 ISO/IEC JTC 1/SC 29/AG 5 N00075

**ISO/IEC JTC 1/SC 29/AG 5**

**MPEG Visual Quality Assessment**

**Convenorship: DE**

**Document type:** Output Document

**Title:** **Visual quality comparison of ECM/VTM encoding**

**Status:** Approved

**Date of document:** 2022-12-22

**Source:** ISO/IEC JTC 1/SC 29/AG 5

**Expected action:** Information

**Action due date:** None

**No. of pages:** 9 (without cover page)

**Email of Convenor:** wien@lfb.rwth-aachen.de

**Committee URL:** <https://isotc.iso.org/livelink/livelink/open/jtc1sc29ag5>

**INTERNATIONAL ORGANISATION FOR STANDARDISATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/AG 5**

**MPEG VISUAL QUALITY ASSESSMENT**

**ISO/IEC JTC 1/SC 29/AG 5 N00075**

**Mainz, DE – October 2022**

|  |  |
| --- | --- |
| **Title** | **Visual quality comparison of ECM/VTM encoding** |
| **Source** | AG 5 MPEG Visual Quality Assessment |
| **Status** | Approved |
| **Serial Number** | 22142 |
| **Authors** | M. Wien, J.-R. Ohm, V. Baroncini |

# Abstract

The results of subjective assessments comparing the compression performance of the ECM-6 to the VTM-11 with ECM-comparable settings are reported. The rate points for the tests were selected from a set of ECM simulations. The corresponding VTM bitstreams were generated using a one-time QP switch (QP+=1) within the sequence to match the VTM bitrate to the ECM. The switching points were chosen such that the ECM rate should never be higher than the VTM rate and that the distance should not exceed about 2% of the VTM rate. Two tests were conducted on the provided test set: An expert viewing test at the Mainz meeting site and a laboratory test with naïve test subjects by VABTech in Rome. It is reported that generally, the tendency of both tests is congruent with a clear visual benefit of the ECM when compared to the VTM in a significant number of test cases. For the laboratory results, reported BD rate savings indicate a benefit of about 38% for the UHD test sequences and about 32-33% for the HD test sequences on the given test set.

# Introduction

This document reports the results of a subjective assessment of the compression performance of the ECM (Enhanced Compression Model, an exploration software on technology targeting higher video compression capability) compared to the VTM (VVC Test Model reference software) as an anchor. ECM-6, which was the latest version available for the task and VTM-11 with ECM-comparable settings were used for this purpose. Algorithm descriptions related to these software packages can be found in [1][2].

Tests are presented for the test configurations UHD-RA, HD-RA, and HD-LD, using the random-access (RA) and low-delay (LD) configuration files available with the corresponding VTM and ECM software. The rate points for the tests were selected from a set of ECM simulations [3]. The corresponding VTM bitstreams were generated using a one-time QP switch (QP+=1) within the sequence to match the VTM bitrate to the ECM. The switching points were chosen such that the ECM rate should never be higher than the VTM rate and that the distance should not exceed about 2% of the VTM rate.

Two tests were conducted on the provided test set: An expert viewing test at the Mainz meeting site and a laboratory test with naïve test subjects by VABTech in Rome.

# Test setup

## Logistics

### On-site

At the meeting site, two identical setups were employed. These included a PC with a Decklink video board for HDMI connection and SSD drives capable of stable playout of the raw YUV data at the required frame rate.

| **Test Site** | **On-site** |
| --- | --- |
| **Display, size, connection  (resolution setting)** | 2× LG 65” E9, HDMI (3840×2160) |
| **Viewing distance** | 3 viewers at 1.5H |
| **Viewing angle** | ±75°, 90° (at screen center) |
| **Total number of viewers** | 24 |

Participating experts confirmed visual acuity and normal colour vision (4 females, 20 males).

### GBTech

| **Test Site** | **VABTech** |
| --- | --- |
| **Display, size, connection  (resolution setting)** | LG 65” CX6LA (3840×2160) |
| **Viewing distance** | 1.5H |
| **Viewing angle** | 90° (at screen center) |
| **Total number of viewers** | 24 |

The viewers were checked for acuity and color blindness (18 females, 6 males), age between 19 and 24.

## Test sequences and quantizer settings

For evaluation a set of sequences known from the JVET common testing conditions (CTC) and from the VVC verification tests was selected. For the verification test sequences, the results from the verification test reports were used to determine the potentially applicable quantization parameters. For all sequences, the resulting quality range for the suggested QP values was inspected.

All bitstreams were produced crosschecked by JVET volunteers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Class** | **Sequence name** | **fps** | **Frames** | **md5 hash** |
| UHD RA | Campfire | 30 | 300 | 63d3d9f9e4e8b5c344e89840e84e6428 |
| UHD RA | CatRobot1 | 60 | 600 | 03a89792693fd9ecfd72ef2590025e97 |
| UHD RA | DaylightRoad2 | 60 | 600 | bf1d22643afb41b1521749d2749fb5f0 |
| UHD RA | DrivingPOV3 | 60 | 600 | e81b65724c4235128b2749ccb3b0fb4a |
| UHD RA | Marathon2 | 30 | 300 | c065dfb87be3b2e2ab0ce35094fd4eb4 |
| UHD RA | MountainBay2 | 30 | 300 | f27b6b70244fb083baac546958fcf696 |
| UHD RA | TallBuildings2 | 30 | 300 | 9a0a3f261d004fa86754751c82fb8b47 |
| HD RA | BQTerrace | 60 | 600 | efde9ce4197dd0b3e777ad32b24959cc |
| HD RA | Cactus | 50 | 500 | be1c5d02a8fb298e26e5c5b890451413 |
| HD RA | Meridian2 | 60 | 600 | 473757104c47837efe3fc3b85aef642a |
| HD LD | Beatriz | 50 | 500 | fe74cd5046fa033b4f743f42b29e69cd |
| HD LD | EuroTruckSimulator2 | 60 | 600 | f6850dbfff967945c0a273a374f28abd |
| HD LD | DOTA2 | 60 | 600 | be1c5d02a8fb298e26e5c5b890451413 |

