IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 July 2022, accepted 21 July 2022, date of publication 1 August 2022, date of current version 4 August 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3195527

RESEARCH ARTICLE

AVQBits–Adaptive Video Quality Model Based on Bitstream Information for Various Video Applications

RAKESH RAO RAMACHANDRA RAO, STEVE GÖRING, AND ALEXANDER RAAKE, (Member, IEEE)
Audiovisual Technology Group, Technische Universität Ilmenau, 98693 Ilmenau, Germany
Corresponding author: Rakesh Rao Ramachandra Rao (rakesh-rao.ramachandra-rao@tu-ilmenau.de)

We acknowledge support for the publication costs by the Open Access Publication Fund of the Technische Universität Ilmenau.

ABSTRACT The paper presents *AVQBits*, a versatile, bitstream-based video quality model. It can be applied in several contexts such as video service monitoring, evaluation of video encoding quality, of gaming video QoE, and even of omnidirectional video quality. In the paper, it is shown that *AVQBits* predictions closely match video quality ratings obained in various subjective tests with human viewers, for videos up to 4K-UHD resolution (Ultra-High Definition, 3840 x 2180 pixels) and framerates up 120 fps. With the different variants of *AVQBits* presented in the paper, video quality can be monitored either at the client side, in the network or directly after encoding. The no-reference *AVQBits* model was developed for different video services and types of input data, reflecting the increasing popularity of Video-on-Demand services and widespread use of HTTP-based adaptive streaming. At its core, *AVQBits* encompasses the standardized ITU-T P.1204.3 model, with further model instances that can either have restricted or extended input information, depending on the application context. Four different instances of *AVQBits* are presented, that is, a Mode 3 model with full access to the bitstream, a Mode 0 variant using only metadata such as codec type, framerate, resoution and bitrate as input, a Mode 1 model using Mode 0 information and frame-type and -size information, and a Hybrid Mode 0 model that is based on Mode 0 metadata and the decoded video pixel information. The models are trained on the authors' own AVT-PNATS-UHD-1 dataset described in the paper. All models show a highly competitive performance by using AVT-VQDB-UHD-1 as validation dataset, e.g., with the Mode 0 variant yielding a value of 0.890 Pearson Correlation, the Mode 1 model of 0.901, the hybrid no-reference mode 0 model of 0.928 and the model with full bitstream access of 0.942. In addition, all four *AVQBits* variants are evaluated when applying them out-of-the-box to different media formats such as 360° video, high framerate (HFR) content, or gaming videos. The analysis shows that the ITU-T P.1204.3 and Hybrid Mode 0 instances of *AVQBits* for the considered use-cases either perform on par with or better than even state-of-the-art full reference, pixel-based models. Furthermore, it is shown that the proposed Mode 0 and Mode 1 variants outperform commonly used no-reference models for the different application scopes. Also, a long-term integration model based on the standardized ITU-T P.1203.3 is presented to estimate ratings of overall audiovisual streaming Quality of Experience (QoE) for sessions of 30 s up to 5 min duration. In the paper, the *AVQBits* instances with their per-1-sec score output are evaluated as the video quality component of the proposed long-term integration model. All *AVQBits* variants as well as the long-term integration module are made publicly available for the community for further research.

INDEX TERMS Bitstream video quality models, quality of experience (QoE), quality assessment, HTTP-based adaptive streaming (HAS), hybrid models, video quality, 360°, HFR, gaming, overall integral quality.

The associate editor coordinating the review of this manuscript and approving it for publication was Zhen Ren.

I. INTRODUCTION

The increase in both affordable capture technology and the average bandwidth for internet connections has resulted in

video becoming the most dominant of all the data that is uploaded, shared, and streamed on the internet. For example, Cisco's global forecast for 2022 indicate that video traffic accounts for 82% of all consumer traffic in 2022 up from 75% in 2017 [1]. In addition, both production and streaming of higher resolution videos (UHD-1/4K and above) is increasing [2], [3] and is being supported by many leading over-the-top (OTT) streaming providers such as Netflix, YouTube, Amazon Prime Video etc. Moreover, the overall volume of video traffic is not limited to traditional 2D video content and there is a significant increase to stream gaming [4], [5] or 360° videos.

The preferred streaming technology for most of this video traffic is HTTP-based adaptive streaming (HAS). A typical HAS session can be characterised by variations in audio and video quality, also referred to as ''quality switches'', initial loading delay and stalling events due to rebuffering. Various implemetations of this technology are in use, such as, e.g. Apple's HLS (HTTP Live Streaming) or Microsoft Smooth Streaming (MSS). Furthermore, advances in video compression have taken place, with the development of video codecs such as H.264 [6], VP9 [7], HEVC [8], EVC [9], LCEVC [10], AV1 [11], and VVC [12] to compress videos efficiently while still being able to maintain a good quality for a specific target bitrate.

All these developments are targeted towards delivering video at the best quality for the lowest possible bandwidth, to increase the Quality of Experience (QoE) for the viewers. In this paper, the encoding video quality is addressed as one of the main factors for video QoE. In this context, different strategies have been proposed to select the optimal encoding settings, e.g. ''fixed'' bitrate ladder [13], per-title encoding [14], and shot-based encoding [15] from Netflix, or context-aware encoding [16]. This necessitates the development of models that can provide accurate prediction of video quality. Those models can be used to evaluate encoding strategies or to monitor quality on the client side.

In general, based on the input information used for quality assessment, video quality models can be classified into several categories [17]–[22], for example, metadata-based, pixel-based, bitstream-based, or hybrid models. For the first type, namely, metadata-based models, the input information available is limited to video resolution, video bitrate, video framerate, and video codec. Since no information related to the underlying video content is available to the model, these models are content-agnostic. The second type, pixel-based models, have access to pixel information to estimate video quality. Based on the availability of pixel information of a reference (undistorted source) video, pixel-based models can be further sub-divided into three categories: Full-Reference (FR) (e.g. VMAF [23]), Reduced-Reference (RR) (e.g. ITU-T P.1204.4 [24]), and No-Reference (NR) (e.g. Deviq [25]) models.

The third type of models is referred to as bitstream-based, which are usually NR models that rely only on the encoded bitstream, typically without a full decoding, to estimate visual quality. Based on the extent to which bitstream information is used, bitstream models can be categorized into the following modes of operation [22]: Mode 0 (e.g. ITU P.1203.1 Mode 0 [26]), Mode 1 (e.g. ITU P.1203.1 Mode 1 [26]), and Mode 3 (e.g. ITU-T P.1203.1 Mode 3 [26], ITU-T P.1204.3 [27]) models. The fourth type of video quality models are hybrid models. Here, usually bitstream and pixel information are combined to estimate video quality.

In addition to measuring video quality, for a holistic evaluation of QoE of a HAS session, the impact of the four main factors, namely, audio and video quality, initial loading, and stalling have to be considered [28]–[35]. Similar to the estimation of short-term video quality, the integral quality of a streaming session can be evaluated, e.g., using subjective tests. Due to the longer viewing time per session, the number of videos that can be assessed in a given test is even more limited than for short-sequence tests, making these tests even more resource-demanding. Therefore, dedicated models have been developed for the case of long-term integral quality estimation, complementing or replacing subjective tests. In this context, ITU-T P.1203.3 [36], [37] was the first standardized model to assess the integral QoE of HAS sessions, integrating the effects of audio and video quality over time, initial loading delay and stalling events occurring during the session under consideration. The development of P.1203 mainly focused on the long-term QoE prediction (see also [38]) for Full-HD and H.264 encoded videos.

To cover newer videos codecs and higher resolutions, newer video quality short-term models were poposed in a subsequent standardization competition in ITU-T Study Group 12 (SG12). One of the standardized models is the high-accuracy, bitstream-based video quality model P.1204.3 (Mode 3) [22], [27], [39]. It is the candidate by the authors of this paper winning the bitstream-model competition in ITU-T SG12. In its standardized form, ITU-T P.1204.3 shows highly accurate prediction results for short term video quality prediction, even in comparison to FR models such as VMAF, see [22], [39]. P.1204.3 is one variant of the novel video quality model *AVQBits* (Audiovisual Video Quality based on Bitstreams) presented in this paper. The model instance ''*AVQBits*|*M*3 / P.1204.3'' has only been trained and validated for traditional 2D video contents up to 4K/UHD-1 resolution and requires full access to the bitstream of the encoded video, currently with implementations for H.264, HEVC/H.265 and VP9. In this paper, further instances of *AVQBits* are introduced and evaluated for 2D video and for other video applications.

With the different *AVQBits* instances, the following research questions are addressed:

- How can *AVQBits* be used in cases where only a limited amount of input data is available for precise video quality estimation?
- Can the different *AVQBits* model instances be applied for video quality assessment of other application scopes than traditional 2D videos?

- Is it possible to predict the overall QoE of a longer ($\geq$ 1 *min*) HAS session with the introduced *AVQBits* model instances?

To address the question of different input data, four different *AVQBits* model versions are proposed in this paper. This includes a type Mode 0, Mode 1 and hybrid no-reference Mode 0 model, as well as the bitstream-based and standardized P.1204.3 version of *AVQBits*. The models are trained on four databases created by the author group during the competition "P.NATS Phase 2" for developing different types of video quality models. The "P.NATS Phase 2" project was run as collaboration between ITU-T SG12 and the Video Quality Experts Group (VQEG).[1] The *AVQBits* models are validated on the publicly available AVT-VQDB-UHD-1 dataset [40].

Besides the 2D video used in training and validation of the models with up to 4K/UHD-1 resolution, the performance of the four *AVQBits* variants is characterized for the following scenarios and databases:

- Gaming video; the models are evaluated on four different databases, namely, GamingVideoSet [41], KUGVD [42], CGVDS [43] and a self-developed dataset based on Twitch streams
- 360° video; the public dataset from [44] is used to evaluate the applicability for 360° video quality estimation.
- High Framerate (HFR) video; the performance of the *AVQBits* models on videos with framerate above 60 fps is evaluated using the LIVE-YT-HFR database [45].

Furthermore, to evaluate *AVQBits* for the case of long-term videos of a HAS session, the model instances are evaluated on five databases of longer video sessions between 1 min and 5 min, which have been developed as part of the "P.NATS Phase 2" project, and access to which was provided by the respective parties.

The source code for all the models along with the evaluation are made publicly available along with this paper for further usage and development by the research community.[2]

The paper is organized as follows: Section II provides an overview of the state-of-the-art (SoA), considering the different types of bitstream and hybrid video quality models available for video quality assessment of traditional 2D video, gaming, 360° and HFR video. Following this, the algorithmic description of the four *AVQBits* instances is described in detail in Section III. The databases used for evaluation of *AVQBits* for different application scopes are presented in Section V. In Section VI, the details of the model training procedure of all the proposed extensions are described. The detailed evaluation and performance characterization of the proposed models for different application scopes and the comparison with SoA video quality models are provided in Section VII. Finally, the paper concludes with a review of the modeling approach, results and an outlook for future work.

Before, proceeding further with the elaboration of the different sections, a brief overview of the notations used to represent different model instances proposed in this paper is provided.

A. NOMENCLATURE

To facilitate readers to follow the naming scheme of the different model instances, the following notations are introduced.

- *AVQBits*: This refers to the complete set of models.
- *AVQBits*$|M3$: This is used to denote the Mode 3 instance of *AVQBits* which is the same as ITU-T Rec. P.1204.3.
- *AVQBits*$|M1$: The Mode 1 instance of *AVQBits*.
- *AVQBits*$|M0$: This notation indicates the Mode 0 model instance.
- *AVQBits*$|H0$: Hybrid models of *AVQBits* in general are denoted by this term.
- *AVQBits*$|H0|f$: This is used to indicate that the hybrid model instance of *AVQBits* uses a fixed, pre-defined codec to generate the as input to the specific *AVQBits*$|H0$ instance of *AVQBits*.
- *AVQBits*$|H0|s$: The hybrid model instance of *AVQBits* which uses the same video codec for feature extraction as was initially used to encode the stream to be evaluated.

It should be noted that the model categories such as Mode 0, Mode 1, Mode 3 and Hybrid No-Reference Mode 0, are inspired from ITU-T Rec. P.1203 [35] and P.1204 [46].

II. RELATED WORK

The SoA discussion presented in this chapter is organized in different sections, reflecting the four *AVQBits* variants proposed in this paper, and the application scenarios. The sections provide an in-depth overview of the bitstream-based and hybrid models for video quality assessment, models for holistic QoE assessment for an HAS session, quality assessment of gaming, 360° and HFR videos. Partly complementary, recent model overviews can be found, for example, in [21] and [22].

A. BITSTREAM-BASED MODELS

Different bitstream models using a diverse set of bitstream features to estimate video quality have been reported in the literature. Many early Mode 3-type models that have been proposed were mainly focused on non-reliable transport and lower resolutions ($< 1080p$) [47]–[57].

One of the first Mode 3-type models that focused on reliable transport is the extension of P.1201.2 for progressive download for H.264 encoded videos. As with H.264 encoded videos, Izumi *et al.* [58], developed a Mode 3 based model using QP and spatial features based on coding units to estimate the quality of H.265 [8] encoded bitstreams. Also,

[1] https://www.its.bldrdoc.gov/vqeg/projects/audiovisual-hd.aspx

[2] https://github.com/Telecommunication-Telemedia-Assessment/p1204_3_extensions
https://github.com/Telecommunication-Telemedia-Assessment/bitstream_mode3_p1204_3

Huang *et al.* [59] proposed an approach to estimate the quality of H.265 encoded videos in terms of PSNR that can be used either as a bitstream-based or a pixel-based method. The model includes QP and transform coefficients as features and has been trained on the LIVE dataset [60] and validated on the SJTU dataset [61]. Both model variants show a good performance in terms of PCC.

ITU-T Rec. P.1203 [35] is the first standardized model for a holistic evaluation of HAS-type video streaming. This recommendation consists of three different modules corresponding to video quality [26], audio quality [62] and the overall quality integration [36]. The video quality models in ITU-T Rec. P.1203.1 [26] are further divided into four different modes of operation, depending on the input information available for quality estimation, namely, Mode 0, 1, 2 and 3 [38]. These models have been specifically developed for the HAS scenario and are applicable for videos encoded with H.264 for resolutions up to 1080p and framerates up to 30 fps. The reference implementation of this model is publicly available[3] [37]. The Mode 3 model corresponding to the ITU-T Rec. P.1203.1 has been further extended to be applicable to H.265 encoded videos of resolution up to UHD-1/4K by Lebreton and Yamagishi [63].

Furthermore, He *et al.* [64] present a model for quality assessment of H.264 and H.265 encoded bitstreams. This model has QP, skip ratio, motion, bitrate and framerate as the features and shows a performance comparable to the ITU-T Rec. P.1203.1 Mode 3 model. In addition, early models for reliable transport and HAS have been proposed by [65] and [66]. The different approaches and models related to holistic QoE evaluation where the cumulative effects of HAS-specific distortions such as momentary audio and video quality and quality switches, and stalling on quality perception will be discussed in a subsequent Section II-C.

1) MODE 0

The previous section described the SoA of bitstream models, with a focus mainly on Mode 3 type models. This section briefly summarizes the SoA of Mode 0 models. The most notable Mode 0 model for quality monitoring of video streaming is the ITU-T P.1203.1 Mode 0 model [26]. As mentioned before, this model is applicable for H.264 encoded videos for resolutions of up to 1080p and framerates up to 30 fps. A first extension of this model for newer codecs such as H.265 and VP9 was provided by a proprietary implementation from TU Ilmenau which has been made publicly available[4] [37], [38]. This extension used VMAF as groundtruth to derive the mapping coefficients for the newer codecs.

Furthermore, Rao *et al.* [67] propose an extension of this model to newer codecs such as H.265, VP9, AV1, and also for videos up to a resolution of UHD-1/4K and framerate up to 60 fps. However, this extension was based on only two subjective tests with limited encoding settings unlike the original standardized Mode 0 model in ITU-T Rec. P.1203, which was developed based on a large scale dataset containing 17 training and 13 validation databases. Furthermore, Lebreton and Yamagishi [63] in addition to the extending the scope of the Mode 3 model as mentioned above have also extended the application scope of the ITU-T P.1203.1 Mode 0 model for H.265 encoded videos for resolution up to UHD-1/4K.

To shorten the development time and the associated subjective quality assessment tests needed for such newer extensions, Yamagishi *et al.* [68] proposed a generic method to derive coefficents for metadata-based models for adaptive bitrate streaming services. The proposed method uses full-reference model scores as groundtruth to estimate new coefficients.

2) MODE 1

A Mode 1 model has access to frame type and frame size information for quality estimation, in addition to the metadata on bitrate, resolution and framerate, as for Mode 0 models. This additional access to the frame type and frame size information allows the quality estimation process to be content-dependent. As with the Mode 0 model, the ITU-T Rec. P.1203.1 Mode 1 model [26] is the first standardized model of this type for the HAS scenario and has been trained on the same 17 databases and validated on the same 13 databases as the Mode 0 model.

Another example of a Mode 1 model is the Bitstream-based Quality Prediction of Gaming Video (BQGV) [43]. It has been developed along the lines of P.1203.1 Mode 1. It has been shown to have a good performance both in terms of PCC and RMSE. As it is the case for the P.1203.1 Mode 1 model, too, this model is applicable to videos of resolutions up to FHD (1920 $\times$ 1080 pixels).

B. HYBRID MODELS

Similar to the pixel-based models, hybrid models can be classified into different categories depending on the access to the reference video for quality estimation. These include hybrid-FR, hybrid-RR and hybrid-NR models which have complete, partial and no access to the reference video, respectively. Furthermore, each of the categories can be divided into Mode 0, 1 and 3 based models, depending on the amount of bitstream information available as input.

Yamagishi *et al.* [69] present a hybrid-NR model for the IPTV scenario using information from packet headers and pixel-based spatial and temporal information for quality estimation [70]. The model is applicable for H.264 encoded videos of resolutions up to 1440 $\times$ 1080 and framerates up to 30 fps.

Another example of a hybrid model for non-reliable transport is the model proposed by Farias *et al.* [71]. Like the model presented in [69], this model, too, is applicable only for H.264 encoded videos, in light of the video technology

[3] https://github.com/itu-p1203/itu-p1203

[4] https://github.com/Telecommunication-Telemedia-Assessment/itu-p1203-codecextension

primarily used at the time. Similarly, the ITU-T J.343 series of recommendations also propose standardized hybrid models of all types, for the case of non-reliable transport.

Moreover, Osamu *et al.* [72] propose a mode 3 hybrid-NR model where the QP is used as the bitstream feature, along with the pixel-based spatial and temporal information to calculate video quality [70]. This model is again restricted to videos encoded with H.264 only.

More recently, hybrid models have been developed also for the HAS scenario. One example is the recently standardized ITU-T Rec. P.1204.5 which is a Mode 0 hybrid-NR model. It was developed as part of the same modelling competition as the bitstream-instance of *AVQBits*, ITU-T Rec. P.1204.3 discussed in this paper. Like all P.1204 models, the P.1204.5 model is applicable to videos encoded with H.264. H.265 and VP9 with resolutions up to UHD-1/4K and framerates up to 60 fps. As it has been indicated in [22], the model shows a good performance in comparison with SoA models.

Another Mode 0 hybrid-NR model called ''hyfu'' has been developed by Göring *et al.* [73] as part of a larger framework for pixel-based video quality models using machine learning. Accordingly, at its core, ''hyfu'' is a random forest (RF) based model. The model has been trained on four databases and validated independently on the four tests of the AVT-VQDB-UHD-1 database [40]. The application scope of ''hyfu'' is the same as ITU-T Rec. P.1204.5. The results show that the model performs well in comparison with the SoA models in terms of both PCC and RMSE.

C. INTEGRAL QUALITY

In general, a typical HAS session is characterized by various factors such as initial loading delay, momentary audio and video quality and quality switches, and stalling. A holistic QoE evaluation model has to consider all these factors, while also taking into account the time at which these changes occur in a video viewing session (see also [74]). ITU-T Rec. P.1203.3 is the first standardized model that incorporates all these factors. Here, ITU-T Rec. P.1203.1 and P.1203.2 are used to compute the video and audio quality, respectively, of each segment at a per-second level. In the integration module P.1203.3, the per-second audio and video quality values are further aggregated with regard to their time of occurence, the longest quality change, and the total number of quality changes, to obtain the final audiovisual quality of the video. A second component called ''stalling quality'' that handles the impact of initial loading delay and stalling is computed using the number of stalls, average stalling duration, and average interval between stalls as features. Then, the overall audiovisual quality and the stalling quality are integrated to obtain the initial overall quality. Besides a parametric, curve-fitting-based model component, a RF-based approach is used to compute the overall quality using features such as per-second video and audio quality scores, stalling ratio, stalling frequency, duration before the last stalling event etc. The final overall integral quality is the convex linear combination of the initial overall quality and RF-based overall quality. This model is applicable to videos of durations between 1 and 5 minutes and the implementation is publicly available.[5]

As the ITU-T Rec. P.1203 model only covers H.264 encoded videos, Lebreton and Yamagishi [63] have further extended ITU-T Rec. P.1203 for H.265 encoded videos of resolutions up to UHD-1/4K. For this purpose, six subjective evaluations with varying encoding conditions involving up to 192 participants in total were used.

In addition to this, other models for holistic QoE evaluation have been proposed [66], [75], but unlike the ITU-T Rec. P.1203.3 these models have not been trained and validated on large-scale databases.

D. VIDEO QUALITY MODELS FOR GAMING VIDEOS

Besides traditional 2D video, there has been a significant increase in gaming video streaming. As a result, a number of video quality models dedicated to gaming video quality evaluation have been proposed in the literature. The focus has mainly been the development of no-reference video quality models, due to the lack of high quality reference videos in a gaming video streaming session. NR-GVQM is an example for machine-learning-based NR-models specifically developed for video quality estimation of gaming videos [76], which uses a support vector regression (SVR) for prediction. VMAF was used as the groundtruth for model training, and hence this model can be viewed as a no-reference counterpart to VMAF. The model was trained and validated with the GamingVideoSet [41]. The model shows a good performance in terms of PCC on the validation set.

Another NR-based gaming video quality model was proposed by Göring *et al.* [77] referred to as ''nofu''. It is based on a number of features that are integrated using a random forest model. It is trained on the GamingVideoSet [41] dataset and shown to perform well based on 10-fold cross-validation.

