A Bayesian Quality-of-Experience Model for Adaptive Streaming Videos

ZHENGFANG DUANMU and WENTAO LIU, University of Waterloo, Canada
DIQI CHEN, Chinese Academy of Sciences, China
ZHUORAN LI and ZHOU WANG, University of Waterloo, Canada
YIZHOU WANG and WEN GAO, Peking University, China

The fundamental conflict between the enormous space of adaptive streaming videos and the limited capacity for subjective experiment casts significant challenges to objective Quality-of-Experience (QoE) prediction. Existing objective QoE models either employ pre-defined parametrization or exhibit complex functional form, achieving limited generalization capability in diverse streaming environments. In this study, we propose an objective QoE model, namely, the Bayesian streaming quality index (BSQI), to integrate prior knowledge on the human visual system and human annotated data in a principled way. By analyzing the subjective characteristics towards streaming videos from a corpus of subjective studies, we show that a family of QoE functions lies in a convex set. Using a variant of projected gradient descent, we optimize the objective QoE model over a database of training videos. The proposed BSQI demonstrates strong prediction accuracy in a broad range of streaming conditions, evident by state-of-the-art performance on four publicly available benchmark datasets and a novel analysis-by-synthesis visual experiment.

CCS Concepts: • **Information systems** → **Multimedia streaming**; • **Human-centered computing** → **User models**; • **Mathematics of computing** → **Approximation;**

Additional Key Words and Phrases: Quality-of-experience assessment, adaptive video streaming, quadratic programming

ACM Reference format:
Zhengfang Duanmu, Wentao Liu, Diqi Chen, Zhuoran Li, Zhou Wang, Yizhou Wang, and Wen Gao. 2022. A Bayesian Quality-of-Experience Model for Adaptive Streaming Videos. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 3s, Article 141 (December 2022), 24 pages.
https://doi.org/10.1145/3491432

1 INTRODUCTION

Video traffic in various content distribution networks is expected to occupy 71% of all consumed bandwidth by 2021 and exceed 82% by 2022 [10]. The explosion of data volume introduced by video streaming will quickly drain available network bandwidth in the next decade. Concurrent

This work is supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada under the Discovery Grant, Canada Research Chair program, and Alexander Graham Bell Canada Graduate Scholarship program.

Authors' addresses: Z. Duanmu, W. Liu, Z. Li, and Z. Wang, University of Waterloo, Canada; D. Chen, Chinese Academy of Sciences, China; Y. Wang and W. Gao, Peking University, China.

1551-6857/2022/12-ART141 $15.00
https://doi.org/10.1145/3491432

with the scarcity of network resources is the steady rise in user demands on video quality. With the emergence of new technologies such as 4K, high dynamic range, wide color Gamut, and high frame rate, viewers' expectation on video quality has been higher than ever. The diversity of streaming environments and complexity of the human **Quality-of-Experience (QoE)** response have posed significant challenges to optimal content distribution services.

Adaptive bitrate (ABR) algorithms are the primary tools for modern Internet **over-the-top (OTT)** video streaming services. In dynamic adaptive streaming environment, ABR achieves player-driven bitrate adaptation by providing video streams in a variety of bitrate and quality levels and breaking them into small HTTP file segments. Throughout the streaming process, the video player at the client adaptively switches among the available streams by selecting segments based on playback rate, buffer condition, and instantaneous throughput, primarily to optimize viewers' QoE [6, 34, 37, 49, 69, 84].

With many ABR algorithms at hand, it becomes pivotal to measure their performance to guide the network resource allocation. The most straightforward and reliable way to measure viewers' QoE is to conduct a user study. However, subjective testing is expensive, inconvenient, and time-consuming. Most importantly, it cannot be integrated into an ABR system to perform real-time bitrate selection. Therefore, the development of an accurate objective QoE model lies in the root of ABR systems. In general, the QoE modeling problem is very challenging due to the fundamental conflict between the enormous size of streaming video space and the limited number of videos available for observation. To overcome the curse of dimensionality problem, all existing QoE models share a common two-stage structure: (1) quality-related features are extracted from a test video; (2) a regression model, namely, the QoE function, is applied to map quality relevant features to a scalar QoE score.

Over the past decade, a significant effort has been devoted into the development of objective QoE models. Albeit the diversity of QoE model implementations, recent studies have gradually converged to a three-dimensional chunk-level feature representation including visual quality, rebuffering duration, and quality adaptation [6, 21, 47, 59, 84]. The major difference among QoE prediction schemes lies in the instantiation of the second stage. Depending on the underlying assumptions about subjective quality integration mechanism, existing QoE models can be roughly categorized into two classes. The first approach makes strong *a priori* assumptions about the mapping between a set of quality-related features and the subjective QoE rating. Specifically, most QoE models [6, 21, 31, 46, 66, 69, 82–84] in this category employ a pre-defined QoE function without the access to the training data. A common drawback of the approach is that the model configuration is often selected on the basis of mathematical convenience rather than as a reflection of any prior beliefs. The second type of model [1, 19, 24, 59, 68] takes a data-driven approach by leveraging sophisticated machine learning models such as random forest [9] and neural network [29], effectively imposing a non-informative prior to the model parameters. However, even with feature extraction, the dimensionality of latent space grows linearly with the number of segments, suggesting that the quality representation of each video still lies in a very high dimensional space. In contrast, the largest streaming video dataset in the literature only contains a thousand subject-rate datapoints, which can be deemed extremely sparsely distributed in the latent space. As a result, these purely data-driven models usually suffer from significant overfitting problem. The tradeoff between model underfitting and overfitting is known as the bias-variance tradeoff in the machine learning literature [13].

In general, the key to the problem is to adaptively fuse appropriate prior knowledge about subjective QoE and the likelihood of observing the training data [7]. This line of thought naturally gives rise to a novel Bayesian framework for robust QoE prediction. We show that each of the existing categories of QoE models corresponds to a special case in the Bayesian framework, which inevitably results in sub-optimal performance. Furthermore, the proposed framework also provides us a principled way to integrate prior knowledge of the **human visual system (HVS)** and a limited number of training samples.

Even given such a unified framework, the QoE modeling problem is still non-trivial. In particular, traditional prior distributions rely on a number of strong assumptions and generalizations, strictly restricting the space of feasible solution. As a result, existing prior models cannot make efficient use of the training data. However, simply removing these assumptions would degenerate the maximum *a posteriori* approach to the maximum likelihood estimator, which suffers from the overfitting problem with limited training samples. Therefore, a meaningful prior probability model for the HVS configuration is of central importance for this application. Following this direction, we perform a comprehensive analysis to the HVS properties based on a plethora of subjective QoE studies, from which we derive a system of linear inequalities. We further show that a family of objective QoE models lies within a convex set that results from a positive cone in a functional space. This gives us both guidance on the form of our model as well as constraints.

Building upon the Bayesian framework and the well-characterized prior knowledge of HVS properties, we derive a new QoE model named **Bayesian Streaming Quality Index (BSQI)**. We demonstrate that the model parameter estimation problem can be formulated as a quadratic programming problem. Using a variant of projected gradient descent, we optimize the proposed model over a database of training samples with limited adaptation patterns. The resulting model is computationally efficient, mathematically well-behaved, and perceptually grounded.

To demonstrate the effectiveness of the new design principle, we compare BSQI to 14 objective QoE models on four benchmark datasets covering a broad set of video contents, encoder configurations, network conditions, ABR algorithms, and viewing devices. BSQI rivals or outperforms the best existing scheme in all considered scenarios, with an average improvement of 4% in prediction accuracy. We show that the proposed model is superior on average and in extreme cases via a set of intuitive examples. We have made the implementation of all objective QoE models available at https://github.com/zduanmu/ksqi to facilitate future objective QoE research.

In summary, this article makes the following key contributions:

- A Bayesian framework that unifies a wide spectrum of objective QoE models.
- Mathematical analysis on the space of QoE functions for adaptive streaming videos;
- Design of a Bayesian objective QoE model combining the constraints from our analysis and human annotated data in a principled way;
- An open-source implementation and comprehensive evaluation of objective QoE models.

2 A BAYESIAN REVIEW OF OBJECTIVE QOE MODELS

The goal of objective QoE models is to predict the subjective quality rating y given a streaming video $\mathbf{x}$. The QoE prediction problem can be formulated as a Bayesian inference problem, where the objective is to determine the probability distribution with a parametric model $p(y|\mathbf{x};\boldsymbol{\theta})$, which may be followed by a decision making process that generates a deterministic estimate of y. Directly establishing a quantitative relationship between a streaming video $\mathbf{x}$ and its QoE y is challenging due to the enormous space of $\mathbf{x}$ and small capacity for subjective measurements. To alleviate the problem, all existing objective models employ a common two-stage QoE prediction framework, which sequentially transforms an input video to a low-dimensional latent representation and builds a connection between the latent variable and subjective QoE rating. From a Bayesian perspective, this approach corresponds to a hierarchical probability graphic model of the form

$$p(y|\mathbf{x};\boldsymbol{\theta}) = \int p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_1)p(y|\mathbf{z};\boldsymbol{\theta}_2)\mathrm{d}\mathbf{z}, \tag{1}$$

where $\mathbf{z}$, $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_1)$, and $p(y|\mathbf{z};\boldsymbol{\theta}_2)$ are the low-dimensional latent variable, a feature extractor, and a regression model, respectively. In this Bayesian network, the objective QoE model parameters $\boldsymbol{\theta}$ consist of both the parameters of the feature extractor and the regression model (i.e., $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2\}$). In general, the marginalization over the latent variable $\mathbf{z}$ in Equation (1) is not tractable to compute exactly. As a result, most methods consider an approximation that makes use of a point estimate instead of performing the integration over $\mathbf{z}$ in Equation (1). We will separately review feature extractors and regression models that are commonly used in the existing objective QoE models, with an emphasis on their underlying assumptions.

2.1 Feature Extractors

The feature extractor $p(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_1)$, which qualitatively determines which piece of information in a streaming video is relevant to the QoE, plays a central role in the objective QoE model. In the past decade, a wide variety of feature extraction schemes have been proposed, which can be roughly categorized into two classes.

Assuming there exists a causal relationship between impairments in the communication pipeline and the QoE, earliest feature extraction schemes attempted to identify a set of objective performance measures that correlate well with the subjective quality evaluation. A unique property of this approach is that the extracted features $\mathbf{z}$ do not depend on the visual signal $\mathbf{x}$ or its pristine counterpart, but are functions of the distortion process. As a result of the decoupling between the visual content and viewers' QoE, the distortion process-based feature extractors are often referred to as the **Quality-of-Service (QoS)**-based approach. There have been a wide variety of QoS-based methods ranging from generic network-level features such as bit error rate, packet loss rate, network jitter, round-trip time, and average bandwidth [39, 50, 57, 61, 75] to application-specific features such as QP, encoding bitrate, rebuffering duration, and bitrate variation [46, 83, 84]. Although this approach has achieved promising results in individual reports, it often struggles to deliver competitive performance in a more comprehensive benchmark including more diverse source contents, video encoders, and ABR algorithms [20]. The fundamental limitation of QoS approach resides in the conditional independence between the service performance and the QoE score y given the visual signal $\mathbf{x}$. In particular, a naïve subject can consistently assess the quality of a streaming video without access to the underlying transmission channel.

Motivated by the limitations of the QoS-based approach, the second type of feature extraction methods tackles the problem from a different perspective by simulating the properties of the HVS. These artificial visual models typically implement well-known HVS functionalities such as contrast sensitivity function, luminance masking, contrast masking, and saliency detection [17, 76]. The final output of the computational models encodes certain quantities that are hopefully measured by the HVS. Typical examples of the approach include References [1, 3, 6, 21], which are built upon SSIMplus [64] or VMAF [42] as the chunk-level feature extractor. There perceptual feature extractors have demonstrated a much better performance on benchmarks covering diverse content, encoders, and viewing conditions [16].

