each viewer, given by

\[ Z_{mij} = \frac{X_{mij} - \mu(X_i)}{\sigma(X_i)}. \]

Here, \( X_{mij} \) and \( Z_{mij} \) denote the raw rating and Z-score of \( m \)-th PVS at \( j \)-th quality level, from \( i \)-th viewer, respectively. \( X_i \) includes all ratings from \( i \)-th viewer. \( \mu(\cdot) \) and \( \sigma(\cdot) \) represent the operators to calculate the mean and the standard deviation of a given set.

Then we make use of the fact that the fact that a video coded at a lower spatial resolution under the same frame rate and QP would not have a rating higher than a video coded at a higher spatial resolution under the same frame rate and QP. We marked the abnormal value of each participant for each video. If the participant’s abnormal value proportion of the whole MOS data is larger than 15%, we will remove all the data by this participant. We replace the remaining pairs of the abnormal value with the average. Then we repeat the same procedure for the scenarios with various frame rate and QP.

Finally, we scale the mapped Z-scores back to [1 10] by the following form:

\[ X_{mij} = \left( \text{MEDIAN}(X_{mij}^I) - \text{MEDIAN}(X_{mij}^I) \right) / \left( Z_{mij} - Z_{i,\text{min}} \right) + \text{MEDIAN}(X_{mij}^I), \]

where \( \text{MEDIAN}(\cdot) \) represents the median operator. \( X_{mij}^I \) and \( X_{mij}^I \) are the set of maximum and minimum ratings from all \( I \) subjects, respectively. \( Z_{i,\text{max}} \) and \( Z_{i,\text{min}} \) denote the maximum and minimum Z-scores of \( i \)-th subject. With this scaling, the ratings from all participants have a common rage of \( \text{MEDIAN}(X_{mij}^I) \) and \( \text{MEDIAN}(X_{mij}^I) \). In our subjective test data, \( \text{MEDIAN}(X_{mij}^I) = 10 \), and \( \text{MEDIAN}(X_{mij}^I) = 1 \).

After these steps, we can obtain about 15 effective ratings for each videos on average. We average all the 15 ratings to obtain the final MOS through above subjective assessments for analytical model derivation in subsequent sections.

IV. Q-STAR+: A PERCEPTUAL QUALITY MODEL FOR VIEWPORT-BASED IMMERSIVE VIDEO

As discussed previously, we have refined the challenging quality modeling problem of immersive content to study the corresponding salient viewport video. Such viewport video is rendered on a HMD display (shown in Fig. 1(a)). Each viewport perceived by the HVS can be seen as the projection from a spherical view to a 2-D conventional view on a flat screen. Inevitably, all of those existing objective quality metrics, including Q-STAR [21], PSNR, MS-SSIM (Multi-scale SSIM) [1], VIFP (Visual Information Fidelity in Pixel Domain) [2], FSIM (Feature Similarity Index) [3] and VSI (Visual Saliency-Induced Index) [4] are first taken into consideration to see whether they can predict the measured MOS accurately.

Figure 6 presents the prediction accuracy of aforementioned metrics, in terms of the PCC and SRCC respectively detailed in Table I. Meanwhile, we also plotted the model curves via least square error (LSE) fitting for each metric and calculated rRMSE referring to the ITU-T J.149 [40] and ITU-T P.1401 [41].

All of these results revealed that existing metrics can not capture the subjective quality accurately at various STAR combinations for viewport videos viewed by wearing a HMD, despite that these metrics belong to the full-reference (FR) category (except Q-STAR as a reduced-reference (RR) index using extracted content features). An intuitive assumption for such phenomenon is the dramatic change of the viewing environment. Conventional methods are applicable to image/videos displayed on flat screens standing away from the subject about 3×W to 6×W [42]. Here, W is the width of the displayed content. But, viewport videos rendered using the HMD, are presented at much closer distance (i.e., 3-4 centimeters) between the display screen and our eyes. In the meantime, the HMD is mounted with our head tightly to prevent unexpected lighting resources from outside, so as to mimic a truly immersive space. With such short viewing distance, distortion (or perceptual sensation) would be amplified (revolutionized) in anyway. But we believe that the statistical trends of respective perceptual quality versus
frame size \(s\), quality versus frame rate \(t\), and quality versus quantization stepsize \(q\), would be kept. For instance, we would definitely feel the quality improvement when increasing the \(s\), until the saturation moment that frame re-sampling induced visual impairment is not perceivable\(^7\). Such trend can be well captured via an inverted exponential function of \(s\) \([22], [34]\). Similar hypotheses can be applied to explain the impacts of \(t\) and \(q\) on the perceptual quality. Therefore, we expected that the overall quality can be presented as a product of three separable exponential functions of \(q\), \(s\) and \(t\), respectively.