Using a similar approach, Barman *et al.* [42] develop two NR model instances, namely, ''NR-GVQSI'' and ''NR-GVQSE'', with NR-GVQSI using subjective MOS and NR-GVQSE using VMAF as the training groundtruth. The models were trained and tested using two different datasets: KUGVD [42] and GVS [41]. The model was trained on GVS using a 10-fold cross validation strategy. Additionally, it was tested on KUGVD for its performance using MOS scores. Both models are shown to perform well in terms of PCC and RMSE considering subjective (MOS) ratings.

In addition to other machine-learning-based models, deep learning approaches have been explored to develop gaming video quality models. One example of such a model is the NDNetGaming model proposed by Utke *et al.* [78], which shows a good performance in terms of PCC on the KUGVD [42] dataset. A further extension of the NDNetGaming model called ''DEMI'' has been presented by Zadtootaghaj *et al.* [79]. ''DEMI'' incorporates a more

[5]https://github.com/itu-p1203/itu-p1203

sophisiticated pooling of the per-frame quality scores to obtain the per-segment quality score, in addition to other improvements. This model is developed to be applicable to non-gaming videos, too. The performance of the model was evaluated on the CGVDS [43] dataset, showing the model to be on par with or better than SoA models.

Although a bigger focus has been on pixel-based NR models for gaming video quality assessment, some studies have investigated bitstream-based models for gaming video quality prediction. One example of a gaming specific bitstream-based model is the BQGV proposed by [43], which is described in Section II-A2. A 5-fold cross validation approach using the CGVDS [43] dataset was done for performance evaluation, and it has been reported to outperform the ITU-T Rec. P.1203.1 Mode 1 and Mode 3 models. However, it should be noted that the ITU-T Rec. P.1203.1 Mode 1 and Mode 3 models were not retrained for gaming videos in that study.

Moreover, a gaming specific planning model called GamingPara has been presented by Zadtootaghaj *et al.* [43]. The model is shown to outperform ITU-T P.1203.1 Mode 0 on gaming data.

In addition, the standard ITU-T Rec. G.1072 comprises a video quality component that can be used to evaluate gaming video quality. It is based on retraining the video quality component of the IPTV-related planning model described in ITU-T G.1071 [80].

E. VIDEO QUALITY MODELS FOR 360° VIDEOS

Like for gaming video quality assessment, pixel-based models have been the main focus of the quality assessment of 360° videos. For example, variants of PSNR to take into account the possibility of viewing of 360° in all directions have been proposed. S-PSNR [81], a sphere-based PSNR computation and WS-PSNR [82], a position-weighted PSNR have been proposed as quality metrics to ultimately increase compression efficiency while maintaining a similar quality.

Tran *et al.* [83] conducted a performance evaluation of 360° video quality metrics considering different variants of PSNR including S-PSNR and WS-PSNR, among others. They concluded that the traditional approach of calculating PSNR was the most appropriate for 360° video.

More perception-oriented, traditional 2D video quality models such as VMAF have also been evaluated for quality assessment of 360°. For example, Fremerey *et al.* [44] evaluated the applicability of both the original version of VMAF and the centre-cropped version of VMAF [84] for 360° video quality evaluation and reported good performance in terms of PCC. Also Orduna *et al.* [85] report similar results for VMAF as reported by Fremerey *et al.* [44] for 360° video quality evaluation. Furthermore, extensions to VMAF to make it more suitable for 360° video quality evaluation have been proposed. To this aim, Croci *et al.* [86] present a Voronoi-based extension of VMAF. In addition to the Voronoi-based extension of VMAF, the study also presents Voronoi-based extensions for PSNR, SSIM and MS-SSIM and report that the Voronoi-based extensions generally outperform their traditional counterparts for 360° video.

More sophisticated models based on neural network approaches have also been proposed. For example, Li *et al.* [87] present a viewport-based convolutional neural networks (V-CNN) to estimate 360° video quality and is shown to outperform the SoA models. The model is also capable of predicting viewport saliency.

In addition to the mentioned pixel-based models, bitstream and hybrid models could also estimate 360° video quality. One example is Yao *et al.* [88], who propose a series of bitstream-based and hybrid models using QP as the bitstream feature and additional features such as spatial genre (simple versus complex), temporal genre (slow- versus fast-paced) and projection scheme. The described models are reported to outperform S-PSNR-I and V-PSNR based on a three-fold cross-validation approach. Moreover, Fremerey *et al.* [44] presented lightweight metadata-based and hybrid models for the quality assessment of 360° videos. The hybrid model calculates spatial and temporal information (SI, TI, cf. [70]) as input features, in addition to metadata such as bitrate, framerate and resolution. Both presented models show performance comparable with the SoA models such as VMAF, ADM2, WS-SSIM, and VIF.

Besides the aforementioned models, extensions to existing bistream models were proposed to accommodate 360° video-specific transmission aspects such as tile-based streaming. In this regard, Koike *et al.* [89] introduced a tile-based extension of the recently standardized ITU-T Rec. P.1204.3 (i.e., the model addressed in the present paper as the bitstream-based instance of the proposed *AVQBits* model), and report good performance in comparison with subjective test results. Also, Yang *et al.* [90] propose a full-reference quality assessment method for panoramic videos which outperforms traditional models and metrics such as PSNR, SSIM and VQM.

F. QUALITY ASSESSMENT OF HFR VIDEOS

The UHD-1/4K and UHD-2/8K standards cover higher framerates compared to traditional cinema or TV, which usually has 24 fps or 30 fps. In the following section, the SoA will be briefly analyzed considering the video quality assessment or prediction of videos with a higher framerate of $> 60\,fps$. A study on the impact of framerate on perceived quality was conducted by Mackin *et al.* [91] in which videos with framerates varying from $15\,Hz$ to $120\,Hz$ were analyzed. The subjective evaluations conducted using these videos show a significant relationship between framerate and perceived video quality. Further, it was observed that the effect of framerate on perceived video quality is content dependent. The study also reports diminishing improvements in terms of quality as framerates increase.

Furthermore, Mackin *et al.* [92] develop a high-framerate video quality database, BVI-HFR, containing videos captured at a framerate of $120\,fps$. Based on their tests they

conclude that models such as FRQM [93] which explicitly account for temporal distortions are more accurate in predicting video quality as compared to traditional metrics such as PSNR.

In addition to this, Madhusudana *et al.* [94] conduct a large-scale study on the subjective and objective quality of high framerate video with framerates up to 120 fps. An evaluation of existing FR and NR models has been performed, and it has been reported that the GSTI [95] model outperforms all the SoA models including VMAF.

Furthermore, Lee *et al.* [96] conducted a subjective and objective assessment of the video quality of space-time subsampled videos. The evaluation shows that the VSTR model proposed by Lee *et al.* [97], which is specifically developed to take into account the joint perceptual effects of spatio-temporal subsampling and compression, outperforms all the considered SoA models including VMAF.

G. SUMMARY

To sum up, it can be concluded that a large number of video quality models have been proposed in the literature for quality evaluation of videos for particular application scopes covering traditional 2D video, gaming video, 360° video and HFR video. Except for VMAF, none of the presented models scale well across different application scopes. VMAF is an FR model and may not be suitable in all scenarios, because of its rather high computational complexity and also the lack of reference videos in some applications, for example, as in case of gaming video streaming. This necessitates the development of lightweight models which can be used for different application scopes and can also be adapted based on the available input information. With this goal, this paper proposes different instances of *AVQBits*, one of which being the standardized ITU-T Rec. P.1204.3, and further instances addressing different types of available model input information. In total, four different *AVQBits* instances are presented, three being bitstream-based, with one being the P.1204.3 model, and one further, hybrid model. All four *AVQBits* instances are evaluated and characterized when applying them out-of-the-box to different application scopes. Furthermore, a QoE integration model for longer HAS session scenarios is presented which can use any of the *AVQBits* model instances as the underlying video quality estimation module.

III. MODEL DESCRIPTION

This section is focused on demonstrating the versatility of the *AVQBits* model in terms of scalability and adaptability regarding the available input information, starting with the standardized ITU-T P.1204.3 model. In the paper, model instances of two different types are introduced, namely, bitstream-based and hybrid. For the bitstream domain, the focus is on the ITU-T P.1204.3 standard, which is a Mode 3 model with access to full bitstream information, referred to as *AVQBits*$|M3$ in the following. Two further *AVQBits* instances are considered for application scenarios where the full bitstream information is not available. For these cases, Mode 0 and Mode 1 variants of *AVQBits* are proposed (*AVQBits*$|M0$, *AVQBits*$|M1$). To describe all *AVQBits* instances, the *AVQBits*$|M3$ algorithm with its full Mode 3 bitstream access forms the starting point. The Mode 0 and 1 instances are implemented by synthetically generating missing model input information based on the Mode 0 or 1 type information available, as will be outlined in subsequent sections. For the case that only Mode 0 type metadata is available, but a more accurate video quality estimation is sought than what can be achieved with a Mode 0 model, a hybrid no-reference Mode 0 model instance of *AVQBits* is proposed (*AVQBits*$|H0$). It has access to Mode 0 metadata and the decoded pixel information. The pixel information is used as an additional input by converting the degraded video into a "quality-equivalent" bitstream using an external video encoder, and then applying the existing and unchanged full-bitstream-based *AVQBits*$|M3$.

The general model structure of the proposed *AVQBits* model is shown in Figure 1. The approach is centred around the full-bitstream-based video quality model by the authors [39] standardized as ITU-T P.1204.3, i.e. *AVQBits*$|M3$. For example, in case of a Mode 0 or Mode 1 model, the required parts of the full-bitstream *AVQBits*$|M3$ model are adapted to handle the input and use the underlying other components for the final prediction. For the hybrid case, in a first iteration a quality-equivalent video bitstream mimicking the original bitstream is created.

To enable reproducibility, an open-source reference implementation of all the proposed models is made publicly available with this paper.[6]

A. MODE 3 MODEL – ITU-T P.1204.3

All *AVQBits* instances are based on the Mode 3 *AVQBits*$|M3$ model (ITU-T Rec. P.1204.3). Hence, its algorithm is described here first, followed by the different further *AVQBits* instances. An overview of ITU P.1204.3 model is shown in Figure 2, which highlights the individual components of the *AVQBits* general structure. It should be noted that the model is developed for two different target device categories, namely, PC/TV and Mobile/Tablet (MO/TA). A detailed description of the model is presented in [39]. This and all further models presented here are applicable to videos encoded with the H.264, H.265 and VP9 codecs. An extension to AV1 is currently underway. For all codecs, a corresponding bitstream parser is used to extract the relevant bitstream information as input to *AVQBits*$|M3$. The model consists of two components, a traditional curve-fitting-based component (referred to as the "Core Model") and a machine-learning component, which are described in more detail in the following sections.

1) CORE MODEL

The "Core Model" is based on the principle of degradation-based modeling, similar to ITU-T Rec. P.1203.1 [38].

[6]https://github.com/Telecommunication-Telemedia-Assessment/p1204_3_extensions

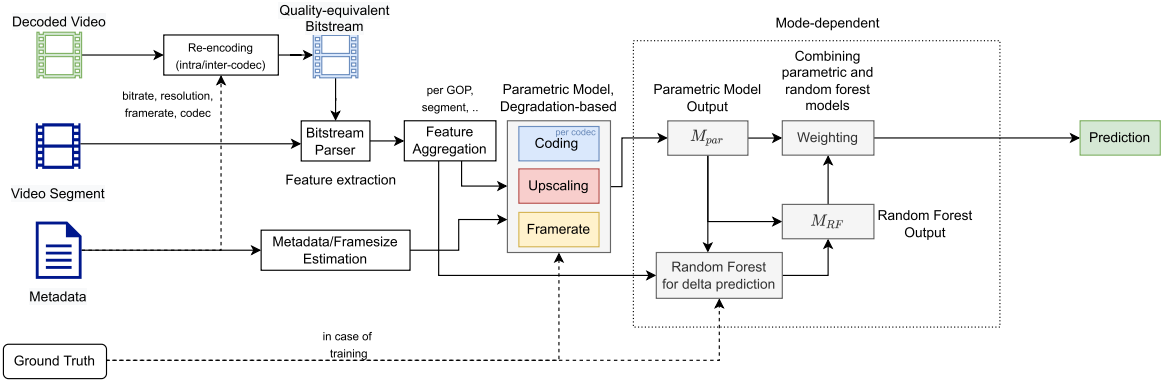

FIGURE 1. General model structure of *AVQBits* including all the four presented models.

It is initially inspired by the so-called E-model for speech quality [98]–[100], also based on the work on modelling television picture quality by [101]. In the core model, three different degradations expressed on a [0, 100] scale are considered: quantization degradation D_q, upscaling degradation D_u and temporal degradation D_t. Values on the 100-scale can be mapped to the 5-point ACR-scale used in subjective test (i.e. the resulting mean opinion score, MOS) using the S-shaped transformation from the E-model [100], as further described below. This way, scale-compression effects of the ACR-scale at the scale ends can be avoided [99], improving predictions especially for the higher-quality range of the scale.

2) QUANTIZATION DEGRADATION: D_q

The observable degradation that results from the chosen quantization settings during the encoding process is termed as "Quantization degradation" (D_q), see also [26], [38]. This type of degradation manifests itself as blockiness or deblocking-filter-related blurring to the end-user. Since this type of degradation is dependent on the specific encoding settings, the "Core Model" handles D_q separately per codec. The number of codec categories is extended from the initial three (H.264, H.265, VP9) to five, by including the bit-depth information and splitting H.264 and H.265 into 8- and 10-bit variants.

D_q is a function of the quantization parameter (QP) used to encode the video, which is extracted as model input information using the respective bitstream parser. To calculate D_q, firstly, *quant*, which is the normalized value of the QP is defined, cf. Equation (1).

$$quant = \frac{QP_{non-Iframes}}{QP_{max}} \tag{1}$$

Here, QP_{max} is codec and bit-depth dependent.

$$QP_{max} = 51, \quad \text{H.264-8-bit and H.265-8-bit} \tag{2}$$
$$QP_{max} = 63, \quad \text{H.264-10-bit and H.265-10-bit} \tag{3}$$
$$QP_{max} = 255, \quad \text{VP9} \tag{4}$$

$QP_{non-Iframes}$ is the average of the QP for all non-I frames for an entire segment.

The resulting $quant \in (0, 1]$. This *quant* value is then used to estimate mos_q, see Equation (5).

$$mos_q = a + b \cdot \exp(c \cdot quant + d) \tag{5}$$

mos_q is used to estimate D_{q_raw}, that uses *RfromMOS* as the mapping function to map the 5-point ACR scale to a 100-point scale similar to the one recommended in ITU-T G.107 [100].

$$D_{q\_raw} = 100 - RfromMOS(mos_q) \tag{6}$$

The final D_q value is the result of constraining D_{q_raw} to [0, 100] as shown in Equation (7).

$$D_q = \max(\min(D_{q\_raw}, 100), 0) \tag{7}$$

3) UPSCALING DEGRADATION: D_u

In addition to the degradations resulting from the chosen encoding settings, there are observable degradations resulting from upscaling the distorted video to the screen resolution during playback, which can be perceived by an end-user as blurriness. This kind of degradation is termed as the upscaling degradation (D_u). Hence, the "Core Model" should be able to account for this upscaling degradation and it is assumed that this degradation is codec-independent. Due to the fact that in real-world streaming scenarios, upscaling is typically performed by the player software, where streaming resolutions lower than the target screen resolution typically are a result of the adaptive streaming of bandwidth-dependent representations, this degradation is assumed to be codec-independent.

$$D_{u\_raw} = x \cdot \log(y \cdot scale\_factor) \tag{8}$$
$$D_u = \max(\min(D_{u\_raw}, 100), 0) \tag{9}$$

Equation (8) shows how D_u is estimated, where D_u is the [0, 100] constrained value of D_{u_raw}, with log being the natural logarithm. The *scale_factor* is calculated according

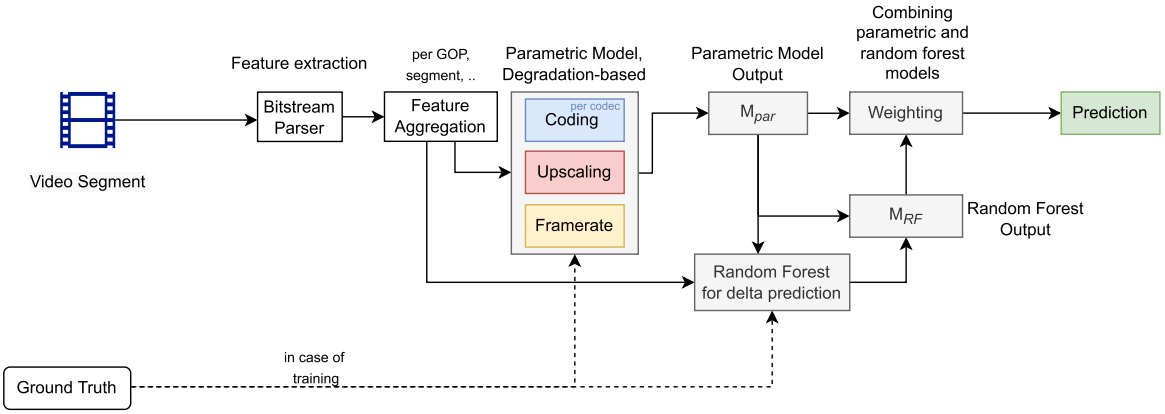

FIGURE 2. General model structure of the *AVQBits*|*M*3 / P.1204.3 model.

to Equation (10)

$$scale\_factor = \frac{coding\_res}{display\_res} \tag{10}$$

as the ratio of coding and display resolution, with $display\_res = 3840 \times 2160$ for PC/TV and 2560×1440 for mobile/tablet. $coding\_res$ is the resolution of the encoded video and is expressed in terms of $height \times weight$. The $scale\_factor$ is always limited to values $\in (0, 1]$.

4) TEMPORAL DEGRADATION: D_t

Finally, the "Core Model" handles the degradations due to the adjustment of the lower framerate representations to the display framerate as temporal degradation (D_t). This type of degradation may be perceivable as jerkiness. Similar to upscaling D_u, it is handled in a codec-independent fashion and is estimated as follows:

$$D_{t_{raw}} = z \cdot \log(k \cdot framerate\_scale\_factor) \tag{11}$$

$$D_t = \max(\min(D_{t\_raw}, 100), 0) \tag{12}$$

The temporal degradation D_t is mainly a function of the *encoded* and the *display* frame rates (the latter assumed to be constant with 60) that are combined in a $framerate\_scale\_factor$, cf. Equation (13), a value scaled in the range (0, 1]:

$$framerate\_scale\_factor = \frac{coding\_framerate}{60} \tag{13}$$

5) PREDICTION

The Equation (14) describes the final prediction of the "Core Model", M_{par}. Here, the described degradation-based approach is shown, using the 100-scale, $M_{p_{[0,100]}}$. The final prediction is further rescaled to a 5-point MOS-scale, Equation (15). Here, *MOSfromR* is the inverse mapping from the 100-point scale to the 5-point scale, similar to the one recommended in ITU-T G.107 [100].

$$M_{p_{[0,100]}} = 100 - (D_q + D_u + D_t) \tag{14}$$

$$M_{p_{[1,4.5]}} = MOSfromR(M_{p_{[0,100]}}) \tag{15}$$

$$M_{par} = scaleto5(M_{p_{[1,4.5]}}) \tag{16}$$

During training of the model, the subjective scores were linearly mapped to a 4.5-point scale from the 5-point scale in order to avoid information loss due to the *RfromMOS* and *MOSfromR* computations, since both of these mapping functions assume that the highest MOS that can be reached is 4.5. Hence, the coefficients predict the video quality scores on a 4.5-scale, denoted as $M_{p_{[1,4.5]}}$. Consequently, as a final step, the predictions on the 4.5-point scale are mapped back to the full 5-point scale range using a simple linear transformation, denoted as *scaleto*5, Equation (16), resulting in the final prediction of the parametric core model M_{par}.

6) MACHINE-LEARNING-BASED VIDEO QUALITY MODEL

The second part of the model is based on a machine-learning approach. It is used to estimate the "residual", that is, the part of the MOS that the parametric "core model" part is unable to predict.

This machine-learning part of the model uses Random Forest (RF) regression as the underlying machine-learning algorithm, and is referred to as M_{RF} in the following. Two different RF models, one each for the PC/TV and MO/TA cases are trained.

In addition to the features the "Core Model" uses, bitstream features such as the average motion per frame, motion in the x-direction (horizontal motion) and frame sizes with frame types are extracted with the bitstream parser and employed as model input. The aggregated features are detailed in [39]. The Random Forest model M_{RF} uses 20 trees with a fixed depth of 8. The final output is calculated as shown in Equation (17):

$$M_{RF} = M_{par} + predicted\_residual \tag{17}$$

Hence, the RF-based quality prediction is the addition of the predicted residual value $predicted\_residual$ to the M_{par} value predicted by the core model.

It should be noted that a more detailed description of the "Core Model" is provided as compared to the RF model as the "Core Model" is specifically re-instantiated for the

development of the *AVQBits|M0* and *AVQBits|M1* models while the RF model is used only in case of *AVQBits|M3* / P.1204.3 and *AVQBits|H0* models.

7) OVERALL VIDEO QUALITY PREDICTION

The overall final video quality prediction is the convex linear combination of the predictions from the parametric M_{par} and machine learning parts M_{RF}. In this case, equal weights, thus $w = 0.5$, are assigned to both of the predictions, shown in Equation 18. Considering Equation (17), it is shown that the RF residual part overall has a weight of 0.5, with the core model prediction being weighted with $0.5 + 0.5 = 1$.

$$Prediction = w \cdot M_{par} + (1 - w) \cdot M_{RF} \tag{18}$$

To enable reproducibility, an open-source reference implementation of the model along with the ffmpeg-based bitstream parser for all three codecs H.264, H.265 and VP9 is made available,[7] including also the trained random forest model.

8) PER-1-SECOND SCORE PREDICTION

In addition to the overall per-segment video quality score, the model also outputs per-1-second scores. The per-1-second score is calculated using Equation (19).

$$per-1-sec-score = \frac{QP_{non-I,per-seg}}{QP_{non-I,per-sec}} \times Prediction \tag{19}$$

where,

- $QP_{non-I,per-seg}$ is the average QP of all non-I frames in a segment
- $QP_{non-I,per-seg}$ is the average QP of all non-I frames for each second
- *Prediction* is per-segment video quality score described in Equation (18)

It should be noted that the per-1-second scores are calculated with a non-overlapping 1-sec window.