It was agreed to perform rate matching for the VTM rate points in this comparison. For this purpose, it was suggested that the ECM rate should not be higher than the VTM rate and that the distance should not exceed about 2% of the VTM rate. It was agreed to keep QP values in the case where the overshoot of the ECM rate was less than 1% of the VTM rate. The selected ECM QPs and the matching VTM QPs are provided in the table below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sequence** | **ECM QP** | **VTM QP** | **VTM QPincrFrame** | **Rate ECM** | **Rate VTM** | **ratio** |
| Campfire | 42 | 43 | 260 | 1384.39 | 1396.63 | 99.12% |
| Campfire | 38 | 39 | 260 | 2545.46 | 2575.79 | 98.82% |
| Campfire | 34 | 35 | 260 | 4537.02 | 4600.17 | 98.63% |
| Campfire | 30 | 30 | 80 | 7595.81 | 7694.10 | 98.72% |
| CatRobot1 | 42 | 43 | 528 | 975.76 | 987.37 | 98.82% |
| CatRobot1 | 38 | 39 | 528 | 1581.51 | 1604.44 | 98.57% |
| CatRobot1 | 34 | 35 | 528 | 2603.26 | 2618.21 | 99.43% |
| CatRobot1 | 30 | 30 | 72 | 4424.86 | 4426.94 | 99.95% |
| DaylightRoad2 | 44 | 45 | 300 | 634.35 | 639.55 | 99.19% |
| DaylightRoad2 | 40 | 41 | 300 | 1076.03 | 1079.68 | 99.66% |
| DaylightRoad2 | 36 | 37 | 360 | 1820.20 | 1823.12 | 99.84% |
| DaylightRoad2 | 30 | 30 | 60 | 4269.29 | 4325.31 | 98.70% |
| DrivingPOV3 | 46 | 47 | 400 | 553.96 | 559.94 | 98.93% |
| DrivingPOV3 | 42 | 43 | n/a | 994.93 | 1017.72 | 97.76% |
| DrivingPOV3 | 38 | 39 | n/a | 1785.76 | 1776.79 | 100.51% |
| DrivingPOV3 | 33 | 33 | 120 | 3828.42 | 3832.26 | 99.90% |
| Marathon2 | 44 | 44 | 90 | 1186.38 | 1193.16 | 99.43% |
| Marathon2 | 40 | 40 | 60 | 2106.88 | 2114.33 | 99.65% |
| Marathon2 | 36 | 36 | 60 | 3645.30 | 3652.54 | 99.80% |
| Marathon2 | 30 | 30 | 120 | 8562.43 | 8601.63 | 99.54% |
| MountainBay2 | 44 | 44 | 150 | 294.69 | 298.38 | 98.76% |
| MountainBay2 | 40 | 40 | 150 | 576.55 | 576.79 | 99.96% |
| MountainBay2 | 34 | 34 | 180 | 1600.41 | 1606.16 | 99.64% |
| MountainBay2 | 30 | 30 | 210 | 3222.79 | 3246.85 | 99.26% |
| TallBuildings2 | 45 | 45 | 160 | 536.80 | 538.79 | 99.63% |
| TallBuildings2 | 42 | 42 | 160 | 827.01 | 828.38 | 99.83% |
| TallBuildings2 | 37 | 37 | 160 | 1617.76 | 1633.04 | 99.06% |
| TallBuildings2 | 33 | 33 | 152 | 2695.72 | 2766.33 | 97.45% |
| Beatriz | 42 | 42 | 144 | 86.98 | 87.07 | 99.89% |
| Beatriz | 38 | 38 | 174 | 143.62 | 145.18 | 98.92% |
| Beatriz | 34 | 34 | 174 | 250.32 | 250.52 | 99.92% |
| Beatriz | 30 | 30 | 300 | 501.76 | 505.63 | 99.24% |
| BQTerrace | 45 | 45 | 400 | 220.75 | 226.35 | 97.53% |
| BQTerrace | 40 | 40 | 400 | 435.32 | 448.30 | 97.10% |
| BQTerrace | 34 | 34 | 400 | 1013.18 | 1046.14 | 96.85% |
| BQTerrace | 29 | 29 | 400 | 2252.13 | 2275.78 | 98.96% |
| Cactus | 46 | 46 | 200 | 222.74 | 224.57 | 99.19% |
| Cactus | 40 | 40 | 208 | 508.29 | 516.66 | 98.38% |
| Cactus | 34 | 34 | 138 | 1157.84 | 1169.14 | 99.03% |
| Cactus | 30 | 30 | 220 | 2046.83 | 2071.43 | 98.81% |
| DOTA2 | 38 | 38 | 296 | 316.38 | 316.38 | 100.00% |
| DOTA2 | 34 | 34 | 256 | 582.31 | 582.12 | 100.03% |
| DOTA2 | 28 | 28 | 248 | 1600.33 | 1600.65 | 99.98% |
| DOTA2 | 24 | 24 | 272 | 3392.32 | 3390.33 | 100.06% |
| EuroTruckSimulator2 | 44 | 44 | 392 | 428.48 | 428.35 | 100.03% |
| EuroTruckSimulator2 | 38 | 38 | 376 | 1472.25 | 1472.58 | 99.98% |
| EuroTruckSimulator2 | 34 | 34 | 368 | 3333.32 | 3334.17 | 99.97% |
| EuroTruckSimulator2 | 30 | 30 | 368 | 8349.28 | 8339.07 | 100.12% |
| Meridian2 | 42 | 42 | 228 | 71.17 | 71.23 | 99.90% |
| Meridian2 | 36 | 36 | 192 | 145.32 | 146.33 | 99.31% |
| Meridian2 | 30 | 30 | 200 | 350.42 | 351.56 | 99.67% |
| Meridian2 | 26 | 26 | 200 | 651.61 | 657.78 | 99.06% |

All sequences under evaluation are of 10sec length. The original files as well as the bitstreams for the tested rate points are available on the JVET ftp site under /ahg/ECM-test/test/.

