|  |
| --- |
| **INTERNATIONAL ORGANIZATION FOR STANDARDIZATION ORGANISATION INTERNATIONALE DE NORMALISATION ISO/IEC JTC 1/SC 29/WG 5 MPEG JOINT VIDEO CODING TEAM(S) WITH ITU-T SG 16** |
| **ISO/IEC JTC 1 / SC 29 / WG 5 N 72** |
| **Online, 7–16 July 2021** |
| |  |  | | --- | --- | | **Title:** | **VVC verification test report for high dynamic range video content** | | **Source:** | **Convenor (Jens-Rainer Ohm)** | | **Type:** | **General** | | **Subtype:** | **N/A** | | **Status:** | **Approved** | | **Date:** | **2021-08-27** | | **Expected Action:** | **Info** | | **Action due date:** | **N/A** | | **No. of pages** | **15** (without this cover page) | | **Email of convenor:** | **ohm @ ient . rwth-aachen . de** | | **Committee URL:** | **https://sd.iso.org/documents/ui/#!/browse/iso/iso-iec-jtc-1/iso-iec-jtc-1-sc-29/iso-iec-jtc-1-sc-29-wg-5** | |

|  |  |
| --- | --- |
| **Joint Video Experts Team (JVET)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29**  23rd Meeting, by teleconference, 7–16 July 2021 | Document: JVET-W2020-v1 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **VVC Verification Test Report for High Dynamic Range Video Content** | | |
| *Status:* | Output document approved by JVET | | |
| *Purpose:* | Report | | |
| *Author(s) or Contact(s):* | Mathias Wien  Vittorio Baroncini | Email: | [wien@lfb.rwth-aachen.de](mailto:wien@lfb.rwth-aachen.de) [baroncini@gmx.com](mailto:baroncini@gmx.com) |
| *Source:* | Verification Test Coordinators | | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Executive Summary

This document reports the verification test results comparing VVC to its predecessor, HEVC, according to the verification test plan JVET-V2021 that was issued at the previous meeting. The compression performance capabilities of the two specifications are compared based on the HEVC reference software HM-16.23 and the VVC reference software VTM-12.0. A visual assessment of VVC compared to HEVC was performed, and the result of a formal subjective assessment evaluation with the participation of naïve test subjects is reported. The assessment includes testing with five sequences with HDR HLG content and five sequences with HDR PQ content. For the HDR PQ test case, test sequences with two different values of Maximum Display Mastering Luminance (1,000 cd/m² and 4,000 cd/m²) were included to reflect different corresponding use cases. The tests were conducted using consumer-type displays in order to match the application scenario of end users. The MOS-over-bitrate plots demonstrate the superior performance of VVC compared to its predecessor, HEVC, with a good coverage of the MOS range from slightly perceptible impairments down to annoying impairments for all sequences. The Bjøntegaard delta rate figures show an overall gain of about 49% for the HLG test sequences and an overall gain of about 52% for the PQ test sequences.

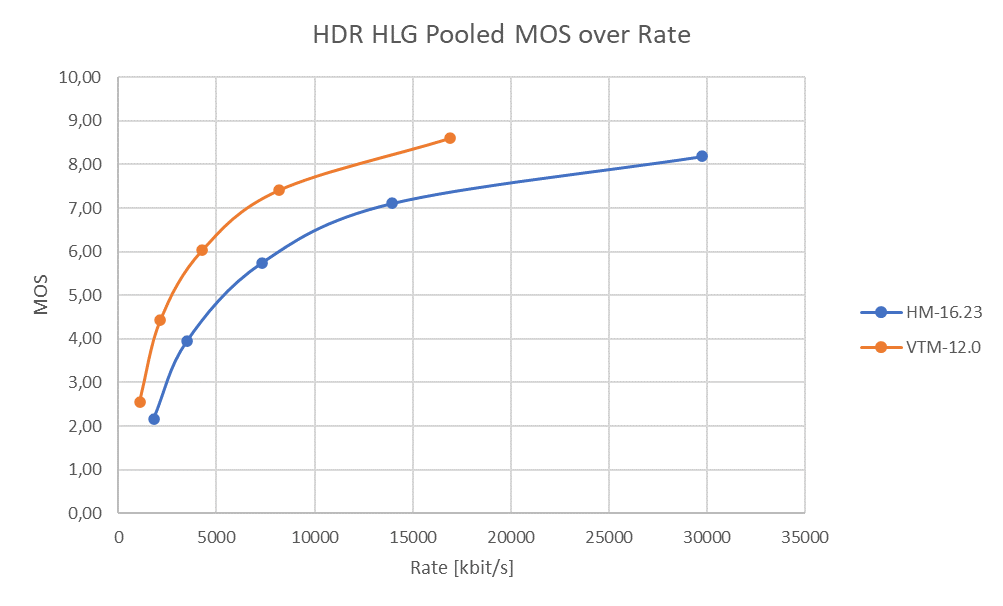
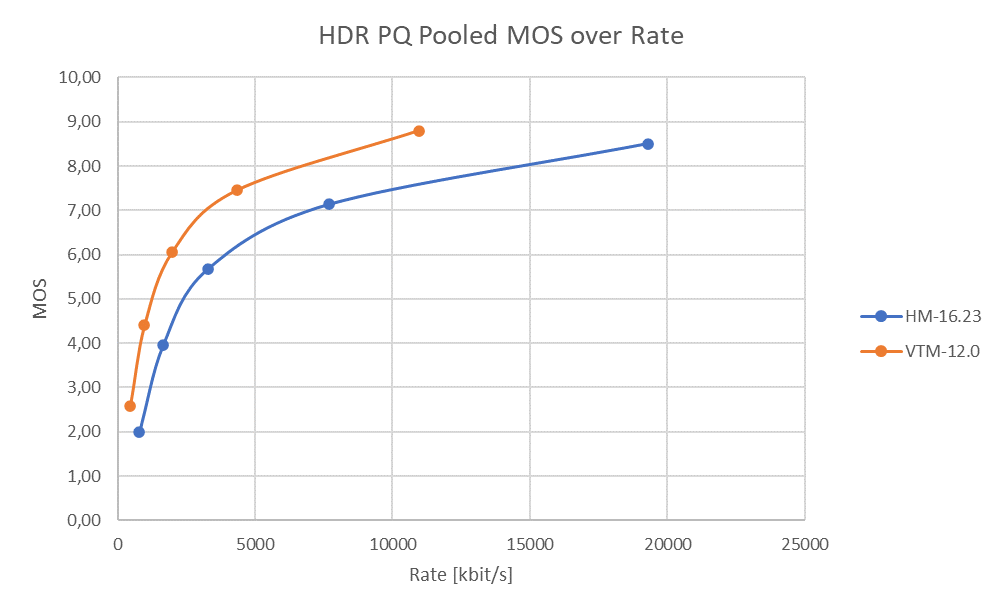
 

Figure 1: Pooled MOS over bit rate plots over the test sets for the HDR HLG and PQ test categories[[1]](#footnote-1)

# Introduction

A major design goal for the development of the VVC standard was to achieve a substantial improvement in compression capability relative to its predecessor, the HEVC standard. This document is the third in a planned series of reports addressing a variety of test categories and embracing some of the available versatile tools provided by the VVC standard. It reports the results of a verification test to confirm that this goal was achieved and to estimate the magnitude of that achievement, following a test plan issued at the previous meeting [1].