B. MODE 0 MODEL

A Mode 0 model is the least complex of bitstream models, both in terms of available input information and computational complexity. It has access to metadata such as bitrate, resolution, framerate and codec information as available input for video quality estimation. The proposed Mode 0 model *AVQBits|M0* instantiates the *AVQBits* model using the same general model structure as outlined for *AVQBits|M3* above and underlying ITU-T Rec. P.1204.3, with some key modifications which are indicated in Figure 3. The traditional curve-fitting-based part of *AVQBits|M3*, referred to as the "Core Model" in Sec. III-A, is exclusively used in *AVQBits|M0*, due to the limited numbers of features available for a Mode 0 model. The RF-based model component of *AVQBits|M3* for the residual is not used in *AVQBits|M0*, and with the purely metadata-based input information, the model is not content-aware. The "Core Model" is made up of three different degradations, namely, coding/quantization degradation, upscaling degradation and temporal degradation. For the Mode 0 instance *AVQBits|M0*, the focus only is on the quantization degradation, because this part is the only one affected by the lack of full-bitstream information.

The quantization degradation in case of *AVQBits|M3* is a function "quantization parameter" (QP), which is codec dependent. Since in a Mode 0 model, there is usually no access to the bit-depth information as input, only three codec categories are defined, namely, H.264, H.265, and VP9, in contrast to five codec categories in *AVQBits|M3*, see Sec. III-A. Accordingly, QP_{max} which is required to define *quant* as proposed in Equation 1 is restricted to one of the following two values based on the used codec.

$$QP_{max} = 63, \quad \text{H.264 and H.265} \tag{20}$$

$$QP_{max} = 255, \quad \text{VP9} \tag{21}$$

Because QP is not accessible as direct input information in case of a Mode 0 model, it is approximated using the available metadata information, namely, bitrate, resolution and framerate, see Equation 22.

$$\begin{aligned} QP_{pred} = {} & a_{qp\_m0} + b_{qp\_m0} \cdot \log(bitrate) \\ & + c_{qp\_m0} \cdot \log(resolution) \\ & + d_{qp\_m0} \cdot \log(framerate) \end{aligned} \tag{22}$$

The resulting *quant* is defined as in Equation 23 and is content agnostic, due to the lack of content-specific features.

$$quant = \frac{QP_{pred}}{QP_{max}} \tag{23}$$

Using *quant* as defined in Equation 23, quantization degradation is calculated as described in Sec. III-A2 (Quantization Degradation: D_q). As a result of using QP_{pred} instead of the actual QP value as in *AVQBits|M3*, the coefficients related to the quantization degradation should also be re-trained by taking into account the QP_{pred} values. The training procedure and the resulting coefficients are detailed in Section VI-B.

1) PER-1-SECOND SCORE PREDICTION

For the *AVQBits|M0* model, no separate windowing approach is used unlike in the case of *AVQBits|M3* / P.1204.3 and hence the per-1-second scores is just equal to the per-segment scores.

C. MODE 1 MODEL

In addition to the metadata such as bitrate, resolution, framerate and codec information, a Mode 1 model has access to framesize and frame type information. This information enables the inclusion of source-, and hence, content-specific features into the model. Like the Mode 0 model *AVQBits|M0*, the Mode 1 model *AVQBits|M1* introduced in this paper is based on the same general model structure as that of

[7] https://github.com/Telecommunication-Telemedia-Assessment/bitstream_mode3_videoparser

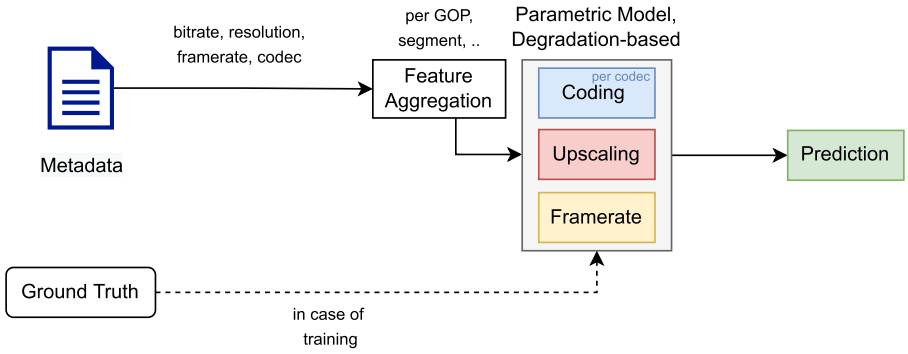

FIGURE 3. General model structure of the *AVQBits|M0* model.

AVQBits|M3 (i.e. ITU-T P.1204.3) and just modifies the information pre-processing for the "Core Model", as seen in Figure 4 which is a reduced variant of the general model structure shown in Figure 1 for clarity. Here too, the focus is on the quantization degradation D_q, as it is the only Mode-dependent part of the model.

As in Mode 0, the *AVQBits|M1* model has been developed for three codecs, namely, H.264, H.265 and VP9, for one single bit-depth, as the exact profile usually cannot be known based on the available input information. Hence, the same QP_{max} values are used for the three categories as for *AVQBits|M0*, cf. Eqs. (20) and (21).

For the purpose of QP estimation, two new features using the framesize and frametype information are defined in Equations (24) and (25).

The feature *fsratio* represents the ratio between the average sizes of I-frames and non-I-frames for a given segment under consideration:

$$fsratio = \frac{1/N_I \sum_i (S_{I,i})}{1/N_{nI} \sum_j (S_{nI,j}))} \tag{24}$$

Here, $S_{I,i}$ is the size of I-frame i, $S_{nI,j}$ is the size of a non-I-frame, that is, P- or B-frame, which are treated alike for this calculation, with index j. N_I is the overall number of I-frames, N_{nI} the overall number of non-I-frames. Like for *AVQBits|M3*, all I-frames i and non-I-frames j belonging to a given segment under consideration are used.

The second feature introduced is the mean size of non-I-frames *ms_nI*.

$$ms\_nI = 1/N_{nI} \sum_j S_{nI,j} \tag{25}$$

As in *AVQBits|M1*, QP_{pred} is calculated according to Equation (26).

$$\begin{aligned} QP_{pred} = {} & a_{qp\_m1} \\ & + b_{qp\_m1} \cdot log(ms\_nI) \\ & + c_{qp\_m1} \cdot log(resolution) \\ & + d_{qp\_m1} \cdot log(framerate) \\ & + e_{qp\_m1} \cdot log(fsratio) \end{aligned} \tag{26}$$

Considering the QP estimation of *qp_m1* following *AVQBits|M0*, the quantization degradation is retrained to be Mode 1 specific. The details of the training procedure and the final coefficients are described in Section VI-C.

1) PER-1-SECOND SCORE PREDICTION

Like the *AVQBits|M0* model, the per-1-second scores for the *AVQBits|M1* model is just equal to the per-segment scores.

D. HYBRID NO-REFERENCE MODE 0 MODEL

As mentioned earlier, a hybrid no-reference Mode 0 model has access to both the metadata and the decoded pixel information of the distorted video to estimate video quality.

The main idea of the proposed model is to create a "quality equivalent bitstream" (QEB) which is similar to the original bitstream using the decoded pixels and the provided metadata. After the QEB is created, the *AVQBits|M3* model (i.e. ITU-T P.1204.3) is applied with slight changes. A somewhat related approach has been used in ITU-T Rec. P.563 [102] to provide a more general description of the received speech quality, which is given by comparing the input signal with a pseudo reference signal generated by a speech enhancer.

The process of creating the QEB is shown in Figure 1 and Figure 5, wherein the provided metadata such as bitrate, resolution, framerate, and codec information is used. The distorted video is re-encoded with the encoding settings corresponding to the metadata following a 1-pass encoding strategy. This is based on the results reported in Stankowski *et al.* [103] that the quality loss across different QP values remains constant for a second round of encoding, which can be compensated by the model due to the use of QP as the feature for quality estimation in the Core Model. Furthermore, the QP that an encoder chooses for a bitrate-resolution during the QEB generation process will be in the same range as that of the initial encoding due to the same bitrate and resolution settings.

In the following, two variants of the hybrid no-reference Mode 0 model *AVQBits|H0* are proposed. These variants are based on the codec used to re-encode the video and are referred to as the "same" and "fixed" codec variants.

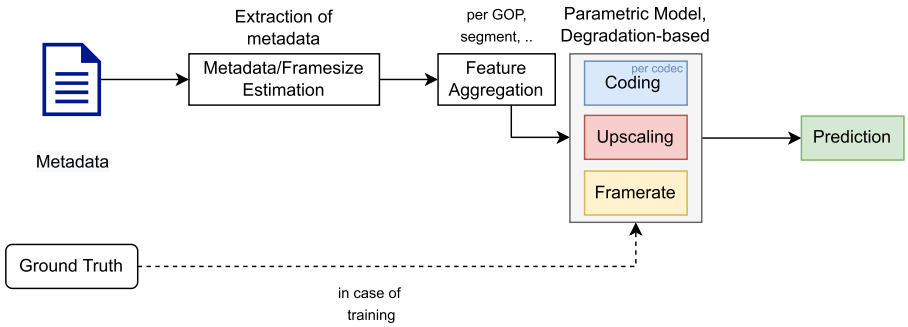

FIGURE 4. General model structure of the *AVQBits*|*M*1 model.

In case of the ''same'' codec variant, hereafter referred to as *AVQBits*|*H*0|*s*, the QEB is created using the codec specified by the metadata. After the QEB is created, the *AVQBits*|*M*3 model ITU-T P.1204.3 is directly applied to estimate the video quality without any modifications.

The second, ''fixed'' codec variant, is referred to as *AVQBits*|*H*0|*f* in the following. By using a fixed, pre-defined codec to create the QEB, no stream-specific encoding and then bitstream parsing is needed, reducing the complexity of the implementation. After the creation of the QEB, the *AVQBits*|*M*3 model ITU-T P.1204.3 estimates the video quality. For the proof-of-concept of *AVQBits*|*H*0|*f* presented in this paper, H.265 is selected as the codec to create the QEB, irrespective of the codec used to generate the original bitstream.

Different codecs have different impact on quality for a given specific setting. Since H.265 is used to create the QEB irrespective of the codec to generate the original bitstream, the initially estimated quality of this QEB may not optimally reflect the impact of the original codec. As a result, to estimate the final quality score, a simple linear mapping function is proposed that takes into account the impact of the original codec on quality to map the initial prediction of the ITU-T P.1204.3 model to the respective codec characteristics.

$$Prediction = a_{cmap} \cdot Prediction_{M3} + b_{cmap} \tag{27}$$

where, $Prediction_{M3}$ is the prediction from the *AVQBits*|*M*3 P.1204.3 model and a_{cmap} and b_{cmap} are codec-specific mapping coefficients. The coefficient values are provided in Section VI-D.

It is noted that besides this instance of the ''fixed'' codec hybrid model variant presented as proof-of-concept, other realizations can be conceived. For example, a more sophisticated codec-specific mapping function can be developed instead of the simple linear mapping as proposed in this work. Further, in principle also another of the three encoders and hence bitstream-parsers can be used to create and analyze the QEB. With H.265, the newest of the three codecs was selected, and currently developed updates of the proposed models can use even newer codecs such as AV1 or VVC.

The impact of the ''fixed'' instead of the ''same'' codec variant on quality prediction accuracy is extensively analyzed in Sec. VII-A.

IV. OVERALL INTEGRAL QUALITY: MODEL DESCRIPTION

Usually there is a tendency to treat QoE as a static event, and the QoE measured for a stimulus of delimited length is assumed to be stable along its duration. However, this rarely happens for stimuli extending over several minutes [74]. This can be well observed in a typical HAS session lasting several minutes, which may include different quality-related events, for example, quality switching, initial loading delay, and stalling. Hence, any model designed to estimate the overall integral quality of a HAS session has to take into account the impact of these events.

ITU-T Rec. P.1203.3 [36] is the first standardized model that takes into account all these factors to predict the QoE of a HAS session [37]. This model takes per-1-second video and audio quality scores, stalling-related information and the device type (either ''PC/TV'' or ''Mobile/Tablet'') as input to calculating the QoE of an HAS viewing session. The main model output (referred to as *O*.46 in [36]) is a final media session quality score on the 5-point ''MOS-scale''. Further, besides the parametric input information, the model produces intermediate values that can be used for HAS-system diagnostics, such as a perceptual stalling indication, audiovisual segment coding quality per output sampling interval, and a final audiovisual coding quality score. The design of the subjective tests conducted to gather groundtruth for model training and validation used a restrospective rating by the participants on a 5-point ACR-scale given at the end of an audiovisual stimulus lasting between $1 - 5\,min$. Due to this test design, it becomes pertinent to address cognitive effects such as the *recency effect* and *primacy effect* (see, e.g., [74] for more details and references around these effects). Accordingly, ITU-T Rec. P.1203.3 considers these cognitive effects as part of the model.

In this paper, a long-term integration model specifically designed for the four types of *AVQBits* models is presented, which is based on ITU-T Rec. P.1203.3 [36]. It relies on the same model structure as P.1203.3, adapting the final

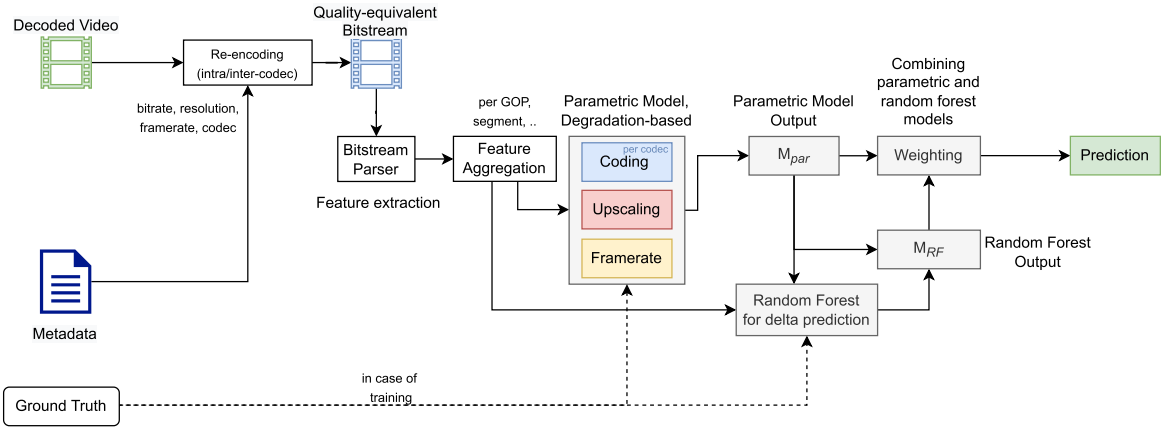

FIGURE 5. General model structure of the *AVQBits*|*H*0 model.

audiovisual coding quality estimation, using more accurate short-term video quality models such as ITU-T P.1204.3. The final audiovisual coding quality score $O.35$ in ITU-T Rec P.1203.3 is estimated following Equation 28.

$$\begin{aligned} O.35 &= O.35_{baseline} - negBias - oscComp \\ &\quad - adaptComp \\ O.35_{baseline} &= \frac{\sum_t w_1(t) \cdot w_2(t) \cdot O.34[t]}{\sum_t w_1(t) \cdot w_2(t)} \\ w_1(t) &= t_1 + t_2 \cdot \exp\left(\frac{\frac{t-1}{T}}{t_3}\right) \\ w_2(t) &= t_4 - t_5 \cdot O.34[t] \end{aligned} \tag{28}$$

Here, $O.34$ is the audiovisual segment coding quality per output sampling interval. The values w_i are weighting coefficients specified in the standard [36], [37]. The three factors *negBias*, *oscComp* and *adaptComp* are used to take into account certain temporal effects related to video-quality fluctuations. In the proposed model, these three factors are ignored, reflecting two assumptions:

1) The per-second and per-segment scores of the ITU P.1204.3 model are generally more accurate than those from the short-term video-quality module variants of ITU-T Rec. P.1203.1 [26], [38], where these were re-engineered from the final, retrospective and longer-session media session quality score ($O.46$) (see, e.g., [38] for more details).
2) This re-engineering may have been impacted by specific ITU-T P.1203.1 implementations, and thus be very specific for data created as part of the ITU-T P.1203 development process, and not optimally suited for the *AVQBits* variants proposed in this paper.

The *AVQBits*|*M*3 model at the starting point of this paper was specifically trained on short-term videos, and hence is capable of more accurately estimating both per-segment and per-1-second video quality scores. As a result, the new, simplified $O.35$ is given by Eq (29).

$$O.35 = O.35_{baseline} \tag{29}$$

It should be noted that no other changes to the model algorithm or coefficients inherited from ITU-T P.1203.3 have been applied.

In the following, the subjective test databases used to train and validate the different model instances will be outlined.

V. SUBJECTIVE VIDEO QUALITY DATASETS

The datasets to train, validate and evaluate the performance of the different *AVQBits* instances for HAS-type encoded video and validate the models for other applications are described in this section.

The following listed datasets are used for the training and validation of the *AVQBits* models, and are described in detail in the remainder of this section.

- PNATS-UHD-1 (Training and validation datasets created as part of the "P.NATS Phase 2 / AVHD" modelling competition cf. also [22])
- AVT-PNATS-UHD-1 (own data contributed to "P.NATS Phase 2 / AVHD" modelling competition by the authors, cf. also [22], proprietary, subset of PNATS-UHD-1)
- AVT-VQDB-UHD-1 (short-term video quality evaluation dataset for 4K/UHD-1, published open source, see [40])
- PNATS-UHD-1-Long (long-sequence database from "P.NATS Phase 2 / AVHD" modelling competition, including all databases submitted by the proponents; proprietary, used with permission)
- GamingVideoSet [41], Kingston University Gaming Video Dataset (KUGVD) [42],Cloud Gaming Video Dataset (CGVDS) [43], Twitch dataset
- 360° Streaming video quality dataset [44]
- LIVE-YT-HFR [45]

AVT-PNATS-UHD-1, AVT-VQDB-UHD-1, the Twitch dataset and the 360° Streaming video quality dataset are proprietary, whereas the others excluding PNATS-UHD-1 are open datasets. To train and validate the *AVQBits*|*M*3 / P.1204.3 model, the PNATS-UHD-1 dataset is used. For the training and validation of the *AVQBits*|*M*0, *AVQBits*|*M*1 and *AVQBits*|*H*0 instances, the AVT-PNATS-

UHD-1 and AVT-VQDB-UHD-1 datasets were used, respectively. The PNATS-UHD-1-long dataset is used to evaluate the proposed long-term integration model. These four are UHD-1 databases, and retrospective ratings on the ACR 5-point-scale were collected from participants for all these tests. For all tests, participants were asked to undergo a simple vision test using Snellen charts as recommended in ITU-T Rec. P.910 [70]. The tests were conducted in controlled lab settings as prescribed by ITU-T Rec. BT.500-13 [104] and ITU-T Rec. P.910 [70] with appropriate lighting conditions and a viewing distance of 1.5×H, with H being the height of the screen. An outlier detection was applied for all tests based on PCC, with a threshold of $PCC = 0.75$ for tests with short-duration (7-9 sec) videos and $PCC = 0.70$ for tests with videos of longer duration ($\geq$1 min). This method for outlier detection was most notably used as part of developing the ITU-T Recommendations P.1203 [34] and P.1204 [22].

The details of the test environment and the outlier detection method for other datasets are described in the following sections.

A. PNATS-UHD-1

The PNATS-UHD-1 dataset was developed as part of the ''P.NATS Phase 2 / AVHD'' modelling competition in ITU-T SG12/Q14. This dataset consists of 26 different tests that were designed and conducted by eight proponents. Of the 26 different tests, 13 were used for training the models submitted to the competition and the remaining 13 were used for model validation. It should be noted that the 13 validation databases were created after model submission. Out of the 13 training tests, nine were created with a PC/TV as viewing device, and four with mobile devices for viewing. For validation, nine tests used a PC/TV as viewing devices, three mobile and one tablet. All tests included PVSs with a duration of 7–9 s. 2464 PVSs were used for training, 2483 for validation, resulting in a total number of 4947 PVSs. Further details on the sources, encoding parameter ranges, number of test subjects etc., is provided in [22].

B. AVT-PNATS-UHD-1

This dataset consists of four different subjective tests that were designed and conducted as part of the P.NATS Phase 2 competition. The tests were targeted to cover a wide range of source contents, with more than 50 different sources used in each of the four tests. Due to the large number of source contents per test, the tests followed a partial-factorial design, with a limited number of PVS to be assessed by participants. A source (in the paper referred to also as ''SRC'') was repeated between 3 and 5 times within a test. Three sources were used across all tests and are referred to as ''common set sources''. In addition to this, five encoding conditions (the processing conditions are also referred to as ''Hypothetical Reference Circuits (HRCs)'' in the paper), have been used across the four tests. These five HRCs in combination with the common sources form the ''anchor conditions'' that can further be used to unify the different tests for the purpose of model training, if desired by a proponent of the aforementioned ITU-T SG12 ''P.NATS Phase 2'' competition.

The SRCs used in all the four tests had a resolution of 3840 × 2160 pixels and a framerate between 24 *fps* and 60 *fps*. To design the HRCs, fixed settings were selected for each HRC, without any adaptation within the short video sequences. The bitrate range was chosen to be 100 *kbps* to 50000 *kbps*, the resolution range between 360*p* and 2160*p* and the framerate values were chosen between 15 *fps* and 60 *fps*. It was ensured that the framerate of the SRC was never lower than the framerate of the corresponding PVS. In all the four tests, the different videos were encoded using one of three codecs, namely, H.264, H.265, and VP9. The encoding has been done with libx264 (ffmpeg), libx265(ffmpeg) and libvpx-vp9 (ffmpeg), respectively. In addition to choosing from different bitrates, resolutions and framerates for each HRC, the design also involved different presets (ultrafast, veryfast, fast, medium, slow, slower, veryslow) for H.264 and H.265 and settings for speed for VP9, different chroma subsampling (YUV420, YUV422), bit-depth (8 and 10 bits), encoding types (1-pass, 2-pass both with and without min max bitrate constraints, HRCs with specific constant rate factor (crf) encoding) and different GOP sizes (auto, 2 s, 5 s). Furthermore, video segments encoded via services such as YouTube, Bitmovin and Vimeo were included to reflect real-world encodings. A 55'' LG OLED55C7D screen was used to present the videos in all the four tests.

In the first test, 52 different SRCs were included and encoded with different HRCs that resulted in a total of 187 PVSs. These 187 PVSs were rated by 27 participants. Following the outlier detection criterion based on PCC described earlier, two outliers were detected and removed from further analysis. The second test covered a total of 53 different SRCs with 187 PVSs created from these, which were rated by a total of 36 participants. Further analysis based on the aforementioned outlier criterion detected two outliers in this test. 52 different sources were used in the third test, and the 185 PVSs resulting from the HRC processing were rated by 30 participants, with five outliers detected. The fourth test had 53 SRCs processed according to different HRCs, resulting in 191 PVSs that were rated by 28 participants. Here, 3 outliers were detected following the PCC based criterion.