Regardless of the underlying assumptions, recent QoE models [3, 6, 21, 24, 33, 83, 84] have agreed that three most relevant features in the QoE of streaming videos are the presentation video quality, rebuffering duration, and quality adaptation. These features are usually computed at chunk-level because (1) the video statistics are usually highly non-stationary, and (2) streaming video distortions are also temporally varying.

2.2 Regression Models

Even with the reduced dimensionality, the design of objective QoE models is still a challenging task, partly because the sequential nature of the streaming video. There have been two distinct approaches to tackle the problem, both of which can be derived from the Bayesian perspective. Given a dataset of observations $\mathcal{D}_z$ comprising N_x latent variables $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_{N_\mathrm{x}})$ and their corresponding target quality scores $\mathbf{y} = (y_1, \ldots, y_{N_\mathrm{x}})$, the objective of the regression model is to obtain a set of parameters $\boldsymbol{\theta}_2$ that optimizes the posterior parameter distribution $p(\boldsymbol{\theta}_2|\mathcal{D}_z)$. The maximum *a posteriori* solution of the QoE function can be expressed as

$$\begin{aligned}\boldsymbol{\theta}_2^* &= \arg\max_{\boldsymbol{\theta}_2} p(\boldsymbol{\theta}_2|\mathcal{D}_z) \\ &= \arg\max_{\boldsymbol{\theta}_2} p(\mathcal{D}_z|\boldsymbol{\theta}_2)p(\boldsymbol{\theta}_2).\end{aligned} \tag{2}$$

Existing QoE models can be categorized based on their assumptions about the prior parameter distribution $p(\boldsymbol{\theta}_2)$ as follows:

- Strong Prior: Given the limited training data in the latent space, the first approach mainly relies on strong prior assumptions about $p(\boldsymbol{\theta}_2)$ to estimate the posterior distribution. To simplify the problem, four basic assumptions are commonly made. The first is that the notion of QoE can be defined locally, and that the overall QoE can be obtained by a linear combination of the chunk-level QoE scores. Typically, one makes a Markov assumption that the chunk-level quality distribution, when conditioned on its previous segment, is independent of the segments beyond the neighborhood. The second is an assumption of temporal homogeneity: The chunk-level QoE distribution is the same across all temporal positions. The two assumptions jointly suggest that

$$p(y|\mathbf{z}; \boldsymbol{\theta}_2) = \mathcal{N}\left(y|\frac{1}{T}\sum_{t=1}^{T} y^t, \beta\right), \tag{3}$$

where $y^t = g(\mathbf{z}^t; \boldsymbol{\theta}_2)$[1] denotes the chunk-level QoE. It should be noted that the mapping between the local latent variables $\mathbf{z}^t$ and the local QoE y^t shares a common functional form across all temporal indices. The third is an independent assumption that the impact of each dimension in $\mathbf{z}^t$ is independent from other dimensions in predicting the local QoE scores. This assumption can be mathematically expressed as $g(\mathbf{z}^t; \boldsymbol{\theta}_2) = \sum_{j=1}^{J} g_j(z^{t,j}; \boldsymbol{\theta}_2^j)$, where $z^{t,j}$ and $g_j(\cdot; \boldsymbol{\theta}_2^j)$ denote the jth dimension in $\mathbf{z}^t$ and the dimensional-specific activation function, respectively. In addition to the three assumptions, most objective QoE models in this category make assumptions about the specific form of $g(\mathbf{z}^t; \boldsymbol{\theta}_2)$ along each dimension. Initial attempts incorporated certain functions with pre-defined parameters. Popular choices of the activation operator include linear function [6, 46, 82, 84], exponential function [21, 31, 66], and logarithmic function [69, 83]. In the case of linear function, the chunk-level QoE can be computed by $y^t = \boldsymbol{\theta}^\top \mathbf{z}^t$, where $\top$ denotes the transpose operator. Since the parameters are fixed, we have $p(\boldsymbol{\theta}_2 = \boldsymbol{\theta}_2^*) = 1$ and $p(\boldsymbol{\theta}_2 = \boldsymbol{\theta}') = 0$ for any function $\boldsymbol{\theta}' \neq \boldsymbol{\theta}_2^*$. Consequently, the posterior distribution $p(\boldsymbol{\theta}_2|\mathcal{D}_z)$ converges to the prior distribution $p(\boldsymbol{\theta}_2)$ for any likelihood function and dataset as long as $p(\mathcal{D}_z|\boldsymbol{\theta}_2^*) > 0$. Recent studies have indicated that these overly simplistic models with manually tuned parameters have achieved limited success in representing the relationship between the latent variables and the subjective QoE [3, 20]. One common drawback of the approach is that the prior distribution is often selected on the basis of mathematical convenience rather than as a reflection of any prior beliefs. The

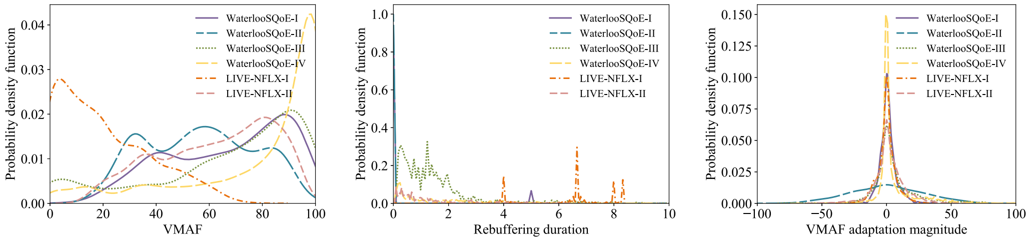
[1] According to the Markov assumption, the chunk-level quality y^t is a function of $\mathbf{z}^{t-1}$ and $\mathbf{z}^t$. However, by the technique of feature enrich, we can denote the feature set $\mathbf{z}^t$ at each time instance t as the aggregation of the previous chunk-level feature and the present chunk-level feature without loss of generality.

resulting strong inductive bias may manifest itself in many ways. For example, the subjective QoE response with respect to each feature can vary significantly from exponential and logarithmic functions. Generally speaking, the problem applies to all handcrafted QoE measures that rely on a pre-defined functional form. Furthermore, the additive assumption is also problematic for QoE modeling, where the impact of one latent variable is hardly independent to the other. In particular, recent experiments have illustrated that the joint impact of conventional feature pairs on the QoE is statistically significant [3, 18, 21]. The assumption becomes increasingly deficient as the dimensionality of the latent space expands.

- Non-informative Prior: Supposing HVS is too complex to understand, the second approach aims to approximate the posterior distribution from the likelihood function $p(\mathcal{D}_z|\boldsymbol{\theta}_2)$. With the emergence of subject-rated QoE databases [3, 4, 16, 18, 20, 21, 24, 26], the data-driven approach has dominated the objective QoE research. A broad range of statistical models such as non-linear auto-regressive model [3], neural network [68], support vector machine [1], random forest [19, 59], and **Long-Short Term Memory (LSTM)** [24] have been utilized to map streaming video features to subjective opinion scores. These models employ a maximum likelihood estimator

$$\boldsymbol{\theta}_2^* = \arg\max_{\boldsymbol{\theta}_2} p(\mathcal{D}_z|\boldsymbol{\theta}_2) \tag{4}$$

to obtain the optimal model parameters, effectively assuming a non-informative prior in the Bayesian inference problem [7]. Although these QoE models can fit arbitrarily complex continuous functions [30], they often suffer from the generalization problem. Specifically, it has been observed that the performance of QoE models trained on one database reduces significantly on other benchmark datasets, largely due to the distribution mismatch in the visual content and the distortion process across datasets [3, 4, 16, 20, 26]. There are at least four sources for the generalization problem. First, in spite of the reduced dimensionality, the latent variable $\mathbf{z}$ still lives in a high dimensional space. Each streaming video is represented by a $Z \times T$-dimensional vector when chunk-level feature extractors are employed, where Z and T represent the number of chunk-level features and the total number of chunks, respectively. However, a typical "large-scale" subjective test allows for a maximum of several hundred or a few thousand test videos to be rated. Given the enormous space of latent variables, a few thousand subject-rated samples are deemed to be extremely sparsely distributed in the space. Second, the learning-based models assume that the training samples and testing samples come from the same distribution. However, the assumption has never been justified in the existing studies and may hardly hold in practice. A motivating example is shown in Figure 1, where the probability density functions of video presentation quality measured by a state-of-the-art video quality assessment model VMAF [42], rebuffering duration and quality adaptation magnitude in six publicly available streaming QoE datasets are presented. Clearly, there is significant variability on the characteristics of streaming videos across different datasets, suggesting that an objective QoE model optimized on a simple dataset such as WaterlooSQoE-I [21] may yield very poor predictions on complex datasets as WaterlooSQoE-III [20], WaterlooSQoE-IV [14], and LIVE-NFLX-II [4], and vice versa. The streaming video probability density estimation is further complicated by the concept drift problem [25], where the characteristics of streaming video changes over time. For example, the drift in streaming video distribution may arise from the advancement of video acquisition [35, 38, 55], compression [11, 54, 71], transmission [6, 34, 37, 49, 69, 84], and reproduction systems [43, 53, 79], and the steady rise in viewers' expectation on video quality [19, 58]. Third, the maximum likelihood estimator generally assumes that each $(\mathbf{z}, y)$ pair in the training set $\mathcal{D}_z$ is independent and identically distributed. In practice, however, the existing QoE



(a) Distribution of VMAF (b) Distribution of rebuffering duration (c) Distribution of adaptation magnitude

Fig. 1. There exists significant variance on the characteristics of streaming videos, evident by the distributions of (a) VMAF, (b) rebuffering duration, and (c) adaptation magnitude in six publicly available datasets.

datasets typically generate multiple streaming videos for each reference video to cover the diversity of distortion processes, suggesting that the training data are not independent and identically distributed. Fourth, the consistency of subjective QoE ratings among streaming video databases is only moderate due to drastically different experimental conditions. Strictly speaking, the quality ratings of a streaming video $\mathbf{x}_t$ collected from a subjective experiment are essentially samples from a context conditional quality distribution $p(y|\mathbf{x}_t, \mathbf{t})$, where $\mathbf{t}$ encodes the information about experiment environment, instruction, training process, presentation order, and experiment protocol. As a result, the subjective quality ratings obtained from different experiments cannot be simply aggregated into a larger QoE dataset $p(y|\mathbf{x}_t)$. These data challenges constantly arise in QoE research and will remain a challenging issue in the future.

One common drawback of both approaches is the lack of perceptually meaningful prior distributions. In particular, none of these models make use of the knowledge about natural videos, distortion processes, and the HVS, despite the plethora of dedicated subjective experiments over the past decade. It remains to be seen how much improvement can be achieved with these informative priors in the Bayesian framework.

3 PRIOR QOE MODEL

In this section, we derive a prior QoE model by analyzing a corpus of subjective QoE experiments. To simplify the discussion, we start with a deterministic formulation of the prior QoE model. In the end of the derivation, we will also present a probabilistic interpretation of the resulting prior model.

3.1 Deterministic View

Formally, the overall QoE can be denoted as $Q(\{p_t, \tau_t, \Delta p_t\}_{t=1}^{T})$, where p_t, τ_t, and $\Delta p_t = p_t - p_{t-1}$ represent the presentation quality, the rebuffering duration, the magnitude of quality adaptation of chunk t, respectively. T denotes the number of chunks in the streaming video. Defining the space of QoE functions helps us build a model of these functions. It not only guides us as to the form such a model should take, but also determines the constraints these functions must satisfy. We begin by summarizing observations from a collection of existing subjective QoE studies and then formulate the domain knowledge to define the space of these functions. For the brevity of math formulation, we will use simplified notations for the rest of this section unless otherwise stated. Specifically, we will omit all the identical variables of the QoE function Q in the same equation and only emphasize the factors that are different. First, various subjective tests [12, 32] have attested that rebuffering duration is negatively correlated with the overall QoE of streaming videos. Formally, we may summarize this observation by

$$Q(\tau_t = \tau^1) \geq Q(\tau_t = \tau^2), \forall \tau^1 \leq \tau^2, t. \tag{5}$$

Note that we have used the simplified notation in Equation (5) to show that the two compared video streams are only different in the rebuffering duration of chunk t.