A subjective evaluation was conducted at two test sites comparing the VVC Main 10 profile to the HEVC Main 10 profile for the HDR test category at UHD resolution for content represented with both HLG and SMPTE ST 2084 transfer functions.

# Verification test logistics

The HDR PQ and HDR HLG subjective test was carried out at the following test sites:

* GBTech, Rome, IT
* Vabtech, London, UK

The tests were conducted using the degradation category rating (DCR) test method [2] with an 11-grade impairment scale [3]. The verification test environment and testing methodology are described in Annex A. The arrangements for the test sites are shown in Table 1.

**Table 1: Test site information and setup**

| **Test Site** | **GBTech** | **Vabtech** |
| --- | --- | --- |
| **Display, size  (resolution setting)** | LG 65” CX6LA (3840x2160) | LG 65” CX6LA (3840x2160) |
| **Viewing distance** | 1 viewer at 1.5H | 1 viewer at 1.5H |
| **Viewing angle** | 90° (at screen center) | 90° (at screen center) |
| **Total number of viewers** | 30 (16 females, 14 males; age 18-24), all screened for visual acuity and normal colour vision. | 29 (20 females, 9 males, age 18-24), all screened for visual acuity and normal colour vision. |

The display used the HDMI interface as input. The HDR capability of the display was enabled by signaling the appropriate InfoFrame as specified in [4]. Specifically, the *EOTF* data byte in the Dynamic Range and Mastering InfoFrame was set to the value of 3 for HLG content, and it was set to the value of 2 for the PQ content. Additionally, for PQ content, the values for the *max\_display\_mastering\_luminance* data bytes and *Maximum Content Light Level* data bytes in the Static Metadata Descriptor were set equal to the “Maximum Display Mastering Luminance” reported for each sequence in Table 4 below.

# Verification test setup

## Verification test content generation

In the test, the HEVC bitstreams were encoded using the HEVC reference software HM16.23 [5]. For VVC, the VTM-12.0 reference software [6] was used. According to the definition of the HDR tests in the verification test plan [1], motion compensated filtering was enabled for both the HM-16.23 and VTM-12.0 bit-streams. For HM-16.23, the random-access configuration provided with the configuration file cfg/encoder\_randomaccess\_main10.cfg in the software package was used. For VTM-12.0, the random-access configuration provided with the configuration file cfg/encoder\_randomaccess\_vtm.cfg in the VTM-12.0 software package was employed. Specific sequence configuration settings as well as settings related to HDR are provided in JVET document JVET-W0041 [7]. All bitstreams have been generated by Sharp Labs of America, and have been independently verified by Ericsson.

## Test method and test design

The test sequences were evaluated using the 11-grade scale as specified in Rec. ITU-R BT.500-14, shown in Figure 2 below.

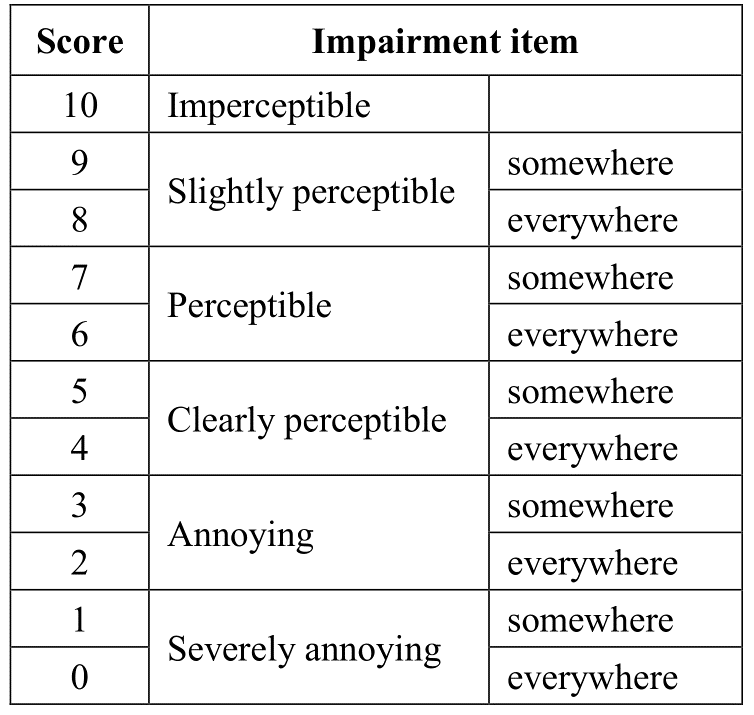


Figure 2: Meaning of the 11 grades numerical scale as specified in Rec. ITU-R BT.500-14 Table 2-4 [2]

A total of four test sessions were designed for the HDR test, two for the PQ and two for the HLG; all the test sessions were designed inserting a “stabilization phase” as suggested in ITU-R BT.500-14, and were not longer than 12 minutes, to avoid fatigue impact.

# Test sequences and rate point selection

According to the verification test plan [1], five bit-rate points for each test sequence were selected for the quality assessment of the test sequences. The bit-rate points were chosen such that the VTM/HM pair for a bit-rate point would represent approximately the same quality while at the same time allowing for approximate bit-rate matching of each HM bit-rate point with the next VTM bit-rate point. Thereby both an assessment of bit-rate savings at similar quality and an assessment of quality improvement at similar bit rates are enabled.

## HDR HLG test sequences

**Table 2: HDR HLG test sequences**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test sequence** | **Resolution** | **fps** | **Frames** | **md5** |
| H3\_AMS01\_3840x2160\_10bit\_420\_HLG (TramBridge) | 3840×2160 | 60 | 600 | 02be2d01e46a4d2316fc0e951f94f742 |
| H3\_AMS06\_3840x2160\_10bit\_420\_HLG (CanalWithTrees) | 3840×2160 | 60 | 600 | d1f11c771febbb8a2bbb7faadc13cbbb |
| RiverByBoat\_3840x2160\_60fps\_10bit\_HLG\_420\_type2 | 3840×2160 | 60 | 600 | 0b5898e124fe38969eb4b7d8747a408a |
| C07\_DramaSea | 3840×2160 | 60 | 600 | 3ced19e6e58b6f4c57ca97c53fe75c92 |
| C16\_PaddockFollow | 3840×2160 | 60 | 600 | 0696df7a9df2ae8be2bc933b4b9a5764 |

**Table 3: QP settings for the HDR HLG test sequences**

|  |  |  |
| --- | --- | --- |
| **Sequence** | **HM QPs** | **VTM QPs** |
| H3\_AMS01\_3840x2160\_10bit\_420\_HLG (TramOnBridge) | 41, 37, 33, 29, 25 | 43, 39, 35, 31, 27 |
| H3\_AMS06\_3840x2160\_10bit\_420\_HLG (CanalWithTrees) | 40, 37, 33, 30, 26 | 42, 39, 35, 32, 28 |
| RiverByBoat\_3840x2160\_60fps\_10bit\_HLG\_420\_type2 | 44, 40, 36, 33, 29 | 47, 43, 39, 35, 31 |
| C07\_DramaSea | 39, 35, 31, 26, 21 | 41, 37, 33, 29, 24 |
| C16\_PaddockFollow | 42, 37, 31, 27, 23 | 45, 40, 35, 30, 26 |