## Test method and test design

The Degradation Category Rating (DCR) method was applied for the subjective evaluation [4]. The test sequences were evaluated using the 11-grade scale as specified in Rec. ITU-R BT.500-14 [5], shown in Figure 1 below.

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Figure 1: Meaning of the 11 grades numerical scale as specified in Rec. ITU-R BT.500-14 Table 2-4

Each basic test cell (BTC) is structured as followed:

**“Original” (1sec) – [uncompressed sequence] (10sec) – “A” (1sec) – [PVS] (10sec) – “Vote <N>” (5sec)**

PVS denotes the processed video sequence under evaluation.

### On-site test

A total of 6 test sessions were designed: three for the UHD sequences and three for the HD sequences (jointly for the LD and RA configurations). All test sessions were designed inserting a stabilization phase as suggested in ITU-R BT.500-14. The session duration was chosen to be no longer than 13 minutes (with a maximum of 24 votes) to avoid a fatigue impact. Furthermore, the test sessions included trapping BTCs where the original uncompressed sequence was shown for evaluation.

The participating experts were trained with one training session for UHD resolution (8 votes) and one session for HD resolution (7 votes). All test sequences under evaluation were occurring at least once in the training sessions, and a selection of rate point representing the expected impairment range was presented for both resolutions. Before the presentation of the training sessions, the experts were instructed on the meaning of the impairment scale. Any occurring requests and questions on the test procedure or the scale were answered. The experts were advised to calibrate their personal voting scale during the training sessions and apply it in the actual test sessions.

The UHD and HD test sessions were both presented with a viewing distance of 1.5H of the UHD display for the center seat. The HD sequences were displayed without scaling in the center of the UHD area with a mid-gray padding around them. Thereby, an effective 3H viewing distance for the HD content was achieved.

### Laboratory test

Two resolutions were tested, HD and UHD, using six test sequences for each resolution, coded at four bit rates. This led to a total of 48 test points for HD and 72 test points for UHD. For HD two test sessions were designed, for UHD three test session were designed; each test session included a three BTC (Basic Test Cell) stabilization phase (stabilization data were discarded). The five test sessions were all made of 27 BTC, for a test length of less than 12 minutes. One reference vs. reference test point was included for each session; this allowed a check of the consistency of the test results and viewers behaviour.

Two training sessions were run, one for each resolution. In the raining sessions all the six contents were shown at different compression rates, to provide an overall indication of the impairments. After a general description of the experiment, the degradation scale was carefully explained, providing inputs also during the training sessions. The test was done using the DSIS test protocol as described in Recommendation ITU-R BT 500-14. An 11 levels degradation scale was used; the meaning of each impairment level was explained according to what specified in Rec. 500-14. The viewers were seated at distance form the screen of 1.5H.

# Results and analysis

## Data processing

### On-site test

In one test session, a playout problem with one of the two PCs occurred for one test sequence. The affected experts were presented the missing BTCs in a separate session to complete their votes.

As a first step, the complete set of votes experts was screened for the trapping sequences. In a total of 6 cases, viewers were voting below score of 8 for the original. The results of corresponding session with that trapping sequence were not regarded for the affected viewer. As a second step, the outlier screening according to ITU-R BT.500-14 A1-2.3.1 [REF] was applied. This triggered for one participant whose results were removed from the test set. As a final processing step, isolated outliers -considered to be obvious errors- were removed.

### Laboratory test

A raw data statistical analysis (Pearson correlation) was done to verify the viewers’ behavior; no viewer was discarded. Only few (6 out of 1920 scores) outliers in raw data were discarded.

## Subjective results

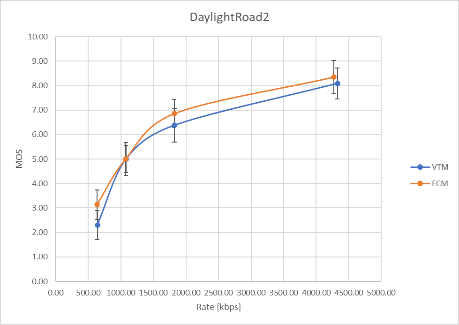
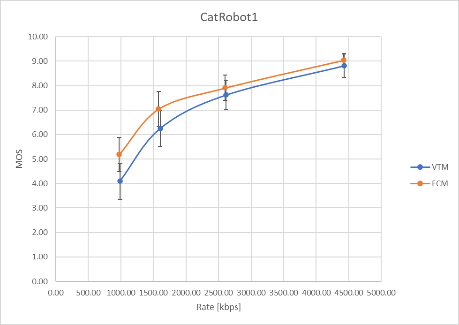
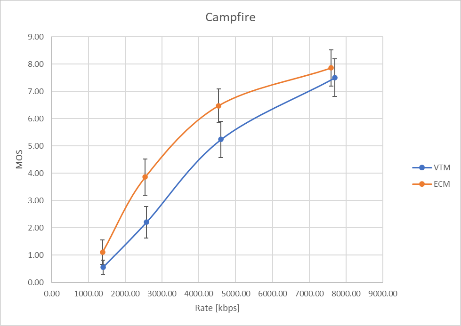
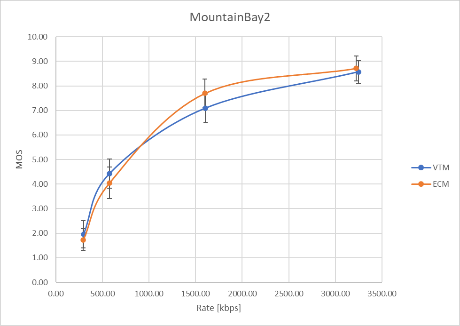
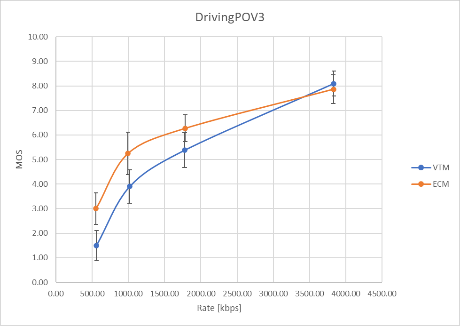
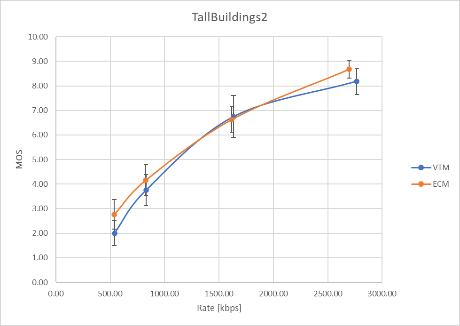
The measured MOS values of the reconstructed video on the 11-grade scale are plotted over the bit rate of the corresponding bitstream. The ±95% confidence intervals for the MOS values are indicated.

