Empowerment of Atypical Viewers via Low-Effort Personalized Modeling of Video Streaming Quality

LEONARDO PERONI, IMDEA Networks Institute and UC3M, Spain
SERGEY GORINSKY, IMDEA Networks Institute, Spain
FARZAD TASHTARIAN, Alpen-Adria Universität Klagenfurt, Austria
CHRISTIAN TIMMERER, Alpen-Adria Universität Klagenfurt, Austria

Quality of Experience (QoE) and QoE models are of an increasing importance to networked systems. The traditional QoE modeling for video streaming applications builds a one-size-fits-all QoE model that underserves atypical viewers who perceive QoE differently. To address the problem of atypical viewers, this paper proposes iQoE (individualized QoE), a method that employs explicit, expressible, and actionable feedback from a viewer to construct a personalized QoE model for this viewer. The iterative iQoE design exercises active learning and combines a novel sampler with a modeler. The chief emphasis of our paper is on making iQoE sample-efficient and accurate. By leveraging the Microworkers crowdsourcing platform, we conduct studies with 120 subjects who provide 14,400 individual scores. According to the subjective studies, a session of about 22 minutes empowers a viewer to construct a personalized QoE model that, compared to the best of the 10 baseline models, delivers the average accuracy improvement of at least 42% for all viewers and at least 85% for the atypical viewers. The large-scale simulations based on a new technique of synthetic profiling expand the evaluation scope by exploring iQoE design choices, parameter sensitivity, and generalizability.

CCS Concepts: • **Networks** → **Application layer protocols**.

Additional Key Words and Phrases: video streaming; personalization; quality of experience; modeling; sample efficiency; accuracy; subjective study; perception dataset; personalized QoE model.

ACM Reference Format:
Leonardo Peroni, Sergey Gorinsky, Farzad Tashtarian, and Christian Timmerer. 2023. Empowerment of Atypical Viewers via Low-Effort Personalized Modeling of Video Streaming Quality. *Proc. ACM Netw.* 1, CoNEXT3, Article 17 (December 2023), 27 pages. https://doi.org/10.1145/3629139

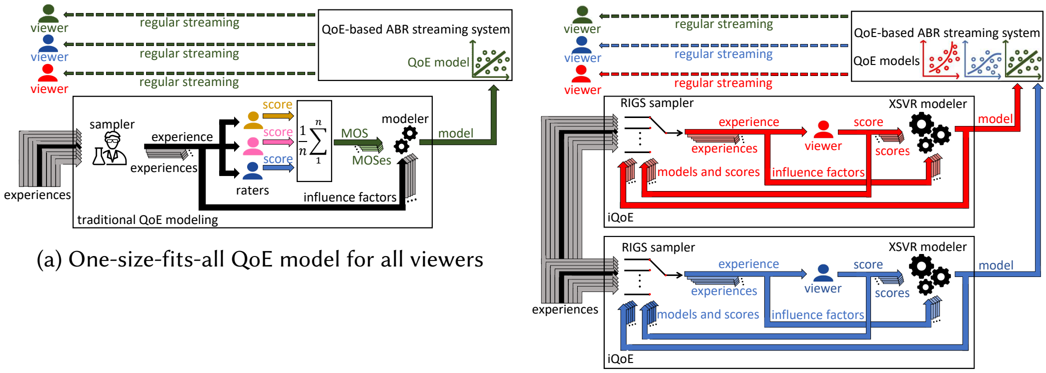
1 INTRODUCTION

A common goal in networked systems is to improve quality of experience (QoE), i.e., user satisfaction with the provided service [16]. QoE forms a basis for various functionalities in networked systems, such as adaptive bitrate (ABR) streaming [2, 64, 87, 93, 106, 107], fair congestion control [69], time-shifted video upload [79], and videoconferencing loss recovery [83]. Although this paper operates in the context of ABR streaming, we believe that our work is pertinent to other kinds of QoE-based systems as well.

Because QoE is a subjective notion, the construction of a QoE model usually relies on subjective tests. Fig. 1a illustrates traditional QoE modeling and utilization of the constructed QoE model by

Authors' addresses: Leonardo Peroni, IMDEA Networks Institute and UC3M, Spain, leonardo.peroni@imdea.org; Sergey Gorinsky, IMDEA Networks Institute, Spain, sergey.gorinsky@imdea.org; Farzad Tashtarian, Alpen-Adria Universität Klagenfurt, Austria, farzad.tashtarian@aau.at; Christian Timmerer, Alpen-Adria Universität Klagenfurt, Austria, christian.timmerer@aau.at.

2834-5509/2023/12-ART17 $15.00
https://doi.org/10.1145/3629139



(a) One-size-fits-all QoE model for all viewers

(b) Personalized QoE models for atypical viewers

Fig. 1. Reliance of QoE-based ABR streaming on: (a) traditional QoE modeling and (b) iQoE.

an ABR streaming system. Raters are the users who participate in the subjective tests. The model construction involves a series of *assessments*. An *experience* constitutes the cornerstone of each assessment and refers to a sequence of video chunks characterized by objective *influence factors*, e.g., stall duration [80]. The *sampler*, commonly a human expert conducting the assessments [94], selects an experience for each assessment from an experience set. Each rater provides an individual *score* for every presented experience [51]. A *mean opinion score (MOS)* averages the individual scores by all raters and represents the QoE perception by the average rater. Based on the MOSes and influence factors of the rated experiences, the *modeler* constructs a *QoE model* by approximating the functional relation between the MOS and influence factors. Existing QoE models differ in their function forms [42, 107] and approximation methods [82, 111].

Viewers are the users who consume regular streaming provided by the *ABR streaming system* that utilizes the QoE model in its ABR algorithm. Because viewers greatly outnumber raters, the traditional QoE modelling eliminates the subjective-test overhead for a vast majority of viewers. On the negative side, the traditional approach builds a one-size-fits-all QoE model that underserves atypical viewers who perceive QoE differently. A viewer might have dissimilar QoE perception compared to not only the average rater but also any rater in the reference group of the QoE model. For example, standard methodologies of subjective tests deliberately exclude from the reference group those viewers who rate experiences with systematic shifts or inversions [53].

In this paper, we define atypical viewers as the 10% of the population who are the furthest from the median QoE perception. Statistical infrequency is a common foundation for defining an atypical person in various scientific disciplines such as psychopathology [40]. While related work demonstrates that humans vary substantially in their QoE perception [26, 41, 44, 49], our paper classifies a viewer as atypical from the statistical perspective alone and does not seek to uncover any physiological, psychological, or other reasons why the viewer is atypical. Selecting 10% or another percentile in our definition of atypical viewers is not particularly important since the results of our study remain qualitatively the same.

We tackle the problem of atypical viewers by developing a novel personalization solution called *iQoE (individualized QoE)*. In iQoE, a viewer acts as the sole rater in building a personalized QoE model. By construction, the personalized QoE model represents the QoE perception of this and only this viewer. iQoE involves the viewer into a short series of assessments and exercises active learning to iteratively select experiences for the assessments. The main focus of this paper is on making iQoE sample-efficient and accurate. The proposed design relies on a novel Randomized

Improved Greedy Sampling (RIGS) strategy and combines our automatic RIGS sampler with an eXtended SVR (XSVR) modeler, which utilizes an extended set of 10 influence factors and Support Vector Regression (SVR) [8].

iQoE represents a dramatic departure from prior work on QoE personalization. One group of existing solutions relies on inferring the QoE perception without explicit feedback from the viewer [21, 48, 116]. These solutions are yet to prove their ability to provide high accuracy. Another approach equips the viewer with a means to personalize parameter values of a generic QoE model [49, 71, 74]. However, the viewer typically does not know which parameter settings yield an accurate personalized QoE model. The conceptual novelty of iQoE is in employing explicit, expressible, and actionable feedback from the viewer. Our extensive evaluations of iQoE via online subjective studies and simulations confirm that iQoE achieves accurate QoE personalization while requiring low effort from the viewer.

While this paper intentionally focuses on the atypical viewers because they tend to benefit the most from QoE personalization, our vision for integrating iQoE into an ABR streaming system imposes no restrictions on who may use this personalization solution, i.e., iQoE is an option available to any viewer. In Fig. 1b, the blue and red viewers opt in and utilize iQoE to construct personalized QoE models that the QoE-based ABR streaming system leverages subsequently to provide personalized streaming to the blue and red viewers. On the other hand, the green viewer does not use iQoE, and the ABR streaming system relies on the one-size-fits-all QoE model for regular streaming to this viewer.

Our paper makes the following contributions:

- In the context of QoE-based ABR streaming, we propose iQoE that employs explicit, expressible, and actionable feedback from a viewer to construct a personalized QoE model for this viewer. The iterative design conducts active learning and combines a new RIGS sampler with an XSVR modeler so as to make iQoE sample-efficient and accurate. Whereas iQoE is an option available to any viewer, atypical viewers are the chief beneficiaries of the proposed personalization solution.
- This paper leverages Microworkers [103] for subjective studies with 120 raters and demonstrates that 50 assessments completed in about 22 minutes empower a viewer to construct a personalized QoE model that, compared to the best of the 10 baseline models, delivers the average accuracy improvement of at least 42% for all viewers and at least 85% for the atypical viewers. The large-scale simulations corroborate the iQoE design choices and clarify method properties.
- The paper collects and makes openly available a QoE-perception dataset of independent value. The dataset contains 14,400 individual scores of experiences characterized by 10 influence factors.

2 BACKGROUND ON QOE MODELING

QoE modeling employs different methodologies and produces QoE models that vary in their scoring scales, influence factors, and intended usage. While subjective assessments traditionally take place in tightly controlled lab environments, online crowdsourcing platforms make it easier to conduct assessments at the cost of weaker control over the experimental settings [84]. A standard scoring scale contains five levels of scores from 1 to 5 [54]. To discern QoE perception at a finer granularity, our paper adopts another standard scale with score levels from 1 to 100, where ranges 1-20, 21-40, 41-60, 61-80, and 81-100 correspond to the bad, poor, fair, good, and excellent QoE, respectively [86].

Prior studies consider a multitude of influence factors that include metrics of video quality and streaming systems. The former class consists of such metrics as the Peak Signal-to-Noise Ratio (PSNR) [50], Structural Similarity Index Measure (SSIM) [102], and Video Multimethod Assessment

Fusion (VMAF) [60]. Within the latter class, system factors from the application layer increasingly attract more attention than network-layer metrics. In particular, stall duration $\mathcal{T}_n$ and bitrate $\mathcal{R}_n$ of chunk n are archetypal influence factors in the ABR streaming systems that partition a video into a sequence of $\mathcal{N}$ chunks and encode each chunk for multiple bitrates [89]. Other examples of system factors are number l and average duration d of stalls during the playback [77].

There is no consensus either on the best way to map influence factors into QoE. Whereas some QoE models are closed-form expressions, construction of QoE models via machine learning (ML) becomes common. This paper considers 10 existing QoE models. For brevity, we label each of the models with a single letter, as specified below. A number of prominent ABR streaming systems rely on different instances of the following general closed-form QoE model:

$$Q_1 = \kappa \sum_{n=1}^{\mathcal{N}} q(\mathcal{R}_n) - \lambda \sum_{n=1}^{\mathcal{N}-1} |q(\mathcal{R}_{n+1}) - q(\mathcal{R}_n)| - \mu \sum_{n=1}^{\mathcal{N}} \mathcal{T}_n, \tag{1}$$

where κ, λ, and μ are tunable parameters, and $q(\cdot)$ denotes a function of the bitrate. We consider **models B** [107], **G** [91], **R** [25], **S** [106], and **V** [46] that instantiate $q(\mathcal{R})$ in Equation 1 as the identity function, $\log(\mathcal{R}/r)$ with r denoting the lowest bitrate, PSNR, SSIM, and VMAF, respectively. Similar to model V, **model N** [12] underlies the SDNDASH architecture. Widely known as the FTW model, **model F** [42] belongs to another type of closed-form QoE models and employs an exponential function with parameters α, β, γ, and δ:

$$Q_2 = \alpha e^{-(\beta d + \gamma) l} + \delta. \tag{2}$$

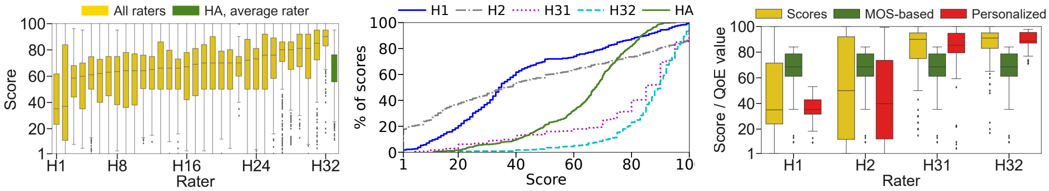
Among the QoE models constructed via ML, **model L** [96] represents state-of-the-art approaches based on deep learning and predicts QoE via a long short-term memory (LSTM) network. Relying on random forests (RF), **model P** [52] refers to the standard P.1203 model. Finally, **model A** [9] denotes the QoE model constructed by Video ATLAS via SVR on VMAF and other influence factors.

QoE models serve various purposes and return values from different ranges. For example, models B, G, and V primarily act as bases for internal improvement of ABR streaming systems [46, 64, 91, 107] and might produce negative values, complicating their value interpretability by humans. On the other hand, models L and P yield values between 1 and 5, as in the standard five-level scale for subjective scores. The heterogeneity of the value ranges undermines direct comparison of QoE models.