## HDR PQ test sequences

**Table 4: HDR PQ test sequences**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Test sequence** | **Resolution** | **fps** | **Frames** | **Transfer Function** | **Maximum Display Mastering Luminance** | **md5** |
| Chimera HDR5 (DrivingPOV\_HDR) | 3840×2160 | 60 | 600 | PQ | 4000 | b75489a8a58a027c11208a993df77253 |
| Meridian HDR2 (Car) | 3840×2160 | 60 | 600 | PQ | 4000 | 9e889c7d78b1b5a0fdebb88871a362dd |
| Meridian HDR5 (Beach) | 3840×2160 | 60 | 600 | PQ | 4000 | 8c2cd7dc85a930380764c14f5c5f63e8 |
| Sparks DirtLot | 3840x2160 | 60 | 600 | PQ | 1000 | 825711f4e73280307b29ca6d842e41b7 |
| Sparks Welding | 3840x2160 | 60 | 600 | PQ | 1000 | 7e0c6bd867c370dde9b11c76ed93409f |

**Table 5: QP settings for the HDR PQ test sequences**

|  |  |  |
| --- | --- | --- |
| **Sequence** | **HM QPs** | **VTM QPs** |
| Chimera HDR5 (DrivingPOV\_HDR) | 42, 35, 31, 25, 20 | 45, 38, 32, 28, 22 |
| Meridian HDR2 (Car) | 35, 31, 27, 24, 20 | 38, 34, 30, 26, 22 |
| Meridian HDR5 (Beach) | 36, 32, 28, 24, 20 | 38, 34, 30, 26, 22 |
| Sparks DirtLot | 42, 36, 31, 25, 20 | 44, 38, 33, 27, 22 |
| Sparks Welding | 42, 36, 31, 25, 20 | 44, 38, 33, 27, 22 |

# Results and analysis

The measured MOS values of the reconstructed video on the 11-grade scale are plotted over the bit rate of the corresponding bitstream. The ±95% confidence intervals for the MOS values are indicated.

For all test categories, the bit rate and MOS differences for all bit-rate points are collected and the Bjøntegaard delta rate relative to HM-16.23 based on bit rate and MOS is reported. The Bjøntegaard delta rate has been computed with the RDPlot tool [15]. In the analysis tables, the bit-rate savings are computed as the difference between the VVC bit-rate point and the corresponding HEVC bit-rate point relative to the HEVC bit rate. Further, the MOS difference is reported as a number if the value is larger than the maximum of the VVC and the HEVC confidence intervals. Otherwise, “< CI” is indicated.

## HRD HLG

### MOS plots

**Table 6: Bitrate, MOS and confidence intervals**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **HM-16.23** |  |  | **VTM-12.0** |  |  |
| **Sequence** | **RP** | **Rate [kbit/s]** | **MOS** | **CI** | **Rate [kbit/s]** | **MOS** | **CI** |
| H3\_AMS01 | R1 | 37880.61 | 8.7 | 0.6 | 23399.52 | 8.9 | 0.5 |
| H3\_AMS01 | R2 | 18431.91 | 7.5 | 0.3 | 11474.43 | 7.7 | 0.4 |
| H3\_AMS01 | R3 | 9308.18 | 6.0 | 0.3 | 6023.96 | 6.5 | 0.3 |
| H3\_AMS01 | R4 | 4789.44 | 4.4 | 0.4 | 3165.08 | 4.8 | 0.4 |
| H3\_AMS01 | R5 | 2467.17 | 2.5 | 0.4 | 1636.40 | 3.0 | 0.3 |
| H3\_AMS06 | R1 | 38139.43 | 7.8 | 0.4 | 22697.30 | 8.3 | 0.3 |
| H3\_AMS06 | R2 | 16508.93 | 6.7 | 0.4 | 10564.61 | 6.9 | 0.3 |
| H3\_AMS06 | R3 | 9133.48 | 5.3 | 0.4 | 6019.55 | 5.6 | 0.3 |
| H3\_AMS06 | R4 | 4167.16 | 3.4 | 0.3 | 2823.51 | 3.9 | 0.4 |
| H3\_AMS06 | R5 | 2320.03 | 1.8 | 0.3 | 1560.59 | 2.4 | 0.5 |
| RiverByBoat | R1 | 43099.51 | 7.8 | 0.4 | 25803.41 | 8.5 | 0.4 |
| RiverByBoat | R2 | 22113.81 | 6.5 | 0.4 | 13816.15 | 7.2 | 0.4 |
| RiverByBoat | R3 | 13407.19 | 5.2 | 0.4 | 7130.70 | 5.6 | 0.4 |
| RiverByBoat | R4 | 6734.77 | 3.7 | 0.3 | 3456.85 | 4.1 | 0.4 |
| RiverByBoat | R5 | 3149.93 | 2.3 | 0.4 | 1524.49 | 2.5 | 0.5 |
| C07\_DramaSea | R1 | 25409.13 | 8.1 | 0.3 | 14085.78 | 8.7 | 0.3 |
| C07\_DramaSea | R2 | 11308.54 | 7.0 | 0.5 | 6022.47 | 7.3 | 0.4 |
| C07\_DramaSea | R3 | 4765.88 | 5.7 | 0.2 | 3025.95 | 5.9 | 0.4 |
| C07\_DramaSea | R4 | 2320.70 | 3.5 | 0.3 | 1584.74 | 4.2 | 0.5 |
| C07\_DramaSea | R5 | 1213.05 | 2.3 | 0.4 | 825.75 | 2.3 | 0.4 |
| C16\_PaddockFollow | R1 | 14752.07 | 8.5 | 0.3 | 7114.76 | 8.7 | 0.4 |
| C16\_PaddockFollow | R2 | 6911.21 | 7.7 | 0.3 | 3681.94 | 8.0 | 0.4 |
| C16\_PaddockFollow | R3 | 3869.57 | 6.6 | 0.4 | 1809.28 | 6.6 | 0.4 |
| C16\_PaddockFollow | R4 | 1664.20 | 4.7 | 0.3 | 924.72 | 5.1 | 0.5 |
| C16\_PaddockFollow | R5 | 880.90 | 2.1 | 0.4 | 480.78 | 2.5 | 0.5 |