\(^7\)Note that all frames will be upscaled to the same display resolution for subjective assessment.

Fig. 6. Illustration of MOS prediction efficiency among various metrics (a) PSNR, (b) MS-SSIM, (c) VIFP, (d) FSIM, (e) VSI (f) Q-STAR, (g) Q-STAR\(^+\) v1 (Frame-Level), (h) Q-STAR\(^+\) v2 (Tile-Level). MOSs are measured through subjective assessments at various STAR combinations. Analytical model curves are presented via least square error fitting.

following the same assumption when developing the Q-STAR model [22], but parameters differ distantly due to the distinct viewing environment.

In the subsequent sections, we first constructed the Q-STAR\(^+\) model as well as its parameter prediction; we then discussed the possible causes why Q-STAR\(^+\) exhibits very different from the original Q-STAR.

A. Overall Q-STAR\(^+\) for Viewport Video

Given that \(q_{\text{min}} = 8\), \(s_{\text{max}} = 1280 \times 960\) and \(t_{\text{max}} = 30\) Hz, we choose to use normalized variables instead, i.e.,

Fig. 7. Illustration of measured NQQ points under different \(s\) and \(t\). Curves are derived at a fixed \(s\) via least squared error (LSE) fitting.
\[ \hat{q} = \frac{q_{\min}}{q} \in (0, 1], \hat{s} = \frac{s}{s_{\max}} \in (0, 1] \text{ and } \hat{t} = \frac{t}{t_{\max}} \in (0, 1]. \]  
First of all, we assume Q-STAR\(^+\) sharing the same decomposition as Q-STAR, i.e.,

\[ Q(s, t, q) = \frac{\text{MOS}(s, t, q)}{\text{MOS}(s_{\max}, t_{\max}, q_{\min})} = \hat{Q}_{\text{NQQ}}(q; t_{\max}, s_{\max}) \hat{Q}_{\text{NQS}}(s; t_{\max}, q) \hat{Q}_{\text{NQT}}(t; s, q). \]  

Specifically, \( \hat{Q}_{\text{NQQ}}(q; t_{\max}, s_{\max}) \) is the normalized quality versus quantization \( q \) (NQQ) at (fixed) maximum frame size and frame rate, which is also noted as \( \hat{Q}_{\text{NQQ}}(q) \), i.e.,

\[ \hat{Q}_{\text{NQQ}}(q; t_{\max}, s_{\max}) = \hat{Q}_{\text{NQQ}}(q) = \frac{1 - e^{-\alpha_q q^{\beta_q}}}{1 - e^{-\alpha_q}}, \]

with \( \alpha_q \) and \( \beta_q \) depending on the content features \( F_c \) only; \( \hat{Q}_{\text{NQS}}(s; t_{\max}, q) \) is the normalized quality versus spatial resolution \( s \) (NQS) at (fixed) maximum frame rate for any given \( q \), noted as \( \hat{Q}_{\text{NQS}}(s; q) \) as well, i.e.,

\[ \hat{Q}_{\text{NQS}}(s; t_{\max}, q) = \hat{Q}_{\text{NQS}}(s; q) = \frac{1 - e^{-\alpha_s s^{\beta_s}}}{1 - e^{-\alpha_s}}, \]
Generally, $\alpha_s$ and $\beta_s$ correlates with weighted $F_c$ and quantization $q$, and

$$\hat{Q}_{\text{NQT}}(t; s, q) = \frac{1 - e^{-\alpha_t t}}{1 - e^{-\alpha_t}},$$

represents the normalized quality with respect to the temporal resolution $t$ (NQT) at any given $s$ and $q$. Theoretically, $\alpha_t$ and $\beta_t$ are the functions of $F_c$, $s$ and $q$. With these parameters, i.e., $\vec{\alpha} = [\alpha_q, \alpha_s(q), \alpha_t(s, q)]$ and $\vec{\beta} = [\beta_q, \beta_s(q), \beta_t(s, q)]$, the overall Q-STAR+ model becomes very complicated. Towards the practical application, we need to simplify the parameters to reduce the model complexity. $\vec{\beta}$ control the shape of the inverted exponential functions of (4), (5) and (6). We first assume all of them are fixed for all content and various STARs, i.e., $\vec{\beta} = [\beta_q, \beta_s, \beta_t]$. As will be revealed in later experiments, this is a reasonable and effective assumption, significantly simplifying the analytical model derivation.