3 MOTIVATION

3.1 Promise of Personalized QoE Modeling

While Section 1 defines atypical viewers as the 10% of the population who are the furthest from the median QoE perception, we examine by how much the atypical viewers deviate in their QoE perception from the one-size-fits-all MOS-based QoE models. Specifically, we analyze the Waterloo-IV dataset that reports individual scores of experiences corresponding to all combinations of five videos of different genres (sports, nature, movies, video games, and slides), two encoders, five ABR algorithms, nine network traces, and three varieties of viewing devices [27]. Each experience in the dataset consists of seven chunks, with the playback of each chunk taking 4 s without stalls. Waterloo-IV employs the 1-100 scoring scale and a reference group of 92 raters. The group excludes five other raters who score experiences negligently or otherwise abnormally. Waterloo-IV is the largest recent dataset of its kind, as other datasets either provide MOSes without individual scores [10, 11, 28, 43, 85] or characterize chunks with smaller sets of influence factors [35, 82]. Our analysis focuses on the 32 raters who use high-definition television (HDTV) devices to watch 450 experiences where 10 influence factors characterize each chunk.



(a) All raters vs. average HA (b) Four atypical raters vs. HA (c) Personalization promise

Fig. 2. Inaccuracy of traditional QoE modeling and promise of personalized modeling for atypical viewers.

Fig. 2a depicts in gold the individual scores by each rater, orders the 32 raters based on the median score, and respectively refers to them as raters H1 through H32. The scores across the reference group are quite distinct in terms of both median and variance, e.g., the gap between the first and third quartiles ranges from 12 to 73. To the right of rater H32, Fig. 2a plots in green the MOSes, i.e., the QoE perception by the average rater labeled as HA. Raters H1, H2, H31, and H32 cover 10% of all 32 raters and comprise the four atypical raters in this population. Their respective median scores of 35 (i.e., poor), 37 (poor), 85 (excellent), and 90 (excellent) are quantitatively far from the median score of 68 (i.e., good) by average rater HA. Fig. 2b zooms in on the scores of all experiences by the average and four atypical raters. The results reveal *substantial numerical differences in the QoE perception between the average and atypical raters.*

To evaluate how the choice of individual scores vs. MOSes as the basis for QoE modeling affects the model accuracy, we consider QoE model A from Section 2 and construct five versions of it: a MOS-based version and, for each of the four atypical raters, a personalized model trained on the individual scores by this rater. We train and test the models on 70%, i.e., 315, and remaining 30%, i.e., 135, of all 450 experiences, respectively. For each atypical rater H1, H2, H31, and H32, Fig. 2c plots the individual scores by this rater, QoE values produced by the MOS-based model, and QoE values produced by the rater's personalized model. The personalized QoE modeling enormously enhances the model accuracy and, on average across the four atypical raters, reduces the numerical gap between the median score and median QoE value by more than 31 times. Hence, *personalized QoE modeling brings promise of significant quantitative improvements in the model accuracy for the atypical raters.*

The above analysis is for the atypical members of the vetted reference group in a subjective study. Atypical viewers who are not raters, such as the abnormal red viewer intentionally excluded from the reference group in Fig. 1, might have even greater numerical gaps from the typical QoE perception and benefit more from personalized QoE modeling.

3.2 Design Goals

Because Section 3.1 indicates that atypical viewers are not only statistically different but also quantitatively far from the typical QoE perception captured by a MOS-based QoE model, we advocate personalized QoE modeling for atypical viewers and establish our design goals in contrast with three alternative means for accuracy improvement of the traditional QoE modeling.

While a vast majority of all viewers in the traditional approach do not exert any modelling effort, one possibility for retaining this attractive property is to form multiple reference groups, build a separate MOS-based QoE model for each group, and associate a viewer with the group that represents the QoE perception by this viewer most accurately. If the viewer and raters of the associated reference group are similar in their QoE perception, the numerical discrepancy between the viewer's QoE perception and MOS-based QoE model of the group is likely to be small. Although the general technique of multiple reference groups works reasonably well in other application domains such as recommendation systems [23, 115], our evaluation in Section 5 demonstrates that this approach does not sufficiently mitigate the inaccuracy of the MOS-based QoE modeling due to the great heterogeneity of QoE perception among humans. In addition, the task of associating each viewer with a representative reference group might be difficult to accomplish without interacting with the viewer. The above discussion leads us to our first goal:

Goal 1. *The construction of an accurate QoE model for an atypical viewer should rely on perception feedback from this viewer.*

The feedback requirement does not necessarily imply that the viewer has to explicitly score experiences as the raters do in the traditional QoE modeling. An intriguing prospect is indirect inference of the viewer's QoE perception, e.g., through automatic monitoring of the viewer's gaze direction, facial expression, engagement, or viewing activities [21, 58, 67, 73, 100]. Unfortunately, such inference techniques might require special equipment or raise privacy issues [73]. Moreover, due to the complexity of overall human behavior, this alternative is yet to prove its suitability for accurate QoE modeling. Hence, we consider only explicit mechanisms for the viewer's feedback:

Goal 2. *The mechanism for perception feedback should be explicit.*

By itself, the feedback explicitness does not assure that the feedback is useful. A possible avenue for personalized QoE modeling is to ask a viewer for personal preferences, e.g., for values of the κ, λ, and μ parameters in Equation 1, and leverage the provided preferences to personalize a generic QoE model [49, 71, 74]. However, a viewer usually does not know how to articulate personal preferences well enough to make the resulting QoE model accurate. Thus, we pursue the following goal:

Goal 3. *The solicited feedback should be of a kind expressible by the viewer and actionable for accurate QoE modeling.*

When combined, Goals 1 through 3 limit the design options to methods that build an accurate QoE model for an atypical viewer by collecting explicit actionable feedback from this viewer. However, to encourage the viewer's participation, the model construction should require only light contributions from the viewer. This leads us to our fourth goal:

Goal 4. *The amount of effort contributed by a viewer into the construction of an accurate personalized QoE model for the viewer should be small.*

4 DESIGN

4.1 iQoE Overview

This section designs iQoE to achieve the goals established in Section 3.2. iQoE meets Goals 1, 2, and 3 by adopting the assessment-based conceptual structure of the traditional QoE modeling and reducing the group of raters to the viewer for whom the method constructs the personalized QoE model, as illustrated in Fig. 1b. Engaging the viewer as the sole rater satisfies Goal 1. The viewer explicitly scores experiences, which conforms to Goal 2, and in the same expressible actionable manner as in the traditional QoE modeling, thereby complying with Goal 3.

The fulfillment of Goal 4 is challenging and constitutes the main focus of this paper. To keep the effort of the viewer low, iQoE limits the viewer's involvement to a short series of H assessments that cumulatively consume a small amount of the viewer's time. This section derives a simple and yet effective iterative design for iQoE where a new automatic RIGS sampler and XSVR modeler require little memory and execute quickly on the client side without causing a perceptible wait for the viewer during the assessment series. iQoE exercises active learning so as to be sample-efficient and accurate. Interactions between RIGS and XSVR steer iQoE to produce an accurate QoE model despite the limited length of the assessment series.

Although the deliberate emphasis of this paper is on the atypical viewers because they constitute the largest beneficiaries of QoE personalization, we design iQoE as an option available to any viewer. By not opting in, a typical or just uninterested viewer avoids any subjective-test overhead. Regular streaming to such viewers continues relying on the one-size-fits-all QoE model.

Turning the viewer into the sole rater during the construction of the personalized QoE model has positive side effects. Since the viewer becomes the only party utilizing the constructed QoE model, iQoE incentivizes the viewer to perform the assessments conscientiously. Also, the same number of assessments by the viewer typically results in a more accurate QoE model than in the traditional MOS-based modeling with many raters. Besides, because the viewer is likely to train the personalized QoE model in the settings of regular streaming, the QoE-model accuracy might increase due to the more specific context than in traditional lab-based subjective tests. That said, we defer to future work a comprehensive study of the impact by different contexts and contents.

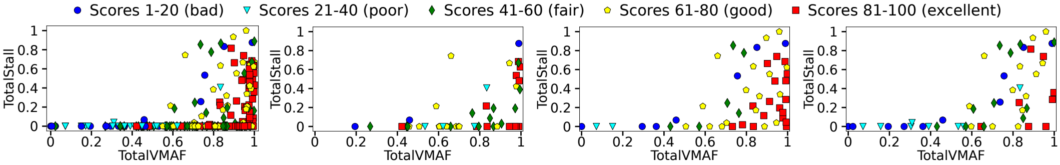
Algorithm 1 *iQoE*(E)

```
 1: $K \leftarrow E$; $J \leftarrow \emptyset$; $M \leftarrow \emptyset$; $Q \leftarrow \emptyset$      ▷ initialization
 2: for $t = 1, \ldots, H$ do
 3:     $e \leftarrow$ RIGS($K, J, M, Q$)                          ▷ sampling
 4:     $s \leftarrow$ the viewer's score of $e$                     ▷ assessment
 5:     $K \leftarrow K - \{e\}$; $J[t] \leftarrow e$; $M[t] \leftarrow s$
 6:     $Q \leftarrow$ XSVR($J, M$)                               ▷ modeling
 7: end for
 8: return $Q$                                          ▷ final QoE model
 9: procedure RIGS($K, J, M, Q$)
10:     if $t \leq h$ then
11:         $e \overset{R}{\leftarrow} K$                           ▷ random sampling
12:     else
13:         $e \leftarrow \text{argmax}_K \min_J D_{jk}$               ▷ IGS sampling
14:     end if
15:     return $e$          ▷ experience for the next assessment
16: end procedure
17: procedure XSVR($J, M$)
18:     if $t \geq h$ then
19:         $Q \leftarrow$ the SVR model trained on $J$ and $M$
20:     end if
21:     return $Q$                                  ▷ current QoE model
22: end procedure
```

Algorithm 1 reports the pseudocode for the iQoE construction of model Q as a personalized QoE model for the viewer. Set E of experiences serves as an input to the algorithm and comprehensively covers the conditions possible during regular streaming. iQoE acquires set E in advance, e.g., by generating it through simulations or as a real dataset supplied by the ABR streaming system. Algorithm 1 tracks the non-rated and rated experiences in set K and array J, respectively, and stores the scores of the rated experiences in array M. Initially, set K contains experience set E while J, M, and Q are empty (Line 1 of Algorithm 1). iQoE performs H iterations (Lines 2–7) to produce the final QoE model (Line 8). Each iteration t uses the RIGS sampler to select experience e for the next assessment (Line 3), obtains score s of this experience by the viewer (Line 4), moves experience e from set K to array J and records score s in the corresponding element of array M (Line 5), and then updates QoE model Q by applying the XSVR modeler to arrays J and M (Line 6). Sections 4.2 and 4.3 elaborate on the designs of our RIGS sampler and XSVR modeler, respectively.

4.2 RIGS Sampler

We strive for an effective simple design of the automatic RIGS sampler. The primary objective is to pick a series of H experiences from set E so that the final QoE model becomes as accurate as possible. On the other hand, the design simplicity is important because of enabling the sampler to select an experience quickly without introducing a discernible wait for the viewer between successive assessments. Set E characterizes each of its experiences e with features from set C. Whereas these $|C|$ features form an *input space*, the viewer's score s_j of rated experience e_j in array J and value $Q(e_k)$ produced by QoE model Q for non-rated experience e_k in set K belong to an *output space* of subjective scores and QoE values. Intuitively, the experiences selected by an effective sampling strategy should provide a balanced coverage of set E in both input and output spaces so that the constructed QoE model successfully deals with both diversities in experiences and



(a) All 315 experiences (b) Selection by RS (c) Selection by GS (d) Selection by IGS

Fig. 3. Selection of 50 experiences by the RS, GS, and IGS samplers from the set of 315 experiences.

their perception by the viewer. The sampler's task to select H out of the $|E|$ experiences faces the following extra complication: while the feature values of all experiences are available in advance, the score of an experience becomes known only after the viewer assesses the experience.

Our derivation of the sampling strategy for RIGS illustrates relevant issues by utilizing again the real Waterloo-IV dataset described in Section 3.1. For clarity, we consider one rater, set E with 315 experiences (which are the same as the training experiences in Section 3.1), and constrain the characterization of each experience to two normalized features, namely *TotalVMAF* (the sum of the VMAF values across all chunks in the experience) and *TotalStall* (the total stall duration divided by the total duration of the experience). Fig. 3a depicts the 315 experiences as points in the two-dimensional input space formed by the TotalVMAF and TotalStall features. The colors of the points expose the ranges of the experience scores in the output space: we color the 1-20 (bad), 21-40 (poor), 41-60 (fair), 61-80 (good), and 81-100 (excellent) score ranges in blue, cyan, green, yellow, and red, respectively. This example sets H equal to 50 experiences.

From the simplicity perspective, the best strategy is *random sampling (RS)* that picks non-rated experiences from set K randomly. While simple, RS might cover the input space unevenly and represent some areas in the space insufficiently. In our example, Fig. 3b shows that most of the 50 experiences selected by RS lie around the TotalStall = 0 or TotalVMAF = 1 lines and that only a handful of the selected experiences represent poor conditions with respect to both stalls and VMAF.