The second assumption is that, given the same rebuffering length, the QoE drop tends to be greater when the presentation quality of the previous chunk is higher, i.e.,

$$\begin{aligned} &Q(p_{t-1} = p^1, \tau_t = 0) - Q(p_{t-1} = p^1, \tau_t = \tau) \leq \\ &Q(p_{t-1} = p^2, \tau_t = 0) - Q(p_{t-1} = p^2, \tau_t = \tau), \forall \tau, p^1 \leq p^2, t. \end{aligned} \tag{6}$$

Such a trend has been observed in recent subjective tests [5, 21] and may be explained by the expectation confirmation theory [58].

The third assumption is elicited from the fact that, given a constant presentation quality and a fixed total duration of rebuffering, the overall QoE degrades as the number of rebuffering occurrences increases [31, 52, 60]. Mathematically, this may be expressed as

$$Q(\tau_{t-1} = \tau^1, \tau_t = \tau^2) \leq Q(\tau_{t-1} = 0, \tau_t = \tau^1 + \tau^2), \forall \tau^1, \tau^2, t. \tag{7}$$

The fourth remark is that, given the same rebuffering duration, videos with higher presentation quality consistently deliver higher overall QoE, despite the greater penalty for the rebuffering event [48]. This statement can be formulated as

$$Q(p_t = p^1) \leq Q(p_t = p^2), \forall p^1 \leq p^2, t. \tag{8}$$

We then analyze the functional properties with respect to the quality adaptation. The fifth assumption suggests that people always assign a penalty to presentation quality degradation, reward to quality elevation, and neither penalty nor reward when no quality adaptation occurs [18, 28, 52, 63]. Mathematically, the assumption can be expressed as

$$\begin{cases} Q(\Delta p_t = \delta p^1) \leq Q(\Delta p_t = 0), & \forall \delta p^1 \leq 0, t \\ Q(\Delta p_t = \delta p^2) \geq Q(\Delta p_t = 0), & \forall \delta p^2 \geq 0, t \end{cases}. \tag{9}$$

Further analysis [18, 52, 56, 63] on the relationship between the QoE adjustment and the intensity of quality adaptation Δp indicates that subjects tend to give greater QoE penalty or reward when quality drops or improves by a greater amount. This finding, together with the fifth assumption, prompts our sixth assumption: QoE is monotonically increasing with regards to Δp:

$$Q(\Delta p_t = \delta p^1) \leq Q(\Delta p_t = \delta p^2), \forall \delta p^1 \leq \delta p^2, t. \tag{10}$$

Experiments in Reference [18] find that quality degradation occurring in the high-quality range leads to greater amount of penalty than that occurring in the low-quality range, while quality elevation in the high-quality range results in smaller rewards. Such an observation leads to the seventh assumption that

$$\begin{aligned} &Q(p_t = p^1, \Delta p_t = \delta p) - Q(p_t = p^1, \Delta p_t = 0) \geq \\ &Q(p_t = p^2, \Delta p_t = \delta p) - Q(p_t = p^2, \Delta p_t = 0), \forall \delta p, p^1 \leq p^2, t. \end{aligned} \tag{11}$$

Another commonly observed trend in QoE is that the reward for a positive quality adaptation is relatively smaller than the penalty for a negative one, given the same intensity of quality adaptation and the same average presentation quality [18, 56, 63]. Formally, this can be summarized by

$$\begin{aligned} &Q(p_t = p^1, \Delta p_t = 0) - Q(p_t = p^1, \Delta p_t = -\delta p) \geq \\ &Q(p_t = p^1 - \delta p, \Delta p_t = \delta p) - Q(p_t = p^1 - \delta p, \Delta p_t = 0), \forall p^1, \delta p \geq 0, t. \end{aligned} \tag{12}$$

In summary, we define the space of QoE functions Q as

$$\mathcal{W}_Q := \{Q : \mathbb{R}^{3T} \to \mathbb{R} | Q \text{ satisfying constraints (5) to (12)}\}. \tag{13}$$

The inequality constraints in Equation (13) represent a positive cone [7], which is convex by its definition.

3.2 Probabilistic View

The conversion from the inequality constraints in Equation (13) to its probability representation is straightforward. Let $\boldsymbol{\theta}_2$ denote the parameters of the regression function Q, then the constraint in Equation (5) corresponds to the following prior distribution:

$$p_1(\boldsymbol{\theta}_2) = \begin{cases} \epsilon, & \forall Q(\{p_t, \tau_t, \Delta p_t\}_{t=1}^T; \boldsymbol{\theta}_2) \text{ satisfying (5)} \\ 0, & \text{otherwise} \end{cases}, \tag{14}$$

where ϵ represents certain probability density for each feasible parameter configuration such that $p_1(\boldsymbol{\theta}_2)$ sum to 1. The constraints in Equations (6)–(12) can be transformed into prior probability distributions of $\boldsymbol{\theta}_2$ in a similar fashion, which can be denoted as $p_2(\boldsymbol{\theta}_2)$–$p_8(\boldsymbol{\theta}_2)$, respectively. The simple aggregation of constraints in Equation (13) implicitly assumes the independence of individual assumptions. Therefore, the joint prior probability distribution of QoE models may be obtained by

$$p(\boldsymbol{\theta}_2) = \frac{\prod_{i=1}^8 p_i(\boldsymbol{\theta}_2)}{\int_{\boldsymbol{\theta}} \prod_{i=1}^8 p_i(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}}. \tag{15}$$

4 A BAYESIAN QOE MODEL

Our discussion on the prior QoE models has been encouraging. However, the general form of the QoE function still exhibits a very high dimensionality. To obtain a meaningful approximation, some further assumptions have to be made. In this section, we present the roadmap to design a perceptually grounded objective QoE model.

4.1 Additional Assumptions

The observations from existing psychophysical experiments not only illustrate the feasible functional form of QoE models, but also point out the joint impact among the three-dimensional features in QoE. As a result, we can effectively replace the specific form assumption and the feature-wise independent assumption in the traditional prior model by the HVS imposed constraints in Equation (13). However, existing subjective QoE studies do not provide enough information in the temporal aspects. For example, how an impairment that appears early in a streaming session affects the QoE of subsequent video segments in a long run is still a subject of ongoing research. There have also been limited studies [66] investigating the validity of the temporal homogeneity assumption. In this study, we adopt a conservative approach by inheriting the Markov assumption and the temporal homogeneity assumption. Nevertheless, the proposed Bayesian framework is general enough to incorporate more prior knowledge once they become available.

Mathematically, the Markov assumption and the temporal homogeneity assumption can be jointly expressed by

$$Q\left(\{p_t, \tau_t, \Delta p_t\}_{t=1}^T\right) = \frac{1}{T} \sum_{t=1}^{T} q(p_t, \tau_t, \Delta p_t),$$

where q is the chunk-level QoE function, which is invariant to t. By incorporating these assumptions, we reduce the original problem to the estimation of a three-dimensional function $q(p_t, \tau_t, \Delta p_t)$.

We further assume that the influence of presentation quality, rebuffering, and quality adaptation are additive. Formally, the QoE of chunk t is determined by

$$q(p_t, \tau_t, \Delta p_t) = p_t + S_t + A_t,$$

where S_t and A_t denote the rebuffering QoE function and the adaptation QoE function of chunk t, respectively. We adopt the additive assumption because of its good mathematical property, interpretability, low complexity, and broad acceptance [6, 21, 31, 46, 51, 69, 83, 84]. With the aid of these additional assumptions, we present the roadmap to design a perceptually grounded objective QoE model. For simplicity, we will drop the subscript t in the rest of this section unless otherwise specified. Note that we do not assume the influences of presentation quality, rebuffering, and quality adaptation are independent to each other if the QoE functions S and A vary with respect to the presentation quality p.

4.2 Modeling the Presentation Quality

Traditionally, for the sake of operational convenience, bitrate is often used as the major indicator of video presentation quality [34, 44, 49, 51, 69, 84]. However, bitrate may heavily deviate from perceptual quality. The presentation-quality model should provide meaningful and consistent QoE predictions across video contents, video resolutions, and viewing conditions/devices. To the best of our knowledge, currently the only video QoE models that satisfy such requirements are SSIMplus [64] and VMAF [42]. Both models perform consistently well on various subject-rated video databases [2, 45], making them an appropriate component in BSQI. In the rest of the article, we present our results using VMAF as our presentation quality model, as it is open source and thus facilitates reproducible research. Although the presentation-quality scores are not available to the adaptive streaming player by default, they can either be embedded into the manifest file that describes the specifications of the video or carried in the metadata of the video container. Thanks to the light overhead, the feature embedding technique has been successfully deployed in practical QoE measurement [21, 78] and ABR optimization systems [6, 81].

4.3 Modeling the Rebuffering QoE Function

In general, the rebuffering QoE function S is a function of presentation quality, rebuffering, and quality adaptation. However, we have not found any evidence from the literature that the quality adaptation may change the perception of rebuffering. As a result, the supporting domain of S reduces to $\{(p,\tau)|p \in [0,P], \tau \in [0,\tau_{\max}]\}$, where P indicates the best quality, and $\tau_{\max}$ is the maximum rebuffering duration. By incorporating the inequality constraints in Equation (13), it is easy to derive the theoretic space of the rebuffering QoE function S

$$\begin{aligned}\mathcal{W}_S := \{&S : \mathbb{R}^2 \rightarrow \mathbb{R} | S(p,0) = 0, S(p,\tau^1) \geq S(p,\tau^2),\\ &S(p^1,\tau) \geq S(p^2,\tau), S(p,\tau^1) + S(p,\tau^2) \leq S(p,\tau^1+\tau^2),\\ &S(p^1,\tau) + p^1 \leq S(p^2,\tau) + p^2, \forall p,\tau,\tau^1 \leq \tau^2, p^1 \leq p^2\},\end{aligned} \tag{16}$$

where we have assumed that no rebuffering corresponds to a penalty of 0.

Strictly speaking, $\mathcal{W}_S$ is a space of continuous functions, but we may approximate it in terms of a vector space by densely sampling the supporting domain of S. By uniformly sampling both p and τ, we approximate the function S with a finite-size matrix $\mathbf{S} \in \mathbb{R}^{(N+1)\times(N+1)}$, where an element $s_{i,j}$ denotes the QoE penalty when $(p,\tau) = (\frac{i-1}{N}P, \frac{j-1}{N}\tau_{\max})$. We then vectorize $\mathbf{S}$ as $\mathbf{s} \in \mathbb{R}^{(N+1)^2}$ for the convenience of further formulation. We employ the uniform vectorization for two reasons. First, the exact form of QoE functions (e.g., exponential, logarithmic) cannot be known *a priori*. In this regard, the uniform sampling implicitly serves as a non-informative prior on the form of QoE functions. Our second motivation is closely related to the smoothness assumption, which will be detailed in subsequent discussion. In particular, when the QoE functions are band-limited, they can be fully recovered from these samples when the sampling density is larger than the Nyquist rate. Finally, we are able to approximate the functional space $\mathcal{W}_S$ with a vector space

$$\mathcal{W}_\mathrm{s} := \{\mathbf{s} \in \mathbb{R}^{(N+1)^2} | \mathbf{G}^s\mathbf{s} \leq \mathbf{h}^s, \mathbf{B}^s\mathbf{s} = \mathbf{c}^s\},$$

where $\mathbf{G}^s, \mathbf{h}^s, \mathbf{B}^s$, and $\mathbf{c}^s$ are constructed so all the entries in $\mathbf{s}$ should satisfy the constraints in Equation (16).