The distribution of the mean opinion scores (MOS) is illustrated in Figure 6. It can be observed that there is a tendency towards higher quality. This test design was motivated to yield better distinction for higher quality levels by test participants and also models.

C. AVT-VQDB-UHD-1

AVT-VQBD-UHD-1 [40] is a publicly available dataset[8] created by the authors' group. Like the AVT-PNATS-UHD-1

[8]https://github.com/Telecommunication-Telemedia-Assessment/AVT-VQDB-UHD-1

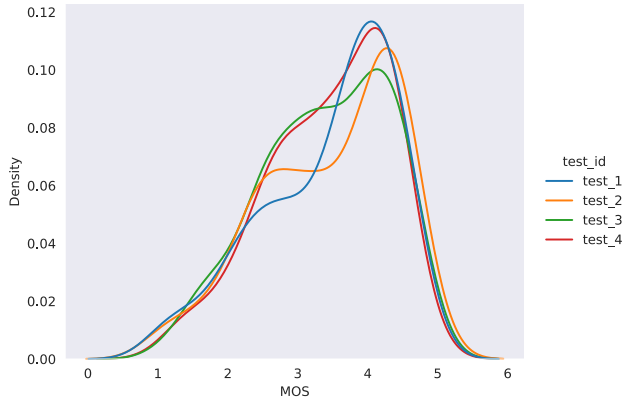

FIGURE 6. MOS distribution of AVT-PNATS-UHD-1 dataset.

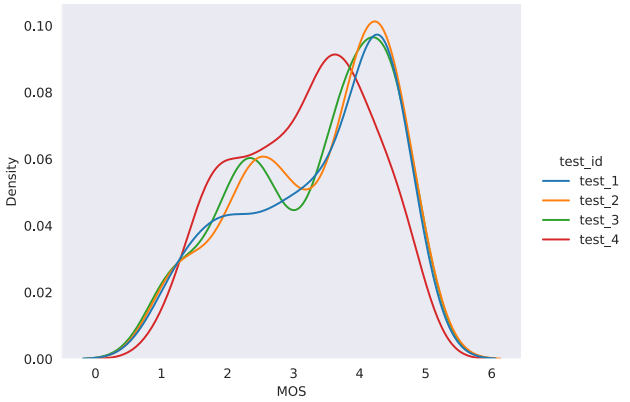

FIGURE 7. MOS distribution of AVT-VQDB-UHD-1 dataset.

dataset, this dataset also consists of four different subjective tests. All the four tests had a full-factorial design. In total, 17 different SRCs with a duration of 7-10 s were used across all the four tests. All the sources had a resolution of 3840 × 2160 pixels and a framerate of 60 fps. For HRC design, bitrate was selected in fixed (i.e. non-adaptive) values per PVS between 200 *kbps* and 40000 *kbps*, resolution between 360*p* and 2160*p* and framerate between 15 *fps* and 60 *fps*. In all the tests, a 2-pass encoding approach was used to encode the videos, with *medium* preset for H.264 and H.265, and the *speed* parameter for VP9 set to the default value "0". As with the tests in AVT-PNATS-UHD-1, the same PCC-based criterion is used for outlier detection. In the following, all the four tests of this dataset are briefly described. The distribution of MOS is illustrated in Figure 7.

1) TEST_1

The HRC design of this test was based on choosing from different bitrates for each of the different resolutions. For this purpose, four different resolutions were considered, namely 360*p*, 720*p*, 1080*p* and 2160*p*. Two bitrates each were selected for 360*p* and 720*p* and three bitrates each for 1080*p* and 2160*p*. Three different codecs, namely, H.264, H.265, and VP9 were used to encode the videos. These HRCs were applied to six different SRCs of 9-10 s duration. This resulted in a total of 180 PVSs. The framerate of all the PVSs was kept at the source framerate of 60 fps. A 65" Panasonic VIERA TX-65CXW804 display was used to present the videos in the test. The 180 PVSs were rated by 29 participants. Following the PCC-based outlier criterion, no outliers were detected.

2) TEST_2

For this test, the HRC design was based on using different bits-per-pixel (*bpp*) settings for different resolutions. Four different *bpp* values were considered, per each of the same four resolutions used also in test_1. As the number of *bpp* – resolution combinations considered was higher than the bitrate – resolution combinations for test_1, only H.264 and H.265 were used to encode the videos. Six SRCs including the three common set SRCs from test_1 were used. The SRCs had a duration of 7-9 s. As in test_1, the framerate of the PVSs was kept at the source framerate of 60 fps. Overall, 192 PVSs were created using the six SRCs and the defined HRCs. The test videos were presented to the participants on a 55" LG OLED55C7D screen. A total of 24 participants rated these PVSs, and no outliers were detected.

3) TEST_3

This test followed the same philosophy for HRC design as test_2, and hence the same *bpp* values and resolutions were used. Also, the same SRCs were employed. Mainly H.265 and VP9 were selected to encode the videos. In test_2, it was observed that some of the PVSs associated with one of the sources (*Dancers_8s*) had uncharacteristically low scores due to encoding errors. The HRCs associated with these PVSs were encoded with H.264, and these PVSs, now correctly encoded, were repeated in this test. The corresponding HRCs associated with H.265 were dropped, to keep the total number of PVSs at 192 as in test_2. 26 participants rated the 192 PVSs presented on a 55" LG OLED55C7D screen. No outliers were detected in this test. As test_2 and test_3 are based on the same design philosophy, these two tests can be combined for further analysis.

4) TEST_4

The objective of this test was to assess the effect of different framerates on perceived video quality. Hence, the HRC design was based on selecting from different framerates for each of the chosen different resolutions. Six different framerates, namely, 15 *fps*, 24 *fps*, 30 *fps*, and 60 *fps* were used across six different resolutions between 360*p* and 2160*p*. Only H.264 was selected to encode videos for this test. Eight SRCs with a duration of 7-9 s each were used, with no overlap between sources from the other tests. The selected HRCs in combination with these eight SRCs resulted in a total of 192 PVSs. These PVSs were presented on a 55" LG OLED55C7D screen. 25 participants took part in the test, with two outliers being detected.

D. PNATS-UHD-1-LONG

Similar to the AVT-PNATS-UHD-1 dataset, see Sec. V-B, this dataset was also developed as part of the P.NATS Phase 2 competition. The dataset consists of five different tests with videos of 1 to 5 min duration. The tests were designed based on the "immersive" paradigm [105] in which the participants never view the same source stimulus more than once. Especially for tests with long sequences, participants are likely to get bored with multiple viewings of the same sequence. All tests used a retrospective rating of integral quality of the simulated HAS viewing session on the 5-point ACR scale. All stimuli included audio so as to make the HAS-session more realistic, with audio kept at the highest available quality.

Long-sequence tests test_1 and test_2 involved retrospective rating of videos of 1 min duration. For this purpose, 60 different SRCs in each test were encoded with different HRCs, with each HRC consisting of a combination of different HAS-specific quality-related effects, such as quality switches, initial loading delay, and stalling. This resulted in a total of 60 PVSs in both tests. In test_1, 24 subjects took part, and in test_2, 37 participants. Following the somewhat less constrained criterion of $PCC = 0.7$ for the long-sequence tests, for test_1, no outliers were detected, and 6 outliers for test_2. In test_1, the PVSs were displayed on a mobile screen, with a viewing distance of 6-8H, giving the test subjects some freedom in how they were placing the screen relative to their eyes. In test_2, a TV was used, with a viewing distance of 1.5H. The highest resolution of the PVS in test_1 was restricted to 2560 × 1440 pixels, as this was the display resolution of the mobile. For test_2, the highest resolution of the PVS was equal to the SRC resolution of 3840 × 2160 pixels.

In test_3 and test_4, videos of 2 min duration were rated. 30 different SRCs were used in each test, resulting in 30 PVSs, as each HRC was associated with a different SRC. The number of PVSs were adapted compared to test_1 and test_2, to keep the test duration within 60 min. The PVSs in test_3 were presented on a mobile, and like in test_1, the highest resolution was restricted to 2560 × 1440 pixels. Test_4 again used a TV for viewing, and hence the highest resolution was kept at the source resolution of 3840 × 2160 pixels. 24 participants rated 30 the PVSs in test_3. In test_4, the 30 PVSs were rated by a total of 31 participants. No outliers were detected in the two tests.

Test_5 involved quality assessment of videos of 5 min duration, with 14 different SRCs being used. In total, 14 PVSs were rated by 31 participants, with 5 outliers being detected. As the videos were presented on a mobile screen, the highest resolution of the PVSs was again restricted to 2560 × 1440.

The following laboratories and companies were involved in conducting the subjective tests: Test_1 and test_2 were conducted by Netscout in England, test_3 by SwissQual in Switzerland, test_4 by TU Ilmenau in Germany and test_5 by Ericsson in Sweden.

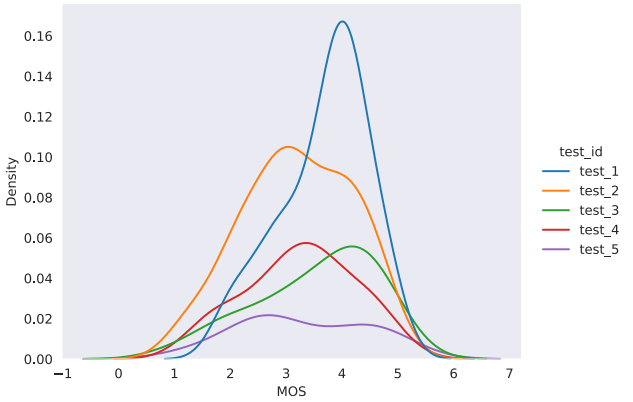

FIGURE 8. MOS distribution of PNATS-UHD-1-Long dataset.

The MOS distribution of all the five tests is shown in Figure 8 and reflects a similar tendency of having more PVSs in the higher quality range as in AVT-PNATS-UHD-1, due to a similar test design philosophy.

E. GAMING DATASETS

This sections briefly describes the four gaming datasets, namely, the GamingVideoSet, KUGVD, CGVDS and Twitch dataset that are used in this paper. It is noted that an evaluation of the *AVQBits*|*M*3 model instance (aka ITU-T Rec. P.1204.3) on this combined dataset has been presented by the authors in [106]. The present paper substantially extends the analysis to the three further instances of the *AVQBits* model.

1) GamingVideoSet (GVS)

This dataset [41] consists of 24 SRCs that have been extracted from 12 different games. The SRCs are of 1920 × 1080 pixel resolution, 30 *fps* framerate and have a duration of 30 *s*. The HRC design included three different resolutions, namely, 480*p*, 720*p* and 1080*p*. 90 PVSs resulting from 15 bitrate-resolution pairs were used for subjective evaluation. A total of 25 participants rated all the 90 PVSs.

2) KINGSTON UNIVERSITY GAMING VIDEO DATASET (KUGVD)

Six SRCs out of the 24 SRCs from the GamingVideoSet were used to develop KUGVD [42]. The same bitrate-resolution pairs from GamingVideoSet were included to define the HRCs. In total, 90 PVSs were used in the subjective evaluation and 17 participants took part in the test.

3) CLOUD GAMING VIDEO DATASET (CGVDS)

This dataset [43] consists of SRCs captured at 60 *fps* from 15 different games. For designing the HRCs, three resolutions, namely, 480*p*, 720*p* and 1080*p* at three different framerates of 20, 30 and 60 *fps* were considered. To ensure that the SRCs from all the games could be assessed by test subjects, the overall test was split into 5 different subjective tests, with a minimum of 72 PVSs being rated in each of the

TABLE 1. HRCs for Tests 1 and 2.

Resolution	Target Bitrate (Mbps)			
1920×1080	0.5	1	3.5	7
3840×1920	1	2	6	12

tests. A total of over 100 participants took part over the five different tests, with a minimum of 20 participants per test.

4) TWITCH DATASET

The Twitch Dataset [106] consists of 36 different games, with 6 games each representing one out of 6 pre-defined genres. The dataset consists of streams directly downloaded from Twitch. A total of 351 video sequences of approximately 50 *s* duration across all representations were downloaded. 90 video sequences out of these 351 video sequences were selected for subjective evaluation. Only the first 30 *s* of the chosen 90 PVSs were considered for subjective testing. Six different resolutions between 160*p* and 1080*p* at framerates of 30 and 60 *fps* were used. 29 participants rated all the 90 PVSs with no outliers being detected following the criterion of $PCC = 0.75$.

F. 360 STREAMING VIDEO QUALITY DATASET

This 360 Streaming Video Quality Dataset [44] consists of a total of three different subjective tests. The playback, subjective score and head-rotation data collection was automated using the publicly available AVTrack [107] software.[9] The participants were instructed that they could freely explore the 360° videos. A criterion based on PCC with a threshold of 0.7 was selected to detect outliers in all the three tests.

test_1 and test_2 had a joint objective of comparing the effect of different Head Mounted Displays (HMDs) on the perceived video quality. Hence, both the tests include the same SRCs and HRCs. Eight SRCs with a resolution of 3840 × 1920 pixels, framerate of 30 *fps* and a duration of 20 *s* were used in these tests. The bitrate and resolutions chosen in the two tests are detailed in Table 1. H.265 was used to encode the videos. A 2-pass encoding approach with the preset of *slow* was choosen. The eight SRCs were encoded with the defined HRCs and resulted in a total of 64 PVSs including high quality audio. In test_1, the videos were presented using an HTC Vive HMD and in test_2, using an HTC Vive Pro. The total test duration of each test was 90 minutes.

In test_1, all the 64 PVSs were rated by a total of 27 participants. 6 outliers were detected following the criterion of $PCC < 0.7$. 27 participants took part in test_2. There were 3 outliers detected in this test.

test_3 focused on the quality assessment of high resolution (> 3840 × 1920) content. For this test, seven SRCs of 7680 × 3840 pixels were selected. The framerate of the selected SRCs was 30 *fps*, and sequence duration was 20 *s*. There was no overlap with the SRCs from test_1 and test_2. The videos were encoded at three different resolutions,

[9]https://github.com/Telecommunication-Telemedia-Assessment/AVTrack360

TABLE 2. HRCs for Test 3.

Resolution	Target Bitrate (Mbps)		
3840x1920	0.5	2	6
5760x2880	1	4.5	13.5
7680x3840	2	8	24

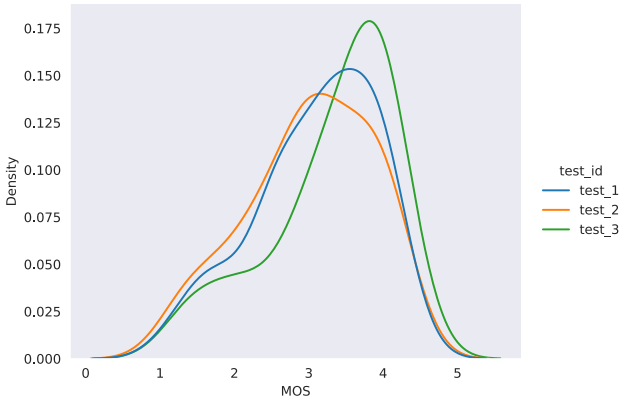

FIGURE 9. MOS distribution of 360_streaming_video_quality_dataset.

namely, 3840, 5760 × 2880 and 7680 × 3840 pixels. Three bitrates each were used for each resolution, the details of which are described in Table 2. As in test_1 and test_2, H.265 was used to encode the videos following a 2-pass encoding approach with *slow* preset. In total, 63 PVSs were rated by 27 participants, with 4 outliers detected according to the criterion of $PCC < 0.7$. The PVSs were presented with an HTC Vive Pro HMD.

The MOS distribution of the three tests is as illustrated in Figure 9.

G. LIVE-YT-HFR

The LIVE-YT-HFR [45] dataset was designed with the objective of analyzing the impact of framerate on perceived video quality, like test_4 of the AVT-VQDB-UHD-1 dataset. For this purpose, 16 SRCs captured at a framerate of 120 *fps* were used. Eleven out of the 16 SRCs are from the BVI-HFR dataset [91]. The SRCs had a duration of 6−10 *s*. They mainly consist of sports content with high motion. Six different framerates were included in the study, namely 24, 30, 60, 82, 98, and 120 fps. All the SRCs were encoded with VP9 at five different CRF values for each framerate, thus resulting in 30 PVSs for each source and in total 480 PVSs. A total of 80 participants took part in the test with each PVS being rated by a minimum of 12 participants.

VI. MODEL TRAINING

This section details the training procedure that was performed to obtain the coefficients for the different *AVQBits* models.

A. AVQBits|M3 /WWWW/P.1204.3 MODEL TRAINING

Initially, the submitted version of the *AVQBits|M3* / P.1204.3 model was trained on the 13 training databases of the PNATS-UHD-1 dataset. After the validation of the

TABLE 3. Quantization-degradation coefficients for *AVQBits|M3* / P.1204.3, PC/TV case.

Codec	a	b	c	d
H.264	4.4344	-1.7058	4.9654	-4.1203
H.264-10bit	4.6467	-0.8091	5.9835	-4.4398
H.265	4.3789	-1.0208	5.7572	-4.5625
H.265-10bit	4.5458	-0.866	6.1116	-3.3828
VP9	4.3404	-0.9961	4.5282	-3.9641

TABLE 4. Quantization-degradation coefficients for *AVQBits|M3* / P.1204.3, MO/TA case.

Codec	a	b	c	d
H.264	4.4365	-1.4909	5.4251	-4.5198
H.264-10bit	4.5399	-0.414	6.2249	-4.2599
H.265	4.3089	-0.6685	6.0551	-4.6974
H.265-10bit	4.9999	-2.6821	1.5069	-1.7664
VP9	4.4024	-1.2504	2.9268	-3.0087

model using the validation databases of the PNATS-UHD-1 dataset, a dedicated model re-optimization was done using a 5-fold cross validation approach. In this step, firstly, five splits of all the 26 databases were created with each split containing 13 training and 13 validation databases. While creating these splits, it was ensured that each of the splits had least similarity with each other and that the overall prediction difficulty of the training databases was similar to the validation databases. In addition to this, for each split, it was also ensured that the databases corresponding to different display types (PC/TV and Mobile/Tablet) have a balanced representation in the training and validation sets. Following this procedure, the coefficients of the D_q, D_u and D_t and the correspond RF components were determined. Tables 3 and 4 present the quantization-degradation–related coefficients of the PC/TV and MO/TA cases respectively. The temporal- and upscaling-related coefficients of the PC/TV and MO/TA cases are presented in Tables 5 and 6 respectively.

A more detailed description of the training procedure can be found in [39] and [22].

B. AVQBits|M0 MODEL TRAINING

For the *AVQBits|M0* model instance, a two-step training procedure was implemented to estimate the coefficients related to QP_{pred} and *quantization degradation*. In the first step, the QP_{pred} prediction module as described in Equation (22) was trained using the true QP values extracted from the 764 PVSs of AVT-PNATS-UHD-1 as ground-truth. The resulting coefficients for determining QP_{pred} are detailed in Table 7.

Following this, the coefficients in Table 7 were used to estimate QP values, and the resulting estimates QP_{pred} were used as input to the *quantization degradation*. The new coefficients were obtained by training the model using the subjective MOS from AVT-PNATS-UHD-1 as ground-truth. The resulting new coefficients of the core model are as shown in Table 8.

To estimate the *upscaling degradation* and *temporal degradation* component of the ''Core Model'', the coefficients for the *AVQBits|M3* / P.1204.3 model reported in Table 5 are used.

TABLE 5. Upscaling- and temporal-degradation coefficients, PC/TV case (valid across all four *AVQBits* instances).

x	y	k	z
-9.5497	1.1999	4.1696	-8.3084

TABLE 6. Upscaling- and temporal-degradation coefficients, MO/TA case (valid across all four *AVQBits* instances).

x	y	k	z
-8.4690	1.1999	4.2701	-6.3648

TABLE 7. QP-Prediction coefficients for *AVQBits|M0*, PC/TV case.

Codec	a_{qp_m0}	b_{qp_m0}	c_{qp_m0}	d_{qp_m0}
H.264	-5.7284	-5.3586	4.1965	5.6231
H.265	-7.6866	-6.0256	4.8298	4.0869
VP9	-140.8384	-46.5290	37.5395	27.5876

TABLE 8. Quantization-degradation coefficients for *AVQBits|M0*, PC/TV case.

Codec	a	b	c	d
H.264	4.7342	-0.9469	4.0831	-2.0624
H.265	4.5731	-0.6835	3.3163	-1.4604
VP9	4.2624	-0.6135	3.2368	-2.2657

TABLE 9. QP-Prediction coefficients for *AVQBits|M1*, PC/TV case.

Codec	a_{qp_m0}	b_{qp_m0}	c_{qp_m0}	d_{qp_m0}	e_{qp_m0}
H.264	28.4333	-7.3951	5.7821	0.2479	-5.4537
H.265	22.3936	-6.5297	5.1573	-0.8999	-2.2889
VP9	92.1245	-51.1209	40.6832	-10.2195	-18.7809

TABLE 10. Quantization-degradation coefficients for *AVQBits|M1*, PC/TV case.

Codec	a	b	c	d
H.264	4.6602	-1.1312	4.2268	-2.4471
H.265	4.5375	-0.6829	3.5053	-1.6074
VP9	4.5253	-1.2635	2.0732	-1.8051

As there were no Mobile/Tablet (MO/TA) databases that were part of the training dataset, AVT-PNATS-UHD-1, a synthetic dataset consisting of *AVQBits|M3* / P.1204.3 was developed to determine the coefficients for the MO/TA case. These coefficients can be found in the reference implementation that is publicly available.

C. AVQBits|M1 MODEL TRAINING

For the *AVQBits|M1* model instance, the same two-step training approach as for Mode 0 was used to determine the coefficients related to QP_{pred} and subsequently the *quantization degradation* component D_q of the ''Core Model'', cf. Equation (5) to Equation (7). The coefficients related to QP_{pred} and the *quantization degradation* D_q (i.e. MOS_q at first, cf. Equation (5)) are presented in Tables 9 and 10, respectively. Similar to the Mode 0 model *AVQBits|M0*, the coefficients for the *AVQBits|M3* / P.1204.3 model detailed in Table 5 are used to estimate the *upscaling degradation* and *temporal degradation*.