The MOS results as well as the bitrates and the PSNR values are provided in the accompanying spreadsheet.

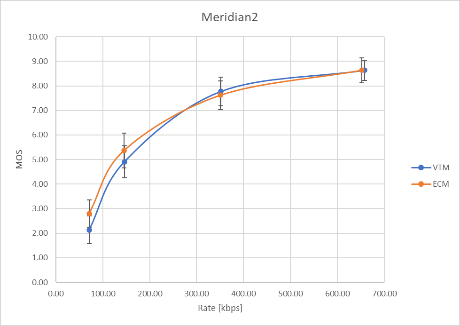
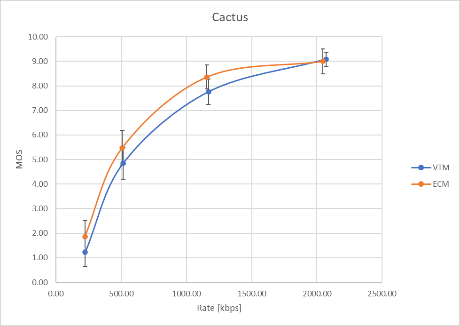
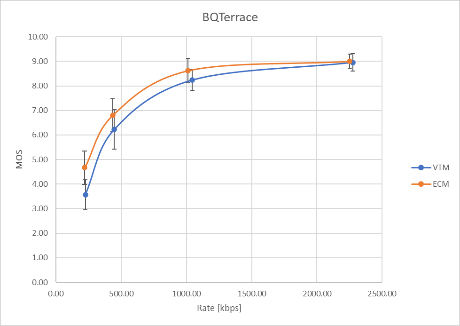
### On-site test

Since the confidence intervals of the expert viewing results overlap for many test points, BD-rate savings are not reported for the MOS-over-rate plots reported below.

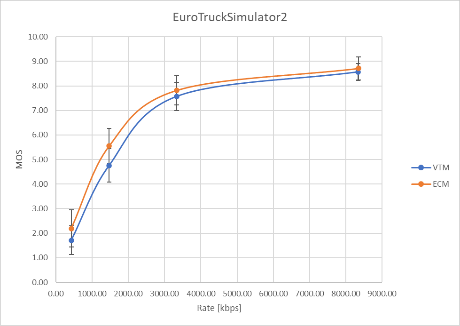
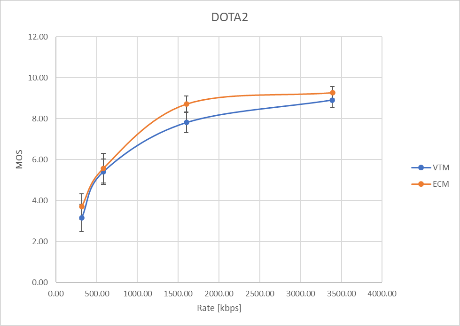
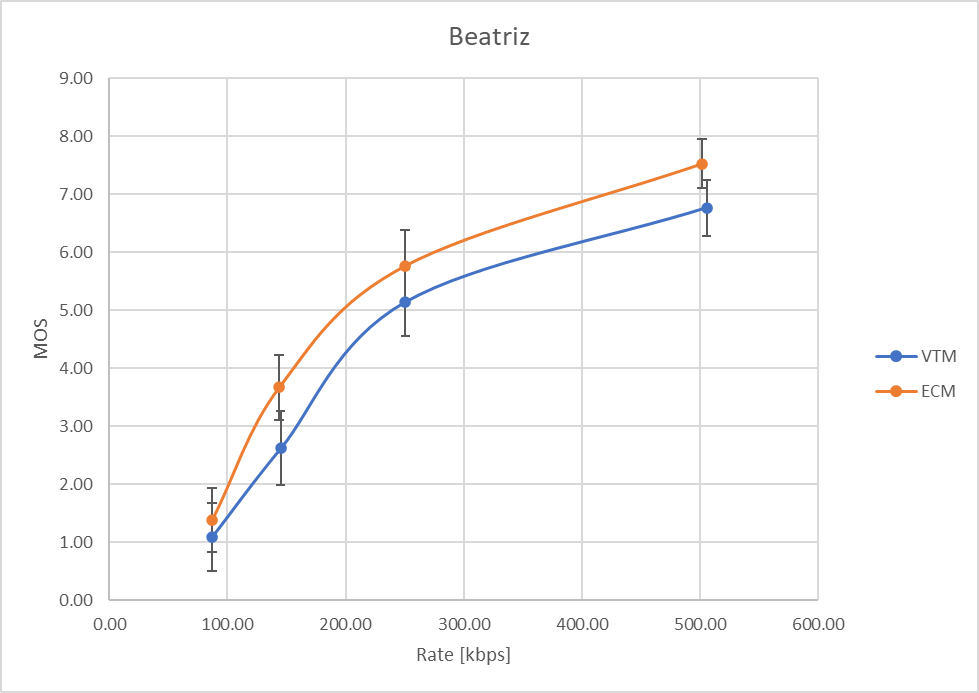
#### UHD-RA