Even though the theoretical space of the rebuffering QoE function is restricted to a positive cone, it is still infinite-dimensional. Ideally, the optimal rebuffering QoE function should be the one that best explains the subjective data and lives in the theoretical space. Specifically, given a training set of M_s video sequences, each of which has C_s chunks, one or more rebuffering events, no adaptation, and a **mean opinion score (MOS)** Q_m to indicate its overall QoE, we want to obtain a vector $\mathbf{s}^* \in \mathcal{W}_\mathrm{s}$ that minimizes the mean squared error between the model prediction and subject-rated data

$$\epsilon_\mathrm{s}^\mathrm{F} := \frac{1}{M_s}\sum_{m=1}^{M_s}\left[Q_m - \frac{1}{C_s}\sum_{c=1}^{C_s}(P_{m_c} + s_{i_{m_c}, j_{m_c}})\right]^2,$$

where P_{m_c} and $s_{i_{m_c}, j_{m_c}}$ denote the presentation quality and rebuffering QoE penalty at chunk c of video m, respectively. However, existing subject-rated streaming video datasets contain very limited samples, which are sparsely distributed in the feature space. In particular, some (p, τ) combinations never appear in the training set, suggesting the optimization problem is ill-conditioned. To obtain a meaningful solution, we impose smoothness prior on the function S. In practice, many subjective experiments have empirically shown the smoothness of the QoE functions [3, 21]. Mathematically, smoothness regularization can be represented as the second-order differences along i and j axes

$$\epsilon_\mathrm{s}^\mathrm{S} := \frac{1}{(N+1)^2}\sum_{i=1}^{N+1}\sum_{j=1}^{N+1}\left[\left(\frac{\partial^2 s_{i,j}}{\partial i^2}\right)^2 + \left(\frac{\partial^2 s_{i,j}}{\partial j^2}\right)^2\right].$$

It is not hard to see that both $\epsilon_\mathrm{s}^\mathrm{F}$ and $\epsilon_\mathrm{s}^\mathrm{S}$ take quadratic forms of $\mathbf{s}$. As a result, we are able to estimate the rebuffering QoE matrix $\mathbf{S}$ by solving the following quadratic programming problem:

$$\begin{aligned} \underset{\mathbf{s}}{\text{minimize}} \quad & L_\mathrm{s} = \epsilon_\mathrm{s}^\mathrm{F} + \lambda\epsilon_\mathrm{s}^\mathrm{S} \\ \text{subject to} \quad & \mathbf{s} \in \mathcal{W}_\mathrm{s}, \end{aligned} \tag{17}$$

where $\lambda > 0$ is a weighting factor. The convexity of $\mathcal{W}_\mathrm{s}$ and the objective function implies that there exists a unique solution for the optimization problem. The problem can be efficiently solved with projected gradient descent-based algorithms such as alternating direction method of multipliers [8].

4.4 Modeling the Adaptation QoE Function

Similarly, the supporting domain of A can be reduced to $\{(p, \Delta p) | p \in [0, P], \Delta p \in [-p, P - p]\}$ because of the limited understanding of the connections between the perception of quality adaptation and rebuffering duration. By incorporating the inequality constraints in Equation (13), we can show that the adaptation QoE function A lies in the space

$$\begin{aligned} \mathcal{W}_A := \quad & \{A : \mathbb{R}^2 \to \mathbb{R} | A(p, 0) = 0, A(p, \delta p^1) \leq A(p, \delta p^2), \\ & A(p, -\delta p) + A(p - \delta p, \delta p) \leq 0, \\ & A(p^1, \delta p) \geq A(p^2, \delta p), \forall p, \delta p, p^1 \leq p^2, \delta p^1 \leq \delta p^2\}, \end{aligned} \tag{18}$$

where we have assumed that no adaptation corresponds to a penalty of 0.

Following the same approach, we work with the discrete version of of A. By uniformly sampling both p and Δp, we approximate the function A with a finite-size matrix $\mathbf{A} \in \mathbb{R}^{(N+1)\times(N+1)}$, where an entry $a_{i,j}$ denotes the QoE change when $(p, \Delta p) = (\frac{i-1}{N}P, \frac{j-i}{N}P)$, and then vectorize $\mathbf{A}$ as $\mathbf{a} \in \mathbb{R}^{(N+1)^2}$. Finally, the vector space of adaptation experience function becomes

$$\mathcal{W}_{\mathrm{a}} := \left\{\mathbf{a} \in \mathbb{R}^{(N+1)^2} \middle| \mathbf{G}^a \mathbf{a} \leq \mathbf{h}^a, \mathbf{B}^a \mathbf{a} = \mathbf{c}^a\right\},$$

where $\mathbf{G}^a, \mathbf{h}^a, \mathbf{B}^a$, and $\mathbf{c}^a$ are according to the constraints in Equation (18).

Given a training set of M_a video sequences, each of which has C_a chunks, no rebuffering events, and a MOS Q_m, we aim to optimize

$$L_{\mathrm{a}} := \epsilon_{\mathrm{a}}^{\mathrm{F}} + \lambda \epsilon_{\mathrm{a}}^{\mathrm{S}},$$

where

$$\epsilon_{\mathrm{a}}^{\mathrm{F}} := \frac{1}{M_a} \sum_{m=1}^{M_a} \left[Q_m - \frac{1}{C_a} \sum_{c=1}^{C_a} (P_{m_c} + a_{i_{mc}, j_{mc}}) \right]^2, \tag{19}$$

and

$$\epsilon_{\mathrm{a}}^{\mathrm{S}} := \frac{1}{(N+1)^2} \sum_{i=1}^{N+1} \sum_{j=1}^{N+1} \left[\left(\frac{\partial^2 a_{i,j}}{\partial i^2} \right)^2 + \left(\frac{\partial^2 a_{i,j}}{\partial j^2} \right)^2 \right]. \tag{20}$$

Here, $s_{i_{mc}, j_{mc}}$ denotes the quality adaptation experience at chunk c of video m. The optimal quality adaptation experience matrix A can be obtained by solving the following quadratic programming problem:

$$\begin{aligned} \underset{\mathbf{a}}{\text{minimize}} \quad & L_{\mathrm{a}} = \epsilon_{\mathrm{a}}^{\mathrm{F}} + \lambda \epsilon_{\mathrm{a}}^{\mathrm{S}} \\ \text{subject to} \quad & \mathbf{a} \in \mathcal{W}_{\mathrm{a}}. \end{aligned} \tag{21}$$

Minimizing the loss functions in Equations (17) and (21) is equivalent to solving the maximum *a posteriori* problem (2), with a Gaussian likelihood function and a prior probability distribution given by the product among a Gaussian distribution over Equation (17), a Gaussian distribution over Equation (20), and a uniform distribution in Equation (15).

4.5 Overall QoE

The optimal solutions of Equations (11) and (12) $\mathbf{s}$ and $\mathbf{a}$ correspond to the vectorized S_t and A_t, respectively. Once we solve the optimization problem on a training set, the segment level QoE takes the form $Q_t = p_t + S_t + A_t$, as introduced at the beginning of Section 3. In practice, one usually requires a single end-of-process QoE measure. We use the mean value of the predicted QoE over the whole playback duration to evaluate the overall QoE. To reduce the memory usage, the end-of-process QoE can be computed in a moving average fashion

$$Y_t = \frac{(t-1)Y_{t-1} + Q_t}{t},$$

where Y_t is the cumulative QoE up to the tth segment in the streaming session.

5 EXPERIMENTS

In this section, we first describe the experimental setups and evaluation criteria. We then compare BSQI with classic and state-of-the-art objective QoE models. Furthermore, we develop an efficient methodology for examining the best-case performance of objective QoE models. Finally, we conduct a series of ablation experiments to identify the contributions of the core factors in BSQI.

Table 1. Comparison of Objective QoE Models

QoE model	Features	Markov	Temporal homogeneity	Independence	Functional form	Training method
Mok2011 [51]	τ	✓	✓	✓	linear	–
FTW [31]	τ	✓	✓	✓	exponential	–
Liu2012 [46]	r, τ	✓	✓	✓	linear	–
Xue2014 [83]	QP, τ	✓	✓	✓	logarithmic	ML
Yin2015 [84]	r, τ, ΔR	✓	✓	✓	linear	–
Spiteri2016 [69]	r, τ	✓	✓	✓	logarithmic	–
Bentaleb2016 [6]	p, τ	✓	✓	✓	linear	–
SQI [21]	p, τ, Δp	✓	✓	✗	exponential	–
P.1203 [59]	r, s, τ, Δr, QP	✗	✗	✗	random forest	ML
VideoATLAS [1]	p, τ, Δp	✗	✗	✗	SVR	ML
NARX-QoE [3]	p, τ, Δp	✗	✗	✗	NARX	ML
TV-QoE [27]	p, τ	✗	✗	✗	HM	ML
LSTM-QoE [24]	p, τ, Δp	✗	✗	✗	LSTM	ML
Bi-LSTM [22]	p, τ, Δp	✗	✗	✗	Bi-LSTM	ML
BSQI	p, τ, Δp	✓	✓	✗	piecewise linear	MAP

Notations: r, bitrate; τ, rebuffering duration; Δr, bitrate variation; p, presentation quality measured by state-of-the-art video quality assessment methods; Δp, quality variation; s, spatial resolution. Abbreviations: QP, quantization parameter; SVR, support vector machine; NARX, nonlinear auto-regressive model; HM, Hammerstein-Wiener model; ML, maximum likelihood; MAP, maximum *a posteriori*.

5.1 Experimental Setup

The experiment assumes the availability of type of viewing device, detailed rebuffering statistics (the duration and the start time of each rebuffering event), detailed adaptation statistics (the start time of each adaptation event), and segment-level video information including bitrate, framerate, spatial resolution, average QP, and segment-level VMAF. The scope is restricted by the available information provided by the benchmark datasets, which will be detailed in subsequent sections. The objective QoE models are evaluated under a comparable setting, where no competing model has the access to the finer (frame) level information.

5.1.1 Objective QoE Models. We evaluate the performance of 15 objective QoE models for adaptive streaming videos. The competing algorithms are chosen to cover a diversity of design philosophies, including eight classic parametric QoE models: FTW [31], Mok2011 [51], Liu2012 [46], Xue2014 [83], Yin2015 [84], Spiteri2016 [69], Bentaleb2016 [6], and SQI [21], six state-of-the-art learning-based QoE models: VideoATLAS [1], NARX-QoE [3], TV-QoE [27], P.1203 [59], LSTM-QoE [24], Bi-LSTM [22], and the proposed BSQI. A description of the existing QoE models is shown in Table 1. The implementation for VideoATLAS is obtained from the original authors, and we implement the other 13 QoE models. We use mode 0 of P.1203 according to the available information in the benchmark datasets. We have made the implementation of the models publicly available at https://github.com/zduanmu/ksqi. For the purpose of fairness, the parameters of all models are optimized on the WaterlooSQoE-I [21] and the WaterlooSQoE-II [18] datasets, except for P.1203 [59] whose training methodology is not specified in the original paper and NARX-QoE [3], TV-QoE [27], LSTM-QoE [24], and Bi-LSTM [22]. Since the WaterlooSQoE-I [21] and the WaterlooSQoE-II [18] datasets do not provide continuous time QoE ratings, we optimize NARX-QoE [3], TV-QoE [27], LSTM-QoE [24], and Bi-LSTM [22] on the LIVE-NFLX-I dataset. The WaterlooSQoE-I dataset contains 60 compressed videos without rebuffering, 60 compressed videos with initial buffering, and 60 compressed videos with rebuffering. The WaterlooSQoE-II dataset involves 588 video clips with variations in compression level, spatial resolution, and frame-rate. For the models with hyper-parameters, we randomly split the datasets into 80% training and 20% validation sets, and the hyper-parameters with the lowest validation loss are chosen. For BSQI, we set the maximum rebuffering duration τ_{max} to 10, while the penalty of a rebuffering event longer than 10 can be easily obtained by extrapolating the rebuffering QoE function **S**. We set the quantization bin number $N = 10$ for both discretized rebuffering and adaptation QoE functions. The maximum presentation quality value $p = 100$ is inherited from state-of-the-art VQA measures SSIMplus and VMAF. Although we can learn an initial buffering experience matrix independent from **S**, it introduces unnecessary model complexity. Instead, we discount the impact of initial buffering with $\frac{1}{9}$ and set the expectation to the initial quality p_{-1} to 80 following the recommendation by Reference [21]. We apply OSQP [70] to solve the quadratic programming problems in Equations (17) and (21). The fidelity-flatness tradeoff parameter $\alpha = 1$ is optimized on the validation set. In the subsequent section, we will also show that BSQI performs consistently over a broad range of α.