One possibility for addressing the above weakness of RS is to adopt *greedy sampling (GS)* that accounts for the distances between experiences in the $|C|$-dimensional input space. With f_{ij} and f_{ik} referring to the values of feature i for rated experience e_j in array J and non-rated experience e_k in set K respectively, GS defines the distance between experiences e_j and e_k as $\sqrt{\sum_{i \in C} (f_{ij} - f_{ik})^2}$, computes for each e_k in K the minimum distance between this e_k and any experience e_j in J, determines the largest of these $|K|$ minimum distances, and iteratively moves the corresponding experience e_k from set K to array J. Whereas GS is likely to strike a better balance than RS in sampling the input space, GS remains oblivious of the output space and might cover it unevenly. Returning to our example, Fig. 3c confirms that while the 50 experiences selected by GS cover the input space more evenly than the 50 RS-selected experiences in Fig. 3b, the coverage of the output space remains insufficiently balanced, e.g., the experiences with excellent (red) scores dominate the experiences with bad (blue), poor (cyan), and fair (green) scores.

A logical fix for the demonstrated drawback of GS is to go for *improved greedy sampling (IGS)* that accounts for distances in the output space as well. IGS redefines the distance metric of GS as $D_{jk} = |s_j - Q(e_k)|\sqrt{\sum_{i \in C} (f_{ij} - f_{ik})^2}$ where Q denotes the current QoE model, and the prepended $|s_j - Q(e_k)|$ factor is the distance in the one-dimensional output space between score s_j of rated experience e_j in array J and QoE value $Q(e_k)$ of non-rated experience e_k in set K. The prepended factor uses $Q(e_k)$ as an estimate for score s_k of experience e_k because s_k becomes known only after the viewer assesses e_k. Similarly to GS, IGS iteratively moves from set K to array J the experience e_k with the largest minimum D_{jk} distance. IGS is a form of active learning since it selects an experience for the next assessment based on the QoE model trained on the previously selected experiences. In our running example, Fig. 3d shows that the 50 experiences selected by IGS provide a good coverage in the input space and cover the output space more evenly that the 50 GS-selected experiences in Fig. 3c, e.g., by reducing the imbalance between the excellent (red) scores and bad (blue), poor (cyan), and fair (green) scores. From the simplicity perspective, IGS has polynomial time complexity of $O(|K||J||C|)$ per iteration and is less attractive than RS.

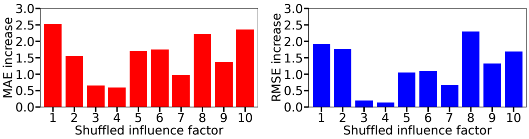


Fig. 4. Importance of the 10 influence factors in XSVR for the atypical raters.

Our RIGS sampler selects a series of H assessments by combining the strengths of RS and IGS. The selection of experiences by RS is always simple and quick. On the other hand, the accuracy advantages of IGS grow as it collects more samples. Early on in the sampling process, the extra work done by IGS to compute the distances between experiences does not yield significant payoffs because the current QoE model remains too inaccurate to make QoE values $Q(e_k)$ an informative prediction for scores s_k of yet non-rated experiences e_k. As the number of samples increases, the QoE model becomes more accurate, and its $Q(e_k)$ values steer IGS to select a more instructive sample for further improvement of the model accuracy. Hence, RIGS starts by quickly picking a series of initial experiences via RS and then switches to selecting the subsequent experiences via IGS. iQoE controls the switching by parameter h: the RIGS sampler selects the first h experiences from set K randomly (Line 11 of Algorithm 1), the QoE model gets updated by the XSVR modeler starting from iteration h of iQoE (Line 19), and RIGS selects experiences $h+1$ through H according to the largest minimum D_{jk} distance (Line 13). In Section 5.2, we analyze sensitivity of iQoE to parameters h and H, show that the default setting of h to 10 experiences is the most beneficial for the final model accuracy, and evaluate RIGS against alternative samplers.

4.3 XSVR Modeler

We also aim for a simple and yet effective design of the XSVR modeler. To build accurate QoE models, XSVR simultaneously considers a broad pragmatic set of influence factors. While many additional influence factors might increase predictive power, e.g., electroencephalographic or other psychophysiological signals [73], we restrict the choice of influence factors to those measurable without special equipment in the viewer's regular settings of video watching and compose an *eXtended (X) set* as a superset of the influence factors in the 10 existing models discussed in Section 2. The X set, which contributes letter X to the XSVR name of our modeler, comprises the following 10 influence factors: (1) representation identifier, (2) stall duration, (3) bitrate, (4) chunk size, (5) frame width, (6) frame height, (7) indicator whether the bitrate is the highest in the bitrate ladder, (8) PSNR capped at 50 [29], (9) SSIM, and (10) VMAF. All 10 influence factors are easily measurable and, in particular, tracked by Waterloo-IV [27]. Because the 10 influence factors separately describe each of the $\mathcal{N}$ chunks in each experience, XSVR employs a total of $10\mathcal{N}$ features, which implicitly also represent inter-chunk influence factors such as bitrate changes and their magnitude.

To understand how much the $10\mathcal{N}$ features contribute to the predictive power of XSVR, we apply the technique of permutation feature importance [15]. Once again, we turn to the Waterloo-IV dataset and revisit four atypical raters H1, H2, H31, and H32 analyzed in Section 3.1. For each of the 10 influence factors, we randomly shuffle the values of a particular influence factor in all $\mathcal{N}$ chunks. As Fig. 4 shows for each shuffled influence factor, the shuffling diminishes the predictive ability of XSVR by increasing, on average across the four atypical raters, both Mean Absolute Error (MAE) and Root-Mean-Square Error (RMSE). Although some of the factors are more important than others, all 10 of them contribute positively to the predictive performance, thereby justifying the usage of all 10 influence factors in XSVR. Appendix A reports on XSVR feature importance in more detail.

The simplicity constraint calls for efficient approximation of the functional relation between QoE and $10\mathcal{N}$ features so that the memory consumption and training overhead of the QoE model on the client device are insignificant. Whereas construction of QoE models increasingly relies on ML techniques [52, 96, 98], we also adopt an ML-based solution but stay away from deep learning due to its high computational complexity. Specifically, the XSVR modeler relies on SVR [8] because of its memory efficiency and general effectiveness on small datasets with high-dimensional input spaces, i.e., in the same sort of situations as ours. Section 5.2 evaluates XSVR design choices, including its reliance on SVR vs. other simple ML-based solutions.

5 EVALUATION

5.1 Subjective Studies

Methodology overview. We conduct the subjective studies in two phases. The first phase directly recruits 34 volunteers from around the globe and all walks of life. The number of 34 raters lies between 15 and 40, i.e., in the range that the International Telecommunication Union suggests for subjective tests in its Recommendation P.910 [54]. To generalize the conclusions to a broader population, the second phase scales the total number of raters in our studies to 120 by leveraging Microworkers, an online crowdsourcing platform [103].

To perform the studies, we develop a real website and deploy it on the Internet. A rater accesses the website via a browser in the rater's regular streaming setting at a time of the rater's choosing. For each rater, the website iteratively draws from a 1,000-element experience set a series of 120 experiences and transfers them one by one to the rater for playback and assessment. Each experience contains four chunks characterized by the 10 influence factors described in Section 4.3. Every chunk has the playback time of 2 s without stalls. We generate the experience set through simulations on the Park platform [65] by using ThroughputRule (TR) [90], BBA [47], and MPC [107] as throughput-based [55, 61, 62], buffer-based [68, 91], and hybrid [24, 46, 64] ABR algorithms, respectively, 102 network traces from three sources [2, 30, 81], a 13-representation bitrate ladder, and *Tears of Steel* in its 4K version[1] as the source video. After watching an experience, the rater scores it on the 1-100 scale. The website allows the rater to pause the series of assessments and resume it later. Upon completion of the 120 assessments, the website asks the rater to fill in a survey. Appendices B and C, respectively, provide additional information on the website design and experience set.

Dataset. We collect and openly release, along with all accompanying code, a dataset consisting of the 1,000-element experience set and 14,400 individual scores provided by the 120 raters [72]. 115 of the raters answer all questions in the post-assessment survey. Among those who answer, 28% and 72% identify themselves as female and male, respectively, from locations in 47 countries on four continents (with 45 home countries). The age ranges from 20 to 63 years old. 64% of the respondents rank their participation in the studies as *pleasant*, with the other three options being *slightly annoying*, *quite annoying*, and *very annoying*. Appendix D describes our dataset in more detail. Due to the scarcity of real data on individual QoE perception, the collected dataset constitutes a contribution of independent importance.

Ethical issues. The Ethical Board of our institute granted full approval to conduct the research. The subjects opted into the studies via informed consent on the front page of the website, with the consent required before any data collection could commence. The studies did not collect any personal identifiers and did not open any opportunities for linking the collected experimental results or demographic statistics with the subjects' actual identities.

Atypical raters. In agreement with the definition of atypical viewers in Section 1, our subjective studies involve 12 atypical raters comprising 10% of all 120 raters. Five and seven of the 12 atypical

[1] https://mango.blender.org/download/ accessed last on October 10, 2023

raters are from the first and second phases of the studies, respectively. Hence, both direct recruitment and crowdsourcing contribute atypical raters to the studies.

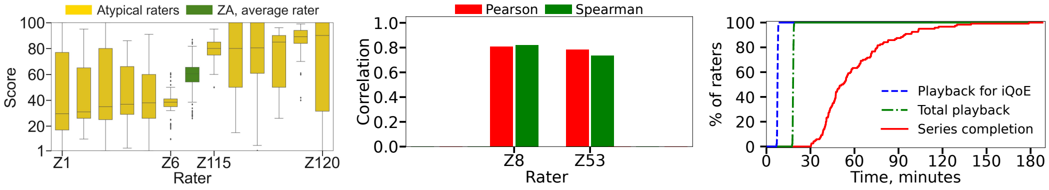
Training and testing of the baselines and iQoE. The subjective studies configure the h and H parameters of iQoE to their default settings of 10 and 50 experiences, respectively, and consider the 10 models from Section 2 as baselines. We randomly pick 50 experiences from the experience set to train baselines, and reuse 10 of these 50 experiences as the iQoE's initial h experiences across all 120 raters. iQoE relies on RIGS to select the subsequent $H - h = 40$ experiences for each rater, and these 40 experiences are generally different across the 120 raters. Hence, 90 out of the 120 experiences assessed by a rater support model training, with each particular QoE model trained on 50 experiences. We use the MOSes and individual scores by the rater, respectively, to train eight parameterized baselines and iQoE via regression. These parameterized baselines are models B, G, R, S, V, N, F, and A, and the regression returns values for their parameters, e.g., parameters κ, λ, and μ for the former five baselines and parameters α, β, γ, and δ for model F. The regression-based training produces QoE models that predominantly compute QoE values on the 1-100 scale. On rare occasions when a QoE model computes a value below 1 or above 100, we treat the spillover as a prediction error without adjusting the out-of-scale value. Models L and P come without a publicly available training module and compute QoE values on the 1-5 scale. We use these two models in their public configurations trained by the models' proposers and linearly map the computed QoE values into the 1-100 scale. We test all 10 baselines and iQoE on the same set of 30 experiences, which accurately represent the 1,000-element experience set. A shuffle of the 90 training and 30 testing assessments throughout the 120-assessment series ensures that the rater is unaware whether the current assessment is for training or testing.

Metrics. We measure accuracy of QoE model Q by means of MAE $= \sum_{e_k \in K} |Q(e_k) - s_k|/|K|$ and RMSE $= \sqrt{\sum_{e_k \in K} (Q(e_k) - s_k)^2/|K|}$, respectively, where K refers to a set of rated experiences, and s_k and $Q(e_k)$ denote the rater's score and QoE value of experience e_k, respectively. As the variance in the individual errors increases, RMSE exceeds MAE by a larger amount.

Results. To analyze the collected dataset and evaluate iQoE against alternatives, we examine the following questions: *(1)* How heterogeneous is the QoE perception in the dataset? *(2)* How consistent is the QoE perception over time? *(3)* How much of the rater's time does iQoE take to construct the personalized QoE model? *(4)* How does iQoE perform compared to the MOS-based baselines? *(5)* Is iQoE superior to using multiple reference groups? *(6)* How does iQoE perform compared to personalized versions of the baselines?

(1) Perception heterogeneity. We arrange the 120 raters in nondecreasing order of their median scores and accordingly label the raters as Z1 through Z120. Raters Z1 through Z6 and Z115 through Z120 comprise the 12 atypical raters. Fig. 5a depicts the individual and median scores of the 12 atypical raters and average rater ZA. The median scores of the atypical raters consist of six poor scores between 29.5 and 38.5, three good scores of 80, 80, and 80.5, and three excellent scores of 85, 89, and 90. In contrast, the median score of average rater ZA, i.e., median MOS, is a good score of 61. Fig. 5a corroborates that *the QoE perception by atypical viewers differs dramatically from the average QoE perception by all viewers.*

(2) Scoring consistency over time. Raters Z8 and Z54 are the two raters who heed our request to repeat the same assessment series in the same settings, e.g., with respect to the viewing device and browser, 20 days after the original studies. For each of the two raters, Fig. 5b presents the Pearson and Spearman correlations between the scores by the rater in the original and repeated series. For both raters, the scores exhibit high correlation in regard to either values or ranks: the Pearson and Spearman coefficients exceed 0.81 and 0.73 for raters Z8 and Z54, respectively. The



(a) Atypical raters vs. average ZA (b) Scoring consistency over time (c) Playback and completion times

Fig. 5. Results of the subjective studies.