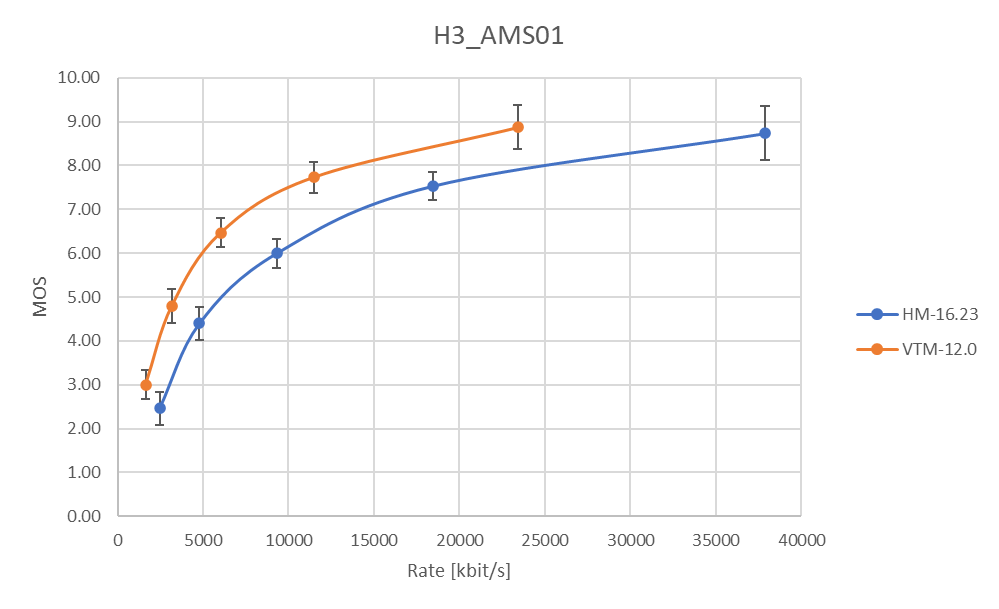
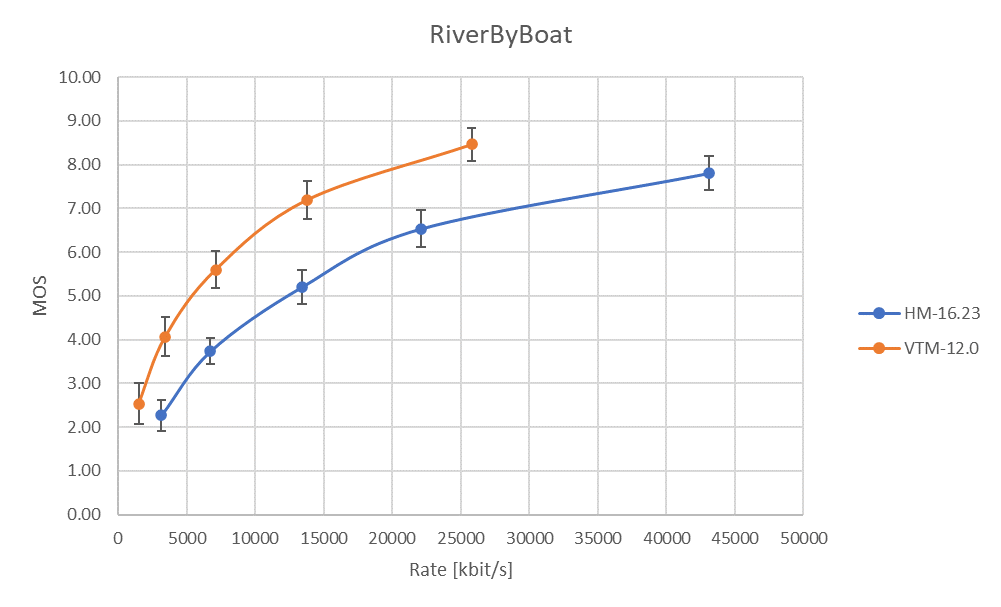
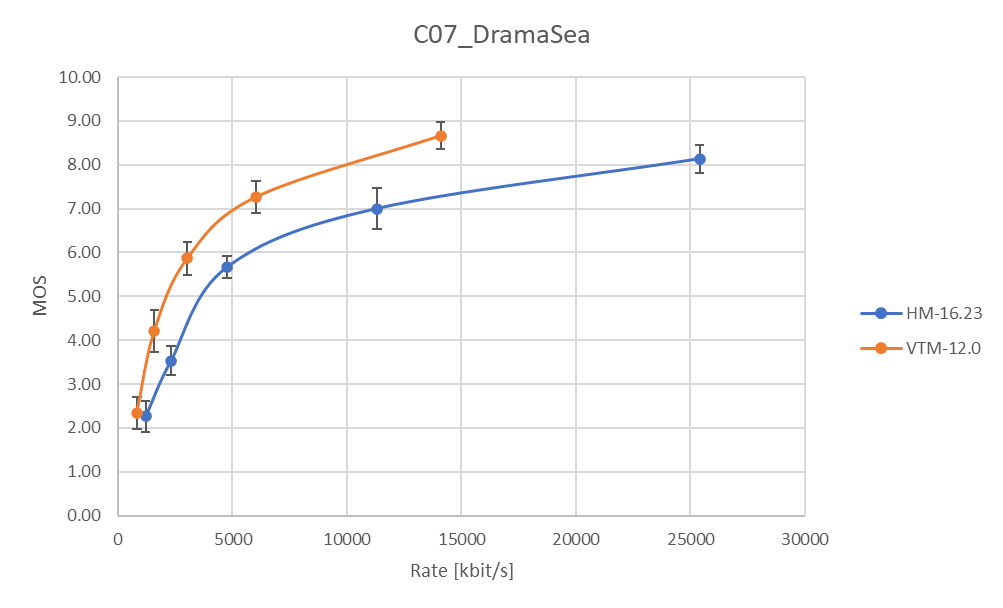
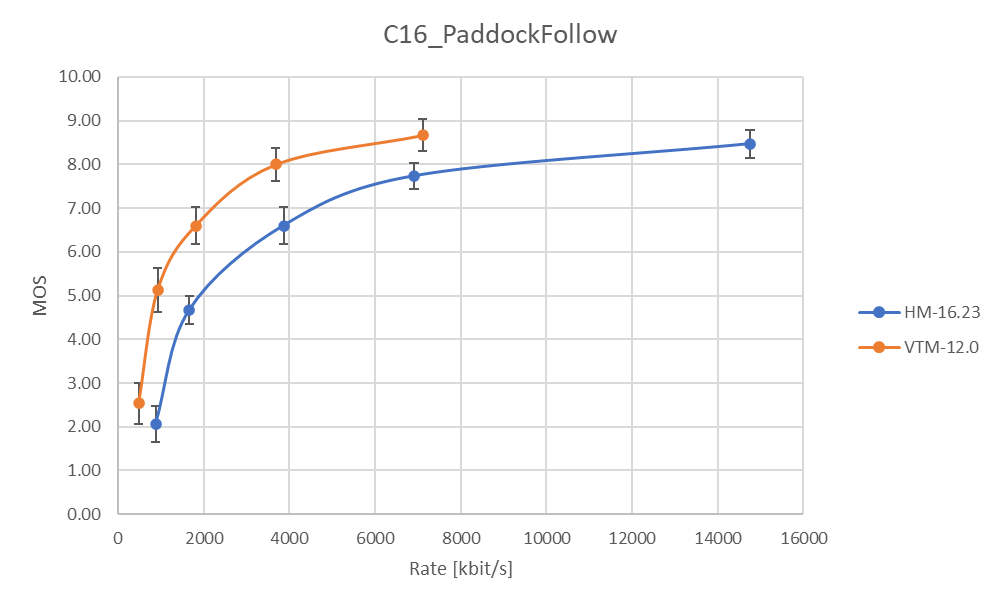
    