Parameters in $\vec{\alpha}$ are used to control the decay speed for individual STAR variables. Following the above decomposition, we first exemplify the NQQ, NQS and NQT shown in Figs. 7, 8 and 9. Under the acceptable fitting error, NQT can be captured by a single curve for any $s$ and $q$. Hence, original joint resolution and content dependent $\alpha_t(s, q)$ is reduced to content dependent parameter only, i.e., $\alpha_t(s, q) = \alpha_t$.

We further examine the $\alpha_s(q)$ for NQS, as presented in Fig. 10, where $\alpha_s(q)$ is linearly related to QP for individual content. Therefore, we propose to model the $\alpha_s(q)$ a product of respective content dependent (e.g., $\tilde{\alpha}_s$) and quantization dependent (e.g., $L_q(QP)$) components, via

$$\alpha_s(q) = \tilde{\alpha}_s \cdot L_q(QP),$$

with

$$L_q(QP) = v_1 QP + v_2,$$

where $QP = 4 + \log^2 q$ [35], and $v_1 = -0.1317$, $v_2 = 6.3227$ are fixed for all contents. This is different from the Q-STAR model where a piece-wise linear functions were used for $QP > 28$ and $QP \leq 28$. In practice, HMD display screen presents much higher pixel density per inch, offering the user more accurate response when varying the pixel amplitude (via quantization), compared with the conventional LCD/LED flat panel. For the scenario that the image quality under QP 28 is not differentiated on conventional display, we still could perceive the quantization-induced quality variations as indicated by (8).

Given that we only need to model the NQQ at $s_{\text{max}}$ and $t_{\text{max}}$ following the decomposition in (3), it is expected to have the $\alpha_q$ predicted using weighted features extracted from the underlying content.

Finally, we come to the conclusion that $\alpha_q$, $\tilde{\alpha}_s$, and $\alpha_t$ can be estimated via the extracted features, and $\beta_q$, $\beta_s$, and $\beta_t$ are fixed, for all STAR combinations.

1) Viewport-Frame Level Content Feature Extraction: The following features are extracted from the original viewport videos at frame level, i.e.,

- $\sigma_{\text{DFD}}$ is the Standard deviation of displaced frame difference calculated via block based motion-compensation between successive frames. For the sake of simplicity, we fixed the block size at $16 \times 16$.
- $\mu_{\text{DFD}}$ is the Mean of frame difference calculated between co-located pixel values between consecutive frames.
- $\sigma$ is the content Contrast derived from each frame.
- $G_m$ is the the average of the mean magnitude of the outputs from four Gabor filters [43] in each frame.

We then applied the features $\sigma_{\text{DFD}}$, $\eta(\mu_{\text{DFD}}, \sigma)$, and $G_m$ with $\eta(\mu_{\text{DFD}}, \sigma) = \mu_{\text{DFD}}/\sigma$, to construct the feature prediction set $F_c$. Note that this minimal feature set is determined via the leave-one-out cross validation criterion for content generalization [22], [44].

We followed the same procedures in [22], [34] to derive the parameters via the LSE fitting, i.e.,

$$\vec{\beta} = [\beta_q, \beta_s, \beta_t] = [0.916, 1.345, 0.404],$$

and parameter $\vec{\alpha}$ predictor is given by

$$\vec{\alpha} = H F_c^T,$$

with $\vec{\alpha} = [\alpha_q, \tilde{\alpha}_s, \alpha_t]^T$, $F_c = [1, \sigma_{\text{DFD}}, \eta(\mu_{\text{DFD}}, \sigma), G_m]^T$, and

$$H = \begin{bmatrix} 0.9178 & 0.077 & 7.5913 & 0.1267 \\ 1.4498 & 0.056 & -0.7993 & -0.0219 \\ 3.011 & 0.025 & -2.559 & 0.038 \end{bmatrix},$$

as the weighted transform matrix, yielding the individual model parameters and performance shown in Table II for validation content.
2) Tile-Level Content Feature Derivation: In practice, instantaneous viewport adapts in immersive environment due to the head and/or body movements. It is inevitably required to calculate the parameters for next viewport in advance to determine the actual function forms of our Q-STAR+ model for network constrained optimization. Therefore, we proposed to do tile based content feature extraction. Specifically, each immersive frame is divided into \( n \times m \) non-overlapped tiles. Each viewport or FoV consists of one or more tiles. A generalized scenario is that current viewport across multiple tiles, shown in Fig. 11.