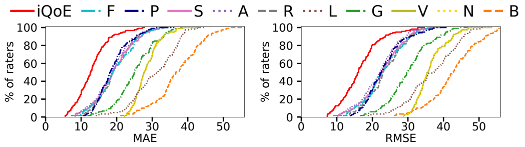


Fig. 6. iQoE vs. MOS-based QoE modeling.

Table 1. Average iQoE gains over the 10 baselines.

		A	S	R	P	F	G	N	V	L	B
All raters	MAE	1.54	1.58	1.61	1.63	1.64	2.17	2.41	2.41	2.67	3.23
	RMSE	1.42	1.47	1.53	1.57	1.52	2.03	2.43	2.43	2.51	2.89
Atypical raters	MAE	2.18	2.19	2.24	2.06	2.1	2.59	2.87	2.87	2.81	3.55
	RMSE	1.89	1.93	2.02	1.92	1.85	2.39	2.86	2.86	2.67	3.18

detected consistency of QoE perception over time indicates that *personalized QoE models preserve their accuracy over time without a need for frequent retraining.*

(3) Time to construct the personalized QoE model. We analyze the collected dataset to estimate the amount of time it would take for a viewer to construct the personalized QoE model. For each rater, our subjective studies use 50 out of the 120 rated experiences to train the personalized QoE model. Fig. 5c shows that the total playback time of the 50 experiences including the stalls varies across all 120 raters from 6.8 to 8.4 minutes. For the entire series of 120 rated experiences, the total playback time including the stalls ranges from 17.5 to 19 minutes, and the overall completion time spreads from 30.4 to 188 minutes, with the median value of 53 minutes. While the completion time excludes all intervals between an explicit pause and subsequent resumption of the series by the rater, the other contributors to the extra delay include downloading an entire experience before the browser starts its playback, rewatching of an experience by the rater, reflecting on an appropriate score for an experience, and being distracted by unrelated tasks without explicitly pausing the series. Because the ratio of the playback times is close to the $120/50 = 2.4$ ratio of the experience counts, we use this 2.4 factor to estimate the median completion time for the 50 assessments to be around 22 minutes, which is relatively low. Hence, our estimates indicate that *the amount of the viewer's time taken by iQoE to construct the personalized QoE model is affordable.*

(4) iQoE vs. MOS-based modeling. To evaluate iQoE against the 10 baselines, Fig. 6 plots the distributions of their accuracy. iQoE significantly outperforms all baselines and provides MAE and RMSE values as low as 5.5 and 7.3, respectively. For 20% of all 120 raters, iQoE provides MAE and RMSE of at most 9.3 and 11.8, whereas model A is the best among all baselines with the corresponding MAE and RMSE of 14.3 and 17.6, meaning that the MAE and RMSE gain by iQoE over the best baseline is 1.53 and 1.48 in MAE and RMSE, respectively. Table 1 shows that the average gains by iQoE over the 10 baselines across all 120 raters range from 1.54 (model A) to 3.23 (model B) in MAE and from 1.42 (model A) to 2.89 (model B) in RMSE. The average gains for the 12 atypical raters are higher and vary from 2.06 and 1.85 (models P and F, respectively) to 3.55 and 3.18 (model B) in MAE and RMSE, respectively. Hence, while iQoE improves the model accuracy for all raters, iQoE provides larger QoE gains to the atypical raters. Our results show that *the average accuracy improvement of iQoE over the MOS-based baselines is at least 42% for all raters and at least 85% for the atypical raters.*

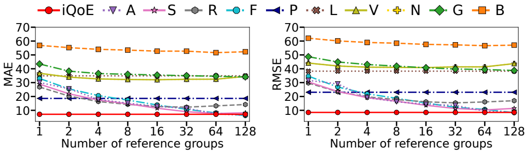


Fig. 7. iQoE vs. multiple reference groups.

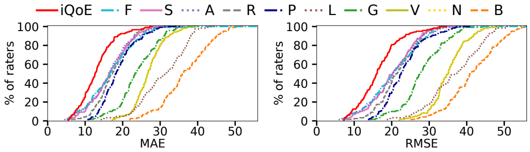


Fig. 8. QoE vs. personalized baselines.

(5) iQoE vs. multiple reference groups. According to Section 3.2, multiple reference groups pose an alternative to the iQoE approach of allowing each viewer to act as a rater in the construction of the viewer's personalized QoE model. The alternative requires more reference groups, with the raters of each group having less heterogeneous QoE perception, and associates every viewer with the most representative group. For this series of experiments, we conduct subjective tests with eight extra raters on Microworkers so as to increase the total number of raters to 128, which is a power of two and supports recursive partitioning of the rater population into equal-sized groups all the way down to the single-rater groups. We rearrange the expanded set of 128 raters in nondecreasing order of their median scores and correspondingly relabel the raters as S1 through S128. Under the new labeling, atypical rater Z120 becomes atypical rater S128. In this experimental series, we use rater S128 as the viewer and consider eight different partitions of the 128-rater population into one, two, four, eight, 16, 32, 64, and 128 groups where the number of raters in each group equals 128, 64, 32, 16, eight, four, two, and one, respectively. The partitions and the viewer's association with the most representative group follow the order of the median scores. Specifically, as the number of reference groups increases from 1 to 128, we associate viewer S128 with reference groups {S1, ..., S128}, {S65, ..., S128}, {S97, ..., S128}, {S113, ..., S128}, {S121, ..., S128}, {S125, ..., S128}, {S127, S128} and {S128}, respectively. In the latter partition with the single-rater groups, the viewer and rater are the same, implying that atypical rater S128 acts as the sole rater in building the personalized QoE model.

Fig. 7 explores how much an increase in the number of reference groups helps the baselines to bridge the accuracy gap with iQoE. Even in the single-rater partition, baseline models B, G, V, N, L, P, and R still fail to close the gap in either MAE or RMSE. Personalized models S, A, and F perform the best among the baselines. Because these models match the iQoE accuracy only when the viewer becomes the sole rater, Fig. 7 supports the conclusion that *iQoE produces more accurate QoE models compared to the approach of multiple reference groups, unless the latter reduces itself to personalized QoE modeling.*

(6) iQoE vs. personalized baselines. To investigate the what-if scenario that personalizes the baseline QoE models, we return to our default setting with 120 raters Z1 through Z120 and train the baselines on the raters' individual scores rather than the MOSes. Fig. 8 reports the accuracy distributions for the personalized baselines vs. iQoE. As Section 5.1 mentions earlier, the training modules of models L and P are not publicly available. Fig. 8 includes the results for models L and P for completeness. Although the personalization of the baselines reduces the accuracy gap, iQoE still outperforms all personalized baselines. In particular, the average gain by iQoE over the best baseline, which is model A, across all 120 raters is 30% and 27% in MAE and RMSE, respectively. Thus, we conclude that *iQoE derives its accuracy advantage from both modeling personalization and its specific design which combines the RIGS sampler and XSVR modeler.*

5.2 Simulations

Methodology. While Section 5.1 leverages crowdsourcing to scale up the number of raters, the method of subjective studies imposes other limitations on the evaluation scope. For example, it is

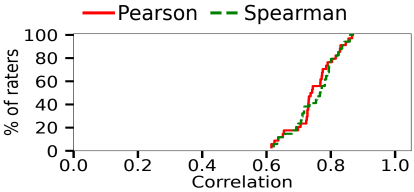


Fig. 9. Synthetic vs. real.

Table 2. Accuracy of sampler-modeler combinations.

	MAE				RMSE			
	XSVR	XGB	RF	GP	XSVR	XGB	RF	GP
RIGS	**4.3±0.3**	5.3±0.2	6.6±0.3	9.9±0.3	**6.4±0.3**	7.4±0.2	8.2±0.4	11.9±0.4
IGS	4.6±0.2	6.4±0.1	7.5±0.2	10.6±0.1	6.5±0.3	8.5±0.2	9.1±0.3	12.4±0.2
RS	4.7±0.2	4.5±0.2	4.7±0.3	9.1±0.3	7.8±0.4	7.7±0.4	7.6±0.5	12.1±0.5
GS	5.7±0.8	8.1±0.6	9.3±0.7	10.9±0.4	7.9±1.0	10.7±0.6	11.6±0.7	12.6±0.4
UC	9.7±1.9	7.1±0.4	7.7±0.3	14.4±2.0	13.2±1.8	9.8±0.4	10.5±0.2	17.2±1.8
QBC	5.0±0.2	4.9±0.3	6.4±0.4	9.0±0.5	8.1±0.4	8.4±0.6	8.0±0.4	12.3±0.8

incredibly difficult for a real rater to evaluate a series of 1,000 experiences. Because simulations are a common way to enhance the scope of real-world experiments, this section develops and applies a new simulation technique of *synthetic profiling* where a large number of *synthetic raters* quickly evaluate experience series of an (almost) arbitrary length. Another aspiration for the synthetic raters is to accurately represent the QoE perception of real raters. The proposed simulation technique utilizes the proliferation of parameterized QoE models and refers to them as *profiles*. Whereas the structure of a profile is predetermined, the profile produces a different instance with different parameter values when trained on a different dataset, e.g., the individual scores of experiences by a real rater. Each instance is a specific QoE model, which automatically and quickly produces a specific QoE value when presented with values of the influence factors. Hence, we utilize each instance to act as a *synthetic rater*. The training of p profiles on the individual scores by g real raters has the multiplicative effect of creating $p \times g$ synthetic raters.

This paper applies synthetic profiling to $p = 8$ profiles and $g = 32$ real raters to create $p \times g = 256$ synthetic raters. We take the 32 real Waterloo-IV raters from Section 3.1, models B, G, R, S, V, N, F, and A from Section 2 as the eight profiles, and utilize the 256 synthetic raters to conduct large-scale simulations where each synthetic rater assesses 1,000 experiences. We train (create) every synthetic rater and test (collect QoE values from) the synthetic rater on 70%, i.e., 700, and remaining 30%, i.e., 300, of all 1,000 experiences, respectively. In contrast, we train iQoE in its default settings on only $H = 50$ experiences. Appendix E provides additional information on our simulation methodology.

Results. To validate the technique of synthetic profiling, we examine the Pearson and Spearman correlations between the scores by the 256 synthetic raters and the 34 real raters from the first phase of the subjective studies in Section 5.1. The scores are also for the same 120 experiences as in Section 5.1. For each real rater, we find the most correlated synthetic rater, and Fig. 9 plots the distribution of these closest correlations across all real raters. Both Pearson and Spearman coefficients are high, above 0.7 for 80% of the real raters, and suggest that *the synthetic raters represent real raters accurately*. For brevity, the rest of this section refers to synthetic raters as raters.

Our large-scale simulations with 256 raters corroborate the conclusion of the subjective studies in Section 5.1 that the personalized QoE models built by iQoE greatly improve on the accuracy of the MOS-based baselines. Appendices F, G, and H report the simulation results in more detail.

(1) iQoE design choices. We consider Uncertainty Clustering (UC) [88], Query By Committee (QBC) [17], IGS [104], GS [108], and RS as alternative samplers and eXtreme Gradient Boosting (XGB) [22], Gaussian Processes (GP) [78], and RF [39] as alternative simple modelers.

Table 2 shows that, with 50 assessments by the 256 raters, the tandem of RIGS and XSVR is the most accurate among the 6 × 4 sampler-modeler combinations and enables iQoE to achieve average MAE and RMSE of 4.3 and 6.4, respectively, which are remarkably low for the 1-100 scale. With XSVR fixed as the modeler, Fig. 10 shows that iQoE with its RIGS sampler consistently outperforms the RS+XSVR, GS+XSVR, UC+XSVR, QBC+XSVR and IGS+XSVR alternatives, e.g., reduces average RMSE to 7 after 39 assessments compared to 77, 57, 190, 96, and 43 assessments by the counterparts. The assessment effort decreases, respectively, by a factor of 1.97, 1.46, 4.87, 2.46, and 1.10. Fig. 10

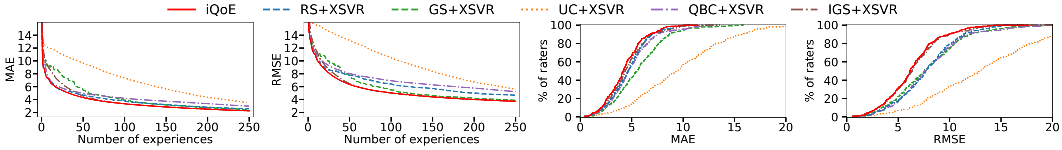


Fig. 10. Evaluating the sampler design choice of iQoE.

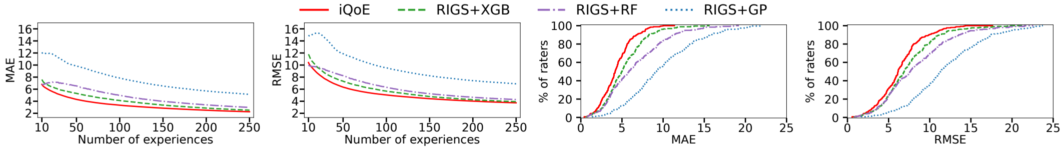


Fig. 11. Evaluating the modeler design choice of iQoE.

