**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION  
ORGANISATION INTERNATIONALE DE NORMALISATION  
ISO/IEC JTC 1/SC 29/AG 5  
MPEG VISUAL QUALITY ASSESSMENT**

**ISO/IEC JTC 1/SC 29/AG 5 N141**

**Nov. 2024, Kemer, TR**

|  |  |
| --- | --- |
| **Title** | **Report on results of expert viewing for dual-layer VVC coding** |
| **Source** | **Test coordinators** |
| **Author** | **Mathias Wien, Adam Wiekowski, Philippe de Lagrange** |
| **Status** | **Output document** |

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

# Abstract

This document reports the results of an on-site expert viewing performed at the 36th JVET meeting in Kemer, TR.

The evaluation targets at the spatial scalability category of the test plan document JVET-AI2021, comparing the compression performance of dual layer (DL) coding with a downscaled base layer at scaling ratio 2 to single layer coding at the low resolution with upscaling to the target resolution using the VVC interpolation filters (UP) as well as single layer coding at the target resolution (SL).

A number of 23 experts participated in the expert viewing. With a comparably low number of valid scores, the results should be considered as an indication of tendencies. They are not regarded as a necessarily conclusive result.

With respect to the goals of this experiment, the following observations are made:

* For all sequences except DrivingPOVLogo, UP is rated higher than SL at least in the lower quality range. This is particularly obvious for TallBuildings, CorsairJoy2, and NxBoxeLogo. This tends to verify that adaptive resolution coding makes sense for VVC.
* For CorsairJoy, the DL approach shows somewhat superior performance compared to SL and UP in the mid-high rate range. At the very low bitrates, UP seems to prevail. At the highest rate, SL starts to be competitive. It is noted that the original and the upscaled BL original show a very similar MOS value. It is concluded that the viewers had difficulties to see a difference between the two versions.
* For DrivingPOVLogo, a clearer difference between the two uncompressed versions of the sequence is observed. The single layer curve shows best performance for all but the lowest rate point. DL performs second in the mid-low rate range but is outperformed by UP at the mid-high and highest rate points. It seems that the visual improvements observed for these points come at a (too) high bitrate cost in the investigated configuration.
* The NxBoxeLogo sequence seems to have been difficult to assess, since all scores are relatively high. The sequence shows a typical behavior with UP being best at lowest rates, then DL taking over, followed by SL. It is noted that the curves converge to the saturation range at or after the second-highest rate point. It is further noted that the lower points of this sequence have been rated higher than those reported in JVET-AF311 in Hannover. This might be attributed to the different viewing conditions at the two meetings.
* For Procession, it is noted that the upscaled BL original appears to be very close to transparency. Accordingly, the UP and SL versions perform very similarly. The DL curve shows competitive performance at the lower rate range.
* TallBuildings2 shows an interesting behavior since here, the UP variant seems to be quite clearly superior to SL. DL slightly outperforms SL for all but the highest rate point. The performance of the UP configuration might indicate that still, there might be an issue with the overall configuration of the reference software.

It is recommended to take these results into account by reconsidering the test setup with respect to bitrate ranges to be tested for certain resolutions and encoder configurations.

The impact of the changes from AJ0225 and AJ0067 may be assessed by comparing the scores reported above to those reported in JVET-AF0311. The following observations are made:

* For CorsairJoy: a clear improvement on the low rate point tested in Hannover (QP 34+41). DL is now above UP (so as the next higher point) whereas it was clearly below in JVET-AF0311. The order (from better to worst) is now DL/UP/SL instead of UP/DL/SL.
* For DrivingPOVLogo: the QPs are slightly different. A slight improvement for DL (e.g. the 2rd lowest rate point (lowest in JVET-AF0311)).
* For NxBoxeLogo: the 3rd rate point is improved for DL
* For TallBuildings: The curves show a similar tendency, otherwise difficult to compare

# Introduction

This document reports the results of an on-site expert viewing performed at the 36th JVET meeting in Kemer, TR. The evaluation targets at the spatial scalability category of the test plan document JVET-AI2021 [1], comparing the compression performance of dual layer (DL) coding with a downscaled base layer at scaling ratio 2 to single layer coding at the low resolution with upscaling to the target resolution using the VVC interpolation filters (UP) as well as single layer coding at the target resolution (SL).

A similar test has been reported in JVET-AF0311 [2]. A main motivation for this viewing test was to assess the effect of the changes proposed in JVET-AJ0255 and JVET-AJ0067 [3][4]:

* Modifications to the RDO for the enhancement layer, related to the selection of inter-layer coding modes, attempting to avoid visual degradation compared to an upscaled base layer.
* Configuration changes for enhancement layer encoding: keeping the inter-layer reference in L0 only, giving more weight to intra picture, and increasing deblocking

# Test setup

## Logistics

The setup at the meeting site 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)** | TCL 85C855, 2×TCL 75C855 (all miniLED); VESTEL 55” OLED, HDMI (3840×2160), 10bit input |
| **Viewing distance** | Viewers sitting at 1.5H (4 viewers for the 85” display, 3 viewers for the 75” displays, 2 viewers for the 55” display) |
| **Viewing angle** | ±75°, 90° (at screen center) |
| **Total number of viewers** | 23 (8 female, 15 male) |

Participants confirmed visual acuity and normal colour vision.

The display settings were browsed, and any enhancement features were switched off. For one TCL 75C855 and the VESTEL 55” displays, an internal frame rate up-conversion was switched on in the test. On the VESTEL display no disabling option was found.