Figure 3: Collection of the MOS-over-rate plots for HM and VTM for the HDR HLG test sequences

### Analysis

**Table 7: Bit-rate savings and MOS deltas for the bit-rate points and Bjøntegaard delta rate relative to HM-16.23 based on bit rate and MOS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **VTM / HM** |  | **Rate Diff.** | **ΔMOS** |  | **BD-Rate** | **VTM-12.0** |
| H3\_AMS01 | R1 | -38.2% | < CI |  | H3\_AMS01 | -45% |
| H3\_AMS01 | R2 | -37.7% | < CI |  | H3\_AMS06 | -44% |
| H3\_AMS01 | R3 | -35.3% | 0.5 |  | RiverByBoat | -48% |
| H3\_AMS01 | R4 | -33.9% | 0.4 |  | C07\_DramaSea | -53% |
| H3\_AMS01 | R5 | -33.7% | 0.5 |  | C16\_PaddockFollow | -56% |
| H3\_AMS06 | R1 | -40.5% | 0.5 |  | **Overall** | **-49%** |
| H3\_AMS06 | R2 | -36.0% | < CI |  |  |  |
| H3\_AMS06 | R3 | -34.1% | < CI |  |  |  |
| H3\_AMS06 | R4 | -32.2% | 0.5 |  |  |  |
| H3\_AMS06 | R5 | -32.7% | 0.6 |  |  |  |
| RiverByBoat | R1 | -40.1% | 0.7 |  |  |  |
| RiverByBoat | R2 | -37.5% | 0.7 |  |  |  |
| RiverByBoat | R3 | -46.8% | < CI |  |  |  |
| RiverByBoat | R4 | -48.7% | < CI |  |  |  |
| RiverByBoat | R5 | **-51.6%** | < CI |  |  |  |
| C07\_DramaSea | R1 | -44.6% | 0.5 |  |  |  |
| C07\_DramaSea | R2 | -46.7% | < CI |  |  |  |
| C07\_DramaSea | R3 | -36.5% | < CI |  |  |  |
| C07\_DramaSea | R4 | -31.7% | 0.7 |  |  |  |
| C07\_DramaSea | R5 | -31.9% | < CI |  |  |  |
| C16\_PaddockFollow | R1 | **-51.8%** | < CI |  |  |  |
| C16\_PaddockFollow | R2 | -46.7% | < CI |  |  |  |
| C16\_PaddockFollow | R3 | **-53.2%** | < CI |  |  |  |
| C16\_PaddockFollow | R4 | -44.4% | < CI |  |  |  |
| C16\_PaddockFollow | R5 | -45.4% | 0.5 |  |  |  |

The results reported in Table 7 indicate significant compression performance improvements for VVC compared to its predecessor HEVC. As can be seen from Figure 3, the MOS values of VVC are always comparable to or higher than the MOS values of the corresponding HEVC rate points. Table 7 reveals that for 56% of the rate points, the MOS values for HEVC and VVC are within the confidence interval, and for 44% of the rate points, the MOS values for VVC are superior to HEVC. Therefore, the estimated Bjøntegaard delta rate savings reported in Table 7 are superior to the rate savings reported for the single rate points.

## HRD HLG

### MOS plots

**Table 8: Bitrate, MOS and confidence intervals**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **HM-16.23** |  |  | **VTM-12.0** |  |  |
| **Sequence** | **RP** | **Rate [kbit/s]** | **MOS** | **CI** | **Rate [kbit/s]** | **MOS** | **CI** |
| DrivingPOV\_HDR | R1 | 49094.45 | 8.3 | 0.2 | 24128.22 | 8.7 | 0.4 |
| DrivingPOV\_HDR | R2 | 18344.12 | 7.3 | 0.2 | 8232.80 | 7.6 | 0.3 |
| DrivingPOV\_HDR | R3 | 7034.90 | 5.8 | 0.3 | 4470.62 | 6.4 | 0.4 |
| DrivingPOV\_HDR | R4 | 3992.06 | 4.4 | 0.4 | 1934.62 | 4.7 | 0.5 |
| DrivingPOV\_HDR | R5 | 1686.31 | 1.6 | 0.5 | 902.95 | 2.9 | 0.5 |
| Car | R1 | 7979.81 | 8.5 | 0.3 | 4291.67 | 8.7 | 0.4 |
| Car | R2 | 3654.84 | 6.9 | 0.3 | 2000.00 | 7.4 | 0.4 |
| Car | R3 | 2209.40 | 5.5 | 0.3 | 1056.78 | 6.0 | 0.3 |
| Car | R4 | 1198.89 | 4.0 | 0.3 | 591.88 | 4.5 | 0.3 |
| Car | R5 | 658.60 | 2.4 | 0.3 | 344.55 | 2.8 | 0.3 |
| Beach | R1 | 10281.20 | 8.1 | 0.3 | 5982.14 | 8.7 | 0.4 |
| Beach | R2 | 4452.02 | 6.9 | 0.3 | 2642.77 | 7.1 | 0.2 |
| Beach | R3 | 2098.14 | 5.5 | 0.3 | 1277.40 | 5.7 | 0.6 |
| Beach | R4 | 1096.59 | 3.7 | 0.4 | 669.31 | 4.1 | 0.5 |
| Beach | R5 | 591.99 | 2.3 | 0.4 | 374.36 | 2.6 | 0.3 |
| Sparks DirtLot | R1 | 19384.04 | 8.7 | 0.4 | 11180.78 | 9.0 | 0.3 |
| Sparks DirtLot | R2 | 6043.00 | 7.0 | 0.3 | 3580.00 | 7.5 | 0.3 |
| Sparks DirtLot | R3 | 1997.90 | 5.1 | 0.6 | 1263.14 | 5.7 | 0.2 |
| Sparks DirtLot | R4 | 901.26 | 3.2 | 0.3 | 598.17 | 3.7 | 0.4 |
| Sparks DirtLot | R5 | 385.00 | 1.7 | 0.4 | 254.50 | 2.3 | 0.4 |
| Sparks Welding | R1 | 34211.82 | 8.8 | 0.4 | 22728.99 | 8.8 | 0.4 |
| Sparks Welding | R2 | 14754.68 | 7.6 | 0.3 | 9776.90 | 7.7 | 0.5 |
| Sparks Welding | R3 | 5864.37 | 6.5 | 0.2 | 4069.17 | 6.6 | 0.3 |
| Sparks Welding | R4 | 2617.98 | 4.5 | 0.3 | 1853.46 | 5.1 | 0.4 |
| Sparks Welding | R5 | 1200.99 | 1.9 | 0.4 | 657.01 | 2.3 | 0.4 |