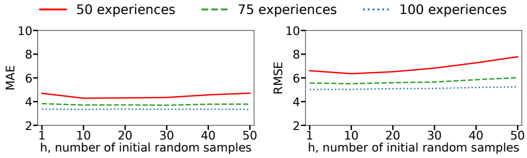


Fig. 12. Sensitivity of iQoE to the h parameter.

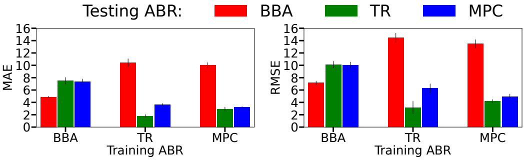


Fig. 13. iQoE generalizability.

also plots the MAE and RMSE distributions for all individual raters after 50 assessments and also backs the choice of RIGS for iQoE. For example, while iQoE attains MAE of 6 for 85% of the raters, this percentage is 77%, 59%, 24%, 73%, and 79% for RS+XSVR, GS+XSVR, UC+XSVR, QBC+XSVR, and IGS+XSVR, respectively. Fig. 11 examines different modelers in conjunction with RIGS. iQoE in its reliance on XSVR consistently delivers lower average MAE and RMSE than the RIGS+XGB, RIGS+RF, and RIGS+GP alternatives. Overall, *the simulations support the adoption of RIGS and XSVR by iQoE.*

(2) Parameter Sensitivity. Fig. 12 analyzes sensitivity of iQoE to its parameter h, which controls the switch from RS to IGS in RIGS. With $H = 50$ experiences, MAE and RMSE reach their minimum values around $h = 10$ experiences, justifying this default setting for h. With H set to 75 or 100 experiences, the impact of parameter h on the model accuracy is less pronounced. Despite the modest accuracy improvements, we consider RS a valuable contribution to RIGS because RS is also simpler than IGS. Appendix G and its Fig. 17 show that the model accuracy is largely insensitive to the balance between the training and testing sets unless the training set becomes quite small.

(3) iQoE generalizability. Not only size but also composition of the experience set affects the constructed QoE models. We create three experience sets by using BBA, TR, or MPC as the ABR algorithm in the Park platform. After training QoE models on each of the three sets, we test every QoE model on all three sets and ensure that the training and testing portions of the sets never overlap. This setup corresponds to a scenario where the current ABR algorithm of a QoE-based streaming system relies on a QoE model built with an outdated ABR algorithm. Fig. 13 reports MAE and RMSE for all combinations of the training and testing ABR algorithms. Training with BBA yields excellent generalizability. With TR or MPC as the training ABR algorithm, the generalizability is noticeably weaker. These results suggest that if the eventual ABR algorithm of the QoE-based streaming system is unknown at the time of constructing the QoE model, BBA constitutes a reasonable choice as the ABR algorithm for generating the experience set.

(4) iQoE overhead. We track the iQoE processing and memory overhead imposed mostly by the RIGS sampler and XSVR modeler. Appendix H and its Fig. 18 demonstrate that *the overhead is*

negligible. In particular, iQoE does not introduce a perceptible wait for the rater between successive assessments.

6 IQOE INTEGRATION INTO VIDEO STREAMING SYSTEMS

6.1 Integration into a Video Streaming Platform

While the main focus of this paper is on personalizing QoE models and making iQoE sample-efficient and accurate, we now discuss iQoE integration into video streaming systems. For concreteness, we start by considering a specific hypothetical *video streaming platform*, or a *platform* for short.

Video streaming platform. The video-on-demand platform extensively employs the cloud and, in particular, uses Amazon Elastic Compute Cloud (EC2) [4] to encode ingested video content into multiple representations and train its proprietary cloud-based ABR algorithm. The ABR algorithm leverages Reinforcement Learning (RL) to maximize a one-size-fits-all QoE model combining VMAF, VMAF stability, and client-side stall duration as influence factors [46]. The platform quickly adapts to specific network conditions by means of transfer learning [105], keeps its vast library of videos in Amazon Simple Storage Service (S3) [7], and utilizes Amazon Relational Database Service (RDS) [6] to store the viewer's account information and large amounts of other structured data about viewing behaviors and preferences. By relying on Amazon Redshift [5], the platform analyzes the large-scale structured data to personalize content recommendations, trending lists, ad selection, etc. [3, 33, 34, 92]. For efficient low-latency distribution of the video content to viewers worldwide, the platform employs content delivery network (CDN) services from Akamai [1].

The platform's client-side app. The platform has its own *client-side application*, or simply *app*, available on smartphones, tablets, laptops, and other device types. The standalone app provides the viewer with an interface to the platform's cloud-based services. The supported functionalities include authenticated access to the viewer's account [70, 97], retrieval of personalized content recommendations, viewing history, and other particulars of the viewer's profile. The app also allows the viewer to supply feedback, e.g., to rate videos and submit reviews, which the platform utilizes as an input to its cloud-based recommendation engine. During regular streaming of a video to the viewer, the app informs the cloud-based ABR algorithm about the client-side stall duration and estimated throughput in real time by appending these data in the Common Media Client Data (CMCD) format [13] to the Uniform Resource Locator (URL) of each Hypertext Transfer Protocol Secure (HTTPS) message requesting a video chunk from Akamai. The CDN scalably communicates the client-side data to the EC2 instance running the ABR algorithm of the session.

iQoE integration. iQoE represents an addition to the platform's vast personalization portfolio. The platform deploys iQoE as part of its regular updates of the client-side app and cloud-based infrastructure. Whereas the platform already stores historical streaming traces in S3 and RDS for the advanced data analytics in Redshift to identify popular content, preferred genres, etc., the platform reuses the historical traces to compile the experience set that iQoE utilizes later to construct personalized QoE models for viewers. The platform composes the experience set offline, characterizes the video chunks of each experience in the set with their precomputed values of QoE influence factors, e.g., VMAF, and stores the experience set in S3. With 100 GB allocated to the experience set, i.e., 10 times more than in Section 5, the storage requirement remains a tiny fraction of the space consumed by the platform's current personalization tasks.

Viewer's role in the iQoE integration. After the app incorporates iQoE, the app's interface offers the viewer the option to build a personalized QoE model via iQoE. If the viewer exercises this option, the app constructs the QoE model as described in Section 4, i.e., by downloading from the cloud-stored experience set, playing back, and collecting the score for each of the experiences chosen by iQoE for this viewer. The app uploads the constructed QoE model to the EC2 instance that trains a personalized ABR algorithm based on the personalized QoE model. The QoE and ABR personalization incur acceptable storage and bandwidth overheads: whereas the personalized QoE and ABR models, respectively, consume about 20 KB in the viewer's device and 3 MB in the cloud, the total amount of data communicated between the cloud and app during the QoE modeling is around 500 MB. The platform offers the viewer the trained personalized ABR algorithm as one of ABR options, including the one-size-fits-all ABR algorithm, for the viewer's subsequent streaming sessions. The app enables the viewer's profile to store up to four personalized ABR algorithms so as to accommodate different genres and contexts, e.g., streaming to a smartphone or HDTV device.

While our paper deliberately differentiates between the typical and atypical viewers because the latter benefit more from QoE personalization and, thus, are more likely to adopt iQoE, the platform permits any viewer to take advantage of iQoE. To clarify the relative utility of iQoE for the viewer, the app's interface offers a similarity check between the viewer's personalized QoE model and the platform's one-size-fits-all QoE model.

Relationship with regular streaming. While the viewer dedicates time and effort to train the personalized QoE model via iQoE, the training of the ABR algorithm in EC2 occurs concurrently with regular streaming. Whenever the viewer wants to stream a video, the viewer launches the regular streaming by selecting one of the ABR algorithms available in the viewer's profile.

Application-layer operation. iQoE and the platform as a whole operate on the application layer and communicate over HTTPS. Before or after incorporating iQoE, the platform does not explicitly deal with network resource allocation or its fairness. The different application-layer transmission patterns under the personalized ABR algorithms contribute to the increasing diversity in network traffic, e.g., caused by the BBR [18] and CUBIC [36] congestion control algorithms which have different levels of aggressiveness.

Privacy. The platform's personalization of QoE modeling and ABR streaming strives for the same main goal as the platform's personalization efforts in general, i.e., enhancing the user experience. On the other hand, any personalization intrinsically raises concerns about privacy, albeit seemingly to a smaller extent for personalized QoE models than content preferences. The platform handles the extended set of concerns through its traditional methods of data protection and privacy control.

6.2 Extensions to Other Streaming Systems

Alternative implementations of ABR streaming. While Section 6.1 discusses integration of iQoE into a hypothetical video streaming platform that places the ABR logic in the cloud in alignment with current industry trends, iQoE also integrates easily with the traditional client-side ABR designs. The client-side ABR implementation diminishes the privacy concerns because the viewer retains the personalized QoE model. However, the local personalization of the RL-based ABR algorithm increases the load on the viewer's device. By adopting instead a control-theoretic ABR logic, e.g., BOLA [91], with the personalized QoE model as the optimization objective, the system decreases the client-side overhead.

Live streaming. In comparison to the video-on-demand streaming in Section 6.1, live streaming imposes different requirements, such as lower end-to-end latency. The system design also changes, e.g., the client discards late frames instead of stalling the playback until the late chunk arrives. Live-streaming systems, e.g., those leveraging WebRTC [19], integrate iQoE by using QoE models with different influence factors, such as the frame rate [38]. Necessary modifications also include techniques suitable for measuring QoE factors in real time, e.g., PSNR instead of VMAF for video quality. In live streaming, the personalized QoE models offered by iQoE provide a promising foundation for not only ABR decisions but also dynamic construction of bitrate ladders [95].

Volumetric video streaming, Virtual Reality (VR), and Augmented Reality (AR). iQoE integration into these emerging applications is conceptually similar to the discussed above. The main difference arises due to the need for distinct QoE models that account for dissimilar application-specific influence factors. Motion-to-photon latency, viewport drift, and point density exemplify the new relevant QoE influence factors [63, 110]. The design of QoE models for volumetric video streaming, VR, and AR is a vibrant research problem without many definitive conclusions so far.

Fairness of network resource allocation. An intriguing possibility is to leverage iQoE to improve fairness of network sharing, especially because the resource allocation in the current Internet falls far short of theoretical ideals such as max-main fairness [14]. Whereas [56, 69, 109] apply max-min fairness to QoE rather than flow rates, the personalized QoE models built by iQoE represent an alternative to the one-size-fits-all QoE model as the basis for QoE fairness. As a word of caution, QoE fairness is a controversial objective because of diminishing the incentives for an application to achieve high QoE by utilizing the available network bandwidth more efficiently.

7 RELATED AND FUTURE WORK

While [26, 41, 44, 49] show great **heterogeneity of QoE perception among humans**, our new dataset corroborates these findings. Unlike prior approaches to **QoE personalization** through indirect inference [21, 58, 67, 73, 100] or control knobs for a generic QoE model [49, 71, 74], iQoE is a novel personalization method that leverages a limited amount of explicit expressible feedback. Whereas [20, 59, 66] apply **active learning** to traditional MOS-based modeling, the focus of our work is on personalized QoE modeling. Future improvement of iQoE might benefit from **additional influence factors** that include personal traits [32, 101, 117], sensitivity [113], emotions [37, 45], and interests [31]. Although this paper deals mostly with QoE, our plans are to expand the work into **other aspects of video streaming** such as power consumption on mobile devices [57, 112], efficiency and fairness of network utilization [75, 76], and cross-layer design [114]. Also, while iQoE does not seem to raise any privacy concerns, this question deserves a deeper investigation.

8 CONCLUSION

One-size-fits-all QoE models built by traditional MOS-based methods misrepresent the QoE perception by an atypical viewer. Seeking to empower the atypical viewers, this paper proposes iQoE, a novel method that utilizes explicit, expressible, and actionable feedback from a viewer to construct a personalized QoE model for this viewer. iQoE combines the RIGS sampler with the XSVR modeler and exercises active learning so as to be sample-efficient and accurate. We use Microworkers to accomplish subjective studies with 120 raters who provide 14,400 individual scores. Based on the subjective studies, an iQoE session of about 22 minutes suffices for constructing an accurate personalized QoE model. Compared to the best of the 10 baseline models, iQoE delivers the average accuracy improvement of at least 42% for all viewers and at least 85% for the atypical viewers. The large-scale simulations support design choices and clarify performance properties of iQoE.

ACKNOWLEDGMENTS

The research is supported in part by the Spanish Ministry of Science and Innovation with grants TED2021-131264B-I00 (SocialProbing) and PID2021-128223OA-I00 (GreenEdge) and the Austrian Federal Ministry for Digital and Economic Affairs, National Foundation for Research, Technology and Development, and Christian Doppler Research Association with project Christian Doppler Laboratory ATHENA (https://athena.itec.aau.at/).

REFERENCES

[1] Akamai. 2023. Content Delivery Network (CDN). [Online] Available: https://www.akamai.com/solutions/content-delivery-network.

[2] Zahaib Akhtar, Sanjay Rao, Bruno Ribeiro, Yun Seong Nam, Jessica Chen, Jibin Zhan, Ramesh Govindan, Ethan Katz-Bassett, and Hui Zhang. 2018. Oboe: Auto-Tuning Video ABR Algorithms to Network Conditions. In *SIGCOMM 2018*.