5.1.2 Benchmark Databases. We compare BSQI with state-of-the-art objective QoE models on four subject-rated adaptive streaming video datasets, including LIVE-NFLX-I [5], LIVE-NFLX-II [4], WaterlooSQoE-III [20], and WaterlooSQoE-IV [14]. The LIVE-NFLX-I dataset consists of 112 streaming videos derived from 14 source content with 8 handcrafted playout patterns. The LIVE-NFLX-II dataset consists of 420 streaming videos generated from content-adaptive encoding profiles, bitrate adaptation algorithms, and network conditions. The WaterlooSQoE-III dataset contains 450 streaming videos of 20 source contents recorded from a set of streaming experiments. The WaterlooSQoE-IV dataset contains 1, 350 highly realistic streaming videos constructed from 5 video contents, 2 video encoders, 9 real-world network traces, 5 ABR algorithms, and 3 viewing devices. The streaming videos in different datasets are of diverse characteristics, since they are generated from different source videos, encoding profiles, adaptive streaming algorithms, and network conditions. We do not evaluate Xue2014 [83] on the LIVE-NFLX-I dataset, because their **quantization parameters (QP)** and encoded representations of the streaming videos are not publicly available. We also do not evaluate NARX-QoE [3], TV-QoE [27], LSTM-QoE [24], and Bi-LSTM [22] on the LIVE-NFLX-I dataset, as it serves as the training set for these models.

5.1.3 Evaluation Criteria. Three criteria are employed for performance evaluation by comparing MOS and objective QoE scores according to the recommendation by the video quality experts group [74]. We adopt **Pearson linear correlation coefficient (PLCC)** to evaluate the prediction accuracy, **Spearman ranking-order correlation coefficient (SRCC)**, and **Kendell rank correlation coefficient (KRCC)** to assess prediction monotonicity. A better objective QoE model should have higher PLCC, SRCC, and KRCC.

5.2 Performance Comparison

Tables 2, 3, and 4 show the PLCC, SRCC, and KRCC on the benchmark datasets, respectively, where top-two best performers are highlighted with bold-face. We have several observations. First, the objective QoE models that employ advanced VQA models as the presentation quality measure generally perform favorably against the conventional bitrate-based QoE models. In particular, Bentaleb2016 significantly outperforms Yin2015, where the only difference between them is the presentation quality measure. Second, although the learning-based QoE models perform competitively on certain test sets, they fail miserably on the other benchmark datasets. Specifically, the SRCC performance degradation of P.1203 and VideoATLAS from one dataset (LIVE-NFLX-II) to another (LIVE-NFLX-I) can be as large as 0.406 (from 0.821 to 0.415) and 0.597 (from 0.673 to 0.076), suggesting that the learning-based models exhibit low generalizability to diverse streaming environments. By contrast, BSQI achieves state-of-the-art performance on all benchmark datasets, thanks to the constraints given by domain knowledge. Third, the classic QoE models with a fixed parametric form cannot faithfully capture the subjective QoE response on streaming videos with complex distortion patterns, evident by the low prediction accuracy on WaterlooSQoE-III. In spite of the authors' effort in designing functional forms to conform known HVS properties [21, 31, 83], the

Table 2. PLCC between the Objective QoE Model Prediction and MOS on the Benchmark Datasets

QoE model	LIVE-NFLX-I	LIVE-NFLX-II	WaterlooSQoE-III	WaterlooSQoE-IV	Average	Weighted Average
Mok2011 [51]	0.292	0.512	0.173	0.046	0.256	0.166
FTW [31]	0.286	0.568	0.323	0.147	0.331	0.263
NARX-QoE [3]	–	0.532	0.323	0.194	0.350	0.284
Xue2014 [83]	–	0.788	0.387	0.166	0.447	0.328
LSTM-QoE [24]	–	0.734	0.456	0.301	0.497	0.414
Bi-LSTM [22]	–	0.702	0.582	0.279	0.521	0.420
Liu2012 [46]	0.524	0.732	0.609	0.282	0.537	0.438
Yin2015 [84]	0.376	0.673	0.722	0.323	0.524	0.466
TV-QoE [27]	–	0.685	0.438	0.422	0.515	0.475
VideoATLAS [1]	0.100	0.644	0.385	0.675	0.451	0.586
P.1203 [59]	0.325	0.817	0.769	0.636	0.637	0.679
Bentaleb2016 [6]	0.741	0.898	0.625	0.682	0.737	0.713
Spiteri2016 [69]	0.612	0.731	**0.809**	0.685	0.709	0.714
SQI [21]	**0.756**	**0.910**	0.673	**0.717**	**0.764**	**0.745**
BSQI	**0.753**	**0.905**	**0.794**	**0.720**	**0.793**	**0.769**

Table 3. SRCC between the Objective QoE Model Prediction and MOS on the Benchmark Datasets

QoE model	LIVE-NFLX-I	LIVE-NFLX-II	WaterlooSQoE-III	WaterlooSQoE-IV	Average	Weighted Average
Mok2011 [51]	0.335	0.516	0.152	0.056	0.265	0.171
FTW [31]	0.325	0.549	0.184	0.082	0.285	0.197
NARX-QoE [3]	–	0.433	0.315	0.132	0.293	0.226
Xue2014 [83]	–	0.778	0.388	0.219	0.462	0.360
Bi-LSTM [22]	–	0.685	0.593	0.255	0.511	0.405
LSTM-QoE [24]	–	0.710	0.488	0.299	0.499	0.415
TV-QoE [27]	–	0.635	0.422	0.395	0.484	0.446
Liu2012 [46]	0.438	0.732	0.598	0.468	0.559	0.539
Yin2015 [84]	0.441	0.686	0.741	0.541	0.602	0.601
VideoATLAS [1]	0.076	0.673	0.469	0.670	0.472	0.603
Spiteri2016 [69]	0.493	0.711	**0.798**	0.662	0.662	0.680
P.1203 [59]	0.415	0.821	**0.797**	0.668	0.675	0.708
Bentaleb2016 [6]	**0.650**	0.883	0.718	**0.692**	0.735	0.730
SQI [21]	0.644	**0.906**	0.690	0.690	**0.735**	**0.732**
BSQI	**0.655**	**0.893**	0.776	**0.699**	**0.756**	**0.747**

QoE functions can vary significantly from exponential and logarithmic functions. However, BSQI does not assume a particular form of QoE functions and instead maximizes the mathematically well-behaveness. In summary, we believe the performance improvement arises because (1) BSQI is equipped with an HVS-inspired VQA measure that generalizes well on a variety of video contents, encoders, and viewing devices; (2) the training procedure optimizes the quality prediction accuracy regularized by the prior knowledge on HVS; and (3) the proposed model does not make inaccurate *a priori* assumptions on the form of QoE functions.

5.3 Best-case Validation

Objective QoE model is not only used to evaluate, but also to optimize a variety of ABR algorithms and systems. A good rule of thumb is that an optimized system is only as good as the optimization criterion used to design it [77]. Conversely, the performance of an objective QoE model can be assessed via synthesizing optimal streaming videos with respect to that objective QoE model followed by visual inspection of the generated stimulus [76, 80]. Specifically, given a set of encoded and segmented videos and a realistic network trace, we can generate an optimal streaming video in terms of each objective QoE model. Subjective evaluation of the synthesized stimuli provides a best-case validation of the underlining objective QoE models. A good objective QoE model should produce perceptually better streaming videos comparing to the other schemes.

Table 4. KRCC between the Objective QoE Model Prediction and MOS on the Benchmark Datasets

QoE model	LIVE-NFLX-I	LIVE-NFLX-II	WaterlooSQoE-III	WaterlooSQoE-IV	Average	Weighted Average
Mok2011 [51]	0.275	0.425	0.112	0.044	0.214	0.137
FTW [31]	0.251	0.425	0.135	0.072	0.221	0.156
NARX-QoE [3]	—	0.455	0.285	0.094	0.278	0.201
Xue2014 [83]	—	0.582	0.262	0.148	0.148	0.253
LSTM-QoE [24]	—	0.632	0.465	0.194	0.430	0.332
Bi-LSTM [22]	—	0.659	0.548	0.223	0.477	0.371
Liu2012 [46]	0.324	0.524	0.434	0.319	0.319	0.378
TV-QoE [27]	—	0.589	0.395	0.346	0.443	0.402
Yin2015 [84]	0.327	0.482	0.543	0.379	0.379	0.427
VideoATLAS [1]	0.050	0.491	0.330	0.480	0.338	0.432
Spiteri2016 [69]	0.376	0.501	0.597	0.461	0.484	0.490
P.1203 [59]	0.300	0.619	**0.604**	0.479	0.501	0.520
Bentaleb2016 [6]	**0.479**	0.712	0.521	0.495	0.552	0.538
SQI [21]	0.475	**0.735**	0.496	**0.504**	**0.553**	**0.543**
BSQI	**0.488**	**0.722**	**0.584**	**0.575**	**0.572**	**0.558**

In this article, we select 12 high-quality videos of diverse complexity to constitute the test sample set. All videos have the length of 30 seconds. Using the source sequences, each video is encoded with two types of encoding strategy including the traditional fixed bitrate encoding [54] and the state-of-the-art per-title encoding suggested by Netflix [11]. In the fixed bitrate encoding, each video is encoded into 10 pre-defined representations. While in the per-title encoding, the number of compressed versions and the choice of encoding configuration depend on the characteristics of source videos. For each video, we select the bitrate-resolution pair such that: (i) At a given bitrate, the produced representation should have as high quality as possible; and (ii) the perceptual difference between two adjacent bitrates should fall just below one just-noticeable different (the difference in VMAF ≈ 10). We segment the test sequences the encoded videos with GPAC's MP4Box [40] with a segment length of 2 seconds for the following reasons: First, 2-second segments are widely used in the development of ABR algorithms [49, 84] and deployment of real-world streaming applications [23, 41], primarily due to its flexibility for stream adaption to bandwidth changes and for its strong impact on reducing the latency of video delivery. Second, it allows us to derive test videos in an efficient way such that they cover a diverse adaptation patterns in a limited time. We randomly selected 12 network traces from both the 3G HSDPA dataset [65] and the 4G Belgium dataset [73] to cover a diversity of network conditions. The HSPDA dataset contains network traces that have significant variability and low average bandwidth, making it a strong test for the QoE models in the complicated scenarios. Traces in the Belgium dataset exhibit higher average throughput and lower standard deviation, which closely represents the realistic streaming environment. We compare BSQI with three objective QoE models that have guided the development of ABR algorithms, including Yin2015, Spiteri2016, and Bentaleb2016. We present results for the offline optimal scheme [49, 69], which is computed using dynamic programming with complete future throughput information. The dynamic programming-based method generates globally optimal streaming videos for the considered QoE models, completely eliminating the influence of inaccurate throughput estimation. For each source video, we randomly select a network trace and optimize the streaming videos with respect to the four objective QoE models. In the end, we obtain a total of 192 streaming videos generated from 24 (source videos, network traces) pairs ×2 encoding strategies × 4 ABR algorithms. An online demonstration of the experiment is available at Reference [15].