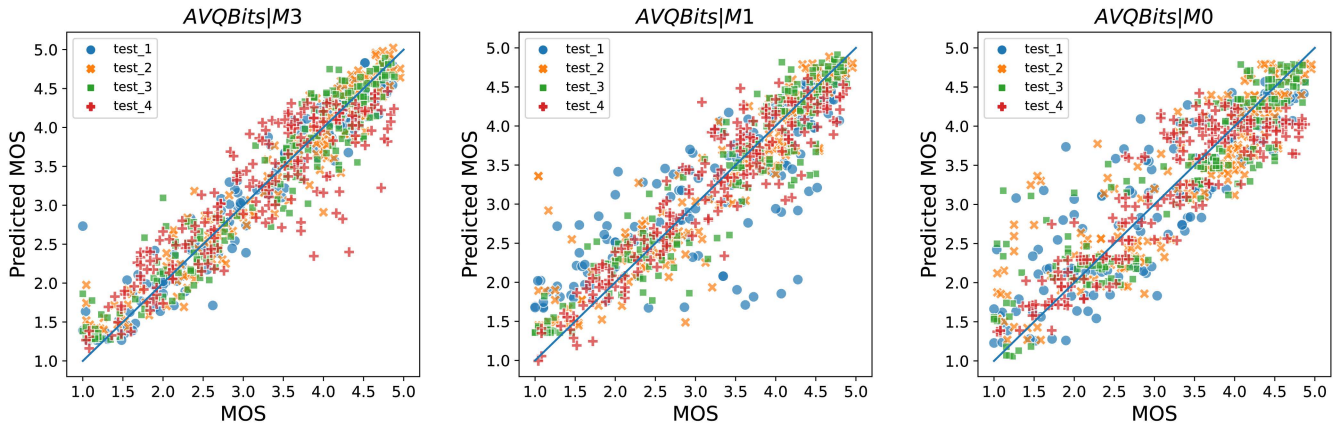

FIGURE 10. Scatter plot of P.1204.3 and its bitstream extensions for AVT-VQDB-UHD-1 dataset.

TABLE 11. Codec mapping coefficients for *AVQBits*|*H*0|*f*, PC/TV case.

Codec	a_{cmap}	a_{cmap}
H.264	0.9053	0.0931
VP9	0.8530	0.6979

A similar approach to determine the MO/TA coeffiencts for the *AVQBits*|*M*0 model was used for the *AVQBits*|*M*1 model and the corresponding coefficients can be found in the publicly available reference implementation.

D. AVQBits|H0 MODEL TRAINING

As discussed in Section III-D, two different variants of the *AVQBits*|*H*0 model are proposed in this paper. The first, *AVQBits*|*H*0|*s*, applies the same encoder used for initially encoding the video to be evaluated, plus Mode 0 data for a quality-equivalent re-encoding of the video. The *AVQBits*|*M*3 / P.1204.3 model is then directly applied to the resulting bitstream, without any further modifications of the model. Hence, no additional training of the *AVQBits*|*H*0|*s* model is needed.

Instead of the original video codec, that has been used for encoding the distorted video, the *AVQBits*|*H*0|*f* model has a fixed video encoder for generating the quality-equivalent bitstream. For the model instance presented in this paper, H.265 is selected. As a result, the prediction from the *AVQBits*|*M*3 / P.1204.3 model requires a codec-specific mapping of the predicted score to represent the quality that would be provided by the originally applied encoder. Hence, a simple mapping function as described in Equation (27) is proposed. MOS scores from AVT-PNATS-UHD-1 are applied as training target to determine the coefficients of the mapping function. The resulting coefficients are presented in Table 11.

VII. EVALUATION

In this section, the performance evaluation of the four different *AVQBits* instances on different application scopes such as traditional 2D video, gaming video, 360° video and HFR video are described. The evaluation is performed on the different publicly available databases described in Sec. V. For all evaluations, the performance measures are computed after performing a linear fit per each database, as recommended in ITU-T Rec. P.1401 [108], to map the objective scores to the subjective scores. This way, a possible test-specific linear bias in comparison to the model is compensated (e.g., range-equalization bias [109]). The scatter plots presented in the following also use the linearly-mapped predictions. The performance of the models are compared with SoA models for each use-case.

A. SHORT-TERM VIDEO QUALITY

The first evaluation of the proposed models is conducted on short videos with $8 - 10\,s$ duration, as this was the primary focus of model development. For this purpose, the publicly available AVT-VQDB-UHD-1 dataset [40] consisting of 756 PVSs is used, see also Sec. V. Note that only 432 PVSs are publicly available due to source copyright issues. However, in this paper the evaluation is performed on the entire dataset consisting of 756 PVSs, as the authors have access to the complete set. Table 12 provides a detailed overview of the performance of the model instances *AVQBits*|*M*3 P.1204.3, *AVQBits*|*M*0, *AVQBits*|*M*1 and the two versions of the Hybrid Mode 0 model *AVQBits*|*H*0. Performance is given in terms of RMSE, PCC, Spearman Rank Correlation (SROCC), Kendall correlation and R^2 Score for the four tests individually and all databases together. As is expected, *AVQBits*|*M*3 / P.1204.3 outperforms all other model instances for all databases combined as it has access to the entire bitstream to estimate video quality. An interesting observation is that the other model instances perform slightly better than *AVQBits*|*M*3 / P.1204.3 for *test_4*. This specific test considers a wide range of framerate variations. It should be noted that such a high variation in framerate between SRC and PVS is rather unrealistic for HAS applications. However, it can be seen from Table 12, that *AVQBits*|*M*3 / P.1204.3 performs significantly better across all databases.

Figures 10 and 11 show the scatter plots for all models. It can be observed that *AVQBits*|*M*3 / P.1204.3 leads to

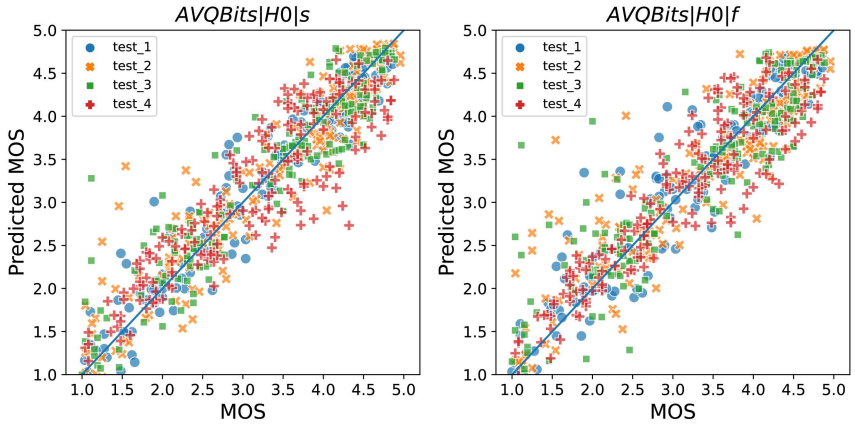

FIGURE 11. Scatter plot of the Hybrid No-Reference extension of P.1204.3 for AVT-VQDB-UHD-1 dataset.

TABLE 12. Performance of the *AVQBits* instances on the AVT-VQDB-UHD-1 dataset (*The RMSE and R^2 numbers for *AVQBits*|*M*3 / P.1204.3 may differ from the ones reported in [39], as here the RMSE and R^2 values after linear mapping whereas in [39] the RMSE and R^2 values were calculated on raw predictions).

Database	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
test_1	$AVQBits\|M3$ / P.1204.3*	0.280	0.968	0.953	0.822	0.937
test_1	$AVQBits\|M1$	0.614	0.836	0.851	0.677	0.699
test_1	$AVQBits\|M0$	0.507	0.891	0.888	0.703	0.795
test_1	$AVQBits\|H0\|s$	0.298	0.964	0.954	0.817	0.929
test_1	$AVQBits\|H0\|f$	0.324	0.957	0.946	0.805	0.916
test_2	$AVQBits\|M3$ / P.1204.3*	0.287	0.966	0.960	0.830	0.934
test_2	$AVQBits\|M1$	0.441	0.918	0.930	0.780	0.844
test_2	$AVQBits\|M0$	0.511	0.889	0.895	0.714	0.790
test_2	$AVQBits\|H0\|s$	0.394	0.936	0.934	0.782	0.875
test_2	$AVQBits\|H0\|f$	0.451	0.915	0.916	0.752	0.837
test_3	$AVQBits\|M3$ / P.1204.3*	0.324	0.957	0.935	0.785	0.917
test_3	$AVQBits\|M1$	0.363	0.946	0.924	0.766	0.895
test_3	$AVQBits\|M0$	0.464	0.911	0.896	0.712	0.830
test_3	$AVQBits\|H0\|s$	0.395	0.936	0.920	0.756	0.877
test_3	$AVQBits\|H0\|f$	0.450	0.916	0.908	0.737	0.840
test_4	$AVQBits\|M3$ / P.1204.3*	0.485	0.876	0.853	0.681	0.767
test_4	$AVQBits\|M1$	0.366	0.931	0.911	0.756	0.867
test_4	$AVQBits\|M0$	0.443	0.897	0.851	0.673	0.805
test_4	$AVQBits\|H0\|s$	0.460	0.889	0.876	0.699	0.790
test_4	$AVQBits\|H0\|f$	0.405	0.915	0.898	0.734	0.837
All	$AVQBits\|M3$ / P.1204.3*	0.370	0.942	0.927	0.768	0.887
All	$AVQBits\|M1$	0.476	0.901	0.900	0.730	0.811
All	$AVQBits\|M0$	0.499	0.890	0.877	0.684	0.792
All	$AVQBits\|H0\|s$	0.408	0.928	0.919	0.755	0.861
All	$AVQBits\|H0\|f$	0.433	0.919	0.909	0.743	0.844

very few outliers as compared to the subjective tests, whereas results for the other instances show a larger number of outliers. Most notably, it can be observed that for Mode 1 *AVQBits*|*M*1, the *Surfing* sequence suffers from under-prediction in a few cases due to a large error in the method used for QP estimation in this case. In addition, it can also be seen that *AVQBits*|*M*0 suffers from slightly more over-prediction, which is a result of the lack of source-specific information for quality estimation, which *AVQBits*|*M*3 and hence also *AVQBits*|*H*0 partly handle in the random forest model part. Also, *AVQBits*|*H*0|*f* would benefit from a more sophisticated codec mapping than the linear one defined in Section III-D to better take into account codec-specific differences.

TABLE 13. Performance comparison of the *AVQBits* instances with SoA models for tests in the AVT-VQDB-UHD-1 dataset without framerate as dependent variable (*The RMSE and R^2 numbers for *AVQBits*|*M*3 / P.1204.3 may differ to the ones reported in [39], as here the RMSE and R^2 values after linear mapping are shown, whereas in [39] the RMSE and R^2 values were calculated on raw predictions).

Model	RMSE	PCC	SROCC	Kendall	R^2 Score
VMAF [23]	0.531	0.880	0.889	0.721	0.774
Brisque [110]	0.653	0.815	0.838	0.653	0.660
NIQE [111]	1.009	0.432	0.445	0.301	0.187
PSNR	1.109	0.131	0.682	0.531	0.017
SSIM [112]	0.956	0.520	0.761	0.569	0.270
MS-SSIM [113]	0.896	0.599	0.752	0.563	0.358
ADM2 [114]	0.580	0.855	0.874	0.698	0.731
VIFP [115]	0.757	0.736	0.756	0.562	0.542
$AVQBits\|M3$ / P.1204.3*	0.306	0.962	0.948	0.804	0.925
$AVQBits\|M1$	0.486	0.901	0.904	0.738	0.812
$AVQBits\|M0$	0.503	0.894	0.891	0.701	0.799
$AVQBits\|H0\|s$	0.373	0.943	0.935	0.778	0.889
$AVQBits\|H0\|f$	0.439	0.920	0.914	0.749	0.846

The performance of the *AVQBits* instances is also compared with that of SoA models. For this purpose, the performance numbers for SoA models on the AVT-VQDB-UHD-1 dataset reported in [22] are used. In Tables 13 and 14, different FR and NR models are compared with the proposed models for tests with and without framerate variation separately. As can be seen from the results, *AVQBits*|*M*3 / P.1204.3 is the best performing model across all tests, with VMAF being the best performing FR model. In spite of the reduced input data for these models, the other *AVQBits* instances are still able to outperform a number of the SoA models. For example, *AVQBits*|*M*0 shows a better performance than Brisque and SSIM, or *AVQBits*|*M*1 shows a better performance than VMAF. The hybrid models also outperform VMAF and generally are surpassed only by *AVQBits*|*M*3 / P.1204.3 in performance. It is noted that the good performance of the *AVQBits* instances other than

TABLE 14. Performance comparison of *AVQBits* instances with SoA models for tests with framerate as independent variable in the AVT-VQDB-UHD-1 dataset (*The RMSE and R^2 numbers for P.1204.3 may differ to the ones reported in [39], as here the RMSE and R^2 values after linear mapping are shown, whereas in [39] the RMSE and R^2 values were calculated on raw predictions).

Model	RMSE	PCC	SROCC	Kendall	R^2 Score
VMAF [23]	0.592	0.807	0.811	0.624	0.652
Brisque [110]	0.641	0.813	0.833	0.646	0.657
NIQE [111]	1.006	0.393	0.387	0.265	0.154
PSNR	1.004	0.313	0.491	0.352	0.000
SSIM [112]	0.871	0.497	0.580	0.418	0.247
MS-SSIM [113]	0.832	0.559	0.581	0.421	0.312
ADM2 [114]	0.598	0.803	0.806	0.615	0.644
VIFP [115]	0.789	0.618	0.612	0.449	0.381
$AVQBits\|M3$ / P.1204.3*	0.485	0.876	0.853	0.681	0.767
$AVQBits\|M1$	0.366	0.931	0.911	0.756	0.867
$AVQBits\|M0$	0.443	0.897	0.851	0.673	0.805
$AVQBits\|H0\|s$	0.460	0.889	0.876	0.699	0.790
$AVQBits\|H0\|f$	0.405	0.915	0.898	0.734	0.837

AVQBits|*M*3 / P.1204.3 may be due to the selected specific encoding settings and their range. In general, *test_4* seems to be the most difficult test in terms of estimating video quality, due to the wide range of framerates included in this test. The comparatively bad performance of VMAF for *test_4* can be attributed to the lack of a sophisticated motion-related feature in the model.

B. OVERALL INTEGRAL QUALITY

As next, the proposed long-term integration model, a simplified version of ITU-T Rec. P.1203.3 [36] is evaluated on PNATS-UHD-1-Long consisting of five tests with PVSs ranging from 1-5 min in duration (cf. Sec. V). As explained in Section IV, for estimating the overall integral quality, the proposed model follows the same architecture as ITU-T Rec. P.1203.3 and takes per-1-second video and audio scores as input, along with stalling-related information. In this evaluation, to estimate the per-1-second video quality scores, the different *AVQBits* instances are considered. The per-1-second audio quality scores are assumed to be 4.5, which is the highest quality estimated by ITU-T Rec. P.1203.2 [62]. This assumption is based on the fact that the audio quality is not varied and the best possible audio quality is used in any of the five tests considered for evaluation.

Table 15 shows the performance numbers for all the tests for the proposed models. It can be concluded that using *AVQBits*|*M*3 P.1204.3 to estimate the per-1-second scores results in very good performance of the proposed long-term integration model. This is due to the high accuracy of the ITU-T P.1204.3 model. The estimation of per-1-second and per-segment quality scores is better as compared to the other instances of *AVQBits*, which use less complex input information without full bitstream access for video quality prediction. Furthermore, it can be observed that the *AVQBits*|*H*0|*s* and *AVQBits*|*H*0|*f* variants show similar performance to the *AVQBits*|*M*3 P.1204.3 in terms of PCC but have a worse performance in terms of RMSE for each of the five tests. This is unlike the short-term video quality prediction where

TABLE 15. Performance of P.1204.3 and its extensions on the PNATS-UHD-1-Long dataset.

Database	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
test_1	$AVQBits\|M3$ / P.1204.3	0.353	0.892	0.823	0.649	0.795
test_1	$AVQBits\|M1$	0.432	0.831	0.822	0.655	0.691
test_1	$AVQBits\|M0$	0.399	0.859	0.863	0.694	0.738
test_1	$AVQBits\|H0\|s$	0.357	0.888	0.870	0.706	0.789
test_1	$AVQBits\|H0\|f$	0.352	0.891	0.868	0.705	0.798
test_2	$AVQBits\|M3$ / P.1204.3	0.559	0.813	0.801	0.610	0.661
test_2	$AVQBits\|M1$	0.624	0.760	0.744	0.562	0.577
test_2	$AVQBits\|M0$	0.658	0.728	0.707	0.514	0.529
test_2	$AVQBits\|H0\|s$	0.640	0.745	0.689	0.511	0.556
test_2	$AVQBits\|H0\|f$	0.650	0.736	0.679	0.498	0.541
test_3	$AVQBits\|M3$ / P.1204.3	0.485	0.888	0.798	0.630	0.788
test_3	$AVQBits\|M1$	0.552	0.852	0.809	0.644	0.725
test_3	$AVQBits\|M0$	0.641	0.794	0.749	0.570	0.630
test_3	$AVQBits\|H0\|s$	0.514	0.873	0.804	0.635	0.762
test_3	$AVQBits\|H0\|f$	0.540	0.858	0.818	0.658	0.737
test_4	$AVQBits\|M3$ / P.1204.3	0.377	0.917	0.899	0.748	0.842
test_4	$AVQBits\|M1$	0.517	0.838	0.798	0.627	0.702
test_4	$AVQBits\|M0$	0.534	0.826	0.782	0.609	0.683
test_4	$AVQBits\|H0\|s$	0.392	0.910	0.878	0.729	0.829
test_4	$AVQBits\|H0\|f$	0.370	0.920	0.878	0.715	0.847
test_5	$AVQBits\|M3$ / P.1204.3	0.386	0.934	0.922	0.796	0.872
test_5	$AVQBits\|M1$	0.700	0.762	0.796	0.641	0.581
test_5	$AVQBits\|M0$	0.832	0.638	0.594	0.464	0.407
test_5	$AVQBits\|H0\|s$	0.502	0.855	0.842	0.684	0.732
test_5	$AVQBits\|H0\|f$	0.500	0.857	0.825	0.658	0.734
All	$AVQBits\|M3$ / P.1204.3	0.479	0.864	0.844	0.660	0.747
All	$AVQBits\|M1$	0.596	0.780	0.787	0.602	0.608
All	$AVQBits\|M0$	0.694	0.686	0.683	0.500	0.471
All	$AVQBits\|H0\|s$	0.570	0.797	0.768	0.584	0.635
All	$AVQBits\|H0\|f$	0.582	0.787	0.756	0.572	0.619

the *AVQBits*|*H*0|*s* and *AVQBits*|*H*0|*f* variants have a similar performance to *AVQBits*|*M*3 P.1204.3 both in terms of PCC and RMSE. This can be attributed to the fact that in case of short-term video quality prediction a simple linear mapping according to ITU-T P.1401 before computing the RMSE would accommodate for the difference in prediction due to the usage of the QEB instead of the original bitstream. Whereas, in the case of overall integral quality prediction, the input consists of per-1-sec scores and these per-1-sec scores are computed on the QEB which may not reflect the true quality directly of the bitstream. A dedicated linear mapping of the per-1-sec scores to take into account the effect of QEB at the per-1-sec level could alleviate such a problem and hence result in a lower RMSE value. Despite this, the overall performance of the *AVQBits*|*H*0|*s* and *AVQBits*|*H*0|*f* variants is significantly better than the *AVQBits*|*M*0 and *AVQBits*|*M*1 models.

The scatter plots illustrated in Figure 12 show that both *AVQBits*|*M*0 and *AVQBits*|*M*1 seem to over-predict in the lower-quality range, which can be attributed to the less accurate per-1-second score estimation by these models. This is assumed to reflect that the quality-impact due to more encoder-demanding video content is less well captured by these models.

TABLE 16. Performance of P.1204.3 and its extensions on different gaming datasets.

Dataset	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
GVS	$AVQBits\|M3$ / P.1204.3	0.45	0.88	0.87	0.69	0.77
GVS	$AVQBits\|M1$	0.42	0.89	0.87	0.71	0.79
GVS	$AVQBits\|M0$	0.69	0.67	0.65	0.49	0.45
GVS	$AVQBits\|H0\|s$	0.48	0.86	0.86	0.69	0.74
GVS	$AVQBits\|H0\|f$	0.62	0.75	0.73	0.56	0.56
KUGVD	$AVQBits\|M3$ / P.1204.3	0.39	0.93	0.92	0.77	0.86
KUGVD	$AVQBits\|M1$	0.50	0.87	0.86	0.69	0.76
KUGVD	$AVQBits\|M0$	0.84	0.59	0.57	0.41	0.35
KUGVD	$AVQBits\|H0\|s$	0.46	0.90	0.89	0.72	0.80
KUGVD	$AVQBits\|H0\|f$	0.65	0.78	0.76	0.58	0.61
CGVDS	$AVQBits\|M3$ / P.1204.3	0.38	0.85	0.84	0.65	0.72
CGVDS	$AVQBits\|M1$	0.36	0.90	0.88	0.70	0.78
CGVDS	$AVQBits\|M0$	0.47	0.78	0.75	0.56	0.60
CGVDS	$AVQBits\|H0\|s$	0.36	0.89	0.88	0.70	0.79
CGVDS	$AVQBits\|H0\|f$	0.38	0.87	0.87	0.68	0.76
Twitch	$AVQBits\|M3$ / P.1204.3	0.40	0.93	0.93	0.77	0.87
Twitch	$AVQBits\|M1$	0.37	0.94	0.93	0.77	0.89
Twitch	$AVQBits\|M0$	0.43	0.92	0.89	0.71	0.85
Twitch	$AVQBits\|H0\|s$	0.31	0.96	0.95	0.82	0.92
Twitch	$AVQBits\|H0\|f$	0.30	0.96	0.95	0.81	0.92
All	$AVQBits\|M3$ / P.1204.3	0.41	0.90	0.90	0.73	0.81
All	$AVQBits\|M1$	0.41	0.90	0.89	0.73	0.81
All	$AVQBits\|M0$	0.60	0.76	0.75	0.56	0.58
All	$AVQBits\|H0\|s$	0.40	0.90	0.90	0.73	0.82
All	$AVQBits\|H0\|f$	0.48	0.86	0.85	0.67	0.73

C. GAMING VIDEO QUALITY

As the first application for extending the initial scope of "traditional" video quality prediction for the *AVQBits* instances, gaming video quality is considered. It should be noted that all model instances of *AVQBits* are used without any retraining to estimate the video quality for the four gaming datasets considered for performance evaluation (see Sec. V). The only difference to the case of "normal" 2D video is that here all databases were created for a full HD ($1920 \cdot 1080$ pixels) display instead of the 4K/UHD-1 target screen resolution used in case of the PC-databases for "normal" video and the initial development of *AVQBits*|*M*3 / P.1204.3 in ITU-T SG12. In this paper, *AVQBits*|*M*3 / P.1204.3 and its extensions are used directly with the target resolution of 4K/UHD-1.