[3] Xavier Amatriain and Justin Basilico. 2015. Recommender Systems in Industry: A Netflix Case Study. In *Recommender Systems Handbook*. Springer.

[4] Amazon Web Services. 2023. Amazon Elastic Compute Cloud (Amazon EC2). [Online] Available: https://aws.amazon.com/ec2/.

[5] Amazon Web Services. 2023. Amazon Redshift. [Online] Available: https://aws.amazon.com/redshift/.

[6] Amazon Web Services. 2023. Amazon Relational Database Service (Amazon RDS). [Online] Available: https://aws.amazon.com/rds/.

[7] Amazon Web Services. 2023. Amazon Simple Storage Service (Amazon S3). [Online] Available: https://aws.amazon.com/s3/.

[8] Mariette Awad and Rahul Khanna. 2015. *Efficient Learning Machines*. Springer.

[9] Christos G. Bampis and Alan C. Bovik. 2017. Learning to Predict Streaming Video QoE: Distortions, Rebuffering and Memory. *arXiv* 1703.00633 (2017).

[10] Christos G. Bampis, Zhi Li, Ioannis Katsavounidis, Te-Yuan Huang, Chaitanya Ekanadham, and Alan C. Bovik. 2021. Towards Perceptually Optimized Adaptive Video Streaming – A Realistic Quality of Experience Database. *IEEE TIP* 30 (2021), 5182–5197.

[11] Christos G. Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan C. Bovik. 2017. Study of Temporal Effects on Subjective Video Quality of Experience. *IEEE TIP* 26, 11 (2017), 5217–5231.

[12] Abdelhak Bentaleb, Ali C. Begen, and Roger Zimmermann. 2016. SDNDASH: Improving QoE of HTTP Adaptive Streaming Using Software Defined Networking. In *MM 2016*.

[13] Abdelhak Bentaleb, May Lim, Mehmet N. Akcay, Ali C. Begen, and Roger Zimmermann. 2021. Common Media Client Data (CMCD): Initial Findings. In *NOSSDAV 2021*.

[14] Dimitri Bertsekas and Robert Gallager. 1992. *Data Networks*. Prentice Hall.

[15] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32.

[16] Kjell Brunnström et al. 2013. Definitions of Quality of Experience. *Qualinet White Paper* (2013).

[17] Robert Burbidge, Jem J. Rowland, and Ross D. King. 2007. Active Learning for Regression Based on Query by Committee. In *IDEAL 2007*.

[18] Neal Cardwell, Yuchung Cheng, C. Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. 2016. BBR: Congestion-Based Congestion Control: Measuring Bottleneck Bandwidth and Round-Trip Propagation Time. *Queue* 14, 5 (2016), 20–53.

[19] Gaetano Carlucci, Luca De Cicco, Stefan Holmer, and Saverio Mascolo. 2016. Analysis and Design of the Google Congestion Control for Web Real-Time Communication (WebRTC). In *MMSys 2016*.

[20] Haw-Shiuan Chang, Chih-Fan Hsu, Tobias Hoßfeld, and Kuan-Ta Chen. 2018. Active Learning for Crowdsourced QoE Modeling. *IEEE TMM* 20, 12 (2018), 3337–3352.

[21] Kuan-Ta Chen, Cheng-Chun Tu, and Wei-Cheng Xiao. 2009. OneClick: A Framework for Measuring Network Quality of Experience. In *INFOCOM 2009*.

[22] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *KDD 2016*.

[23] Sang-Min Choi, Sang-Ki Ko, and Yo-Sub Han. 2012. A Movie Recommendation Algorithm Based on Genre Correlations. *Expert Systems with Applications* 39, 9 (2012), 8079–8085.

[24] Luca De Cicco, Vito Caldaralo, Vittorio Palmisano, and Saverio Mascolo. 2013. ELASTIC: A Client-Side Controller for Dynamic Adaptive Streaming over HTTP (DASH). In *PV 2013*.

[25] Ismael de Fez, Román Belda, and Juan Carlos Guerri. 2020. New Objective QoE Models for Evaluating ABR Algorithms in DASH. *Computer Communications* 158 (2020), 126–140.

[26] Zhengfang Duanmu, Wentao Liu, Zhuoran Li, Diqi Chen, Zhou Wang, Yizhou Wang, and Wen Gao. 2020. Assessing the Quality-of-Experience of Adaptive Bitrate Video Streaming. *arXiv* 2008.08804 (2020).

[27] Zhengfang Duanmu, Wentao Liu, Zhuoran Li, Diqi Chen, Zhou Wang, Yizhou Wang, and Wen Gao. 2020. The Waterloo Streaming Quality-of-Experience Database-IV. IEEE Dataport. https://dx.doi.org/10.21227/j15a-8r35.

[28] Zhengfang Duanmu, Abdul Rehman, and Zhou Wang. 2020. The Waterloo Streaming Quality-of-Experience Database-III. IEEE Dataport. https://dx.doi.org/10.21227/xzt6-p944.

[29] Zhengfang Duanmu, Kai Zeng, Kede Ma, Abdul Rehman, and Zhou Wang. 2017. A Quality-of-Experience Index for Streaming Video. *IEEE JSTSP* 11, 1 (2017), 154–166.

[30] Federal Communications Commission. 2020. Raw Data-Measuring Broadband America. https://www.fcc.gov/oet/mba/raw-data-releases.

[31] Guanyu Gao, Huaizheng Zhang, Han Hu, Yonggang Wen, Jianfei Cai, Chong Luo, and Wenjun Zeng. 2018. Optimizing Quality of Experience for Adaptive Bitrate Streaming via Viewer Interest Inference. *IEEE TMM* 20, 12 (2018), 3399–3413.

[32] Yun Gao, Xin Wei, and Liang Zhou. 2020. Personalized QoE Improvement for Networking Video Service. *IEEE JSAC* 38, 10 (2020), 2311–2323.

[33] Geeksforgeeks. 2023. System Design Netflix – A Complete Architecture. [Online] Available: https://www.geeksforgeeks.org/system-design-netflix-a-complete-architecture/.

[34] Geeksforgeeks. 2023. System Design of Youtube – A Complete Architecture. [Online] Available: https://www.geeksforgeeks.org/system-design-of-youtube-a-complete-architecture/.

[35] Deepti Ghadiyaram, Janice Pan, and Alan C. Bovik. 2019. A Subjective and Objective Study of Stalling Events in Mobile Streaming Videos. *IEEE TCSVT* 29, 1 (2019), 183–197.

[36] Sangtae Ha, Injong Rhee, and Lisong Xu. 2008. CUBIC: A New TCP-Friendly High-Speed TCP Variant. *SIGOPS Oper. Syst. Rev.* 42, 5 (2008), 64–74.

[37] Yixue Hao, Jun Yang, Min Chen, M. Shamim Hossain, and Mohammed F. Alhamid. 2019. Emotion-Aware Video QoE Assessment Via Transfer Learning. *IEEE MM* 26, 1 (2019), 31–40.

[38] Jian He, Mubashir Adnan Qureshi, Lili Qiu, Jin Li, Feng Li, and Lei Han. 2018. Favor: Fine-Grained Video Rate Adaptation. In *MMSys 2018*.

[39] Tin Kam Ho. 1998. The Random Subspace Method for Constructing Decision Forests. *IEEE TPAMI* 20, 8 (1998), 832–844.

[40] Dennis Hocevar. 1979. A Comparison of Statistical Infrequency and Subjective Judgment as Criteria in the Measurement of Originality. *JPA* 43, 3 (1979), 297–299.

[41] Tobias Hoßfeld, Poul E. Heegaard, Martín Varela, and Sebastian Möller. 2016. QoE Beyond the MOS: An In-Depth Look at QoE via Better Metrics and Their Relation to MOS. *Quality and User Experience* 1, 2 (2016), 1–23.

[42] Tobias Hossfeld, Raimund Schatz, Ernst Biersack, and Louis Plissonneau. 2013. Internet Video Delivery in YouTube: From Traffic Measurements to Quality of Experience. In *Data Traffic Monitoring and Analysis*. Springer.

[43] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. 2017. The Konstanz Natural Video Database. http://database.mmsp-kn.de

[44] Tobias Hoßfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS Is Not Enough!. In *QoMEX 2011*.

[45] Shenghong Hu, Min Xu, Haimin Zhang, Chunxia Xiao, and Chao Gui. 2019. Affective Content-Aware Adaptation Scheme on QoE Optimization of Adaptive Streaming over HTTP. *ACM TOMM* 15, 3s (2019), 1–18.

[46] Tianchi Huang, Chao Zhou, Xin Yao, Rui Xiao Zhang, Chenglei Wu, Bing Yu, and Lifeng Sun. 2020. Quality-Aware Neural Adaptive Video Streaming with Lifelong Imitation Learning. *IEEE JSAC* 38, 10 (2020), 2324–2342.

[47] Te Yuan Huang, Ramesh Johari, Nick McKeown, Matthew Trunnell, and Mark Watson. 2014. A Buffer-Based Approach to Rate Adaptation: Evidence From a Large Video Streaming Service. In *SIGCOMM 2014*.

[48] Xinyu Huang, Conghao Zhou, Wen Wu, Mushu Li, Huaqing Wu, and Xuemin Shen. 2022. Personalized QoE Enhancement for Adaptive Video Streaming: A Digital Twin-Assisted Scheme. In *GLOBECOM 2022*.

[49] Liangyu Huo, Zulin Wang, Mai Xu, Yong Li, Zhiguo Ding, and Hao Wang. 2020. A Meta-Learning Framework for Learning Multi-User Preferences in QoE Optimization of DASH. *IEEE TCSVT* 30, 9 (2020), 3210–3225.

[50] Quan Huynh-Thu and Mohammed Ghanbari. 2008. Scope of Validity of PSNR in Image/Video Quality Assessment. *Electronics Letters* 44, 13 (2008), 1–2.

[51] International Telecommunication Union. 2015. Reference Guide to Quality of Experience Assessment Methodologies. Recommendation G.1011.

[52] International Telecommunication Union. 2017. Parametric Bitstream-Based Quality Assessment Of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport. Recommendation P.1203.

[53] International Telecommunication Union. 2019. Methodology for Subjective Assessment of the Quality of Television Picture. Recommendation BT.500-14.

[54] International Telecommunication Union. 2022. Subjective Video Quality Assessment Methods for Multimedia Applications. Recommendation P.910.

[55] Junchen Jiang, Vyas Sekar, and Hui Zhang. 2012. Improving Fairness, Efficiency, and Stability in HTTP-Based Adaptive Video Streaming with FESTIVE. In *CoNEXT 2012*.

[56] Wanchun Jiang, Pan Ning, Zheyuan Zhang, Jintian Hu, Zhicheng Ren, and Jianxin Wang. 2021. Practical Bandwidth Allocation for Video QoE Fairness. In *WASA 2021*.

[57] Jiayi Meng, Qiang Xu, and Y. Charlie Hu. 2021. Proactive Energy-Aware Adaptive Video Streaming on Mobile Devices. In *USENIX ATC 2021*.

[58] Conor Keighrey, Ronan Flynn, Siobhan Murray, Sean Brennan, and Niall Murray. 2017. Comparing User QoE via Physiological and Interaction Measurements of Immersive AR and VR Speech and Language Therapy Applications. In *MM Workshops 2017*.

[59] Muhammad Jawad Khokhar, Thierry Spetebroot, and Chadi Barakat. 2018. An Online Sampling Approach for Controlled Experimentation and QoE Modeling. In *ICC 2018*.

[60] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. 2016. Toward A Practical Perceptual Video Quality Metric. Netflix Technology Blog. https://medium.com/netflix-techblog/toward-a-practical-perceptual-video-quality-metric-653f208b9652.

[61] Zhi Li, Xiaoqing Zhu, Joshua Gahm, Rong Pan, Hao Hu, Ali C. Begen, and David Oran. 2014. Probe and Adapt: Rate Adaptation for HTTP Video Streaming at Scale. *IEEE JSAC* 32, 4 (2014), 719–733.

[62] Chenghao Liu, Imed Bouazizi, and Moncef Gabbouj. 2011. Rate Adaptation for Adaptive HTTP Streaming. In *MMSys 2011*.

[63] Yu Liu, Bo Han, Feng Qian, Arvind Narayanan, and Zhi-Li Zhang. 2022. Vues: Practical Mobile Volumetric Video Streaming through Multiview Transcoding. In *MOBICOM 2022*.

[64] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *SIGCOMM 2017*.

[65] Hongzi Mao et al. 2019. Park: An Open Platform for Learning-Augmented Computer Systems. In *NeurIPS 2019*.

[66] Vlado Menkovski, Georgios Exarchakos, and Antonio Liotta. 2012. Tackling the Sheer Scale of Subjective QoE. In *MobiMedia 2011*.

[67] Ricky K. P. Mok, Edmond W. W. Chan, Xiapu Luo, and Rocky K. C. Chang. 2011. Inferring the QoE of HTTP Video Streaming from User-Viewing Activities. In *W-MUST 2011*.

[68] Christopher Mueller, Stefan Lederer, Reinhard Grandl, and Christian Timmerer. 2015. Oscillation Compensating Dynamic Adaptive Streaming over HTTP. In *ICME 2015*.