All features discussed above related to the corresponding “mean” and “standard deviation (std)” process for each individual tile. For the viewport covering multiple tiles shown in Fig. 11, we proposed to apply the following functions to do weighted “mean” and “std”, i.e.,

\[
\mu_{\text{FoV}} = \sum_{k=0}^{N-1} \frac{s_k}{s_{\text{FoV}}} \mu_k, \quad (12)
\]

\[
\sigma_{\text{FoV}}^2 = \sum_{k=0}^{N-1} \frac{s_k}{s_{\text{FoV}}} \left( \sigma_k^2 + \frac{(N \cdot \mu_k - \mu_{\text{total}})^2}{N^2} \right). \quad (13)
\]

\( N \) is the total number of tiles involed in current viewport (\( N = 9 \) as exemplified in Fig. 11). \( s_k \) is the area size of each tile covered by the current FoV (shadow area in Fig. 11). \( s_{\text{FoV}} \) represents the area size of current viewport/FoV. Here, it is 1280×960. \( \mu_k \) is the precomputed mean of corresponding tile, and \( \mu_{\text{FoV}} \) is the mean of current viewport. \( \sigma_k \) is the standard deviation of a corresponding tile, \( \mu_{\text{total}} \) is the averaged mean of all associated tiles, and \( \sigma_{\text{FoV}} \) is the final standard deviation, for current FoV.

Corresponding parameters \( \tilde{a} \) and performance using tiled features are also presented in Table II. Note that parameters \( \beta \) keeps the identical for both frame-level and tile-level feature calculations.

B. Model Validation

In this section, we evaluated the performance of the Q-STAR+ model by performing validations using different viewport videos. Towards this goal, we selected another six different videos from the VRU and JVET test sequence pool, as shown in Fig. 12.

We prepared the PVSs following the same methodology discussed in Section III-A. One slight difference was that we removed PVSs coded using QP 28, 36 and 44 when \( s = 320 \times 240 \). This is mainly because of their pretty annoying visual quality, and we thought it would not be of great interests to the consumer in practical applications. Eventually, it produced 234 individual PVSs for cross-validation from all six panoramic videos. Note that we used the same video content (“KiteFlite”) for training to let the subjects familiarize themselves with the rating platform and protocol.

We then recruited another group of subjects who did not participate in the previous subjective assessments. In total, we invited 60 people, with 28 male and 32 female, for raw scores collection. The same data post-processing and screening were performed to derive the final MOSs for each PVSs. In the meantime, content features are extracted from the underlying viewport video to constructed Q-STAR+ model via (9), (10), and (3) for MOS calculation analytically. We visualized the MOS prediction performance of proposed Q-STAR+ model, as well as other metrics, in Fig. 14 and Table III, in terms of the PCC, SRCC and rRMSE. Parameters are detailed in Table II. As we can see, our Q-STAR+ model (either frame-level or tile-level based scheme) still presented the state-of-the-art performance on an independent validation dataset.

We observed few dispersed points with lower MOS than predicted in Q-STAR+ plot, shown in Fig. 14(g) and 14(h). All of them belonged to “SongMenRidge” video. This was mainly due to the fact that “SongMenRidge” presented relatively
lower contrast of content brightness. The same observations were found in [45], [46] as well.

V. Model Generalization

Following the above explorations, we have demonstrated that the same mathematical functional forms (i.e., a product of separable exponential functions of respective q, s and t) are applicable to both Q-STAR [22] and Q-STAR+, but the model parameters exhibit very different behaviors which are mainly due to the dramatic changes of viewing environment.

Popular VR device, such as HTC Vive used in this work, offers a virtualized space via a wearable HMD, shown in Fig. 13(a), providing the stunning immersive experience. This HMD system can be further decomposed into (a) a head tracking subsystem to help the interaction and navigation; (b) an customized AMOLED display; (c) and a Fresnel lens in Fig. 13(b). AMOLED display is chosen because of its high-density pixel per inch (PPI) and ultra-low motion-to-photon latency for vivid content rendering. A Fresnel lens, shown in Fig. 13(c)(d) with concentric circles on the surface, can be much thinner and capture more lights from the lighting display, than a comparable conventional lens in Fig. 13(e).