Table 16 provides a detailed view of the performance of the proposed bitstream-based models on all the four considered tests. *AVQBits*|*M*3 / P.1204.3 and *AVQBits*|*M*1 perform on par across all datasets, with Mode 0 being the least well performing model. The good performance of the Mode 1 model indicates that the features related to framesize and frame type can be used to estimate the impact of content improving the estimation of the QP value and bringing it closer to the one of the *AVQBits*|*M*3 model with its full bitstream access. Although Mode 1 performs on par with *AVQBits*|*M*3 / ITU-T P.1204.3 on average, from the scatter plots shown in Figure 13 it can be observed that there is a general tendency of the Mode 1 model to slightly over-predict as compared to ITU-T P.1204.3. Furthermore, it can be observed from the scatter plot associated with Mode 0 in Figure 13 that Mode 0 suffers significantly from the lack of content-related features, leading to cases with a larger prediction inaccuracy. As the main goal of gaming video quality prediction is to run it in an environment with less requirements for computation, the hybrid models may be less practical for real-time monitoring, as they need additional resources. However, it is shown that in case Mode 0 type data and pixel information can be accessed in a practical monitoring scenario, the *AVQBits*|*H*0 models are highly usable. The results show that *AVQBits*|*H*0|*s* performs as well as *AVQBits*|*M*3 / P.1204.3 for all the four considered gaming datasets. The *AVQBits*|*H*0|*f* model variant with less requirements on the set of codecs available during monitoring performs on par with *AVQBits*|*M*3 / P.1204.3 for the CGVDS and Twitch datasets, but less well for GVS and KUGVD. This may be due to the coefficients a_{map} and b_{map} being obtained by training on traditional 2D video datasets. A dedicated retraining of these two coefficients for gaming content may result in improved performance.

TABLE 17. Comparison of performance of P.1204.3 and its extensions with SoA models on different gaming datasets.

Dataset	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
GVS	PSNR	0.63	0.74	0.74	0.57	0.55
GVS	SSIM	0.57	0.80	0.80	0.61	0.62
GVS	VMAF	0.47	0.87	0.86	0.69	0.75
GVS	NIQE	0.64	0.77	0.71	0.53	0.52
GVS	$AVQBits\|M3$ / P.1204.3	0.45	0.88	0.87	0.69	0.77
GVS	$AVQBits\|M1$	0.42	0.89	0.87	0.71	0.79
GVS	$AVQBits\|M0$	0.69	0.67	0.65	0.49	0.45
GVS	$AVQBits\|H0\|s$	0.48	0.86	0.86	0.69	0.74
GVS	$AVQBits\|H0\|f$	0.62	0.75	0.73	0.56	0.56
KUGVD	PSNR	0.62	0.80	0.84	0.67	0.64
KUGVD	SSIM	0.48	0.89	0.91	0.74	0.79
KUGVD	VMAF	0.41	0.92	0.92	0.77	0.85
KUGVD	NIQE	0.55	0.85	0.84	0.66	0.72
KUGVD	$AVQBits\|M3$ / P.1204.3	0.39	0.93	0.92	0.77	0.86
KUGVD	$AVQBits\|M1$	0.50	0.87	0.86	0.69	0.76
KUGVD	$AVQBits\|M0$	0.84	0.59	0.57	0.41	0.35
KUGVD	$AVQBits\|H0\|s$	0.46	0.90	0.89	0.72	0.80
KUGVD	$AVQBits\|H0\|f$	0.65	0.78	0.76	0.58	0.61
CGVDS	PSNR	0.60	0.64	0.65	0.47	0.41
CGVDS	SSIM	0.59	0.67	0.78	0.60	0.45
CGVDS	VMAF	0.38	0.88	0.87	0.69	0.77
CGVDS	NIQE	0.66	0.54	0.56	0.41	0.29
CGVDS	$AVQBits\|M3$ / P.1204.3	0.38	0.85	0.84	0.65	0.72
CGVDS	$AVQBits\|M1$	0.36	0.90	0.88	0.70	0.78
CGVDS	$AVQBits\|M0$	0.47	0.78	0.75	0.56	0.60
CGVDS	$AVQBits\|H0\|s$	0.36	0.89	0.88	0.70	0.79
CGVDS	$AVQBits\|H0\|f$	0.38	0.87	0.87	0.68	0.76
Twitch	NIQE	0.96	0.24	0.11	0.17	0.04
Twitch	$AVQBits\|M3$ / P.1204.3	0.40	0.93	0.93	0.77	0.87
Twitch	$AVQBits\|M1$	0.37	0.94	0.93	0.77	0.89
Twitch	$AVQBits\|M0$	0.43	0.92	0.89	0.71	0.85
Twitch	$AVQBits\|H0\|s$	0.31	0.96	0.95	0.82	0.92
Twitch	$AVQBits\|H0\|f$	0.30	0.96	0.95	0.81	0.92

In addition to this, the performance of the proposed bitstream-based models is compared with SoA models, and the details are reported in Table 17. The performance numbers corresponding to the SoA models relating to the open datasets

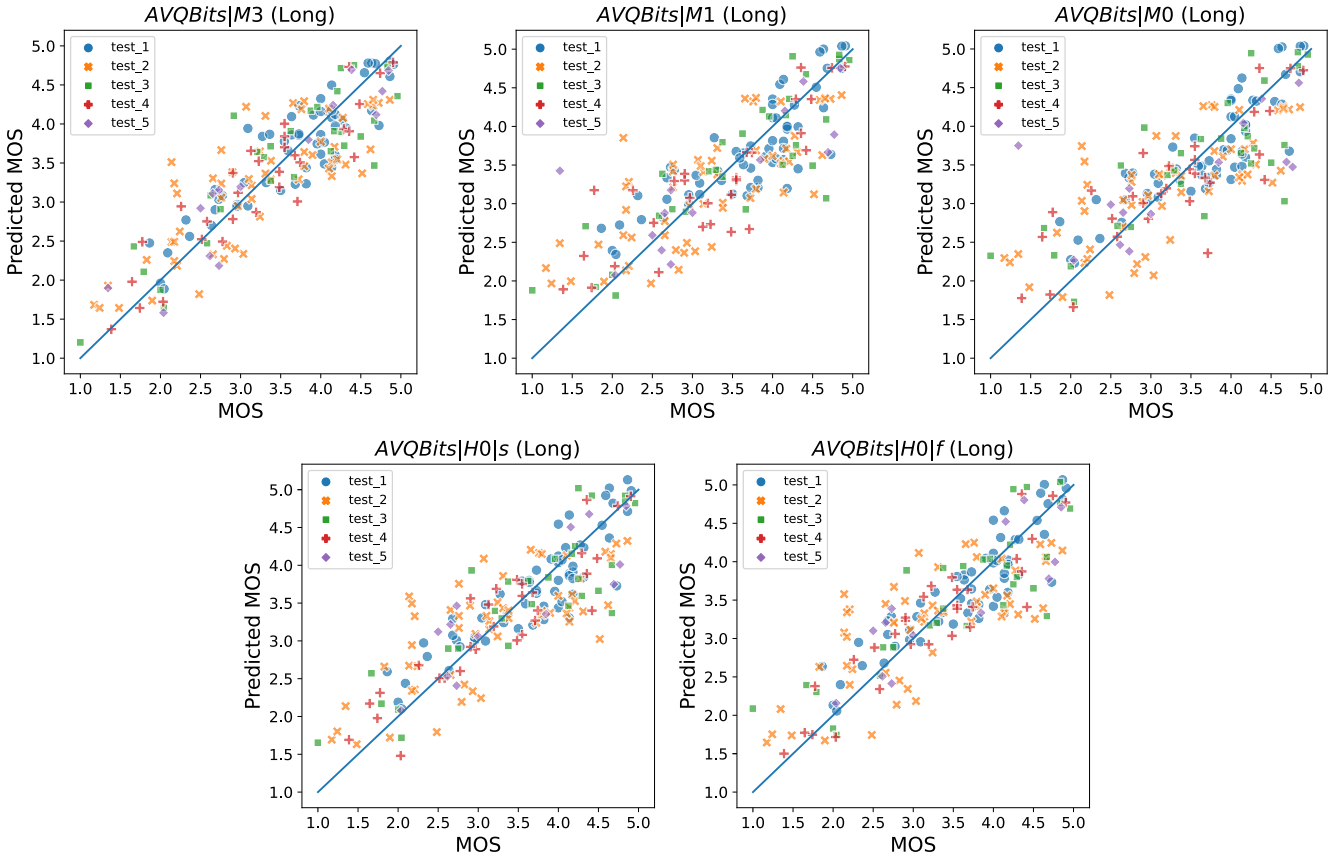

FIGURE 12. Scatter plot of P.1204.3 and its extensions for PNATS-UHD-1-Long dataset.

are directly taken from respective papers. In general, it can be concluded that *AVQBits*|*M*3 / P.1204.3 and *AVQBits*|*M*1 perform on par with VMAF across all datasets. It should be noted that two out of the four datasets, namely, CGVDS and the Twitch dataset use completely different encoding strategies than the ones these models were trained on. For CGVDS, a hardware-accelerated encoder was applied, and the Twitch dataset consists of PVSs with proprietary Twitch encoding. Despite this, the models perform well, indicating the generalizability of the model framework with regard to different encoder implementations and strategies. Although the Mode 0 *AVQBits*|*M*0 model is the least well performing bitstream model, it still outperforms all the considered NR models for all datasets. The performance of the Mode 0 model could be enhanced by retraining it for the gaming-specific use-case. Furthermore, *AVQBits*|*H*0|*s* performs as well as both the best performing pixel and bitstream models. Although the fixed encoder variant *AVQBits*|*H*0|*f* suffers somewhat from a lower performance for the GVS and KUGVD datasets, the average performance across all the four datasets is still competitive in comparison with the SoA models.

A FHD-specific mapping *AVQBits*|*M*3 / P.1204.3 for gaming content has been proposed in [106] with a more detailed handling of the target screen resolution being a topic of ongoing work.

D. 360° VIDEO QUALITY

As the next application scope for evaluation, 360° video quality estimation is considered. For this purpose, the 360 Streaming Video Quality Dataset (cf. Sec. V-F) consisting of three different tests is considered. As in the case of gaming, no retraining of the proposed models has been performed. In addition to this, as with the gaming use-case, in this paper, *AVQBits*|*M*3 / P.1204.3 and its extensions are applied directly with the target resolution of 4K/UHD-1 despite the different tests being considered have different target resolutions. Table 18 provides a detailed view of the performance numbers for all the tests for the proposed bitstream-based models.

In general, it can be observed from Table 18 that *AVQBits*|*M*3 / P.1204.3 performs well for all the tests. Mode 0 (*AVQBits*|*M*0) and Mode 1 (*AVQBits*|*M*1) show satisfactory performance for test_1 and test_2, but perform considerably worse for test_3. The general tendency towards worse performance of these models can be attributed to the fact that the QP estimation is not optimal for 360° video, as encoders may use different strategies in QP selection for specific bitrates. Hence, a more use-case specific QP estimation should be considered to enhance the model performance. Especially for these low-complexity bitstream models, a dedicated model could be used, which is usually even how it is handled for existing 2D video streaming applications, due to the sheer amount of different encoding

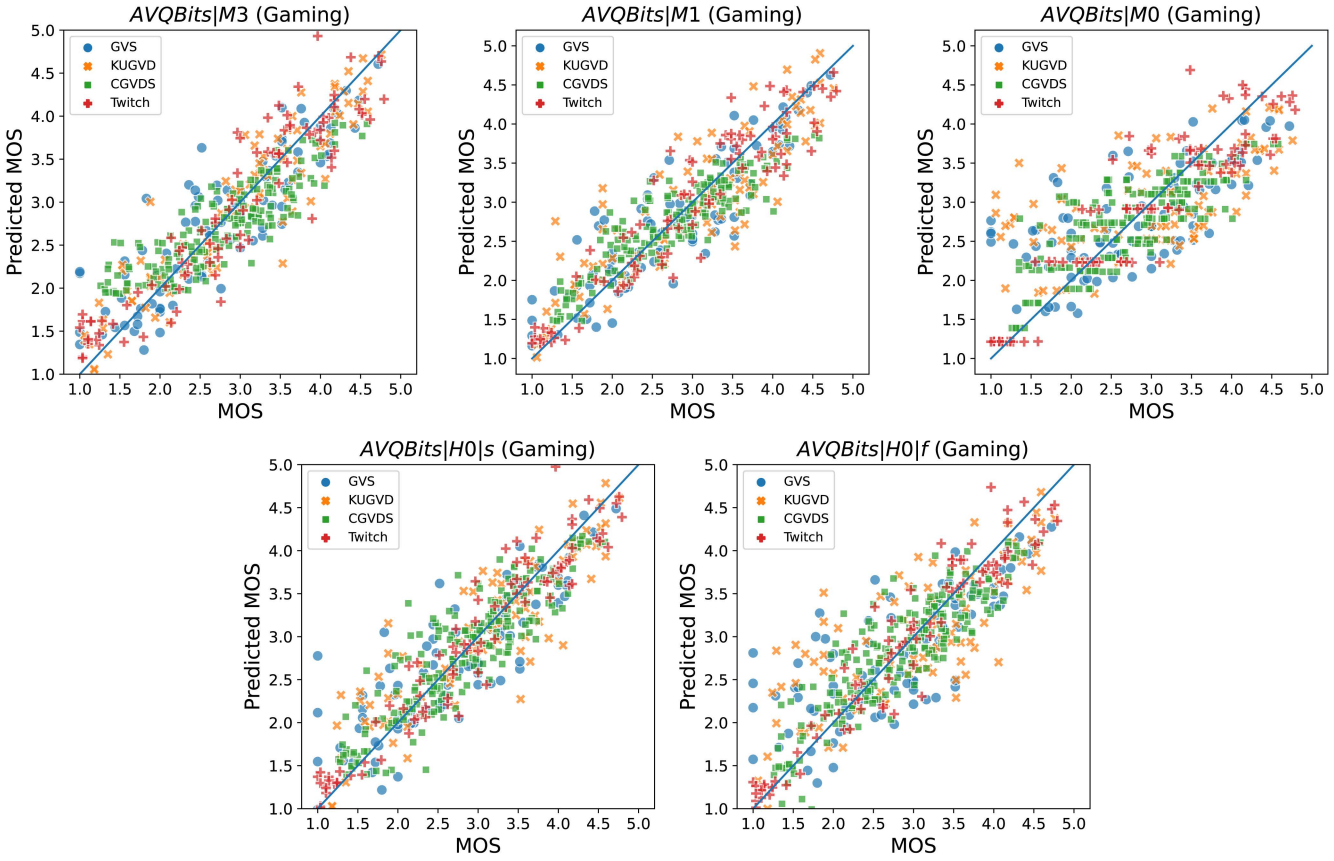

FIGURE 13. Scatter plot of P.1204.3 and its extensions for all Gaming datasets.

TABLE 18. Performance of P.1204.3 and its extensions on the 360 Streaming Video Dataset.

Test	Model	RMSE	PCC	SROCC	Kendall	R^2 Score
test_1	$AVQBits\|M3$ / P.1204.3	0.319	0.917	0.880	0.709	0.841
test_1	$AVQBits\|M1$	0.482	0.798	0.785	0.612	0.637
test_1	$AVQBits\|M0$	0.558	0.717	0.757	0.578	0.514
test_1	$AVQBits\|H0\|s$	0.343	0.903	0.872	0.700	0.816
test_2	$AVQBits\|M3$ / P.1204.3	0.314	0.926	0.917	0.757	0.858
test_2	$AVQBits\|M1$	0.452	0.841	0.849	0.669	0.707
test_2	$AVQBits\|M0$	0.527	0.775	0.834	0.653	0.600
test_2	$AVQBits\|H0\|s$	0.314	0.927	0.916	0.753	0.859
test_3	$AVQBits\|M3$ / P.1204.3	0.495	0.824	0.707	0.504	0.679
test_3	$AVQBits\|M1$	0.780	0.324	0.224	0.134	0.105
test_3	$AVQBits\|M0$	0.770	0.382	0.405	0.267	0.146
test_3	$AVQBits\|H0\|s$	0.395	0.880	0.796	0.596	0.775

TABLE 19. Comparison of performance of P.1204.3 and its extensions with SoA models on the 360 Video Streaming Quality Dataset.

Model	RMSE	PCC	SROCC	Kendall
Hybrid [44]	0.425	0.891	0.890	0.714
Mode 0 [44]	0.503	0.832	0.865	0.680
VMAF_cc [84]	0.384	0.898	0.872	0.700
VMAF [23]	0.431	0.870	0.834	0.664
ADM2 [114]	0.494	0.825	0.819	0.640
WS_SSIM	0.500	0.820	0.864	0.671
VIFP [115]	0.554	0.773	0.656	0.502
WS_PSNR	0.598	0.729	0.767	0.582
SSIM [112]	0.622	0.702	0.730	0.563
PSNR	0.762	0.489	0.627	0.469
$AVQBits\|M3$ / P.1204.3	0.377	0.894	0.870	0.679
$AVQBits\|M1$	0.581	0.709	0.677	0.497
$AVQBits\|M0$	0.627	0.658	0.686	0.401
$AVQBits\|H0\|s$	0.356	0.906	0.886	0.695

strategies [116]. The difference in performance for the different bitstream-based models can also be observed in the scatter plots depicted in Figure 14. Here, it can be seen that both *AVQBits*|*M*0 and *AVQBits*|*M*1 suffer from large prediction errors for certain cases. The difference in performance between the proposed models is most prominent for test_3 which involved comparison between 4K, 6K, and 8K 360° videos. It should be noted that the proposed models have only been trained and validated on videos up to 4K/UHD-1 resolution. Furthermore, from the results for the Hybird No-reference Mode 0 model *AVQBits*|*H*0|*s* it can be seen that the model performs well for all the three tests, and is on par with the performance of *AVQBits*|*M*3 / P.1204.3. The *AVQBits*|*H*0|*s* model performs significantly better than Mode 0 and Mode 1 due to its ability to better estimate the complexity of the content compared to either Mode 0 or Mode 1, as it can use the entire bitstream information of the QEB. *AVQBits*|*H*0|*f* is not explicitly considered for evaluation because the codec used to encode videos in the test

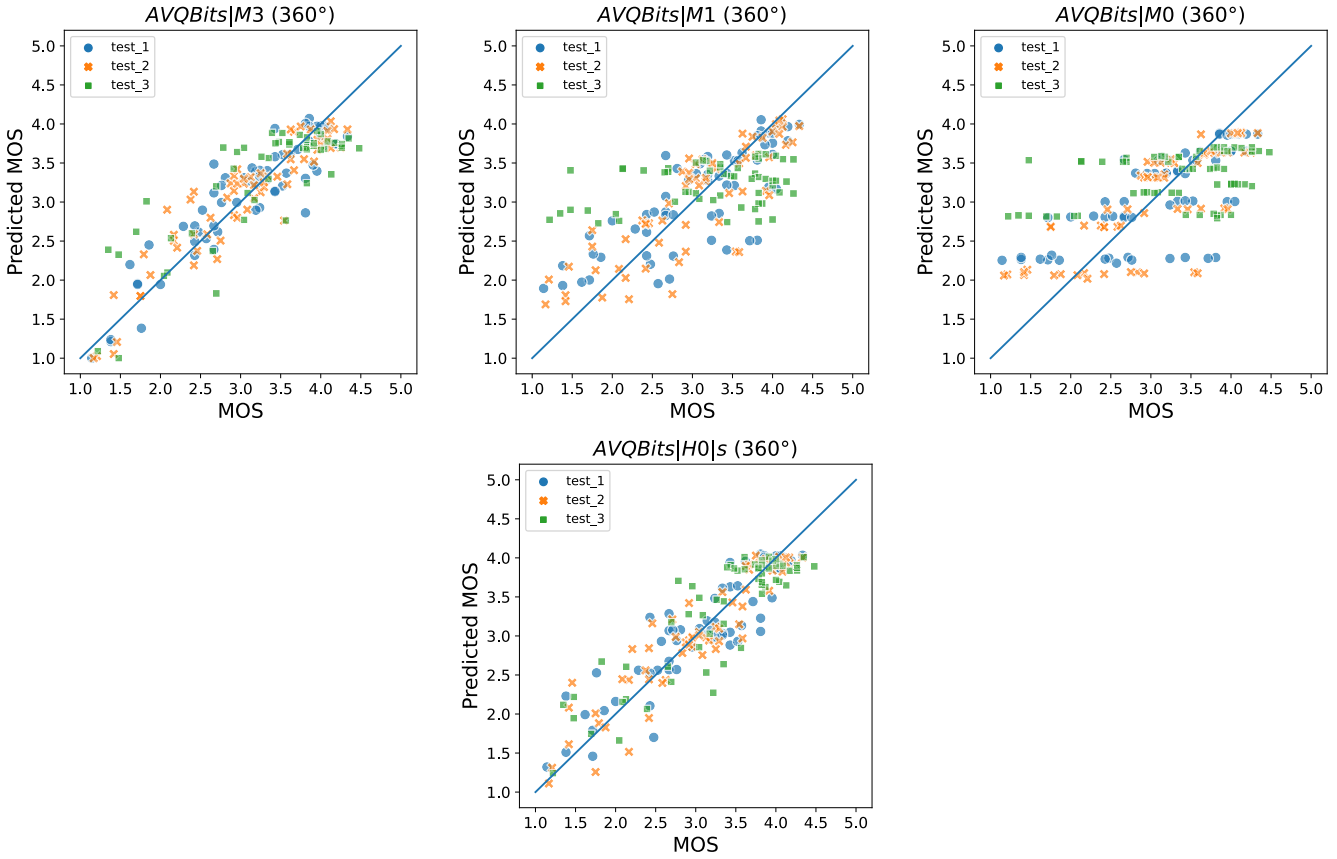

FIGURE 14. Scatter plot of P.1204.3 and its extensions for 360 Streaming Video Quality Dataset.

was H.265, which is the default codec for $AVQBits|H0|f$ and hence both $AVQBits|H0|s$ and $AVQBits|H0|f$ are the same model in this case.