[69] Vikram Nathan, Vibhaalakshmi Sivaraman, Ravichandra Addanki, Mehrdad Khani, Prateesh Goyal, and Mohammad Alizadeh. 2019. End-to-End Transport for Video QoE Fairness. In *SIGCOMM 2019*.

[70] Netflix Inc. 2021. Edge Authentication and Token-Agnostic Identity Propagation. [Online] Available: https://netflixtechblog.com/edge-authentication-and-token-agnostic-identity-propagation-514e47e0b602.

[71] Minh Nguyen, Ekrem Çetinkaya, Hermann Hellwagner, and Christian Timmerer. 2011. WISH: User-Centric Bitrate Adaptation for HTTP Adaptive Streaming on Mobile Devices. In *MMSP 2021*.

[72] Leonardo Peroni, Sergey Gorinsky, Farzad Tashtarian, and Christian Timmerer. 2023. iQoE Dataset and Code. GitHub. https://github.com/Leo-rojo/iQoE_Dataset_and_Code.

[73] Simone Porcu, Alessandro Floris, and Luigi Atzori. 2019. Towards the Prediction of the Quality of Experience from Facial Expression and Gaze Direction. In *ICIN 2019*.

[74] Chunyu Qiao, Jiliang Wang, and Yunhao Liu. 2021. Beyond QoE: Diversity Adaptation in Video Streaming at the Edge. *IEEE/ACM ToN* 29, 1 (2021), 289–302.

[75] Yanyuan Qin, Shuai Hao, Krishna R. Pattipati, Feng Qian, Subhabrata Sen, Bing Wang, and Chaoqun Yue. 2019. Quality-Aware Strategies for Optimizing ABR Video Streaming QoE and Reducing Data Usage. In *MMSys 2019*.

[76] Jason J. Quinlan, Ahmed H. Zahran, K. K. Ramakrishnan, and Cormac J. Sreenan. 2015. Delivery of Adaptive Bit Rate Video: Balancing Fairness, Efficiency and Quality. In *LANMAN 2015*.

[77] Benjamin Rainer and Christian Timmerer. 2014. Quality of Experience of Web-Based Adaptive HTTP Streaming Clients in Real-World Environments Using Crowdsourcing. In *VideoNext 2014*.

[78] Carl Edward Rasmussen and Christopher K. I. Williams. 2005. *Gaussian Processes for Machine Learning*. MIT Press.

[79] Devdeep Ray, Jack Kosaian, K. V. Rashmi, and Srinivasan Seshan. 2019. Vantage: Optimizing Video Upload for Time-Shifted Viewing of Social Live Streams. In *SIGCOMM 2019*.

[80] Ulrich Reiter et al. 2014. Factors Influencing Quality of Experience. In *Quality of Experience: Advanced Concepts, Applications and Methods*. Springer.

[81] Haakon Riiser, Paul Vigmostad, Carsten Griwodz, and Pål Halvorsen. 2013. Commute Path Bandwidth Traces from 3G Networks: Analysis and Applications. In *MMSys 2013*.

[82] Werner Robitza et al. 2018. HTTP Adaptive Streaming QoE Estimation with ITU-T Rec. P.1203 – Open Databases and Software. In *MMSys 2018*.

[83] Michael Rudow, Francis Y. Yan, Abhishek Kumar, Ganesh Ananthanarayanan, Martin Ellis, and K. V. Rashmi. 2023. Tambur: Efficient Loss Recovery for Videoconferencing via Streaming Codes. In *NSDI 2023*.

[84] Anika Seufert, Florian Wamser, David Yarish, Hunter Macdonald, and Tobias Hoßfeld. 2021. QoE Models in the Wild: Comparing Video QoE Models Using a Crowdsourced Data Set. In *QoMEX 2021*.

[85] Zaixi Shang, Joshua Peter Ebenezer, Yongjun Wu, Hai Wei, Sriram Sethuraman, and Alan C. Bovik. 2022. Study of the Subjective and Objective Quality of High Motion Live Streaming Videos. *IEEE TIP* 31 (2022), 1027–1041.

[86] Hamid R. Sheikh, Muhammad F. Sabir, and Alan C. Bovik. 2006. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE TIP* 15, 11 (2006), 3440–3451.

[87] Russell Shirey, Sanjay Rao, and Shreyas Sundaram. 2021. Optimizing Quality of Experience for Long-Range UAS Video Streaming. In *IWQOS 2021*.

[88] Roee Shraga, Gil Katz, Yael Badian, Nitay Calderon, and Avigdor Gal. 2021. From Limited Annotated Raw Material Data to Quality Production Data: A Case Study in the Milk Industry. In *CIKM 2021*.

[89] Iraj Sodagar. 2011. The MPEG-DASH Standard for Multimedia Streaming Over the Internet. *IEEE MM* 18, 4 (2011), 62–67.

[90] Kevin Spiteri, Ramesh K. Sitaraman, and Daniel Sparacio. 2019. From Theory to Practice: Improving Bitrate Adaptation in the DASH Reference Player. *ACM TOMM* 15, 2s (2019), 1–29.

[91] Kevin Spiteri, Rahul Urgaonkar, and Ramesh K. Sitaraman. 2016. BOLA: Near-Optimal Bitrate Adaptation for Online Videos. In *INFOCOM 2016*.

[92] Harald Steck, Linas Baltrunas, Ehtsham Elahi, Dawen Liang, Yves Raimond, and Justin Basilico. 2021. Deep Learning for Recommender Systems: A Netflix Case Study. *AI Mag.* 42, 3 (2021), 7–18.

[93] Yi Sun, Xiaoqi Yin, Junchen Jiang, Vyas Sekar, Fuyuan Lin, Nanshu Wang, Tao Liu, and Bruno Sinopoli. 2016. CS2P: Improving Video Bitrate Selection and Adaptation with Data-Driven Throughput Prediction. In *SIGCOMM 2016*.

[94] Babak Taraghi, Minh Nguyen, Hadi Amirpour, and Christian Timmerer. 2021. INTENSE: In-Depth Studies on Stall Events and Quality Switches and Their Impact on the Quality of Experience in HTTP Adaptive Streaming. *IEEE Access* 9 (2021), 118087–118098.

[95] Farzad Tashtarian, Abdelhak Bentaleb, Hadi Amirpour, Sergey Gorinsky, Junchen Jiang, Hermann Hellwagner, and Christian Timmerer. 2024. ARTEMIS: Adaptive Bitrate Ladder Optimization for Live Video Streaming. In *NSDI 2024*.

[96] Huyen T. T. Tran, Duc V. Nguyen, Nam Pham Ngoc, and Truong Cong Thang. 2021. Overall Quality Prediction for HTTP Adaptive Streaming Using LSTM Network. *IEEE TCSVT* 31, 8 (2021), 3212–3226.

[97] Twitch. 2019. How Twitch Addresses Scalability and Authentication. [Online] Available: https://blog.twitch.tv/en/2019/03/15/how-twitch-addresses-scalability-and-authentication/.

[98] Raza Ul Mustafa, Simone Ferlin, Christian Esteve Rothenberg, Darijo Raca, and Jason J. Quinlan. 2020. A Supervised Machine Learning Approach for DASH Video QoE Prediction in 5G Networks. In *Q2SWinet 2020*.

[99] Talha Waheed, Ihsan Ayyub Qazi, Zahaib Akhtar, and Zafar Ayyub Qazi. 2022. Coal Not Diamonds: How Memory Pressure Falters Mobile Video QoE. In *CoNEXT 2022*.

[100] Shibo Wang, Shusen Yang, Hairong Su, Cong Zhao, Chenren Xu, Feng Qian, Nanbin Wang, and Zongben Xu. 2023. Robust Saliency-Driven Quality Adaptation for Mobile 360-Degree Video Streaming. *IEEE TMC* (2023), 1–18.

[101] Ying Wang, Peilong Li, Lei Jiao, Zhou Su, Nan Cheng, Xuemin Sherman Shen, and Ping Zhang. 2017. A Data-Driven Architecture for Personalized QoE Management in 5G Wireless Networks. *IEEE Wireless Communications* 24, 1 (2017), 102–110.

[102] Zhou Wang, Alan C. Bovik, Hamid Sheikh, and Eero Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE TIP* 13, 4 (2004), 600–612.

[103] Weblabcenter. 2009. Microworkers. [Online] Available: https://www.microworkers.com/.

[104] Dongrui Wu, Chin-Teng Lin, and Jian Huang. 2019. Active Learning for Regression Using Greedy Sampling. *Information Sciences* 474 (2019), 90–105.

[105] Zhiyuan Xu, Dejun Yang, Jian Tang, Yinan Tang, Tongtong Yuan, Yanzhi Wang, and Guoliang Xue. 2021. An Actor-Critic-Based Transfer Learning Framework for Experience-Driven Networking. *IEEE/ACM ToN* 29, 1 (2021), 360–371.

[106] Francis Y. Yan, Hudson Ayers, Chenzhi Zhu, Sadjad Fouladi, James Hong, Keyi Zhang, Philip Levis, and Keith Winstein. 2020. Learning in Situ: A Randomized Experiment in Video Streaming. In *NSDI 2020*.

[107] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. 2015. A Control-Theoretic Approach for Dynamic Adaptive Video Streaming over HTTP. In *SIGCOMM 2015*.

[108] Hwanjo Yu and Sungchul Kim. 2010. Passive Sampling for Regression. In *ICDM 2010*.

[109] Yali Yuan, Weijun Wang, Yuhan Wang, Sripriya Srikant Adhatarao, Bangbang Ren, Kai Zheng, and Xiaoming Fu. 2023. Joint Optimization of QoE and Fairness for Adaptive Video Streaming in Heterogeneous Mobile Environments. *IEEE/ACM ToN* (2023), 1–15.

[110] Anlan Zhang, Chendong Wang, Bo Han, and Feng Qian. 2022. YuZu: Neural-Enhanced Volumetric Video Streaming. In *NSDI 2022*.

[111] Huaizheng Zhang, Linsen Dong, Guanyu Gao, Han Hu, Yonggang Wen, and Kyle Guan. 2020. DeepQoE: A Multimodal Learning Framework for Video Quality of Experience (QoE) Prediction. *IEEE TMM* 22, 12 (2020), 3210–3223.

[112] Jingyu Zhang, Zhi-Jie Wang, Song Guo, Dingyu Yang, Gan Fang, Chunyi Peng, and Minyi Guo. 2018. Power Consumption Analysis of Video Streaming in 4G LTE Networks. *Wireless Network* 24, 8 (2018), 3083–3098.

[113] Xu Zhang, Yiyang Ou, Siddhartha Sen, and Junchen Jiang. 2021. SENSEI: Aligning Video Streaming Quality with Dynamic User Sensitivity. In *NSDI 2021*.

[114] Anfu Zhou, Huanhuan Zhang, Guangyuan Su, Leilei Wu, Ruoxuan Ma, Zhen Meng, Xinyu Zhang, Xiufeng Xie, Huadong Ma, and Xiaojiang Chen. 2019. Learning to Coordinate Video Codec with Transport Protocol for Mobile Video Telephony. In *MOBICOM 2019*.

[115] Renjie Zhou, Samamon Khemmarat, and Lixin Gao. 2010. The Impact of YouTube Recommendation System on Video Views. In *IMC 2010*.

[116] Huadi Zhu, Tianhao Li, Chaowei Wang, Wenqiang Jin, Srinivasan Murali, Mingyan Xiao, Dongqing Ye, and Ming Li. 2022. EyeQoE: A Novel QoE Assessment Model for 360-Degree Videos Using Ocular Behaviors. *ACM IMWUT* 6, 1 (2022), 1–26.

[117] Yi Zhu, Alan Hanjalic, and Judith A. Redi. 2016. QoE Prediction for Enriched Assessment of Individual Video Viewing Experience. In *MM 2016*.

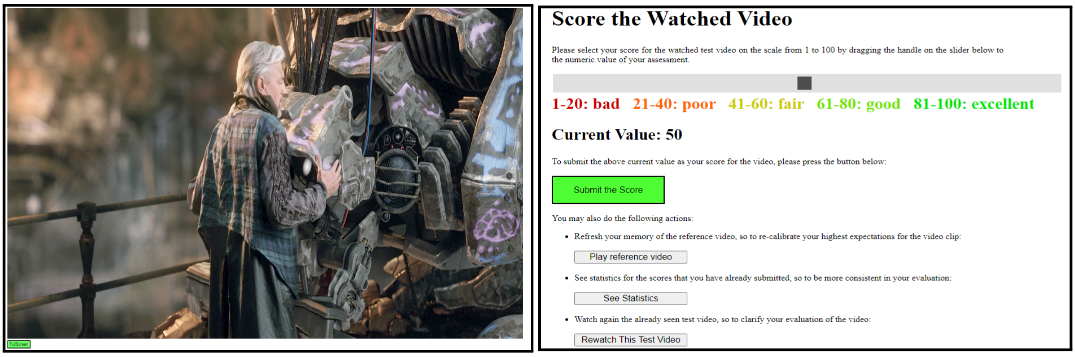
A XSVR FEATURE IMPORTANCE

The technique of permutation feature importance randomly shuffles the values of a feature and measures the impact of the disruption on the predictive power of the model [15]. The measurements reveal the importance of features for the predictive ability. In Section 4.3, we apply this technique to select influence factors for the XSVR modeler. For each of the four atypical raters in the Waterloo-IV dataset, we train XSVR separately and repeat the experiment 30 times. Fig. 4 in Section 4.3 demonstrates that all 10 influence factors of the X set contribute positively to the predictive power of XSVR. Influence factors 8 (PSNR), 1 (representation identifier), 2 (stall duration), and 10 (VMAF) are the most important as their disruption increases average RMSE across the four raters by 2.30, 1.90, 1.79, and 1.62 and average MAE by 2.19, 2.48, 1.58, and 2.33, respectively. On the other hand, influence factors 3 (bitrate) and 4 (chunk size) are relatively the least important, as their random shuffling worsens average RMSE by 0.18 for either of them and average MAE by 0.63 and 0.61, respectively.