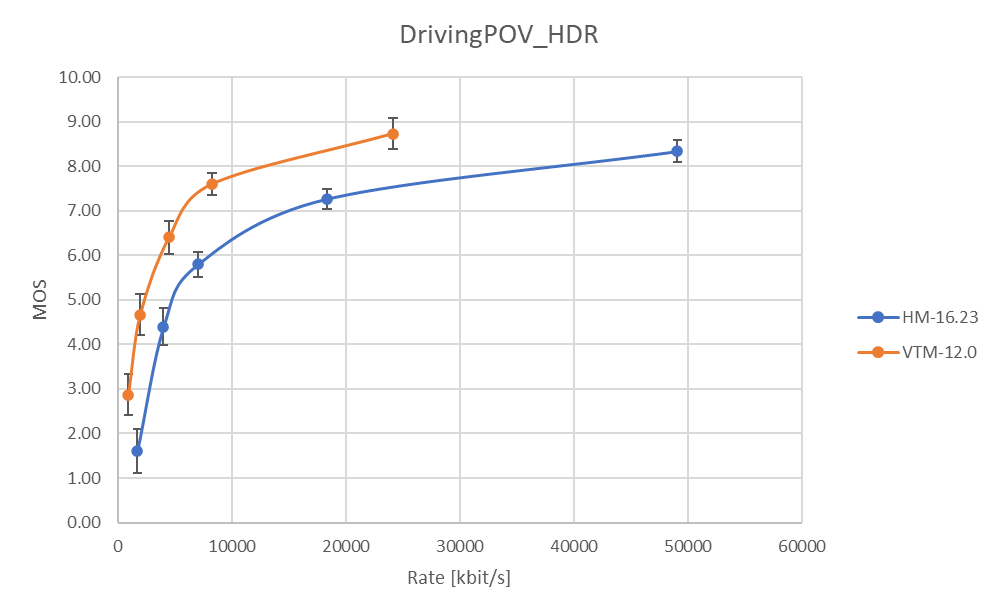
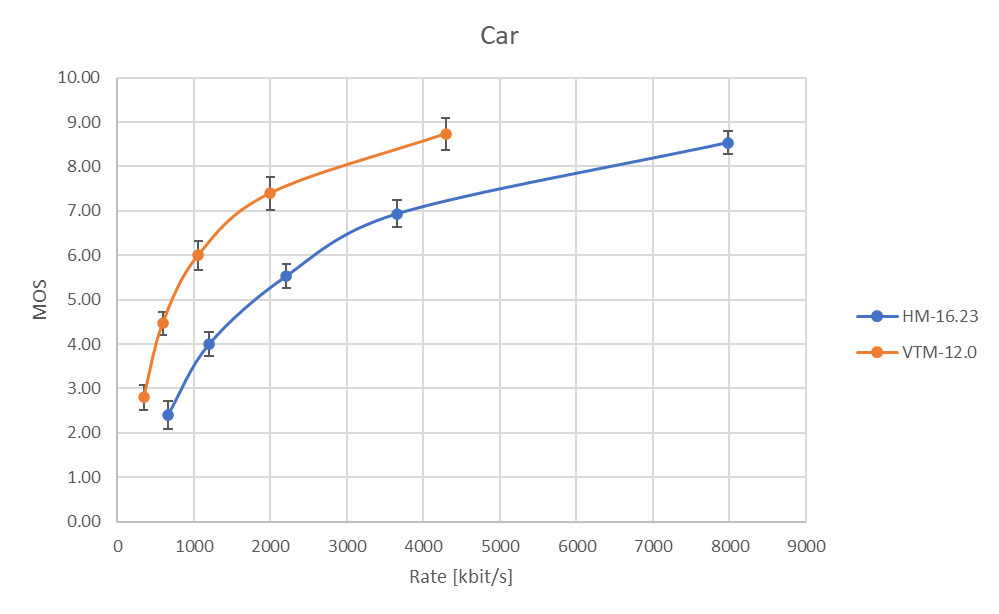
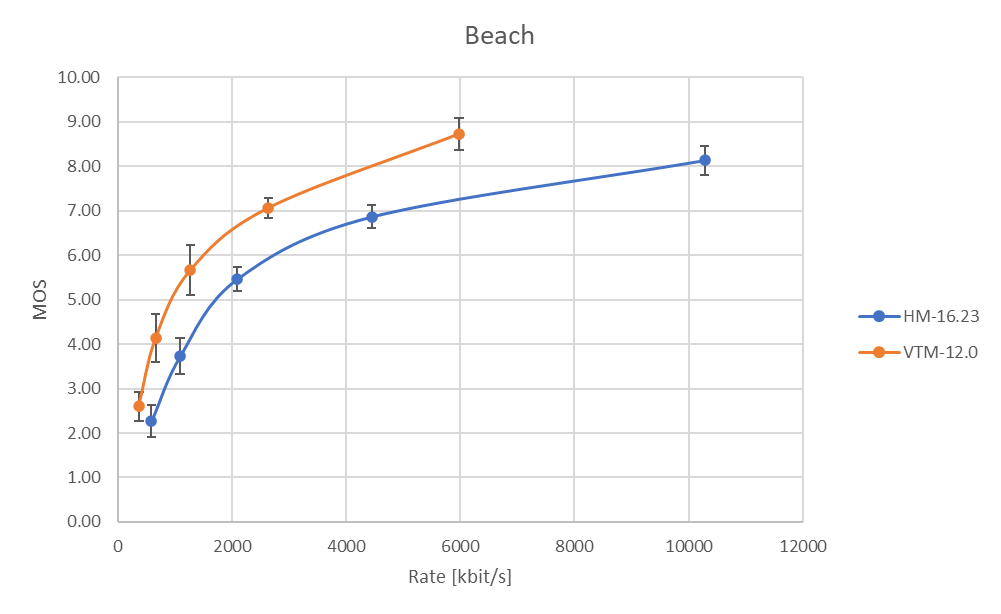
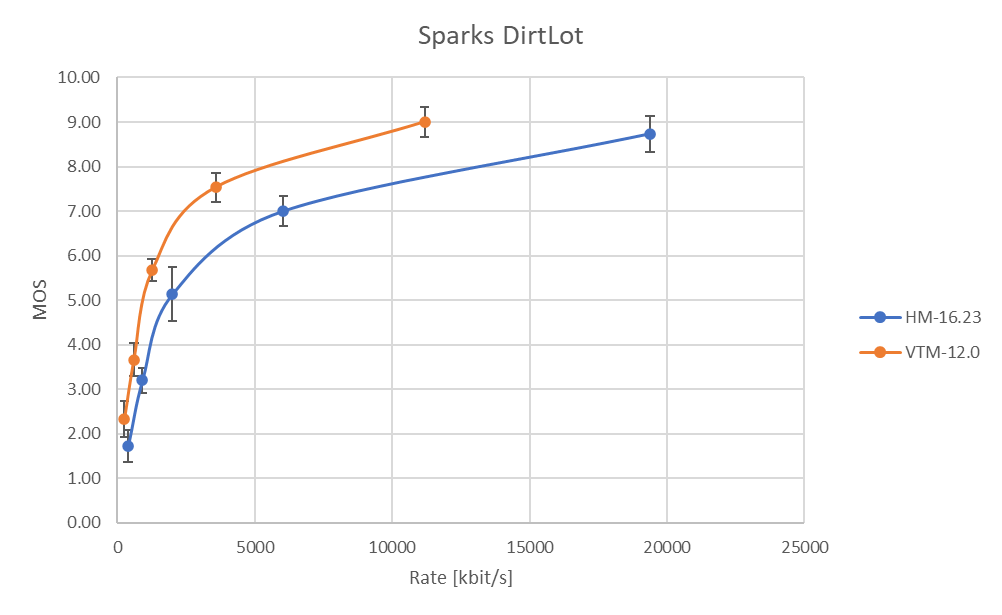
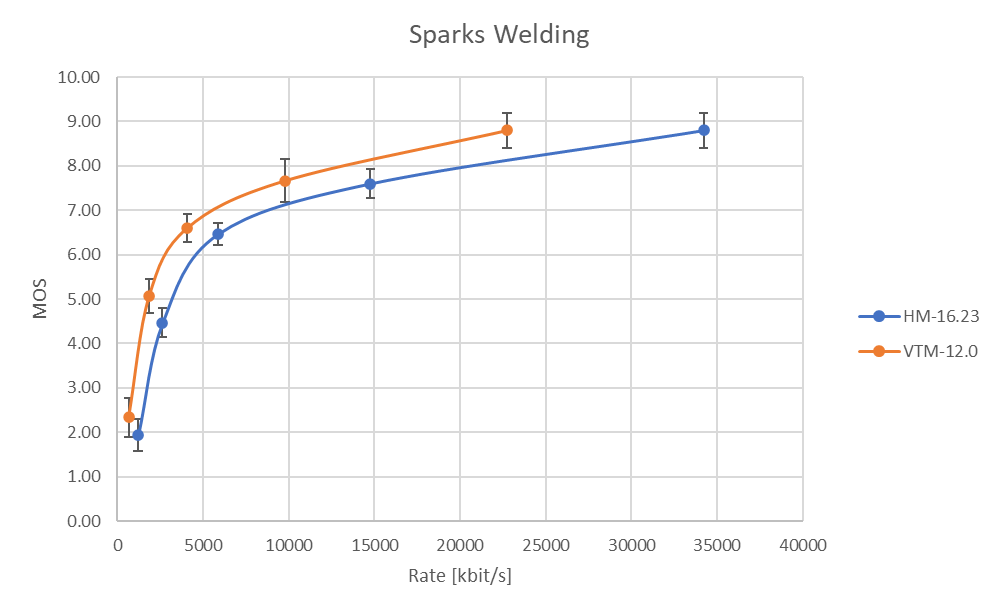
    