In Table 19, a comparison of the proposed bitstream models with a number of SoA models is reported. The performance numbers for the SoA models are taken directly from the work by Madhusudana *et al.* [44]. It can be observed that $AVQBits|M3$ / P.1204.3 performs on par with the best performing FR model, i.e. VMAF. It is further shown that the Mode 0 model proposed in [44], hereafter referred to as "Mode0F", performs better than the Mode 0 model $AVQBits|M0$ proposed in this paper. It should be noted that the Mode0F model was specifically trained for the 360° video use-case and the performance numbers reported in Table 19 are based on a 50:50 training-validation strategy. Moreover, the sources in test_1 and test_2 are the same, which leads to an increase in prediction accuracy of the Mode0F model. Furthermore, it can be seen that the proposed $AVQBits|H0|s$ model outperforms the hybrid model proposed by Fremerey *et al.* [44], despite not being specificaly trained for 360° videos. This is due to the fact that a more holistic approach is proposed in this paper with the QEB, using re-encoded bitstream features that are considerably more indicative of content complexity in the Random Forest part of the underlying $AVQBits|M3$ / P.1204.3 model than the SI and TI information used in [44].

E. HIGH FRAMERATE VIDEOS

The last extended use-case that is considered for the evaluation of the proposed bitstream models in this paper is HFR video. For this purpose the LIVE-YT-HFR dataset is used. Although this use-case falls into the broad category of traditional 2D videos, the HFR use-case is still considered as an extended application scope as the proposed models have been trained and validated only for video of framerate up to 60 fps. As was the case with gaming and 360° video, no retraining was performed on the proposed bitstream-based models for the specific use-case.

Table 20 compares the performance of the proposed *AVQBits* models with SoA models for each framerate. The performance numbers for the SoA models are taken directly from the work by Fremerey *et al.* [45]. In general, it can be observed that $AVQBits|M3$ / P.1204.3 model performs on par with VMAF for all framerates. The performance is similarly good for the hybrid models $AVQBits|H0|s$ and $AVQBits|H0|f$, although $AVQBits|H0|f$ with its fixed encoder shows a slightly worse performance. The results for this model variant could be enhanced by a dedicated retraining of the a_{cmap} and b_{cmap} for HFR specific content. The Mode 0 ($AVQBits|M0$) and Mode 1 ($AVQBits|M1$) models show similar performance to that of SSIM, MS-SSIM, ST-RRED and FRQM. It can also be seen that prediction accuracy in terms of both PCC and SROCC is significantly

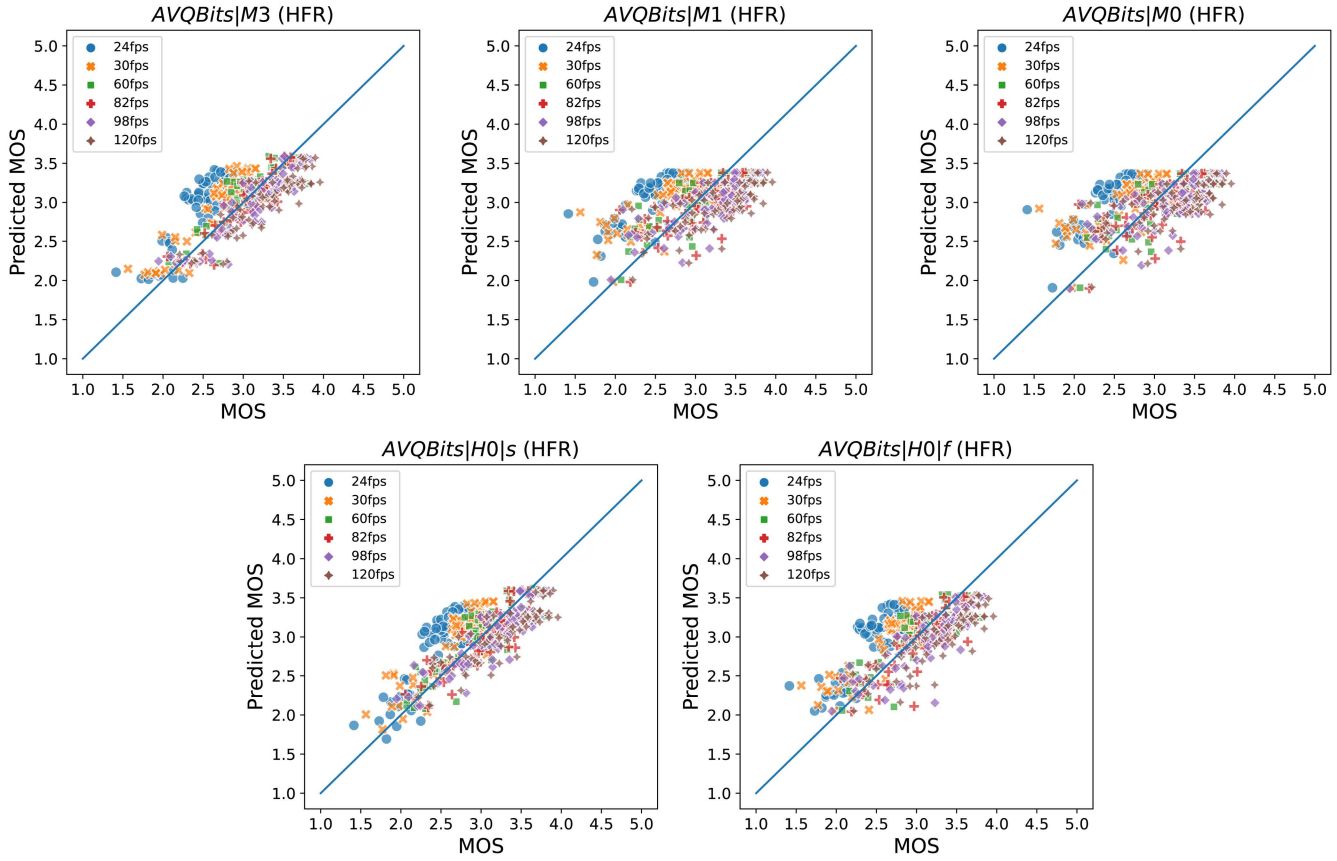

FIGURE 15. Scatter plot of P.1204.3 and its extensions for LIVE-YT-HFR dataset.

TABLE 20. Comparison of performance of P.1204.3 and its extensions with SoA models on the LIVE-YT-HFR dataset (The performance numbers for the SoA models are taken directly from the work by Madhusudana *et al.* [45]).

Model	24fps		30fps		60fps		82fps		98fps		120fps		Overall	
	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC	SROCC	PCC
PSNR	0.4101	0.3647	0.4414	0.4179	0.6202	0.5719	0.6878	0.6431	0.7171	0.6489	0.6019	0.5937	0.6950	0.6685
SSIM [112]	0.1277	0.0949	0.1108	0.0816	0.2123	0.1845	0.2079	0.2430	0.3876	0.3964	0.7485	0.6726	0.4494	0.4526
MS-SSIM [113]	0.2221	0.1500	0.1929	0.1112	0.2516	0.1900	0.2906	0.2549	0.4237	0.4007	0.6165	0.5843	0.4898	0.4673
FSIM [117]	0.3670	0.3038	0.3208	0.2638	0.2472	0.2615	0.3225	0.3055	0.3861	0.2646	0.3056	0.1178	0.4469	0.4435
ST-RRED [118]	0.1541	0.0369	0.1188	0.0307	0.5062	0.4457	0.3394	0.3271	0.4962	0.4556	0.6745	0.5906	0.5531	0.5107
SpEED [119]	0.2591	0.1237	0.2278	0.0896	0.1824	0.1110	0.2955	0.2425	0.4118	0.3295	0.6827	0.6097	0.4861	0.4449
FRQM [93]	0.1556	0.2089	0.0983	0.0854	0.0947	0.0309	0.0137	0.0035	0.0317	0.0100	-	-	0.4216	0.4520
VMAF [23]	0.1743	0.2669	0.2855	0.3740	0.5408	0.6015	0.6820	0.7390	0.8214	0.8128	0.7943	0.7844	0.7303	0.7071
deepVQA[120]	0.1144	0.0495	0.1353	0.1059	0.2527	0.1652	0.1803	0.1515	0.2816	0.2654	0.6865	0.6209	0.3463	0.3329
GSTI [95]	0.4554	0.5827	0.5079	0.6664	0.6853	0.7507	0.7584	0.8194	0.7886	0.7953	0.7508	0.7258	0.7983	0.7917
$AVQBits\|M3$ / P.1204.3	0.5395	0.7806	0.6244	0.8619	0.7722	0.8921	0.8325	0.9145	0.8548	0.9125	0.8752	0.9184	0.7118	0.7805
$AVQBits\|M1$	0.4990	0.6870	0.5433	0.7223	0.7086	0.7638	0.7129	0.7326	0.7417	0.7477	0.7476	0.7218	0.4809	0.5528
$AVQBits\|M0$	0.5053	0.6643	0.5453	0.6953	0.6961	0.7320	0.6788	0.6893	0.7261	0.7061	0.7273	0.6832	0.4947	0.5538
$AVQBits\|H0\|s$	0.5535	0.7784	0.6459	0.8326	0.7794	0.8633	0.8133	0.8814	0.8519	0.8843	0.8849	0.8855	0.7324	0.7887
$AVQBits\|H0\|f$	0.5680	0.7806	0.6362	0.8303	0.7789	0.8591	0.7832	0.8572	0.8219	0.8424	0.8421	0.8498	0.6740	0.7242

worse for lower framerates than for higher framerates for all the proposed models. This is due to the fact that the *temporal degradation* component of the "Core Model" considers 60 *fps* as the maximum framerate as that was the framerate of the used display for subjective testing for both AVT-PNATS-UHD-1 and AVT-VQDB-UHD-1. The temporal degradation associated with the perceived video quality is then estimated relatively to 60 *fps* thereby underestimating the impact of lower framerates on perceived video quality when viewed on a display with higher framerate such as 120 *fps*. The models show a significantly better performance at higher framerates ($\geq$ 60 *fps*), as the effect of temporal degradation on perceived video quality decreases at higher framerates. This is consistent with findings presented in [91]. Figure 15 illustrates the scatter plots for the different *AVQBits* variants on the LIVE-YT-HFR dataset.

VIII. DISCUSSION AND CONCLUSION

The paper presents an information-adaptive, bitstream-based video quality model, *AVQBits*. With its four variants, it can operate on bitstream-based input information of different complexity. The so-called Mode 3 variant of the model, *AVQBits*|*M*3, has been standardized as ITU-T Rec. P.1204.3 [27]. The model algorithm and performance of *AVQBits*|*M*3 have been described in [39] and [22]. The present paper complements this work in three directions: (1) introducing three further *AVQBits* variants that adapt the model algorithm according to the input information available in a given context; (2) proposing a long-term integration module that allows accurate predictions for typical HTTP-based adaptive video streaming sessions; (3) evaluating all four short-term video model types, out-of-the-box, for application scenarios the models have not initially been developed for, and showing their competitive performance in comparison with SoA models. The newly introduced, light-weight Mode 0 model *AVQBits*|*M*0 uses metadata such as framerate, resolution and bitrate for prediction. The Mode 1 model *AVQBits*|*M*1 delivers more precise predictions by additionally using frame-type and frame-size information. Both models employ the available input information to predict *QP* as the main information used in the Mode 3 model *AVQBits*|*M*3 / P.1204.3. The third type of video quality models presented in this paper is hybrid, that is, metadata- / Mode 0 and pixel-based, *AVQBits*|*H*0. The underlying approach is to re-encode the decoded pixel-information into a ''Quality-equivalent Bitstream'', using the metadata for selecting encoder settings. Two versions of this algorithm are presented, one using the same codec as the one initially employed (*AVQBits*|*H*0|*s*), and one using a fixed encoder, in this case HEVC/H.265 (*AVQBits*|*H*0|*f*).

All model variants were trained for traditional 2D videos and specific encoder implementations. Based on the evaluation on a dataset different from the training data, it is shown in this paper that all models have highly competitive performance for ''normal'' 2D video, also in comparison to SoA models. To analyze the generalizability of the introduced models, all models were evaluated on different further application scopes (contribution (3) mentioned above). For this purpose, gaming video, 360° video and HFR video have been considered. Within this evaluation on extended application scopes, other use-cases, e.g., different encoder implementations have also been addressed. The evaluation for each case is based on multiple subjective tests to investigate the impact of different designs on the prediction accuracy of the models. For this purpose, four different datasets for gaming, three tests from the 360 Streaming Video Quality Dataset for 360° video and a dataset consisting of HFR videos are considered.

The evaluation of the proposed models also in comparison to SoA models shows that the *AVQBits*|*M*3 / P.1204.3 bitstream Mode 3 model performs very well in terms of both RMSE and PCC for all use-cases. The performance of this model may further be improved by including use-case-specific features, e.g. for 360° or gaming videos, updating the *temporal degradation* component of the ''Core Model'' in case of HFR videos. The Mode 1 model *AVQBits*|*M*1 shows very good performance for the gaming video use-case. The somewhat reduced performance for 360° video and HFR can be attributed to multiple factors, which are, for example, the lack of a use-case-specific retraining and missing of specific features. More sophisticated features using framesize and frame type information to take into account the effect of encoding of different contents may be used to extend the models to improve prediction performance.

The performance of the Mode 0 model *AVQBits*|*M*0 varies more strongly in terms of accuracy for different use-cases and also for individual tests within a specific case. This is along expected lines since a Mode 0 model is content-agnostic and hence its generalizability is very limited. However, it should be mentioned, that a Mode 0 model is also more applicable for monitoring and real-time evaluation of video quality, for example in case that no access to encoded bitstream information is available and a light-weight model is sought. Furthermore, to improve the performance of such metadata-based models, use-case specific coefficient sets may be helpful. Moreover, for both Mode 0 and Mode 1 models, the QP prediction may be modified and made use-case specific to cover different encoding strategies and will be considered for further development. The performance of both variants of the Hybrid Mode 0 model, *AVQBits*|*H*0|*s* and *AVQBits*|*H*0|*f*, is comparable to the *AVQBits*|*M*3 / P.1204.3 both in terms of PCC and RMSE for all the considered use-cases.

The comparison of the proposed models with SoA models for different application scopes shows that for all use-cases *AVQBits*|*M*3 / P.1204.3 either performs on par with or outperforms the best performing FR model, i.e. VMAF. Although the performance of the Mode 0 and Mode 1 models varies for different use-cases, the performance of these models is better than SoA NR models in general and also comparable to FR models other than VMAF.

Additionally, the hybrid model *AVQBits*|*H*0|*s* shows great potential not only for ''normal'' 2D video, but also for other application scopes, with similar or only slightly worse performance as compared to the Mode 3 *AVQBits*|*M*3 / P.1204.3 model. Also the codec-agnostic *AVQBits*|*H*0|*f* model shows good performance across the different use-cases. A more use-case-specific parameter-handling for this generic hybrid codec will be considered in future work, thus even better ensuring that there is no need to develop bitstream parsers for newer codecs in the case of having the decoded pixels accessible.

As a further contribution, a long-term integration model is proposed in this paper, which is the item (3) mentioned in the beginning of this section. A simplified version of the P.1203.3 model [36] has been proposed for this purpose. Hence, in addition to evaluating the applicability of the proposed models for different short-term video application scopes, their usage as the video quality estimation component of the proposed long-term integration model has also been

investigated. The per-1-second quality scores estimated by the *AVQBits*|*M*3 / P.1204.3 model are used, and shown to yield high accuracy. Also all other model variants show good performance when using their predictions in terms of per-1-second scores in conjunction with the proposed integration module. As expected, Mode 0 (*AVQBits*|*M*0) works least well, Mode 1 (*AVQBits*|*M*1) shows intermediate and the hybrid model *AVQBits*|*H*0|*s* shows strong performance, almost as good as that of the Mode 3 model *AVQBits*|*M*3.

IX. OUTLOOK

As outlined above, the focus of the present paper was on presenting the overall *AVQBits* framework and the different models of four model types, bitstream Mode 3 (*AVQBits*|*M*3 / P.1204.3), Mode 1 (*AVQBits*|*M*1) and Mode 0 (*AVQBits*|*M*0), as well as hybrid Mode 0 (*AVQBits*|*H*0|*s* and *AVQBits*|*H*0|*f*). The models were initially developed for 4K/UHD-1 with a maximum framerate of 60 fps, and shown to yield competitive performance also in comparison to SoA models. The performance evaluation of the models for other application scopes than the one they were initially developed for was performed with the models *out-of-the-box*. Although the proposed models have been evaluated for different scenarios showing competitive performance, the models can be enhanced to increase the prediction accuracy, adjusting them to the specific use-cases. As a starting point for future work, retraining the coefficients of the Mode 0 and Mode 1 models *AVQBits*|*M*0 and *AVQBits*|*M*1 with several application-specific databases will be performed. Furthermore, a dedicated HFR variant of *AVQBits*|*M*3 / P.1204.3 will be developed (see Sec. III-A), from which the hybrid models *AVQBits*|*H*0|*s* and *AVQBits*|*H*0|*f* can directly be derived. To this aim, the existing Mode 3 bitstream features will be adapted to be more frame-rate specific. For example, scaling the motion vectors in a more precise way to the actual speed of motion in pixels per time is expected to improve the specificity of the motion complexity information utilized in the Random Forest part of the model. Here, new and complementary bitstream features can be considered in addition. Already the retraining of the Random Forest component with a mixed dataset comprising some of the data from the LIVE-YT-HFR ([45], see Sec. V-G) and the AVT-PNATS-UHD-1 dataset (see Sec. V-B) is expected to result in an improved handling of low and high framerates. It is noted that the subjective test data initially used for the *AVQBits* model development did not comprise many combinations of low or high framerates with other diverse video settings in terms of resolution or bitrate.

Moreover, improvements of the D_t component in the Core model that captures frame-rate specific effects will be considered for all model variants. This way, also the content-agnostic Mode 0 model can be further improved for HFR video quality estimation. Similarly to the Mode 3 feature adaptations for the Random Forest model component to HFR, improvements are conceivable for the 360° video case. Here, modifications to motion-related features as well as other features that better handle the specific projection geometry will be addressed, and are expected to lead to an even better prediction performance. In addition to this, further investigations of the proposed models on newer applications such as quality evaluation of user generated content, encoding optimization, newer video codecs, or point cloud compression will be considered.

ACKNOWLEDGMENT

The authors would like to thank their colleagues, namely Peter List, Bernhard Feiten, and Ulf Wüstenhagen from Deutsche Telekom AG, for their invaluable input during the P.NATS Phase 2 competition and the course of this work. Further, they wish to thank Werner Robitza for fruitful exchanges on modeling approaches during the P.NATS Phase 2 times. They wish to extend their thanks to their colleagues from Ericsson (David Lindero, Gunnar Heikkilä, and Jörgen Gustafsson), SwissQual (Silvio Borer), and Netscout (Simon Broom) for sharing their long-duration video quality databases from ITU-T P.NATS Phase 2 for model development and evaluation.

REFERENCES

[1] ùCisco. *Cisco Visual Networking Index: Forecast and Trends, 2017–2022*. Accessed: Apr. 6, 2022. [Online]. Available: https://twiki.cern.ch/twiki/pub/HEPIX/TechwatchNetwork/HtwNetworkDocume%nts/white-paper-c11-741490.pdf
[2] Cisco. *Global—2021 Forecast Highlights*. Accessed: Apr. 6, 2022. [Online]. Available: https://www.cisco.com/c/dam/m/en_us/solutions/service-provider/vni-fore% cast-highlights/pdf/Global_2021_Forecast_Highlights.pdf
[3] *Cisco Annual Internet Report (2018–2023) White Paper*, Cisco, San Jose, CA, USA, 2020.
[4] S. Gnome. *Twitch Statistics and Analytics*. Accessed: Apr. 6, 2022. [Online]. Available: https://sullygnome.com/
[5] S. Charts. *All Streaming Data in One Place*. Accessed: Apr. 6, 2022. [Online]. Available: https://streamscharts.com/
[6] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.
[7] D. Mukherjee, J. Han, J. Bankoski, R. Bultje, A. Grange, J. Koleszar, P. Wilkins, and Y. Xu, "A technical overview of VP9—The latest open-source video codec," *SMPTE Motion Imag. J.*, vol. 124, no. 1, pp. 44–54, Jan. 2015.
[8] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[9] MPEG. *ISO/IEC 23094–1:2020 Information Technology—General Video Coding—Part 1: Essential Video Coding*. Accessed: Apr. 6, 2022. [Online]. Available: https://www.iso.org/standard/57797.html
[10] MPEG. *Low Complexity Enhancement Video Coding*. Accessed: Apr. 6, 2022. [Online]. Available: https://www.lcevc.org/
[11] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi, and C. H. Chiang, "An overview of core coding tools in the AV1 video codec," in *Proc. Picture Coding Symp. (PCS)*, Jun. 2018, pp. 41–45.
[12] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021.
[13] A. Inc. *Hls Authoring Specification for Apple Devices*. Accessed: Apr. 6, 2022. [Online]. Available: https://developer.apple.com/documentation/http_live_streaming/hls_autho%ring_specification_for_apple_devices
[14] N. T. Blog. *Per-Title Encode Optimization*. Accessed: Apr. 6, 2022. [Online]. Available: https://netflixtechblog.com/per-title-encode-optimization-7e99442b62a2

[15] I. Katsavounidis. *Dynamic Optimizer—A Perceptual Video Encoding Optimization Framework*. Accessed: Apr. 6, 2022. [Online]. Available: https://netflixtechblog.com/dynamic-optimizer-a-perceptual-video-encodi%ng-optimization-framework-e19f1e3a277f

[16] Y. Reznik, X. Li, K. Lillevold, R. Peck, T. Shutt, and P. Howard, "Optimizing mass-scale multi-screen video delivery," *SMPTE Motion Imag. J.*, vol. 129, no. 3, pp. 26–38, 2020, doi: 10.5594/JMI.2020.2973561.

[17] S. Winkler and P. Mohandas, "The evolution of video quality measurement: From PSNR to hybrid metrics," *IEEE Trans. Broadcast.*, vol. 54, no. 3, pp. 660–668, Sep. 2008.

[18] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Trans. Broadcast.*, vol. 57, no. 2, pp. 165–182, Jun. 2011.

[19] A. Raake, J. Gustafsson, S. Argyropoulos, M.-N. Garcia, D. Lindegren, G. Heikkilä, M. Pettersson, P. List, and B. Feiten, "IP-based mobile and fixed network audiovisual media services," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 68–79, Nov. 2011.