B WEBSITE DESIGN

Each rater accesses the first page of the website through a directly provided link and accepts terms and conditions before commencing a session. Then, the website creates a new folder tracking the session status and assigns a random unique identifier to the session. The subsequent page offers the rater to watch a reference experience. This reference experience has the maximum quality among all elements of the experience set and helps the rater to calibrate the highest expectations for the experiences watched during actual assessments.

After this introductory stage, the website takes the rater to a *current status page* where the rater can monitor the progress of the session, watch the reference experience, monitor own scoring history, pause the session, or watch a new experience. Selecting the latter option opens a *playback window* that automatically starts playing back the new experience to the rater. Because modern browsers require an explicit command from the viewer to enlarge video dimensions, we do not put the playback window into the full-screen mode automatically. Instead, the playback window contains a button for full-screen toggling. We remove the control bar from the playback window to prevent unwanted behaviours, such as skipping an experience to the end without watching the experience. Fig. 14a shows a screenshot of the playback window with the full-screen button in the bottom left corner and a scene from *Tears of Steel*. Once the experience playback ends, the browser opens a *scoring page* where the rater can provide a score of the experience by sliding a bar along the 1-100 scale, watch the reference experience, monitor own scoring history, or rewatch



Score the Watched Video

Please select your score for the watched test video on the scale from 1 to 100 by dragging the handle on the slider below to the numeric value of your assessment.

1-20: bad 21-40: poor 41-60: fair 61-80: good 81-100: excellent

Current Value: 50

To submit the above current value as your score for the video, please press the button below:

You may also do the following actions:

- Refresh your memory of the reference video, so to re-calibrate your highest expectations for the video clip:

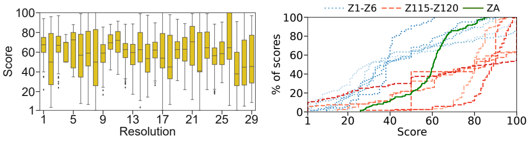
- See statistics for the scores that you have already submitted, so to be more consistent in your evaluation:

- Watch again the already seen test video, so to clarify your evaluation of the video:

(a) Playback window (b) Scoring page

Fig. 14. Screenshots of website pages.

the latest experience. Fig. 14b depicts the scoring page. Upon the score submission, the browser redirects the rater to a *current status page* which displays the updated status of the rater's session. This cycle repeats until the playback window finishes playing back the last experience of the series. At each iteration, the website saves information about the rater's session in .txt and .npy files and the current QoE model in the .pkl format. The website implementation and deployment use Python 3.7 with Flask 1.1.2 for the backend, HyperText Markup Language (HTML), JavaScript, and Cascading Style Sheets (CSS) for the frontend, and Apache 2.4.29 server on an Ubuntu 18.04.6 machine.



(a) Screen resolution (b) 12 atypical raters

Fig. 15. Extra insights into the collected dataset.

The rater's machine installs cookies to enable the rater to resume a previously paused session within three days. At the end of the assessment series, the website asks the rater to fill in a form with the following information: the rater's home country, the country where the rater takes the assessment series, gender, age, viewing device, level of satisfaction with the assessment series, and any optional suggestions.

C EXPERIENCE SET

To generate the 1,000-element experience set for our subjective studies in Section 5.1, we utilize the Park platform [65]. Park simulates ABR streaming over a network trace that specifies dynamic network bandwidth available for streaming. We experiment with TR [90], BBA [47], and MPC [107] and select 34 real-world network traces from each of the FCC [30], Norway [81], and Oboe [2] datasets, i.e., 102 network traces in total. The used bitrate ladder consists of 13 levels where the resolution and bitrate range from 320×180 and 235 Kbps to 3,840×2,160 and 16,800 Kbps, respectively [26]. To engender experiences with diverse scenes and increase the rater's involvement, we adopt *Tears of Steel* without its closing credits as the source video. The running time of the video is about 10 minutes. We segment the video into 294 chunks by applying the FFmpeg and MP4Box tools and transcode each chunk as per the used bitrate ladder. Our application of Park to the segmented video produces 306 experiences. Under the constraints that an extracted experience has stalls only between its chunks and that the total stall duration of the experience does not exceed 50% of the original running time without stalls, we randomly extract from the 306 long experiences a set of 1,000 short experiences. Each of these short experiences consists of four chunks and has the total playback time of 8 s without stalls.

D DATASET DETAILS

We report more details of our dataset described in Section 5.1. The answers to the post-assessment survey indicate that 94% and 6% of the respondents complete the assessment series on a personal computer and phone, respectively. 96% and 4% of the respondents view the experiences in Google Chrome and Mozilla Firefox, respectively. We also track the screen resolution of each respondent's viewing device, detect 29 different resolutions altogether, and label them with the following resolution identifiers: (1) 360×640, (2) 360×800, (3) 384×857, (4) 393×873, (5) 412×915, (6) 1,028×578, (7) 1,093×615, (8) 1,241×697, (9) 1,280×720, (10) 1,280×800, (11) 1,360×768, (12) 1,366×768, (13) 1,440×900, (14) 1,455×819, (15) 1,512×982, (16) 1,536×864, (17) 1,536×960, (18) 1,595×897, (19) 1,600×900, (20) 1,680×1050, (21) 1,707×1067, (22) 1,728×1,117, (23) 1,856×1018, (24) 1,920×1,080, (25) 1,920×1,200, (26) 1,928×970, (27) 2,048×1,152, (28) 2,560×1,440, and (29) 3,840×2,160. Resolutions 12, 16, and 24 are the most popular and account for 22%, 17%, and 18% of all resolutions, respectively. Fig. 15a depicts the respondents' scores grouped according to the screen resolution and reveals that the

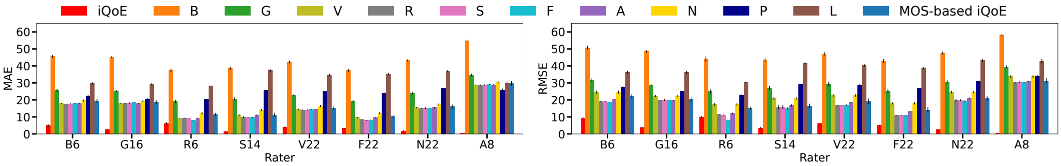


Fig. 16. iQoE vs. baseline MOS-based QoE models in the simulations.

resolution affects the QoE perception relatively mildly, with somewhat lower scores for the three highest screen resolutions. Fig. 15b shows that 12 atypical raters Z1 through Z6 and Z115 through Z120 of our dataset differ significantly in their QoE perception from average rater ZA.

E SIMULATION METHODOLOGY

To denote a synthetic rater, we combine the label of the respective synthetic profile with the number in the name of the real rater on whose individual scores we train this synthetic rater. For example, synthetic rater F32 corresponds to synthetic profile F and real rater H32. In the simulations, each experience contains seven chunks, and the playback of each chunk takes 4 s without stalls.

Because a profile aggregates the QoE perception by the raters in the reference group behind the respective QoE model, the usage of multiple diverse profiles allows the technique of synthetic profiling to realistically enhance the heterogeneity of the QoE perception in the simulations. The heterogeneity of the profiles also creates complications because the models vary in their ranges of produced scores. To emulate scores given by real raters, we define a scoring function for each synthetic rater by mapping the respective QoE model into the same scoring scale. Without loss of generality, we employ the 1-100 scale and map QoE models Q into scoring functions S through the following equation:

$$S = 1 + \frac{99}{1 + e^{-(Q-\sigma)\rho}} \tag{3}$$

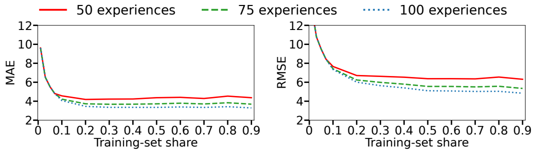
where least-squares minimization between Q values and assessment scores by the corresponding raters configures parameters σ and ρ.

We implement iQoE in Python 3.7. The implementations of RS, GS, UC, QBC, IGS and RIGS utilize the modAL framework. The implementations of SVR, RF, GP, and XGB are from the scikit-learn library with their optimal hyperparameters determined by grid search. The experiments run on an Intel i7 machine that has six cores, 2.6-GHz CPUs, 16-GB RAM, and Windows 10, with each experiment repeated five times by shuffling the experience set to improve generalizability.

F EVALUATION OF IQOE VS. MOS-BASED MODELING

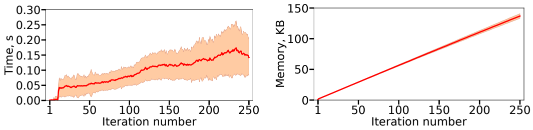
In this set of simulations, each of the 256 raters from Section 5.2 scores the 700 experiences in the training set. The respective 700 MOSes guide the construction of the eight MOS-based models, while models L and P use their default configurations. iQoE trains its personalized models on rater-specific sets of 50 experiences.

Fig. 16 evaluates the iQoE and MOS-based models in worst-case scenarios for iQoE. From each of the eight rater profiles, we pick the rater whose QoE model has the largest MAE within that profile, e.g., rater B6 within profile B. Even in these worst-case scenarios for iQoE, the personalized QoE modeling decreases both MAE and RMSE substantially compared to the baseline MOS-based models. On average over all 256 raters, iQoE reduces MAE by a factor of 7.63, 7.38, 4.53, 3.57, 2.79, 2.60, 2.60, 2.55, 2.43, and 2.34 in comparison to MOS-based models L, B, P, G, F, N, V, S, R, and A, respectively. The corresponding reduction factors for RMSE are 5.74, 5.81, 3.54, 3.04, 2.33, 2.24, 2.24, 2.14, 2.03, and 1.95. Hence, in the simulations, iQoE improves the model accuracy over the MOS-based models by a factor of at least 2.34 and 1.95 in MAE and RMSE, respectively.



(a) MAE sensitivity (b) RMSE sensitivity

Fig. 17. iQoE sensitivity to the training-set share.



(a) Execution time (b) Memory consumption

Fig. 18. iQoE processing and memory overhead.

Fig. 16 also plots, as the rightmost bar in each bar group, the model accuracy of a *MOS-based iQoE* variant that constructs a QoE model based on the MOSes instead of the individual scores by a rater. Despite being trained on 50 experiences only, the MOS-based iQoE variant provides a similar model accuracy as the best of the 10 existing MOS-based QoE models.

G SENSITIVITY TO THE TRAINING SHARE OF THE EXPERIENCE SET

By default, our evaluation uses 70% of the 1,000-element experience set for training, i.e., the training-set share equals 0.7. Fig. 17 shows that average MAE and RMSE of the personalized QoE models are largely insensitive to the training-set share. Only when the training-set share decreases below 0.2 (i.e., 200 experiences), the inaccuracy starts to ramp up. This result opens the *possibility to train iQoE on a smaller experience set*, which has a positive effect of reducing the computational and storage overhead.

H IQOE OVERHEAD

iQoE employs an iterative design where each iteration entails selection of an experience by RIGS, assessment of the experience by the rater, and update of the QoE model by XSVR. We evaluate overhead for the automated part of the iQoE method per iteration, from obtaining the rater's score of an experience until selecting the next experience for the rater.

We measure execution time of every iQoE iteration for each rater. Considering all 256 raters, Fig. 18a plots the average and standard deviation of the per-iteration execution time as a function of the iteration number. The execution time grows in a nearly linear pattern from iteration 11 to iteration 241, which reflects the linear increase in the number of the accumulated assessment scores. At iteration 11, the execution time exhibits a perceptible step-up because RIGS switches from RS to IGS. At iteration 241, the execution time slightly decreases due to issues related to XSVR training. At iteration 50, which corresponds to the default number of 50 experiences, the average execution time reaches 54 ms, and the respective wait of the rater for the next experience still remains insignificant. Because each experience lasts 28 s, 50 experiences take at least 23 minutes, not accounting for unavoidable extra delays on the rater's side, such as additional time to provide the scores. The automated portion of the iQoE algorithm consumes a total of 1.82 s, which constitutes at most 0.13% of the total time to construct the QoE model.

Because memory pressure might degrade video streaming performance [99], we also track the memory consumed by the personalized QoE model during its construction. For all 256 raters, Fig. 18b depicts the average and standard deviation of the memory consumption, which grows in a nearly linear fashion due to the increasing complexity of the refined QoE model. In the default setting with 50 experiences, the personalized QoE model consumes 20 KB on average, which is an extreme small amount by current standards. Overall, the results confirm that *the time and memory overhead of constructing an accurate personalized QoE model via iQoE is affordable.*

Received July 2023; revised September 2023; accepted October 2023.