Figure 4: Collection of the MOS-over-rate plots for HM and VTM for the HDR PQ test sequences

### Analysis

**Table 9: Bit-rate savings and MOS deltas for the bit-rate points and Bjøntegaard delta rate relative to HM-16.23 based on bit rate and MOS**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **VTM / HM** |  | **Rate Diff.** | **ΔMOS** |  | **BD-Rate** | **VTM-12.0** |
| DrivingPOV\_HDR | R1 | **-50.9%** | 0.4 |  | DrivingPOV\_HDR | -59% |
| DrivingPOV\_HDR | R2 | **-55.1%** | 0.3 |  | Car | -58% |
| DrivingPOV\_HDR | R3 | -36.5% | 0.6 |  | Beach | -47% |
| DrivingPOV\_HDR | R4 | **-51.5%** | < CI |  | Sparks DirtLot | -53% |
| DrivingPOV\_HDR | R5 | -46.5% | 1.3 |  | Sparks Welding | -42% |
| Car | R1 | -46.2% | < CI |  | **Overall** | **-52%** |
| Car | R2 | -45.3% | 0.5 |  |  |  |
| Car | R3 | **-52.2%** | 0.5 |  |  |  |
| Car | R4 | **-50.6%** | 0.5 |  |  |  |
| Car | R5 | -47.7% | 0.4 |  |  |  |
| Beach | R1 | -41.8% | 0.6 |  |  |  |
| Beach | R2 | -40.6% | < CI |  |  |  |
| Beach | R3 | -39.1% | < CI |  |  |  |
| Beach | R4 | -39.0% | < CI |  |  |  |
| Beach | R5 | -36.8% | < CI |  |  |  |
| Sparks DirtLot | R1 | -42.3% | < CI |  |  |  |
| Sparks DirtLot | R2 | -40.8% | 0.5 |  |  |  |
| Sparks DirtLot | R3 | -36.8% | < CI |  |  |  |
| Sparks DirtLot | R4 | -33.6% | 0.5 |  |  |  |
| Sparks DirtLot | R5 | -33.9% | 0.6 |  |  |  |
| Sparks Welding | R1 | -33.6% | < CI |  |  |  |
| Sparks Welding | R2 | -33.7% | < CI |  |  |  |
| Sparks Welding | R3 | -30.6% | < CI |  |  |  |
| Sparks Welding | R4 | -29.2% | 0.6 |  |  |  |
| Sparks Welding | R5 | -45.3% | < CI |  |  |  |

The results reported in Table 9 indicate significant compression performance improvements for VVC compared to its predecessor HEVC. Similar to the results for the HDR HLG test sequences, the MOS values of VVC are always comparable to or higher than the MOS values of the corresponding HEVC rate points. About half of the rate points have a distance within the confidence interval or show a superior MOS value, respectively. Again, this behavior induces the higher Bjøntegaard delta rate savings as reported in Table 9 relative to the rate savings reported for the single rate points.

# Conclusions

The compression performance capabilities of the two specifications are compared based on the HEVC reference software HM-16.23 and the VVC reference software VTM-12.0. A visual assessment of VVC compared to HEVC was performed, and the result of a formal subjective assessment evaluation with the participation of naïve test subjects is reported. The tests were conducted using consumer-type displays in order to match the application scenario of end users. The assessment includes testing with five sequences with HDR HLG content and five sequences with HDR PQ content. For the HDR PQ test case, test sequences with two different values of Maximum Display Mastering Luminance (1,000 cd/m² and 4,000 cd/m²) were included to reflect different corresponding use cases. The MOS-over-bitrate plots demonstrate the superior performance of VVC compared to its predecessor, HEVC, with a good coverage of the MOS range from slightly perceptible impairments down to annoying impairments for all sequences. The Bjøntegaard delta rate figures show an overall gain of about 49% for the HLG test sequences and an overall gain of about 52% for the PQ test sequences.