We perform a subjective user study that adopts the pairwise comparison methodology in which a pair of streaming videos generated from the same video contents and network traces are presented to human viewers. The subjective experiment is set up as a normal indoor home setting with an ordinary illumination level, with no reflecting ceiling walls and floors. A customized

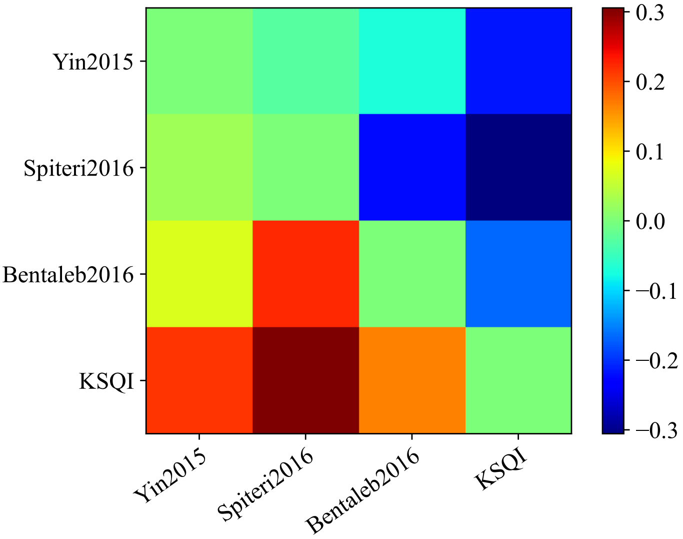


Fig. 2. Pairwise comparison matrix **R**. Each entry indicates the preference of the row model against the column model. $\mathbf{R} - \mathbf{R}^T$ are drawn here for better visibility.

interface is created to render a pair of 1,920 × 1,080 videos side-by-side on a 27-inch 4K monitor. The display is calibrated in accordance with the recommendations of ITU-R BT. 500 [36]. For each video pair, the subjects are forced to choose which one has a better perceptual quality. A total of 15 naïve subjects, including 7 males and 8 females aged between 18 and 55, participate in the subjective experiment. Visual acuity and color vision are confirmed from each subject before the subjective test. A training session is performed, during which, three video pairs that are different from the videos in the testing set are presented to the subjects. We used the same methods to generate the videos used in the training and testing sessions. Therefore, subjects knew what distortion types would be expected before the test session, and thus learning effects are kept minimal in the subjective experiment. For each subject, the whole study takes three hours, which is divided into six sessions with a five-minute break in between.

The results of the subjective experiment can be summarized as a 4 × 4 matrix **R**, where $r_{i,j}$ represents the probability of QoE model i better than QoE model j. Figure 2 shows the result matrix **R**, where the higher value of an entry (warmer color), the stronger the row model against the column model. It is obvious that BSQI performs favorably to the competing models. We further aggregate the pairwise comparison results into a global ranking via the maximum likelihood method for multiple options [48, 62, 72]. Let $\boldsymbol{\mu} = [\mu_1, \mu_2, \mu_3, \mu_4] \in \mathbb{R}^4$ be the global ranking score vector; we maximize the log-likelihood of $\boldsymbol{\mu}$

$$\begin{aligned} \underset{\mu}{\arg\max} \quad & \sum_{i,j} r_{i,j} \log(\Phi(\mu_i - \mu_j)) \\ \text{subject to} \quad & \sum_{i} \mu_i = 0, \end{aligned}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. The constraint $\sum_i \mu_i = 0$ is introduced to resolve the translation ambiguity. The optimization problem is convex and enjoys efficient solvers. A larger μ_i means the optimal streaming video in terms of the ith model is perceptually better than the optimal samples generated by other QoE models in general. Figure 3 shows the experimental results. It can be seen that BSQI significantly outperforms the standard QoE

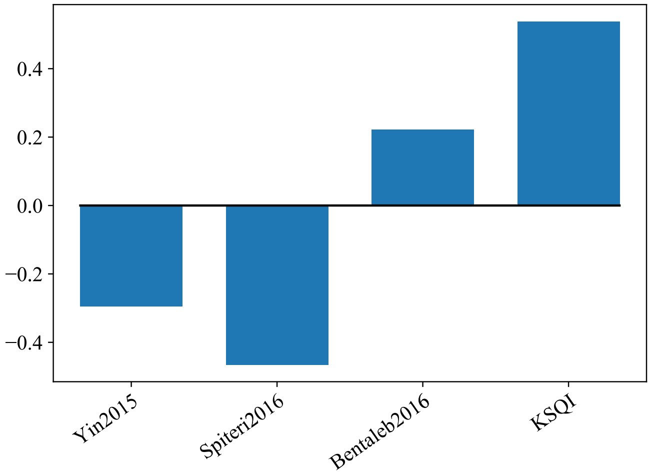


Fig. 3. Global ranking results of the four QoE models.

models. By taking a closer look at the trace-specific experiment results, we find that BSQI consistently delivers the best performance across different experiment setups, although the improvement is less significant on the 4G dataset. We notice that the small performance gain in the 4G experiment arises from the abundant bandwidth resource, especially when the highest resolution of streaming videos is restricted by the pairwise comparison experiment. Note that the maximum width/height of one test stimulus can be at most half of the width/height of the display. We expect a more significant improvement in the realistic setting where 4K, high dynamic range, and high framerate video contents are involved. The results have significant implications on the development of ABR algorithms. Specifically, state-of-the-art ABR algorithms have achieved a performance plateau level and significant improvement has become difficult to attain. However, the enormous difference in perceptual relevance between the bitrate-based QoE model and BSQI suggests that further improvement is attainable simply by adopting perceptually motivated optimization criterion.

5.4 Statistical Significance Test

To ascertain that the improvement of the proposed model is statistically significant, we carry out a statistical significance analysis by following the approach introduced in Reference [67]. First, we linearly scale MOSs in each dataset to the same perceptual scale [0, 100]. Second, a nonlinear regression function is applied to map the objective quality scores to predict the subjective scores independently on the four testing datasets. The prediction residuals of each QoE models from all datasets are aggregated into a vector. We observe that the prediction residuals all have zero-mean, and thus the model with lower variance is generally considered better than the one with higher variance. We conduct a hypothesis testing using F-statistics. Since the number of samples exceeds 50, the Gaussian assumption of the residuals approximately hold based on the central limit theorem [7]. The test statistic is the ratio of variances. The null hypothesis is that the prediction residuals from one quality model come from the same distribution and are statistically indistinguishable (with 95% confidence) from the residuals from another model. After comparing every possible pair of objective models, the results are summarized in Table 6, where a symbol “1” means the row model performs significantly better than the column model, a symbol “0” means the opposite, and a symbol “-” indicates that the row and column models are statistically indistinguishable. It can be observed that the proposed model is statistically better than all other methods on the combination of all existing benchmark datasets.

Table 5. PLCC between the Variants of BSQI Prediction and MOS on the Benchmark Datasets

QoE model	LIVE-NFLX-I	LIVE-NFLX-II	WaterlooSQoE-III	WaterlooSQoE-IV	Average	Weighted Average
BSQI with bitrate	0.622	0.722	0.670	0.618	0.658	0.647
BSQI with log bitrate	0.686	0.715	0.787	**0.738**	0.732	0.741
BSQI with QP	—	0.776	0.416	0.184	0.459	0.343
BSQI with VMAF	**0.753**	**0.905**	**0.794**	0.720	**0.793**	**0.769**

Table 6. Statistical Significance Matrix Based on F-statistics on the Combination of WaterlooSQoE-III, WaterlooSQoE-IV, LIVE-NFLX-I, and LIVE-NFLX-II Datasets

	FTW	Mok2011	NARX-QoE	Liu2012	LSTM-QoE	Bi-LSTM	Yin2015	TV-QoE	P.1203	VideoATLAS	Bentaleb2016	Spiteri2016	SQI	BSQI
FTW	-	-	0	0	0	0	0	0	0	0	0	0	0	0
Mok2011	-	-	-	0	0	0	0	0	0	0	0	0	0	0
NARX-QoE	1	-	-	0	0	0	0	0	0	0	0	0	0	0
Liu2012	1	1	1	-	0	0	0	0	0	0	0	0	0	0
LSTM-QoE	1	1	1	1	-	-	-	-	0	0	0	0	0	0
Bi-LSTM	1	1	1	1	-	-	-	-	0	0	0	0	0	0
Yin2015	1	1	1	1	-	-	-	-	0	0	0	0	0	0
TV-QoE	1	1	1	1	-	-	-	-	0	0	0	0	0	0
P.1203	1	1	1	1	1	1	1	1	-	-	-	0	0	0
VideoATLAS	1	1	1	1	1	1	1	1	-	-	-	-	0	0
Bentaleb2016	1	1	1	1	1	1	1	1	-	-	-	-	0	0
Spiteri2016	1	1	1	1	1	1	1	1	1	-	-	-	0	0
SQI	1	1	1	1	1	1	1	1	1	1	1	1	-	0
BSQI	1	1	1	1	1	1	1	1	1	1	1	1	1	-

A symbol "1" means that the performance of the row model is statistically better than that of the column model, A symbol "0" means that the row model is statistically worse, and a symbol "-" means that the row and column models are statistically indistinguishable.

Table 7. PLCC between the Variants of BSQI Prediction and MOS on the Benchmark Datasets

Constraint #	LIVE-NFLX-I	LIVE-NFLX-II	WaterlooSQoE-III	WaterlooSQoE-IV	Average	Weighted Average
None	0.731	0.903	0.663	0.681	0.745	0.720
(5)	0.743	0.902	0.788	0.718	0.788	0.766
(5)(6)	0.748	0.904	0.780	0.719	0.788	0.765
(5)(6)(7)	0.748	0.896	**0.800**	0.713	0.788	0.764
(5)(6)(7)(8)	**0.753**	**0.905**	0.794	**0.720**	**0.793**	**0.769**
(5)(6)(7)(8)(9)	**0.753**	**0.905**	0.794	**0.720**	**0.793**	**0.769**
(5)(6)(7)(8)(9)(10)	**0.753**	**0.905**	0.793	**0.720**	**0.793**	**0.769**
(5)(6)(7)(8)(9)(10)(11)	**0.753**	**0.905**	0.794	**0.720**	**0.793**	**0.769**
(5)	0.744	0.902	0.788	0.718	0.788	0.766
(6)	0.743	0.906	0.758	0.717	0.781	0.760
(7)	0.743	0.895	0.798	0.713	0.788	0.764
(8)	0.753	0.902	0.787	0.717	0.790	0.766
(9)	0.745	0.884	0.770	0.691	0.773	0.744
(10)	0.745	0.884	0.770	0.692	0.773	0.744
(11)	0.745	0.884	0.770	0.691	0.773	0.744
(12)	0.746	0.884	0.770	0.692	0.773	0.744
BSQI	**0.753**	**0.905**	0.794	**0.720**	**0.793**	**0.769**

5.5 Ablation Experiment

We conduct a series of ablation experiments to single out the core contributors of BSQI. We first take bitrate [46, 84], logarithmic bitrate [69], and QP [83] as the video presentation quality measure as opposed to VMAF and then train the QoE model with the proposed optimization framework. To map the range of video presentation quality measure into the same perceptual scale [0, 100], we apply a linear transform to the alternative measures before the training stage. From Table 5, we observe that BSQI achieves the best performance with the adoption of state-of-the-art video quality measure such as VMAF.