With such close-up rendering from the HMD display to our eyes, we expect more sensitive impacts when adapting the spatial resolution s. In another words, users could perceive visual quality impairment much easier, when decreasing the s on a HMD screen (cp. Fig. 13(a)(b)), than a conventional flat screen (i.e., Smartphone, TV panel, etc). Mathematically, we would have $\beta_{q+} > \beta_q$ and $\alpha_{s+} < \alpha_s^9$, yielding a much larger decay speed of $\tilde{Q}_{NQS}(s; q)$ with respect to s, as illustrated in Fig. 15(a).

A compact Fresnel lens can improve the luminance of displayed content, with the sacrifice of picture quality. It weakened the visual sensitivities on pixel degradations induced by the quantization. Therefore, we would have $\alpha_{q+} > \alpha_q$ and $\beta_{q+} < \beta_q$, presenting a slower decay speed of $\tilde{Q}_{NQQ}(q)$ with respect to q. Figure 15(b) exemplified the analytical curves using the actual Q-STAR and Q-STAR+ with expected outcomes following above theoretical discussions.

A wearable HMD was tightly mounted with user’s head to build up a truly immersive space for users to dive in. It typically took a while to familiarize themselves in such virtualized reality [16], [27], and gradually lose the connections with the natural reality outside. For conventional video displayed on a flat screen, our HVS and associated subsystems in brain could perceive the information from the “virtual world” (a.k.a., video content rendering) and surrounding reality environment. We could well capture the motions in rendered video, with the intuitive understanding of the coordinates system of the natural reality. But, when users deeply dived into the virtualized reality by wearing the HMD, they lose the coordinate grounds of natural reality (i.e., relative speed), thus impairing the motion sensations subjectively. Therefore, we expected a reduced decay speed of $\tilde{Q}_{NQT}(t; s, q)$ with respect to t, for Q-STAR+ over Q-STAR shown in Fig. 15(c), with $\alpha_{t+} > \alpha_t$ and $\beta_{t+} < \beta_t$.

Table IV and V listed parameters for Q-STAR and Q-STAR+, respectively, evidencing our previous plausible inferences discussion. Note that parameter set $\alpha$ were calculated using the extracted features via (10). It also further support our motivations that the same functional forms are generalized

---

**TABLE II**

The Parameters (predicted) and Performance of Q-STAR+ model on validation content

<table>
<thead>
<tr>
<th>Frame-Level</th>
<th>DrivingInCityCity</th>
<th>Fengjing1</th>
<th>Gaslamp</th>
<th>Hangpin3</th>
<th>SkateboardTrick</th>
<th>SongMenRidge</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha_q$</td>
<td>5.07</td>
<td>4.23</td>
<td>5.15</td>
<td>3.77</td>
<td>6.89</td>
<td>6.07</td>
</tr>
<tr>
<td>$\alpha_s$</td>
<td>3.18</td>
<td>3.67</td>
<td>4.12</td>
<td>3.78</td>
<td>3.01</td>
<td>3.69</td>
</tr>
<tr>
<td>$\alpha_t$</td>
<td>3.19</td>
<td>2.94</td>
<td>2.67</td>
<td>3.11</td>
<td>2.11</td>
<td>2.42</td>
</tr>
<tr>
<td>PSNR</td>
<td>0.960</td>
<td>0.962</td>
<td>0.985</td>
<td>0.975</td>
<td>0.933</td>
<td>0.981</td>
</tr>
<tr>
<td>SRC</td>
<td>0.963</td>
<td>0.965</td>
<td>0.985</td>
<td>0.987</td>
<td>0.950</td>
<td>0.991</td>
</tr>
<tr>
<td>rRMSE</td>
<td>6.9%</td>
<td>9.2%</td>
<td>4.3%</td>
<td>9.4%</td>
<td>10.2%</td>
<td>8.0%</td>
</tr>
</tbody>
</table>