#### HD-RA



#### HD-LD



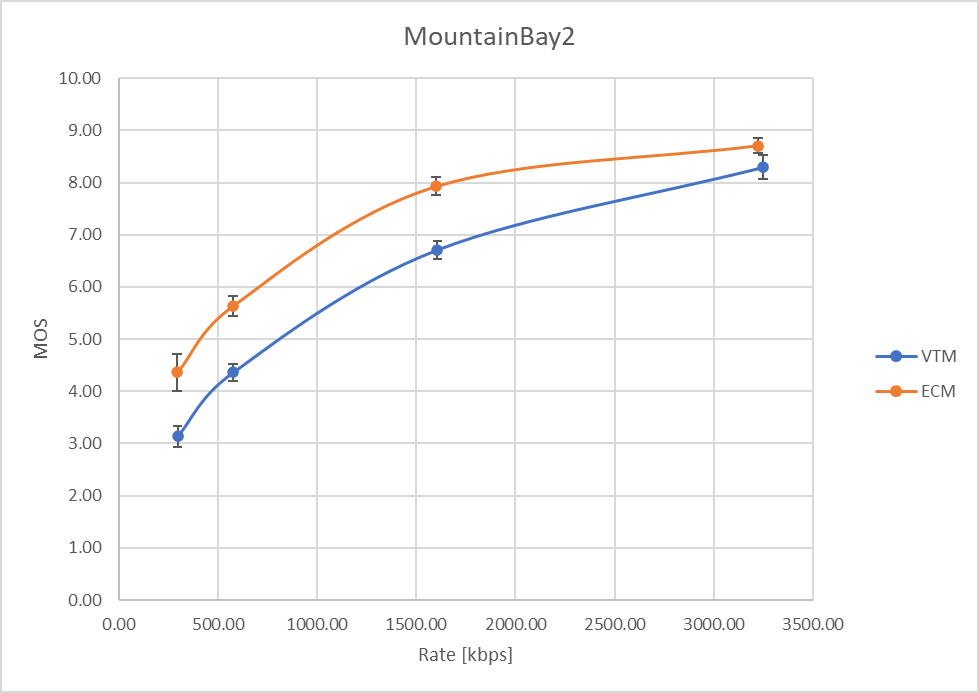
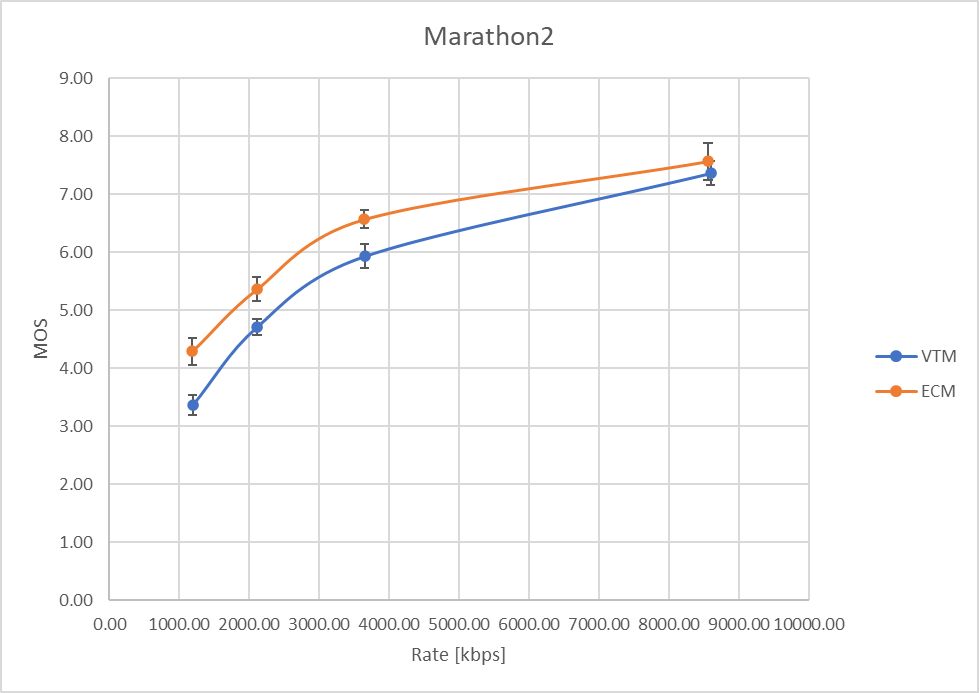
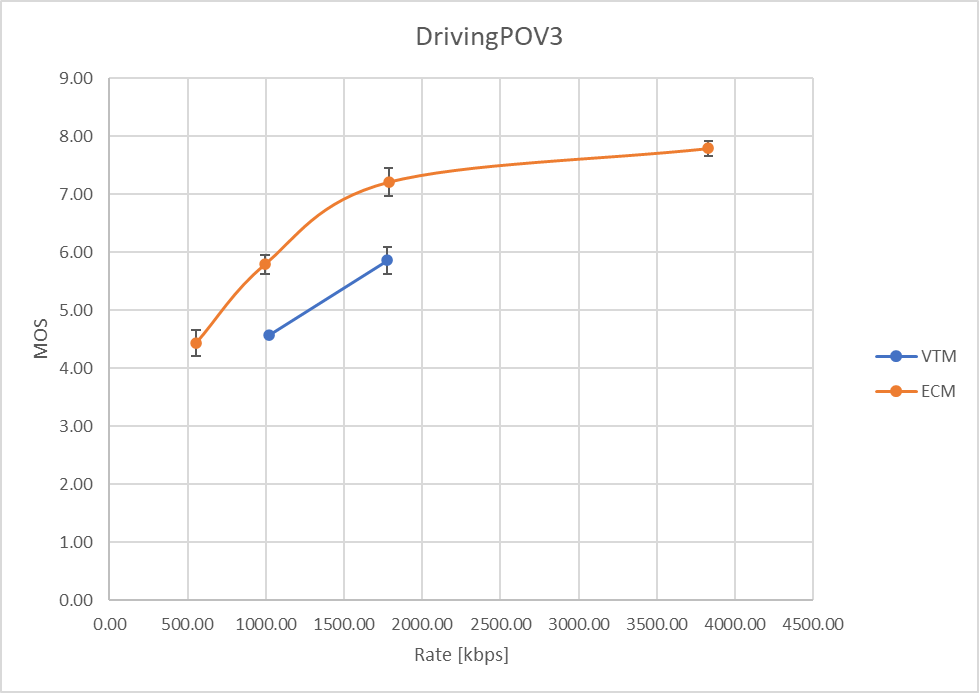