Next, we analyze the impact of the knowledge-imposed constraints on the quality prediction performance. We start from a baseline model by solving the problems in Equations (17) and (21) with no constraints and gradually increase the number of constraints. We then investigate the validity

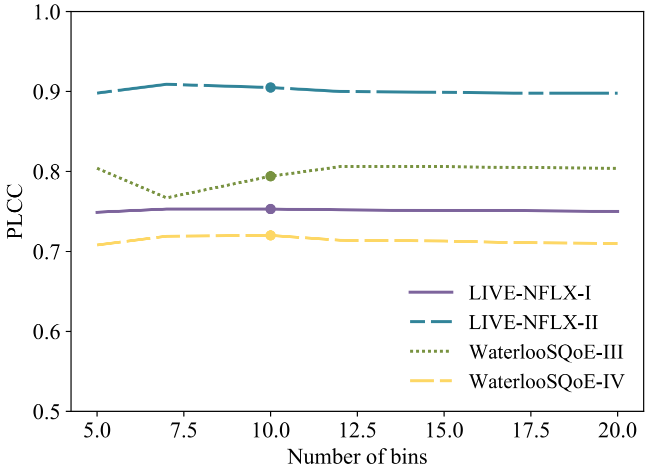


Fig. 4. Performance of BSQI with different number of bins.

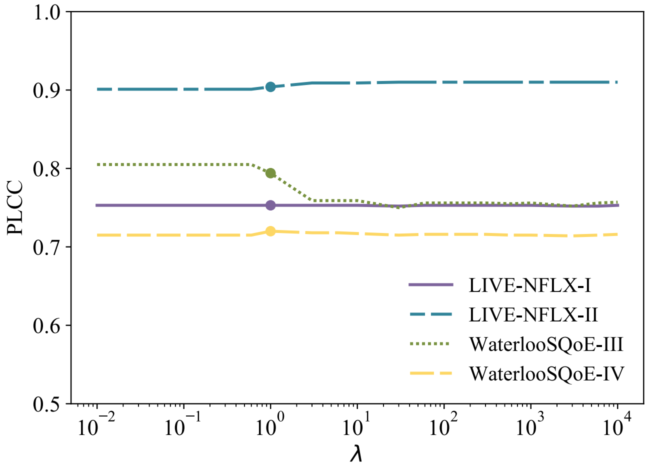


Fig. 5. Performance of BSQI with different λ.

of each observation by imposing only one constraint in a variant model. The results are listed in Table 7, from which the key observations are as follows: First, the performance of BSQI generally improves with respect to the number of imposed constraints, advocating the effectiveness of prior knowledge in regularizing the objective QoE functions. Second, while some of the constraints do not improve the performance of BSQI by themselves, the joint model achieves state-of-the-art performance. This suggests that the constraints may be complementary to each other. Third, the constraint (7) has drastically different impacts on the LIVE-NFLX-II dataset and the WaterlooSQoE-III dataset, suggesting that the validity of the constraint may be influenced by other factors. A careful investigation may further improve the performance of the proposed QoE model.

5.6 Impact of Step Sizes

In previous experiments, we set the bin sizes of video presentation quality and rebuffering duration to 10 and 1, respectively. To investigate the impact of step sizes, we train several variants of BSQI, where the number of bins ranges from 5 to 20. We show the experimental results in Figure 4.

Theoretically speaking, the performance of BSQI should increase monotonically with respect to the precision of feature representations. However, the observation does not echo our expectation, which may be a consequence of insufficient training data and intrinsic noise in the subjective opinion scores. Nevertheless, BSQI is generally very robust to a broad range of bin sizes.

5.7 Impact of the Weighting Parameter

The parameter λ in BSQI determines the tradeoff between fidelity and smoothness of the QoE functions. Although the optimal parameter is obtained from cross-validation in previous experiments, we also perform an experiment to investigate the impact of λ. Specifically, we train several versions of BSQI, where λ ranges from 0.01 to 10,000. The results are shown in Figure 5, from which we can observe that the performance of BSQI is generally insensitive to λ.

6 CONCLUSIONS

We propose a novel objective QoE model for adaptive streaming videos, namely, BSQI, by regularizing a non-parametric model with known HVS properties. BSQI outperforms the existing objective QoE models by a sizable margin over a wide range of video contents, encoding configurations, network conditions, and viewing devices, which we believe arises from a perceptually motivated video quality representation, a knowledge-constrained optimization framework, and a non-parametric model of QoE functions.

The proposed model may be improved in many ways. First, BSQI is readily extendable when new knowledge of HVS properties is acquired. With proper modifications of the non-parametric functions, we may incorporate more features such as motion strength [47] into the QoE model. Second, there may be better ways to combine the video presentation quality, rebuffering experience, and quality adaptation experience. For example, we can jointly model all influencing factors by escalating the dimensionality of the non-parametric model. Third, how to integrate the QoE model into the adaptive bitrate selection algorithm for optimal playback control is another challenging problem that is worth further investigations.

APPENDIX

A PROOF OF CONVEXITY

Let f and g be two different functions in the theoretical QoE function space $\mathcal{W}_Q$. We aim to show that $\forall\lambda \in [0,1]$, the function $\lambda f + (1-\lambda)g$ also satisfies the constraints (5)–(12). Let us consider the constraint in Equation (5). Based on our assumption, we have $f(\tau_t = \tau^1) \geq f(\tau_t = \tau^2), \forall\tau^1 \leq \tau^2, t$ and $g(\tau_t = \tau^1) \geq g(\tau_t = \tau^2), \forall\tau^1 \leq \tau^2, t$. It follows that

$$\begin{aligned}(\lambda f + (1-\lambda)g)(\tau_t = \tau^1) &= \lambda f(\tau_t = \tau^1) + (1-\lambda)g(\tau_t = \tau^1) \\ &\geq \lambda f(\tau_t = \tau^2) + (1-\lambda)g(\tau_t = \tau^2) \\ &= (\lambda f + (1-\lambda)g)(\tau_t = \tau^2).\end{aligned} \tag{22}$$

Therefore, the functional space given by constraint (5) represents a convex set. Similarly, we can show that each of the linear inequalities (6)–(12) also constitutes a convex set. It follows that the intersection of two/more convex sets is also a convex set, which completes the proof.

REFERENCES

[1] C. G. Bampis and A. C. Bovik. 2017. Learning to predict streaming video QoE: Distortions, rebuffering and memory. *ArXiv preprint arXiv:1703.00633* (Mar. 2017).

[2] C. G Bampis, A. C. Bovik, and Z. Li. 2018. A simple prediction fusion improves data-driven full-reference video quality assessment models. In *Picture Coding Symposium*. IEEE, 298–302.

[3] C. G. Bampis, Z. Li, and A. C. Bovik. 2017. Continuous prediction of streaming video QoE using dynamic networks. *IEEE Sig. Process. Lett.* 24, 7 (Jul. 2017), 1083–1087.

[4] C. G. Bampis, Z. Li, I. Katsavounidis, T. Y. Huang, C. Ekanadham, and A. C. Bovik. 2018. Towards perceptually optimized end-to-end adaptive video streaming. *ArXiv preprint arXiv:1808.03898* (Aug. 2018).

[5] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik. 2017. Study of temporal effects on subjective video quality of experience. *IEEE Trans. Image Process.* 26, 11 (Nov. 2017), 5217–5231.

[6] A. Bentaleb, A. C. Begen, and R. Zimmermann. 2016. SDNDASH: Improving QoE of HTTP adaptive streaming using software defined networking. In *ACM International Conference on Multimedia*. ACM, 1296–1305.

[7] C. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin.

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. 2011. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends® Mach. Learn.* 3, 1 (July 2011), 1–122.

[9] L. Breiman. 2001. Random forests. *Mach. Learn.* 45, 1 (Oct. 2001), 5–32.

[10] Cisco Mobile VNI. 2017. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper.* Retrieved from https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html.

[11] J. De Cock, Z. Li, M. Manohara, and A. Aaron. 2016. Complexity-based consistent-quality encoding in the cloud. In *IEEE International Conference on Image Processing*. IEEE, 1484–1488.

[12] F. Dobrian, V. Sekar, A. Awan, I. Stoica, D. Joseph, A. Ganjam, J. Zhan, and H. Zhang. 2011. Understanding the impact of video quality on user engagement. *ACM SIGCOMM Comput. Commun. Rev.* 41, 4 (Aug. 2011), 362–373.

[13] Pedro Domingos. 2000. A unified bias-variance decomposition. In *International Conference on Machine Learning*. 231–238.

[14] Z. Duanmu, D. Chen, Z. Li, W. Liu, Z. Wang, Y. Wang, and W. Gao. 2019. *Waterloo Streaming Quality-of-Experience Database IV*. Retrieved from http://ece.uwaterloo.ca/~zduanmu/waterloosqoe4.

[15] Z. Duanmu, W. Liu, D. Chen, Z. Li, Z. Wang, Y. Wang, and W. Gao. 2019. *Pairwise Comparison of Objective QoE Models via Analysis-by-synthesis*. Retrieved from http://ivc.uwaterloo.ca/research/KSQI/demo/.

[16] Z. Duanmu, W. Liu, Z. Li, D. Chen, Z. Wang, Y. Wang, and W. Gao. 2020. Assessing the quality-of-experience of adaptive bitrate video streaming. *arXiv preprint arXiv:2008.08804* (2020).

[17] Z. Duanmu, W. Liu, Z. Wang, and Z. Wang. 2021. To Appear. Quantifying visual image quality: A Bayesian view. *Ann. Rev. Vis. Sci.* (Sep. 2021, To Appear).

[18] Z. Duanmu, K. Ma, and Z. Wang. 2017. Quality-of-experience of adaptive video streaming: Exploring the space of adaptations. In *ACM International Conference on Multimedia*. ACM, 1752–1760.

[19] Z. Duanmu, K. Ma, and Z. Wang. 2018. Quality-of-experience for adaptive streaming videos: An expectation confirmation theory motivated approach. *IEEE Trans. Image Process.* 27, 12 (Dec. 2018), 6135–6146.

[20] Z. Duanmu, A. Rehman, and Z. Wang. 2018. A quality-of-experience database for adaptive video streaming. *IEEE Trans. Broadcast.* 64, 2 (June 2018), 474–487.

[21] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang. 2017. A quality-of-experience index for streaming video. *IEEE J. Select. Topics Sig. Process.* 11, 1 (Sep. 2017), 154–166.

[22] Tho Nguyen Duc, Chanh Minh Tran, Phan Xuan Tan, and Eiji Kamioka. 2019. Bidirectional LSTM for continuously predicting QoE in HTTP adaptive streaming. In *International Conference on Information Science and Systems*. 156–160.

[23] Encoding.com. 2016. *Microsoft Smooth Streaming*. Retrieved from https://www.encoding.com/microsoft-smooth-streaming/.

[24] N. Eswara, S. Ashique, A. Panchbhai, S. Chakraborty, H. P. Sethuram, K. Kuchi, A. Kumar, and S. S. Channappayya. 2020. Streaming video QoE modeling and prediction: A long short-term memory approach. *IEEE Trans. Circ. Syst. Video Technol.* 30, 3 (Mar. 2020), 661–673.

[25] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv.* 46, 4 (Apr. 2014), 1–37.

[26] Deepti Ghadiyaram, Janice Pan, and A. C. Bovik. 2017. A subjective and objective study of stalling events in mobile streaming videos. *IEEE Trans. Circ. Syst. Video Technol.* 29, 1 (2017), 183–197.

[27] D. Ghadiyaram, J. Pan, and A. C. Bovik. 2018. Learning a continuous-time streaming video QoE model. *IEEE Trans. Image Process.* 27, 5 (May 2018), 2257–2271.