**TABLE III**

Performance Comparisons of various objective metrics as well as our proposed Q-STAR+. (Model parameters are content predicted.)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>DrivingInCityCity</th>
<th>Fengjing1</th>
<th>Gaslamp</th>
<th>Hangpin3</th>
<th>SkateboardTrick</th>
<th>SongMenRidge</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSNR</td>
<td>0.55</td>
<td>0.58</td>
<td>21.50%</td>
<td>17.50%</td>
<td>15.69%</td>
<td>18.40%</td>
</tr>
<tr>
<td>MS-SSIM [1]</td>
<td>0.69</td>
<td>0.71</td>
<td>17.50%</td>
<td>17.50%</td>
<td>15.69%</td>
<td>18.40%</td>
</tr>
<tr>
<td>VIFP [2]</td>
<td>0.76</td>
<td>0.76</td>
<td>15.69%</td>
<td>15.69%</td>
<td>15.69%</td>
<td>15.69%</td>
</tr>
<tr>
<td>FSIM [3]</td>
<td>0.67</td>
<td>0.71</td>
<td>18.40%</td>
<td>18.40%</td>
<td>18.40%</td>
<td>18.40%</td>
</tr>
<tr>
<td>VSI [4]</td>
<td>0.70</td>
<td>0.70</td>
<td>18.01%</td>
<td>18.01%</td>
<td>18.01%</td>
<td>18.01%</td>
</tr>
<tr>
<td>RR Q-STAR [22]</td>
<td>0.80</td>
<td>0.81</td>
<td>17.77%</td>
<td>17.77%</td>
<td>17.77%</td>
<td>17.77%</td>
</tr>
<tr>
<td>Q-STAR+ v1</td>
<td>0.95</td>
<td>0.95</td>
<td>8.31%</td>
<td>8.31%</td>
<td>8.31%</td>
<td>8.31%</td>
</tr>
<tr>
<td>Q-STAR+ v2</td>
<td>0.95</td>
<td>0.95</td>
<td>8.33%</td>
<td>8.33%</td>
<td>8.33%</td>
<td>8.33%</td>
</tr>
</tbody>
</table>

**TABLE IV**

Parameter $\beta$ comparison of Q-STAR and Q-STAR+.

<table>
<thead>
<tr>
<th></th>
<th>$\beta_s$</th>
<th>$\beta_q$</th>
<th>$\beta_t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q-STAR</td>
<td>0.74</td>
<td>1</td>
<td>0.63</td>
</tr>
<tr>
<td>Q-STAR+</td>
<td>1.345</td>
<td>0.916</td>
<td>0.404</td>
</tr>
</tbody>
</table>

---

9HTC Vive HMD display presents 447 PPI.
to both Q-STAR and Q-STAR+. Due to the different viewing environment, transform matrix for parameter prediction (e.g., (10)) differs, but the methodology keeps the same.

VI. PERCEPTUAL OPTIMIZED VIEWPORT VIDEO STREAMING APPLICATION

Viewport-dependent immersive video streaming is preferred in practice [14] for constrained network. Herein, Fanyi et al. [47]–[49] proposed a two-tier 360° video streaming framework with prioritized buffer control which can effectively accommodate the dynamics in both bandwidth and view direction. In this streaming framework, a base-tier (BT) chunk encoded the entire 360° view span at a low bitrate to provide basic quality. While the enhancement-tier (ET) chunks could provide enhanced quality within a certain viewport, the position and bitrate of ET chunks were determined by the network bandwidth and viewport prediction. But it still lacks of an analytical models to ensure the the optimal perceptual quality.

Therefore, we proposed to use the Q-STAR+ as the measurement of the perceptual quality, together with the R-STAR [50], to optimize the two-tier panoramic video streaming [47]–[49], maximizing the subjective quality under a given network constraint.

A. Analytical R-STAR Model

The rate model proposed in [50] relates the rate with \((q, s, t)\) by

\[
R(q, s, t) = R_{\text{max}} \left( \frac{q}{q_{\text{min}}} \right)^{-a} \left( \frac{t}{t_{\text{max}}} \right)^{b} \left( \frac{s}{s_{\text{max}}} \right)^{c},
\]

where \(q_{\text{min}}, s_{\text{max}}\) and \(s_{\text{max}}\) should be chosen based on the underlying application, \(R_{\text{max}}\) is the rate when coding a video at \(q_{\text{min}}, s_{\text{max}}\) and \(t_{\text{max}}\), and \(a, b\) and \(c\) are the model parameters, characterizing how faster the rate decreased when \(s\) and \(t\) reduce and \(q\) increases. For the sake of simplicity, we measured the actual bit rates corresponding to different STARs used in our subjective quality assessment (i.e., \(s \in \{1280\times960, 960\times720, 640\times480, 320\times240\}\), \(t \in \{7.5\text{Hz}, 15\text{Hz}, 30\text{Hz}\}\), \(q \in \{8, 16, 28, 44\}\)), then obtained the corresponding \(a, b\) and \(c\) for each video by least error fitting, with the PCC and rRMSE also shown in Table VI. We can see that the model (14) is very accurate for compressed rate prediction.