# References

1. M. Wien, V. Baroncini, A. Segall, and Y. Ye, “VVC verification test plan (Draft 6),” Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, output document JVET-V2021, 22nd JVET meeting, by teleconference, Apr. 2021.
2. Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.
3. Recommendation ITU-R BT.500-14 (2019), *Methodologies for the subjective assessment of the quality of television images*.
4. CTA 861-G, *A DTV Profile for Uncompressed High Speed Digital Interfaces*, 2016.
5. HEVC Test Model, <https://vcgit.hhi.fraunhofer.de/jct-vc/HM/-/tree/HM-16.23>, online, accessed Jun. 2021.
6. VVC Test Model, <https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-12.0>, online, accessed Jun. 2021.
7. A. Segall, M. Wien, V. Baroncini, K. Andersson, “AHG4: Status Report on HDR Video Verification Test Preparation,” Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, input document JVET-W0041, 23rd JVET meeting, by teleconference, Jul. 2021.
8. Recommendation ITU-R BT.2100-2 (2018), *Image parameter values for high dynamic range television for use in production and international programme exchange*.
9. SMPTE ST 2084, *High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays*, 2014.
10. SMPTE ST 2036-1, *Ultra High Definition Television – Image Parameter Values for Program Production*, 2014.
11. ETSI TS 101 154, *Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcast and Broadband Applications*, 2019.
12. ITU-T Tech. Paper HSTP-VID-WPOM and ISO/IEC TR 23008-8 (*Eds.*: K. Andersson, F. Bossen, J.-R. Ohm, A. Segall, R. Sjöberg, J. Ström, G. J. Sullivan, A. Tourapis), Working practices using objective metrics for evaluation of video coding efficiency experiments, July 2020.
13. G. Bjøntegaard, “Calculation of average PSNR differences between RD-curves,” ITU-T SG 16 Q.6 input document VCEG-M33, 13th VCEG meeting, Austin, Texas, USA, Apr. 2001.
14. G. Bjøntegaard, "Improvements of the BD-PSNR model," ITU-T SG16 Q.6 document VCEG-AI11, 35th VCEG meeting, Berlin, Germany, July 2008.
15. RDPlot, <https://github.com/IENT/RDPlot>, accessed 2021-06-18.
16. J.-R. Ohm, G. J. Sullivan, “Meeting Report of the 22nd Meeting of the Joint Video Experts Team (JVET), by teleconference, 20–28 April 2021”, Joint Video Exploration Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-V1000, teleconference, April 2021.
17. V. Baroncini and M. Wien, “Report on VVC compression performance verification testing in the SDR HD and 360 Video categories”, Joint Video Exploration Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-V0174, teleconference, April 2021.
18. F. Bossen, X. Li, K. Sühring, K. Sharman, and V. Seregin, “JVET AHG report: Test model software development (AHG3) ”, Joint Video Exploration Team of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-V0003, teleconference, April 2021.

# Acknowledgements

The verification test coordinators wish to thank the organizations who contributed to and supported the VVC verification tests, particularly including Erisccon, GBTech, Sharp Labs of America, and Vabtech.

# Annex A

The same evaluation method as for the HEVC verification tests is adopted for the VVC verification tests. The following description is based on JCTVC-Q1011 [A4] with minor adaptations.

A.1 Test method

The test method adopted for this evaluation is degradation category rating (DCR) [A1].

A.1.1 Degradation Category Rating (DCR)

This test method is commonly adopted when the material to be evaluated shows a range of visual quality that well distributes across all quality scales. All the video material used for these tests consist of video clips of 10 seconds duration.

This method has been used under the schema of evaluation of the quality; for this reason, a quality rating scale made of 11 levels was adopted, ranging from "0" (lowest quality) to "10" (highest quality), see also Figure 2.

The structure of the basic test cell (BTC) of the DCR method was made by two consecutive presentations of the video clip under test; at first the original version of the video clip is displayed, immediately afterwards the coded version of the video clip is presented; then a message displays for 5 seconds asking the viewers to vote. The presentation of the video clips is preceded by a mid-grey screen displaying “Source” for the original and “Test” for the coded version of the sequence under test for one second.



Figure 6 – DCR BTC

A.2 How to express the visual quality opinion with DCR

The viewers were asked to express their vote putting a mark on a scoring sheet.

The scoring sheet for a DCR test is made of a section for each BTC; each section has a box wherein which the viewer shall write the score ranging from 0 to 10. By writing a score of “10”, the subject will express an opinion of “best” quality, while by writing a score of “0” the subject will express an opinion of “worst” quality, as shown in Figure 2.

The vote has to be written when the message "Vote N" appears on the screen. The number "N" is a numerical progressive indication on the screen aiming to help the viewing subjects to use the appropriate box of the scoring sheet.

A.4 Training and stabilization phase

The outcome of a test is highly dependent on a proper training of the test subjects.

For this purpose, each subject has to be trained by means of a short practice (training) session demonstrating the range of qualities to be expected in the test.

The stabilization phase uses the test material of a test session; three BTCs, containing one sample of best quality, one of the worst qualities and one of medium quality, are duplicated at the beginning of the test session. By this way, the test subjects have an immediate impression of the quality range they are expected to evaluate during that session.

The scores of the stabilization phase are discarded.

A.5 The laboratory setup

The laboratories for subjective assessments were arranged according to [A1], except for the selection of the display and the video play-out server. Play-out of the HD video clips was done at the native resolution without upscaling, with the HD video being centered on the UHD screen with a mid-gray surrounding.

The PCs used to play the video sequence supported the display of 10 bit UHD at 30 and 60 frames per second, without any limitation, or without introducing any additional temporal or visual degradation. At GBTech and Vabtech, the connection between the PC and the display was provided by a 10 bit-capable HDMI connection. At RWTH Aachen University, the display was connected via quad-link SDI.

A.6 Viewing environment

The viewing distance was 1.5H, where H is equal to the height of the active part of the screen, depending on the size of the active part of the screen and its native resolution.

The test laboratories were protected from external visual or audio pollution. Internal general light was low (just enough to allow the viewing subjects to fill out the scoring sheets) and a uniform light was placed behind the monitor, in a way no direct light hits the viewing subjects seated in front of the screen; the light behind the monitor must be dimmed to an intensity as specified in Table 4 of Recommendation ITU-T P.911 (“Typical viewing and listening conditions as used in audio-visual quality assessment”). No other light source was admitted, and in particular any light source directed to the screen or creating reflections.

A.7 Overall test effort and subjects’ involvement

Each viewing session did not run for more than 20 minutes and the same viewing subject did not participate to the test run for more than six hours in total. Young people were hired as test subjects, selecting them for an age from 16 to 30, mostly students of scientific faculties. Viewing subjects were compensated for their participation to the testing activities.

A.8 Statistical analysis and presentation of the results

The data collected from the score sheets, filled out by the viewing subjects, were stored in an Excel spread sheet. For each coding condition the Mean Opinion Score (MOS) and associated Confidence Interval (CI) values were computed in the spread-sheets.

The MOS and CI values are used to draw graphs. The graphs are drawn grouping the results for each video test sequence. No graph grouping results from different video sequences is considered.

From the “raw” data subject reliability should be calculated and the method used to assess subject reliability should be reported. Some criteria for subjective reliability are given in [A2] and [A3].

A.9 References

1. Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.
2. *Pseudo Isochromatic Plates*, engraved and printed by *The Beck Engraving Co., Inc*., Philadelphia and New York, United States.
3. KIRK (R.E.): *Experimental Design – Procedures for the Behavioural Sciences*, 2nd Edition, *Brooks/Cole Publishing Co*., California, 1982.
4. Tan, T. K.; Mrak, M.; Baroncini, V.; and Ramzan, N., “Report on HEVC compression performance verification testing,” Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11 output document JCTVC-Q1011, Apr. 2014.

1. Pooling was done by computing the geometric mean of the bitrates and the arithmetic mean of the MOS scores across the test sequences of each test category. [↑](#footnote-ref-1)