[28] M. Grafl and C. Timmerer. 2013. Representation switch smoothing for adaptive HTTP streaming. In *IEEE International Workshop on Perceptual Quality of Systems*. ISCA/DEGA, 178–183.

[29] M. T. Hagan, H. B. Demuth, M. H. Beale, and R. De Jesús. 1996. *Neural Network Design*. Vol. 20. PWS Pub. Boston.

[30] K. Hornik, M. Stinchcombe, and H. White. 1989. Multilayer feedforward networks are universal approximators. *Neural Netw.* 2, 5 (Jan. 1989), 359–366.

[31] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau. 2013. Internet video delivery in YouTube: From traffic measurements to quality of experience. In *Data Traffic Monitoring and Analysis*. Springer, Berlin, 264–301.

[32] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia, and R. Schatz. 2011. Quantification of YouTube QoE via crowdsourcing. In *IEEE International Symosium on Multimedia*. IEEE, 494–499.

[33] T. Huang, C. Zhou, R. Zhang, C. Wu, X. Yao, and L. Sun. 2019. Comyco: Quality-aware adaptive video streaming via imitation learning. In *ACM International Conference on Multimedia*. 429–437.

[34] T. Y. Huang, R. Johari, N. McKeown, M. Trunnell, and M. Watson. 2015. A buffer-based approach to rate adaptation: Evidence from a large video streaming service. *ACM SIGCOMM Comput. Commun. Rev.* 44, 4 (Feb. 2015), 187–198.

[35] I. Ishii, T. Tatebe, Q. Gu, Y. Moriue, T. Takaki, and K. Tajima. 2010. 2000 fps real-time vision system with high-frame-rate video recording. In *IEEE International Conference on Robotics and Automation*. IEEE, 1536–1541.

[36] ITU-R BT.500-12. 1993. Recommendation: Methodology for the Subjective Assessment of the Quality of Television Pictures.

[37] J. Jiang, V. Sekar, and H. Zhang. 2014. Improving fairness, efficiency, and stability in HTTP-based adaptive video streaming with FESTIVE. *IEEE/ACM Trans. Netw.* 22, 1 (Feb. 2014), 326–340.

[38] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. 2003. High dynamic range video. *ACM Trans. Graph.* 22, 3 (July 2003), 319–325.

[39] H. J. Kim, D. G. Yun, H. Kim, K. S. Cho, and S. G. Choi. 2012. QoE assessment model for video streaming service using QoS parameters in wired-wireless network. In *IEEE International Conference on Advanced Communications Technology*. 459–464.

[40] J. Le Feuvre, C. Concolato, and J. Moissinac. 2007. GPAC: Open source multimedia framework. In *ACM International Conference on Multimedia*. ACM, 1009–1012.

[41] S. Lederer. 2015. *Optimal Adaptive Streaming Formats MPEG-DASH & HLS Segment Length*. Retrieved from https://bitmovin.com/mpeg-dash-hls-segment-length/.

[42] Z. Li, A. Aaron, L. Katsavounidis, A. Moorthy, and M. Manohara. 2016. *Toward a Practical Perceptual Video Quality Metric*. Retrieved from http://techblog.netflix.com/2016/06/toward-practical-perceptual-video.html.

[43] Z. Li, Z. Duanmu, W. Liu, and Z. Wang. To Appear. AVC, HEVC, VP9, AVS2, or AV1? - A comparative study of state-of-the-art video encoders on 4K videos. In *International Conference on Image Analysis and Recognition*. AIMI.

[44] Z. Li, X. Zhu, J. Gahm, R. Pan, H. Hu, A. C. Begen, and D. Oran. 2014. Probe and adapt: Rate adaptation for HTTP video streaming at scale. *IEEE J. Select. Areas Commun.* 32, 4 (Apr. 2014), 719–733.

[45] W. Liu, Z. Duanmu, and Z. Wang. 2018. End-to-end blind quality assessment of compressed videos using deep neural networks. In *ACM International Conference on Multimedia*. ACM, 546–554.

[46] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang. 2012. A case for a coordinated internet video control plane. *ACM SIGCOMM Comput. Commun. Rev.* 42, 4 (Sep. 2012), 359–370.

[47] Y. Liu, S. Dey, F. Ulupinar, M. Luby, and Y. Mao. 2015. Deriving and validating user experience model for DASH video streaming. *IEEE Trans. Broadcast.* 61, 4 (Dec. 2015), 651–665.

[48] K. Ma, Z. Duanmu, Z. Wang, Q. Wu, W. Liu, H. Yong, H. Li, and L. Zhang. 2019. Group maximum differentiation competition: Model comparison with few samples. *IEEE Trans. Patt. Anal. Mach. Intell.* (2019), DOI: 10.1109/TPAMI.2018.2889948

[49] H. Mao, R. Netravali, and M. Alizadeh. 2017. Neural adaptive video streaming with pensieve. In *ACM SIGCOMM*. ACM, 197–210.

[50] A. M. Mishra. 2001. *Quality of Service in Communications Networks*. Wiley.

[51] R. K. Mok, X. Luo, E. W. Chan, and R. K. Chang. 2012. QDASH: A QoE-aware DASH system. In *ACM Conference on Multimedia Systems*. ACM, 11–22.

[52] A. K. Moorthy, L. K. Choi, A. C. Bovik, and G. De Veciana. 2012. Video quality assessment on mobile devices: Subjective, behavioral and objective studies. *IEEE J. Select. Topics Sig. Process.* 6, 6 (Oct. 2012), 652–671.

[53] R. M. Nasiri, J. Wang, A. Rehman, S. Wang, and Z. Wang. 2015. Perceptual quality assessment of high frame rate video. In *IEEE International Conference on Multimedia and Signal Processing*. IEEE, 1–6.

[54] Netflix Inc. 2015. *Per-title Encode Optimization*. Retrieved from http://techblog.netflix.com/2015/12/per-title-encode-optimization.html.

[55] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan. 2005. Light Field Photography with a Hand-held Plenoptic Camera. Computer Science Technical Report.

[56] P. Ni, R. Eg, A. Eichhorn, C. Griwodz, and P. Halvorsen. 2011. Flicker effects in adaptive video streaming to handheld devices. In *ACM International Conference on Multimedia*. ACM, 463–472.

[57] J. Nightingale, Q. Wang, C. Grecos, and S. Goma. 2014. The impact of network impairment on quality of experience (QoE) in H.265/HEVC video streaming. *IEEE Trans. Consum. Electron.* 60, 2 (2014), 242–250.

[58] R. L. Oliver. 1980. A cognitive model of the antecedents and consequences of satisfaction decisions. *J. Market. Res.* 17, 4 (Nov. 1980), 460–469.

[59] ITU-T P.1203. 2017. *Parametric Bitstream-based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport*. Retrieved from https://www.itu.int/rec/dologin_pub.asp?lang=e&id=T-REC-P.1203-201710-I!!PDF-E&type=items.

[60] R. Pastrana-Vidal, J. C. Gicquel, C. Colomes, and H. Cherifi. 2004. Sporadic frame dropping impact on quality perception. In *Human Vision and Electronic Imaging IX*. SPIE, 182–194.

[61] P. Paudyal, F. Battisti, and M. Carli. 2016. Impact of video content and transmission impairments on Quality of Experience. *Multim. Tools Applic.* 75, 23 (2016), 16461–16485.

[62] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, and F. Battisti. 2015. Image database TID2013: Peculiarities, results and perspectives. *Sig. Process.: Image Commun.* 30 (Jan. 2015), 57–77.

[63] A. Rehman and Z. Wang. 2013. Perceptual experience of time-varying video quality. In *IEEE International Workshop Quality of Multimedia Experience*. IEEE, 218–223.

[64] A. Rehman, K. Zeng, and Z. Wang. 2015. Display device-adapted video quality-of-experience assessment. In *SPIE*. SPIE, 939406.1–939406.11.

[65] H. Riiser, P. Vigmostad, C. Griwodz, and P. Halvorsen. 2013. Commute path bandwidth traces from 3G networks: Analysis and applications. In *ACM Conference on Multimedia Systems*. ACM, 114–118.

[66] D. Z. Rodríguez, Z. Wang, R. L. Rosa, and G. Bressan. 2014. The impact of video-quality-level switching on user quality of experience in dynamic adaptive streaming over HTTP. *EURASIP J. Wirel. Commun. Netw.* 2014, 1 (2014), 1–15.

[67] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.* 15, 11 (Nov. 2006), 3440–3451.

[68] K. Singh, Y. Hadjadj-Aoul, and G. Rubino. 2012. Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC. In *IEEE Consumer Communications & Networking Conference*. IEEE, 1–6.

[69] K. Spiteri, R. Urgaonkar, and R. K. Sitaraman. 2016. BOLA: Near-optimal bitrate adaptation for online videos. In *IEEE International Conference on Computer Communications*. IEEE, 1–9.

[70] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. 2017. OSQP: An operator splitting solver for quadratic programs. *ArXiv preprint arXiv:1711.08013* (Nov. 2017).

[71] L. Toni, R. Aparicio-Pardo, K. Pires, W. Simon, A. Blanc, and P. Frossard. 2015. Optimal selection of adaptive streaming representations. *ACM Trans. Multim. Comput., Commun., Applic.* 11, 2s (Feb. 2015), 1–43.

[72] K. Tsukida and M. R. Gupta. 2011. How to Analyze Paired Comparison Data. University of Washington, Technical Report. UWEETR-2011-0004.

[73] J. van der Hooft, S. Petrangeli, T. Wauters, R. Huysegems, P. R. Alface, T. Bostoen, and F. De Turck. 2016. HTTP/2-based adaptive streaming of HEVC video over 4G/LTE networks. *IEEE Commun. Lett.* 20, 11 (Aug. 2016), 2177–2180.

[74] VQEG. 2000. Final Report from the Video Quality Experts Group on the Validation of Objective Models of Video Quality Assessment. Technical Report. Retrieved from http://www.vqeg.org/.

[75] Z. Wang. 2001. *Internet QoS: Architectures and Mechanisms for Quality of Service*. Morgan Kaufmann.

[76] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* 13, 4 (Apr. 2004), 600–612.

[77] Z. Wang and A. C. Bovik. 2009. Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Sig. Process. Mag.* 26, 1 (Jan. 2009), 98–117.

[78] Z. Wang, Z. Duanmu, A. Rehman, and K. Zeng. 2017. Method and system for automatic user quality-of-experience measurement of streaming video. US Patent WO/2017/152274.

[79] Z. Wang and A. Rehman. 2017. Begin with the end in mind: A unified end-to-end quality-of-experience monitoring, optimization and management framework. In *SMPTE Annual Technical Conference and Exhibition*. SMPTE, 1–11.

[80] Z. Wang and E. P. Simoncelli. 2008. Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *J. Vis.* 8, 12 (2008), 8–8.

[81] Z. Wang, K. Zeng, and A. Rehman. 2016. Method and system for smart adaptive video streaming driven by perceptual quality-of-experience estimations. US Patent WO/2016/123721.

[82] K. Watanabe, J. Okamoto, and T. Kurita. 2007. Objective video quality assessment method for evaluating effects of freeze distortion in arbitrary video scenes. In *Image Quality and System Performance IV*, Vol. 64940P. SPIE, 1–8.

[83] J. Xue, D. Zhang, H. Yu, and C. W. Chen. 2014. Assessing quality of experience for adaptive HTTP video streaming. In *IEEE International Conference on Multimedia and Expo Workshop*. IEEE, 1–6.

[84] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli. 2015. A control-theoretic approach for dynamic adaptive video streaming over HTTP. *ACM SIGCOMM Comput. Commun. Rev.* 45, 4 (Apr. 2015), 325–338.

Received 7 April 2021; revised 11 September 2021; accepted 13 October 2021