[20] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: A classification and review of recent approaches," *EURASIP J. Image Video Process.*, vol. 2014, no. 1, pp. 1–32, Dec. 2014.

[21] N. Barman and M. G. Martini, "QoE modeling for HTTP adaptive video streaming—A survey and open challenges," *IEEE Access*, vol. 7, pp. 30831–30859, 2019.

[22] A. Raake, S. Borer, S. Satti, J. Gustafsson, R. Rao, S. Medagli, P. List, S. Göring, D. Lindero, W. Robitza, G. Heikkilä, S. Broom, C. Schmidmer, B. Feiten, U. Wüstenhagen, T. Wittmann, M. Obermann, and R. Bitto, "Multi-model standard for bitstream-, pixel-based and hybrid video quality assessment of UHD/4K: ITU-T P.1204," *IEEE Access*, vol. 8, pp. 193020–193049, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9234526?source=authoralert

[23] (2018). Netflix. *VMAF 4K Included*. [Online]. Available: https://github.com/Netflix/vmaf

[24] *Video Quality Assessment of Streaming Services Over Reliable Transport for Resolutions up to 4K With Access to Full and Reduced Reference Pixel Information*, document ITU-T Rec. P.1204.4, International Telecommunication Union, 2019.

[25] S. Göring, J. Skowronek, and A. Raake, "DeViQ—A deep no reference video quality model," *Electron. Imag.*, vol. 30, no. 14, pp. 1–6, Jan. 2018.

[26] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport—Video Quality Estimation Module*, document ITU-T Rec. P.1203.1, International Telecommunication Union, Geneva, Switzerland, 2019.

[27] *Video Quality Assessment of Streaming Services Over Reliable Transport for Resolutions up to 4K With Access to Full Bitstream Information*, document ITU-T Rec. p.1204.3, International Telecommunication Union, 2019.

[28] M.-N. Garcia, F. De Simone, S. Tavakoli, N. Staelens, S. Egger, K. Brunnstrom, and A. Raake, "Quality of experience and HTTP adaptive streaming: A review of subjective studies," in *Proc. 6th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 141–146.

[29] M. Seufert, S. Egger, M. Slanina, T. Zinner, T. Hoßfeld, and P. Tran-Gia, "A survey on quality of experience of HTTP adaptive streaming," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 469–492, Sep. 2014.

[30] C. G. Bampis, Z. Li, A. K. Moorthy, I. Katsavounidis, A. Aaron, and A. C. Bovik, "Study of temporal effects on subjective video quality of experience," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5217–5231, Nov. 2017.

[31] C. G. Bampis, Z. Li, and A. C. Bovik, "Continuous prediction of streaming video QoE using dynamic networks," *IEEE Signal Process. Lett.*, vol. 24, no. 7, pp. 1083–1087, Jul. 2017.

[32] C. G. Bampis, Z. Li, I. Katsavounidis, and A. C. Bovik, "Recurrent and dynamic models for predicting streaming video quality of experience," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3316–3331, Jul. 2018.

[33] W. Robitza, M. N. Garcia, and A. Raake, "At home in the lab: Assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm," in *Proc. 7th Int. Workshop Quality Multimedia Exper. (QoMEX)*, May 2015, pp. 1–6.

[34] W. Robitza, M.-N. Garcia, and A. Raake, "A modular HTTP adaptive streaming QoE model—Candidate for ITU-T P.1203," in *Proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2017, pp. 1–6.

[35] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport*, document ITU-T Rec. p.1203, International Telecommunication Union, Tech. Rep., 2016.

[36] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport—Quality Integration Module*, document ITU-T Rec. P.1203.3, International Telecommunication Union, Geneva, Switzerland, 2020.

[37] W. Robitza, S. Göring, A. Raake, D. Lindegren, G. Heikkilä, J. Gustafsson, P. List, B. Feiten, U. Wüstenhagen, M.-N. Garcia, K. Yamagishi, and S. Broom, "HTTP adaptive streaming QoE estimation with ITU-T Rec. P. 1203: Open databases and software," in *Proc. 9th ACM Multimedia Syst. Conf.*, Amsterdam, The Netherlands, 2018, pp. 466–471.

[38] A. Raake, M.-N. Garcia, W. Robitza, P. List, S. Goring, and B. Feiten, "A bitstream-based, scalable video-quality model for HTTP adaptive streaming: ITU-T P.1203.1," in *Proc. 9th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2017, p. 1203. [Online]. Available: http://ieeexplore.ieee.org/document/7965631/

[39] R. R. R. Rao, S. Goring, P. List, W. Robitza, B. Feiten, U. Wustenhagen, and A. Raake, "Bitstream-based model standard for 4K/UHD: ITU-T P.1204.3—Model details, evaluation, analysis and open source implementation," in *Proc. 12th Int. Conf. Quality Multimedia Exper. (QoMEX)*, May 2020, p. 1204. [Online]. Available: https://www.researchgate.net/publication/341792225_Bitstream-based_Mode%l_Standard_for_4KUHD_ITU-T_P12043_-_Model_Details_Evaluation_Analysis_and_Open%_Source_Implementation

[40] R. R. R. Rao, S. Goring, W. Robitza, B. Feiten, and A. Raake, "AVT-VQDB-UHD-1: A large scale video quality database for UHD-1," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 1–8. [Online]. Available: https://www.researchgate.net/publication/338201010_AVT-VQDB-UHD-1_A_Lar% ge_Scale_Video_Quality_Database_for_UHD-1

[41] N. Barman, S. Zadtootaghaj, S. Schmidt, M. G. Martini, and S. Moller, "GamingVideoSET: A dataset for gaming video streaming applications," in *Proc. 16th Annu. Workshop Netw. Syst. Support Games (NetGames)*, Jun. 2018, pp. 1–6.

[42] N. Barman, E. Jammeh, S. A. Ghorashi, and M. G. Martini, "No-reference video quality estimation based on machine learning for passive gaming video streaming applications," *IEEE Access*, vol. 7, pp. 74511–74527, 2019.

[43] S. Zadtootaghaj, S. Schmidt, S. S. Sabet, S. Möller, and C. Griwodz, "Quality estimation models for gaming video streaming services using perceptual video quality dimensions," in *Proc. 11th ACM Multimedia Syst. Conf.*, May 2020.

[44] S. Fremerey, S. Goring, R. R. R. Rao, R. Huang, and A. Raake, "Subjective test dataset and Meta-data-based models for 360° streaming video quality," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.

[45] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," 2020, *arXiv:2007.11634*.

[46] *Video Quality Assessment of Streaming Services Over Reliable Transport for Resolutions up to 4K*, document ITU-T Rec. P.1204, International Telecommunication Union, 2019.

[47] A. Raake, M. Garcia, S. Moller, J. Berger, F. Kling, P. List, J. Johann, and C. Heidemann, "TV-model: Parameter-based prediction of IPTV quality," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 03 Mar. 2008, pp. 1149–1152.

[48] *Opinion Model for Video-Telephony Applications*, document ITU-T Rec.G.1070, International Telecommunication Union, Geneva, Switzerland, 2007.

[49] M. Garcia, R. Schleicher, and A. Raake, "Towards a content-based parametric video quality model for IPTV," in *Proc. 3rd Int. Workshop Perceptual Quality Syst. (PQS)*, 2010, pp. 1–5.

[50] M. N. Garcia and A. Raake, "Frame-layer packet-based parametric video quality model for encrypted video in IPTV services," in *Proc. 3rd Int. Workshop Quality Multimedia Exper.*, Sep. 2011, pp. 102–106.

[51] X. Lin, H. Ma, L. Luo, and Y. Chen, "No-reference video quality assessment in the compressed domain," *IEEE Trans. Consum. Electron.*, vol. 58, no. 2, pp. 505–512, May 2012.

[52] M. Garcia, P. List, S. Argyropoulos, D. Lindegren, M. Pettersson, B. Feiten, J. Gustafsson, and A. Raake, "Parametric model for audiovisual quality assessment in IPTV: ITU-T Rec. P. 1201.2," in *Proc. IEEE 15th Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2013, pp. 482–487.

[53] M. Shahid, A. Rossholm, and B. Lovstrom, "A no-reference machine learning based video quality predictor," in *Proc. 5th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 176–181.

[54] A. Rossholm and B. Lovstroem, "A new low complex reference free video quality predictor," in *Proc. IEEE 10th Workshop Multimedia Signal Process.*, Oct. 2008, pp. 765–768.

[55] M. Shahid, A. Rossholm, and B. Lövström, "A reduced complexity no-reference artificial neural network based video quality predictor," in *Proc. 4th Int. Congr. Image Signal Process.*, vol. 1, Oct. 2011, pp. 517–521.

[56] D. C. Mocanu, J. Pokhrel, J. P. Garella, J. Seppänen, E. Liotou, and M. Narwaria, "No-reference video quality measurement: Added value of machine learning," *J. Electron. Imag.*, vol. 24, no. 6, Dec. 2015, Art. no. 061208.

[57] E. Demirbilek and J.-C. Grégoire, "Machine learning based reduced reference bitstream audiovisual quality prediction models for realtime communications," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 571–576.

[58] K. Izumi, K. Kawamura, T. Yoshino, and S. Naito, "No reference video quality assessment based on parametric analysis of HEVC bitstream," in *Proc. 6th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Sep. 2014, pp. 49–50.

[59] X. Huang, J. Sogaard, and S. Forchhammer, "No-reference video quality assessment by HEVC codec analysis," in *Proc. Vis. Commun. Image Process. (VCIP)*, Dec. 2015, pp. 1–4.

[60] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Trans. Image Process.*, vol. 19, no. 6, pp. 1427–1441, Jan. 2010.

[61] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K video sequence dataset," in *Proc. 5th Int. Workshop Quality Multimedia Exper. (QoMEX)*, Jul. 2013, pp. 34–35.

[62] *Parametric Bitstream-Based Quality Assessment of Progressive Download and Adaptive Audiovisual Streaming Services Over Reliable Transport—Audio Quality Estimation Module*, document ITU-T Rec. P.1203.2, International Telecommunication Union, Geneva, Switzerland, 2017.

[63] P. Lebreton and K. Yamagishi, "Transferring adaptive bit rate streaming quality models from H.264/HD to H.265/4K-UHD," *IEICE Trans. Commun.*, vol. 102, no. 12, pp. 2226–2242, 2019.

[64] T. He, R. Xie, J. Su, X. Tang, and L. Song, "A no reference bitstream-based video quality assessment model for H.265/HEVC and H.264/AVC," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2018, pp. 1–5.

[65] K. D. Singh, Y. Hadjadj-Aoul, and G. Rubino, "Quality of experience estimation for adaptive HTTP/TCP video streaming using H.264/AVC," in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Jan. 2012, pp. 127–131.

[66] H. T. T. Tran, N. P. Ngoc, A. T. Pham, and T. C. Thang, "A multi-factor QoE model for adaptive streaming over mobile networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2016, pp. 1–6.

[67] R. R. R. Rao, S. Göring, N. P. P. Vogel, J. J. V. Villarreal, W. Robitza, P. List, B. Feiten, and A. Raake, "Adaptive video streaming with current codecs and formats: Extensions to parametric video quality model ITU-T P. 1203," *Electron. Imag.*, vol. 10, p. 314, Jan. 2019.

[68] K. Yamagishi, N. Egi, N. Yoshimura, and P. Lebreton, "Derivation procedure of coefficients of metadata-based model for adaptive bitrate streaming services," *IEICE Trans. Commun.*, vol. 104, no. 7, pp. 725–737, 2021.

[69] K. Yamagishi, T. Kawano, and T. Hayashi, "Hybrid video-quality-estimation model for IPTV services," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Nov. 2009, pp. 1–5.

[70] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document ITU-T Rec. P. 910, 2008.

[71] M. C. Q. Farias, M. M. Carvalho, H. T. M. Kussaba, and B. H. A. Noronha, "A hybrid metric for digital video quality assessment," in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, Jun. 2011, pp. 1–6.

[72] O. Sugimoto, N. Sei, S. Shigeyuki, and K. Atsushi, "Objective perceptual video quality measurement method based on hybrid no reference framework," in *Proc. 16th IEEE Int. Conf. Image Process. (ICIP)*, Nov. 2009, pp. 2237–2240.

[73] S. Goring, R. R. R. Rao, B. Feiten, and A. Raake, "Modular framework and instances of pixel-based video quality models for UHD-1/4K," *IEEE Access*, vol. 9, pp. 31842–31864, 2021.

[74] B. Weiss, D. Guse, S. Möller, A. Raake, A. Borowiak, and U. Reiter, "Temporal development of quality of experience," in *Quality of Experience*. Berlin, Germany: Springer, 2014, pp. 133–147.

[75] H. T. T. Tran, T. Vu, N. P. Ngoc, and T. C. Thang, "A novel quality model for HTTP adaptive streaming," in *Proc. IEEE 6th Int. Conf. Commun. Electron. (ICCE)*, Jul. 2016, pp. 423–428.

[76] S. Zadtootaghaj, N. Barman, S. Schmidt, M. G. Martini, and S. Moller, "NR-GVQM: A no reference gaming video quality metric," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2018, pp. 131–134.

[77] S. Goring, R. R. R. Rao, and A. Raake, "NOFU—A lightweight no-reference pixel based video quality model for gaming content," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Berlin, Germany, Jun. 2019, pp. 1–6.

[78] M. Utke, S. Zadtootaghaj, S. Schmidt, S. Bosse, and S. Möller, "NDNetGaming-development of a no-reference deep CNN for gaming video quality prediction," *Multimedia Tools Appl.*, vol. 81, pp. 1–23, Jul. 2020.

[79] S. Zadtootaghaj, N. Barman, R. R. R. Rao, S. Goring, M. G. Martini, A. Raake, and S. Moller, "DEMI: Deep video quality estimation model using perceptual video quality dimensions," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, pp. 1–6.

[80] *Opinion Model for Network Planning of Video and Audio Streaming Applications*, document ITU-T Rec.G.1071, International Telecommunication Union, Geneva, Switzerland, 2016.

[81] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *Proc. IEEE Int. Symp. Mixed Augmented Reality*, Sep. 2015, pp. 31–36.

[82] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1408–1412, Sep. 2017.

[83] H. T. T. Tran, N. P. Ngoc, C. M. Bui, M. H. Pham, and T. C. Thang, "An evaluation of quality metrics for 360 videos," in *Proc. 9th Int. Conf. Ubiquitous Future Netw. (ICUFN)*, Jul. 2017, pp. 7–11.

[84] S. Goring, C. Krammer, and A. Raake, "Cencro–Speedup of video quality calculation using center cropping," in *Proc. IEEE Int. Symp. Multimedia (ISM)*, Dec. 2019, pp. 1–8.

[85] M. Orduna, C. Diaz, L. Munoz, P. Perez, I. Benito, and N. Garcia, "Video multimethod assessment fusion (VMAF) on 360 VR contents," *IEEE Trans. Consum. Electron.*, vol. 66, no. 1, pp. 22–31, Feb. 2020.

[86] S. Croci, C. Ozcinar, E. Zerman, J. Cabrera, and A. Smolic, "Voronoi-based objective quality metrics for omnidirectional video," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–6.

[87] C. Li, M. Xu, L. Jiang, S. Zhang, and X. Tao, "Viewport proposal CNN for 360° video quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10169–10178.

[88] S.-H. Yao, C.-L. Fan, and C.-H. Hsu, "Towards Quality-of-Experience models for watching 360° videos in head-mounted virtual reality," in *Proc. 11th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2019, pp. 1–3.

[89] M. Koike, Y. Urata, N. Egi, and K. Yamagishi, "Extension of ITU-T P.1204.3 model to tile-based VR streaming services," in *Proc. IEEE Int. Workshop Tech. Committee Commun. Quality Rel. (CQR)*, May 2021, p. 1204.

[90] S. Yang, J. Zhao, T. Jiang, J. Wang, T. Rahim, B. Zhang, Z. Xu, and Z. Fei, "An objective assessment method based on multi-level factors for panoramic videos," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[91] A. Mackin, F. Zhang, and D. R. Bull, "A study of subjective video quality at various frame rates," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 3407–3411.

[92] A. Mackin, F. Zhang, and D. R. Bull, "A study of high frame rate video formats," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1499–1512, Jun. 2019.

[93] F. Zhang, A. Mackin, and D. R. Bull, "A frame rate dependent video quality metric based on temporal wavelet decomposition and spatiotemporal pooling," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 300–304.

[94] P. C. Madhusudana, X. Yu, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective quality assessment of high frame rate videos," *IEEE Access*, vol. 9, pp. 108069–108082, 2021, doi: 10.1109/ACCESS.2021.3100462.

[95] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Capturing video frame rate variations via entropic differencing," *IEEE Signal Process. Lett.*, vol. 27, pp. 1809–1813, 2020, doi: 10.1109/LSP.2020.3028687.

[96] D. Y. Lee, S. Paul, C. G. Bampis, H. Ko, J. Kim, S. Y. Jeong, B. Homan, and A. C. Bovik, "A subjective and objective study of space-time subsampled video quality," *IEEE Trans. Image Process.*, vol. 31, pp. 934–948, 2021.

[97] D. Y. Lee, H. Ko, J. Kim, and A. C. Bovik, "Video quality model for space-time resolution adaptation," in *Proc. IEEE 4th Int. Conf. Image Process., Appl. Syst. (IPAS)*, Dec. 2020, pp. 34–39.

[98] N. O. Johannesson, "The ETSI computation model: A tool for transmission planning of telephone networks," *IEEE Commun. Mag.*, vol. 35, no. 1, pp. 70–79, Jan. 1997.

[99] S. Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. Berlin, Germany: Springer, 2000.

[100] *The E—Model, a Computational Model for Use in Transmission Planning*, document ITU-T Rec. G.107, International Telecommunication Union, CH–Geneva, 2009.

[101] J. Allnatt, "Subjective rating and apparent magnitude," *Int. J. Man-Machine Stud.*, vol. 7, no. 6, pp. 801–816, Nov. 1975.

[102] *Single-Ended Method for Objective Speech Quality Assessment in Narrow-Band Telephony Applications*, document ITU-T Rec. P.563, International Telecommunication Union, Geneva, Switzerland, 2004.

[103] J. Stankowski, T. Grajek, K. Wegner, and M. Domanski, "Video quality in multiple HEVC encoding-decoding cycles," in *Proc. 20th Int. Conf. Syst., Signals Image Process. (IWSSIP)*, Jul. 2013, pp. 75–78.

[104] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-T. RECOMMENDATION ITU-R BT.500-13, International Telecommunication Union, Tech. Rep., 2014.

[105] M. Pinson, M. Sullivan, and A. Catellier, "A new method for immersive audiovisual subjective testing," in *Proc. 8th Int. Workshop Video Process. Qual. Metrics Consum. Electron. (VPQM)*, 2014.

[106] R. R. R. Rao, S. Goring, R. Steger, S. Zadtootaghaj, N. Barman, S. Fremerey, S. Moller, and A. Raake, "A large-scale evaluation of the bitstream-based video-quality model ITU-T P.1204.3 on gaming content," in *Proc. IEEE 22nd Int. Workshop Multimedia Signal Process. (MMSP)*, Sep. 2020, p. 1204.

[107] S. Fremerey, A. Singla, K. Meseberg, and A. Raake, "AVtrack360: An open dataset and software recording people's head rotations watching 360° videos on an HMD," in *Proc. 9th ACM Multimedia Syst. Conf.*, Jun. 2018, pp. 403–408, doi: 10.1145/3204949.3208134.

[108] *Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*, document ITU-T. P.1401, Int. Telecommunication Union, 2014.

[109] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451, 2008.

[110] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[111] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2013.

[112] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[113] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Jul. 2003, pp. 1398–1402.

[114] S. Li, F. Zhang, M. Lin, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

[115] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[116] W. Robitza, R. R. Ramachandra-Rao, S. Göring, A. Dethof, and A. Raake, "Deploying the ITU-T P.1203 QoE model in the wild and retraining for new codecs," in *Proc. 1st Mile-High Video Conf.*, Mar. 2022, p. 1203, doi: 10.1145/3510450.3517310.

[117] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[118] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 4, pp. 684–694, Apr. 2013.

[119] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "SpEED-QA: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

[120] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *Computer Vision*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham, Switzerland: Springer, 2018, pp. 224–241.

RAKESH RAO RAMACHANDRA RAO received the M.Sc. degree in communications engineering from RWTH Aachen University, in 2017, with a focus on image content analysis and millimeter wave transmission systems. Since 2017, he has been working as an Electrical Engineer with the Audiovisual Technology (AVT), TU Ilmenau. Since 2017, he has also been actively involving in standardization activities on QoE assessment methods with the ITU-T Study Group 12. Before joining AVT, he has worked as an Intern at HEAD Acoustics, where he has worked on reference-based noise estimation. His current research interests include video quality analysis and bitstream-based video quality modeling. His specializations include video quality and image content analysis.

STEVE GÖRING received the B.Sc. and M.Sc. degrees in computer science from TU Ilmenau, in 2012 and 2013, respectively, and the Ph.D. degree in visual quality prediction using machine learning, in 2022. He is currently working as a Computer Scientist with the Audiovisual Technology Group, TU Ilmenau. Before he started working at the Audiovisual Technology Group, in 2016, he was working at the Big Data Analytics Group, Bauhaus University Weimar. His current research interests include data analysis problems for video quality models and video streams. His specializations are data analytics/machine learning, video quality, and distributed communication/information systems.

ALEXANDER RAAKE (Member, IEEE) received the Dr.-Ing. degree from the Faculty of Electrical Engineering and Information Technology, Ruhr-Universität Bochum, in 2005, with the book *Speech Quality of VoIP*. Since 1999, he has been involved with the ITU-T Study Group 12's standardization work on QoS and QoE assessment methods. From 2004 to 2005, he was a Postdoctoral Researcher at LIMSI-CNRS, Orsay, France. From 2005 to 2015, he was a Senior Researcher, an Assistant, and later an Associate Professor at the TU Berlin's An-Institut T-Labs, a joint venture between Deutsche Telekom AG and TU Berlin, heading the Assessment of IP-Based Applications Group. In 2015, he has joined as a Full Professor at TU Ilmenau, where he currently heads the Audiovisual Technology Group. His research interests include audiovisual and multimedia technology, speech, audio, and video signals, human audiovisual perception, and quality of experience. He is a member of the Acoustical Society of America, AES, VDE/ITG, and DEGA.