B. Rate constrained quality and STAR optimization

Note that the rate constraint and FoV or viewport prediction can be obtained from the two-tier 360° video streaming framework [47]–[49]. Here we focus on the optimization of the STAR for ET chunks using analytical Q-STAR+ and R-STAR models, the essence of this problem is to maximize the video quality in user’s current FoV under the bit rate constraint \(R_0\),

<table>
<thead>
<tr>
<th>Seq.</th>
<th>(\alpha_{q+})</th>
<th>(\alpha_a)</th>
<th>(\alpha_{q+})</th>
<th>(\alpha_b)</th>
<th>(\alpha_{q+})</th>
<th>(\alpha_c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DrivingInCity</td>
<td>3.18</td>
<td>5.20</td>
<td>5.07</td>
<td>2.34</td>
<td>3.19</td>
<td>2.23</td>
</tr>
<tr>
<td>Fengjing1</td>
<td>3.68</td>
<td>6.19</td>
<td>4.23</td>
<td>3.04</td>
<td>2.94</td>
<td>1.92</td>
</tr>
<tr>
<td>Gaslamp</td>
<td>4.13</td>
<td>6.01</td>
<td>5.16</td>
<td>3.57</td>
<td>2.67</td>
<td>1.70</td>
</tr>
<tr>
<td>Hangpai3</td>
<td>3.78</td>
<td>7.53</td>
<td>3.77</td>
<td>3.05</td>
<td>3.12</td>
<td>1.15</td>
</tr>
<tr>
<td>SkateboardTrick</td>
<td>3.01</td>
<td>7.01</td>
<td>6.89</td>
<td>2.73</td>
<td>2.11</td>
<td>1.05</td>
</tr>
<tr>
<td>SongMenRidge</td>
<td>3.69</td>
<td>7.61</td>
<td>6.07</td>
<td>3.26</td>
<td>2.47</td>
<td>1.48</td>
</tr>
</tbody>
</table>

Table V: PARAMETER \(\alpha\) COMPARISON OF Q-STAR AND Q-STAR+

Fig. 15. Illustration of NQS, NQT, NQQ models for Q-STAR+ and Q-STAR (parameters of “AerialCity” were used for generating curves. Other video shared the same trends.)
i.e.,

\[
\begin{align*}
\text{Determine } q, s, t \text{ to maximize } Q(q, s, t) \\
\text{subject to } R(q, s, t) &\leq R_0. 
\end{align*}
\]

1) Optimal solution assuming continuous \( s, q, t \): We first assumed \( s, q \) and \( t \) can take on any value in a continuous range, \( s \in (0, s_{\text{max}}] \), \( t \in (0, t_{\text{max}}] \) and \( q \in [q_{\text{min}}, \infty) \), with \( q_{\text{min}} = 8 \), \( t_{\text{max}} = 30Hz \) and \( s_{\text{max}} = 1280 \times 960 \) (as a viewport resolution for the recent mainstream HMD display). With the models in Eq. (3), (14), we could solve (15) either analytically via a Lagrangian optimization method [34], or empirically via brute-force iterations over the limited spaces of \( (s, t, q) \). Results are shown in Fig. 16. We can see that the \( s \) tends to be fixed at its maximal \( s_{\text{max}} \). It is mainly because that \( s \) imposes the most significant impact on the quality model, as shown in Fig. 15. This also supports our observations that \( s \) is the most sensitive factor to the visual quality of viewport dependent immersive content when wearing the HMD.