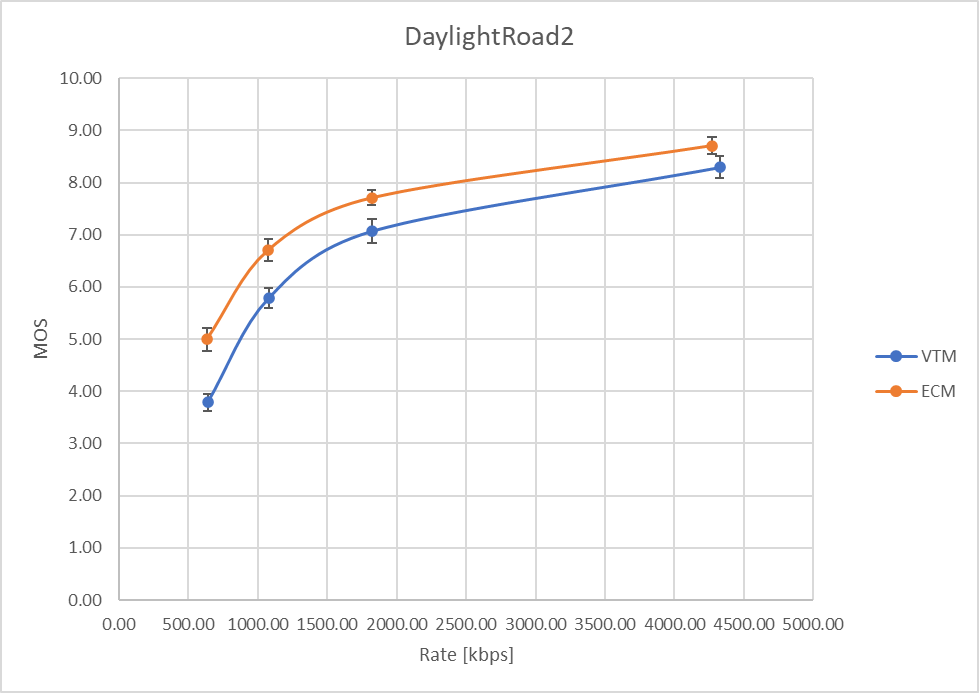
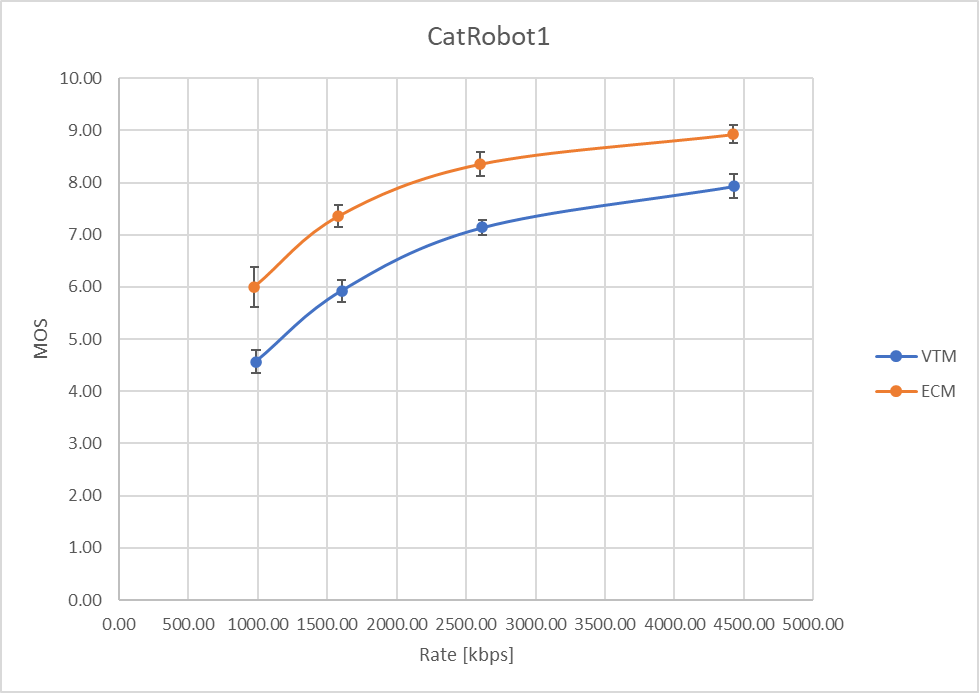
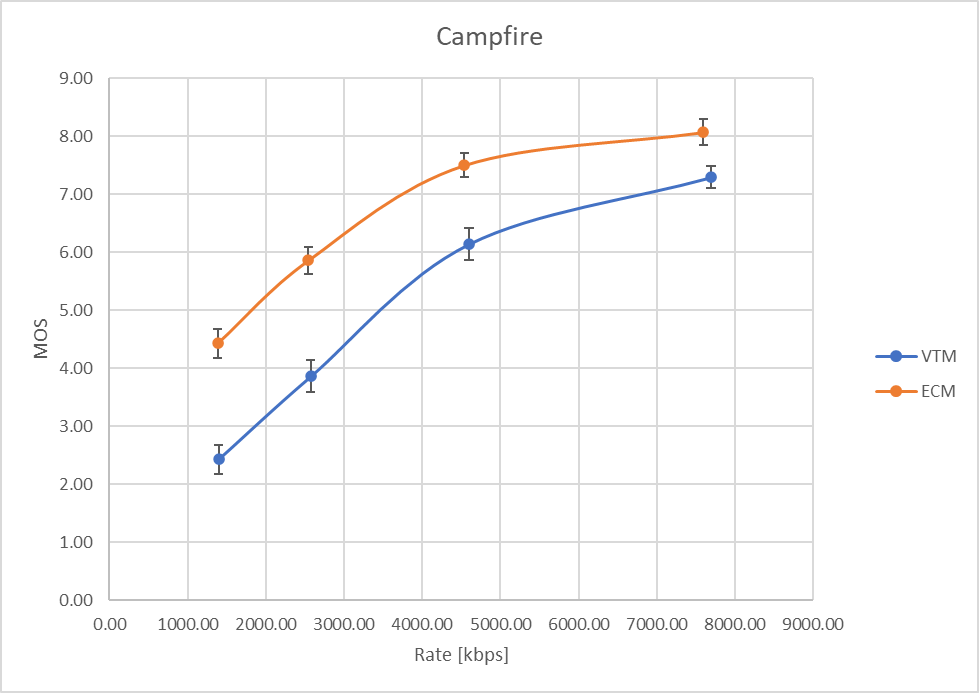
### Laboratory test