2) Optimal solution assuming dyadic \( s \) and \( t \): In practical video encoder, \( t \) and \( s \) only take on limited discrete values, i.e., \( t \in \{7.5 \, Hz, 15 \, Hz, 30 \, Hz\} \), and \( s \in \{320 \times 240, 640 \times 480, 1280 \times 960\} \). We further assume \( q \in [8, 104] \), taking continuous values. For each given rate, we search through all possible \( (s, t, q) \) from the feasible sets, and select the one that leads to the highest quality. The results are shown in Fig. 17. We can see that the the \( s \) also tends to be fixed as same as the continuous value case in Fig. 16. As illustrated in aforementioned examples, our models could be facilitated to guide the perceptual optimized viewport video streaming under any given network bandwidth supply, when having the knowledge of viewport orientations from the existing methods [47]–[49].

**VII. CONCLUSION**

We have explored the impacts of spatial \( s \), temporal \( t \), amplitude \( q \) resolutions on the perceptual quality of a compressed viewport video viewed by wearing a HMD, through well-designed subjective quality assessments and data post-processing, following the standard protocols [39]–[41]. Experimental results have demonstrated that a closed form function of three separable exponential functions of respective \( q \), \( s \) and \( t \), so-called Q-STAR+, provides the state-of-the-art subjective quality prediction performance, in terms of the PCC, SRCC, and rRMSE, in comparison to those famous quality metrics. This Q-STAR+ model is self-adaptive by implementing the parameter prediction using the features extracted from the underlying content.

![Fig. 16. Optimal Q, s, t & q vs. R by assuming continuous s and t](image)

Our Q-STAR+ model could be of great interests to immersive video encoding and streaming scenarios. It made bandwidth-constrained quality optimization analytically tractable, when delivering the viewport-based immersive video.

This Q-STAR+ was motivated by our previous Q-STAR [22] model, but with the focus on the immersive video rendering on a viewport display screen using the popular HMD. We presented that the Q-STAR and Q-STAR+ shared the same analytical forms, and the differences of parameters were mainly because of the dramatic changes of viewing environment (a.k.a., tight HMD for immersive content vs. relative distant flat panel for conventional video).

There are still many research avenues to explore. For instance, we could construct an unified subjective quality framework by combining the peripheral quality models in [25], motion-induced quality adaptation models in [52], for more general applications of immersive video. Together with our

**TABLE VI**

<table>
<thead>
<tr>
<th>Seq.</th>
<th>( a )</th>
<th>( b )</th>
<th>( c )</th>
<th>( r\text{RMSE} )</th>
<th>( \text{PCC} )</th>
<th>( R_{\text{max}} \text{(kbps)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>AerialCity</td>
<td>2.11</td>
<td>0.68</td>
<td>1.05</td>
<td>4.42%</td>
<td>0.981</td>
<td>7939</td>
</tr>
<tr>
<td>BearAttack</td>
<td>2.10</td>
<td>0.78</td>
<td>0.78</td>
<td>1.60%</td>
<td>0.997</td>
<td>7362</td>
</tr>
<tr>
<td>Broadway</td>
<td>1.12</td>
<td>0.59</td>
<td>0.44</td>
<td>1.41%</td>
<td>0.998</td>
<td>10127</td>
</tr>
<tr>
<td>Canolafeld</td>
<td>1.08</td>
<td>0.54</td>
<td>0.57</td>
<td>0.92%</td>
<td>0.999</td>
<td>1875</td>
</tr>
<tr>
<td>Harbor</td>
<td>1.87</td>
<td>0.71</td>
<td>0.74</td>
<td>1.34%</td>
<td>0.998</td>
<td>8523</td>
</tr>
<tr>
<td>Highway</td>
<td>0.94</td>
<td>0.48</td>
<td>0.30</td>
<td>1.49%</td>
<td>0.998</td>
<td>1318</td>
</tr>
<tr>
<td>Jansession</td>
<td>1.72</td>
<td>0.51</td>
<td>1.4</td>
<td>1.76%</td>
<td>0.997</td>
<td>8011</td>
</tr>
<tr>
<td>PoleVault</td>
<td>1.11</td>
<td>0.45</td>
<td>1.4</td>
<td>2.05%</td>
<td>0.996</td>
<td>10050</td>
</tr>
</tbody>
</table>
parallel camera capturing system [7], we could realize the real-time gigapixel video interaction over the Internet. By far, we had only considered the video components of an immersive media. It will be very interesting to study the joint audio-visual quality, attention model, as well as model guided optimization strategy for immersive media (e.g., Cloud VR Gaming [53]).

VIII. ACKNOWLEDGMENT

The authors would like to thank all anonymous reviewers for their constructive comments to improve this manuscript.

REFERENCES