Due to logistic constraints, the data for the tests was delivered to the laboratory before the start of the JVET meeting. However, some flaws in delivered bistreams were only found later and new bitstreams were produced which could only be used in the on-site tests. This applied to the VTM streams of the sequences DrivingPOV3, TallBuildings2, and BQTerrace. As a consequence, the missing test points are not reported in the following.

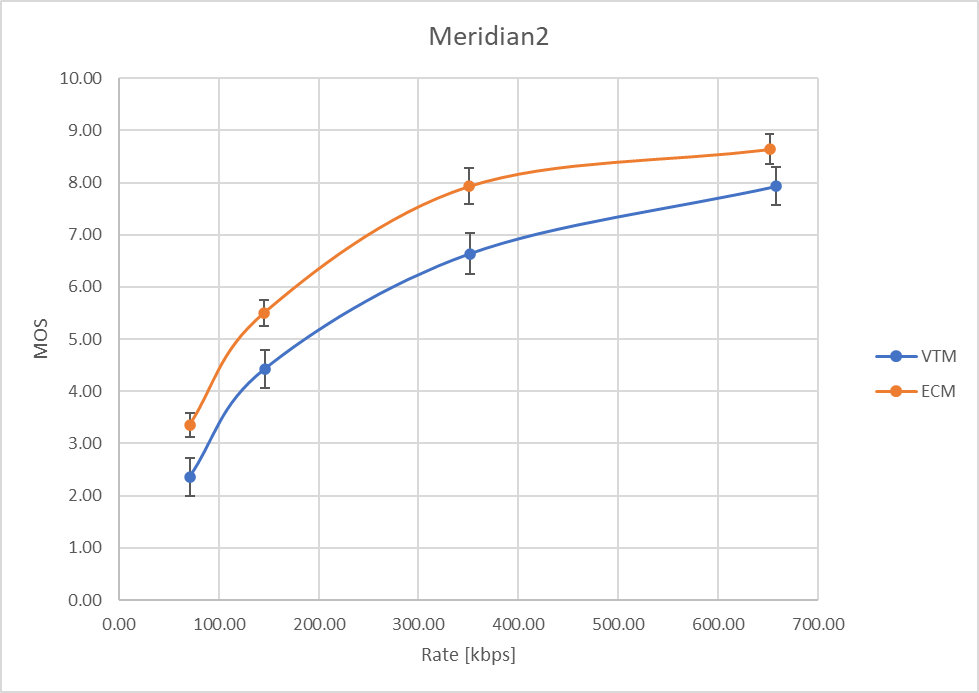
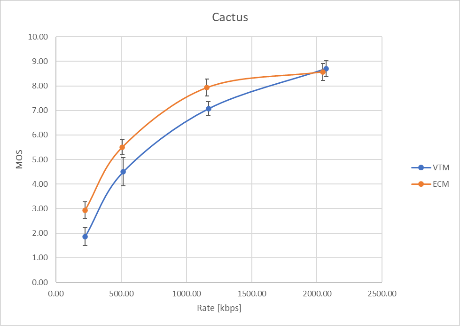
The BD-rate savings for the reported MOS-over-rate points are provided in the table below. The BD numbers were computed using the RDPlot tool [6] on the data provided in the accompanying Excel sheet.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **UHD-RA [%]** | | **HD-RA [%]** | | **HD-LD [%]** | |
| Campfire | -44.45 | Cactus | -27.74 | Beatriz | -28.68 |
| CatRobot1 | -45.44 | Meridian2 | -36.76 | DOTA2 | -32.00 |
| DaylightRoad2 | -30.20 |  | | EuroTruckSimulator2 | -38.82 |
| Marathon2 | -27.23 |  | |
| MountainBay2 | -44.27 |
| **Average** | **-38.32** | **Average** | **-32.25** | **Average** | **-33.17** |

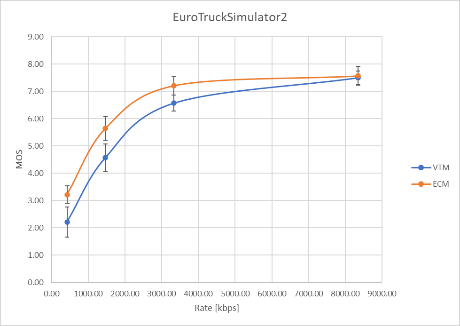
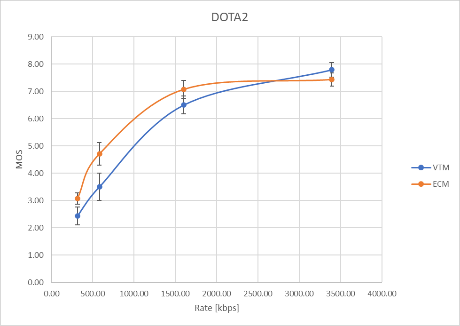
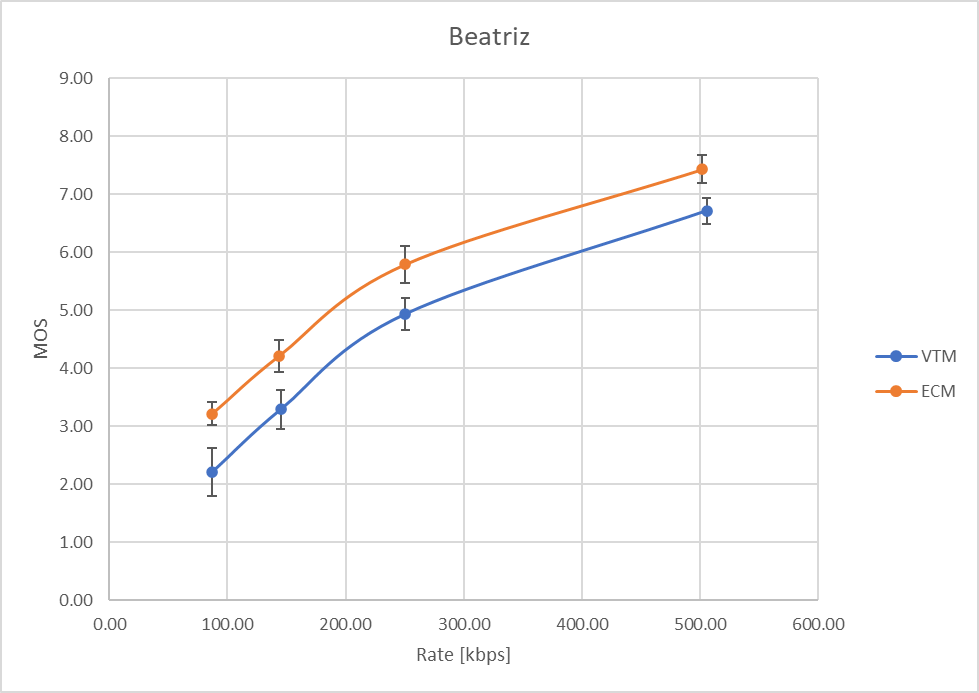
#### UHD-RA



#### HD-RA



#### HD-LD



# Discussion and conclusions

The results of the two tests reveal congruent results for the performance of the ECM and the VTM. Both, the laboratory test conducted with naïve viewers and the on-site test with experts demonstrate a clear visual benefit of the ECM when compared to the VTM for a significant number of cases. For the laboratory results, reported BD rate savings indicate a benefit of about 38% for the UHD test sequences and about 32-33% for the HD test sequences on the given test set.

As already found in previous successful experiments with expert viewing, the MOS scores of the experts show relatively wider confidence intervals. It must be noted that this document reports the first-time experiment with exactly the same content and the same test procedure applied in a laboratory with naïve viewers and on-site with expert viewers. The results confirm the validity of the on-site expert viewing procedure. Give that it would be helpful for decision taking processes to achieve consistent scores with narrow confidence intervals, the results suggest further investigations on improving the expert viewing method in this regard.

# References

1. JVET, VTM software package, <https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tree/VTM-11.0>.
2. JVET, ECM software package, <https://vcgit.hhi.fraunhofer.de/ecm/ECM/-/tree/ECM-6.0>.
3. M. Wien, “AHG4, 7, 12: Report on AHG meetings on ECM performance evaluation preparation,” Doc. JVET-AB0041, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, 28th meeting, Mainz, DE, Oct. 2022.
4. Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.
5. Recommendation ITU-R BT.500-14 (2019), Methodologies for the subjective assessment of the quality of television images.
6. J. Schneider, J. Sauer, M. Wien, “RDPlot - An Evaluation Tool for Video Coding Simulations”, Visual Communications and Image Processing VCIP’21, Munich, Dec. 2021.

# Annex: Expression of thanks

* ECM and VTM encoding and crosscheck:   
  Xiaomi, InterDigital, Bytedance, Qualcomm, Alibaba, Ericsson
* Provision of equipment at meeting site: Fraunhofer HHI
* Laboratory tests: VABTech
* The 24 expert viewers participating in the tests
* A. Wieckowski for support during test preparation