## Test sequences and quantizer settings

For evaluation, a subset of test sequences and quantizer settings listed in JVET-AJ0042 were tested. Five UHD sequences were tested, including the test sequences CorsaireJoy2, TallBuildings2, DivingPOVLogo and NxBoxeLogo which have been assessed in JVET-AF0311 [2]. Additionally, the test sequence procession was included in the test set. Three coding configurations were tested: Single layer (SL) at full resolution, dual layer (DL) with an HD base layer (2× spatial scalability), and upscaled base resolution (UP) where the base layer of the DL streams was upscaled using the VVC upscaling filters. As an additional measurement point, the original test sequences, downscaled using the SHVC downscaling process and re-upscaled using the VVC upscaling filters, were evaluated (origUP). The QP values evaluated in the expert viewing are shown in the table below.

Table – Selected test sequences and QP settings for the viewing.

| **Seq** | **SL QP** | **BL UP QP** | **DL QPs** |
| --- | --- | --- | --- |
| CorsairJoy2 | 33 | 27 | 27, 34 |
| 37 | 31 | 31, 38 |
| 40 | 34 | 34, 41 |
| 43 | 37 | 37, 44 |
| TallBuildings2 | 34 | 28 | 28, 38 |
| 38 | 32 | 32, 42 |
| 42 | 36 | 36, 46 |
| 46 | 40 | 40, 50 |
| DrivingPOV3Logo | 32 | 26 | 26, 33 |
| 37 | 31 | 31, 38 |
| 42 | 36 | 36, 43 |
| 46 | 40 | 40, 47 |
| BoxeLogo | 30 | 24 | 24, 31 |
| 34 | 30 | 30, 37 |
| 38 | 32 | 32, 39 |
| 42 | 37 | 37, 44 |
| Procession | 35 | 29 | 29, 35 |
| 40 | 34 | 34, 40 |
| 44 | 38 | 38, 44 |
| 47 | 41 | 41, 47 |

Compared to previous tests in JVET-AF0311, the enhancement layer QP was kept to BL QP + 7 for most cases, except for:

* TallBuildings, trying + 10, as for this sequence, the EL bitrate is much higher than for e.g. CorsairJoy2
* Procession, trying +6, as for this sequence, the EL bitrate is lower

### Rate distortion plots for the selected test points

Figure 1 shows the RD-curves for the upscaled base-layer, single-layer full resolution and dual layer (dual resolution) encodings, respectively.

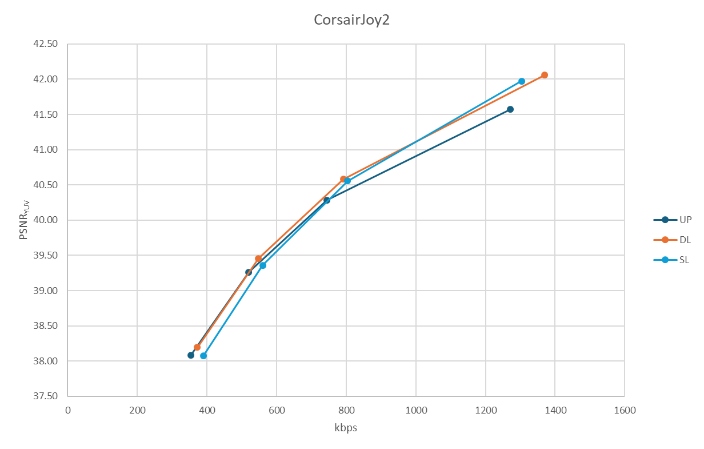
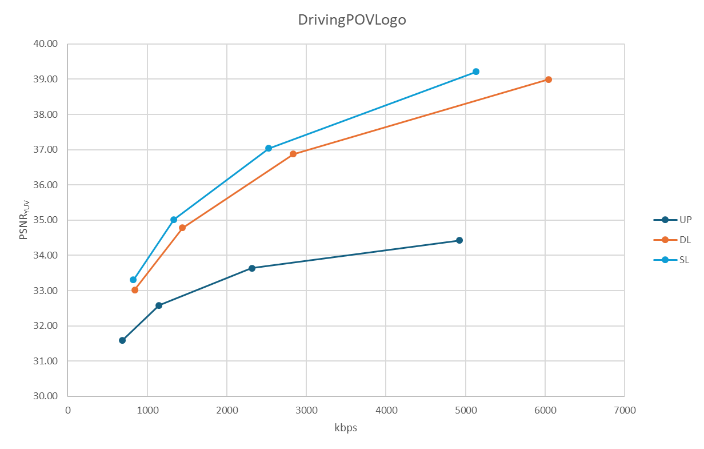
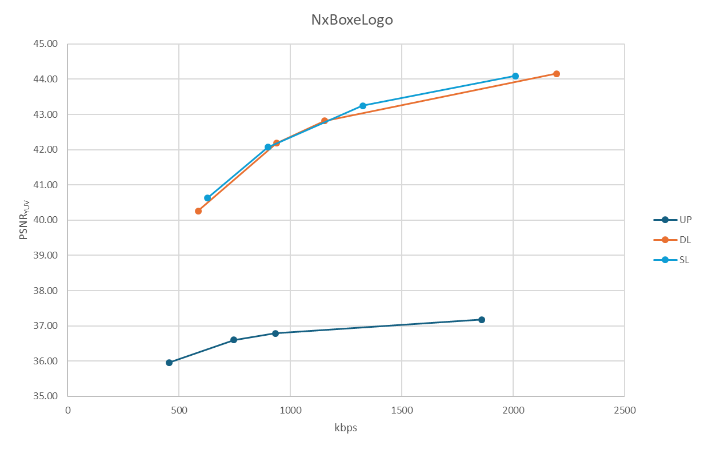
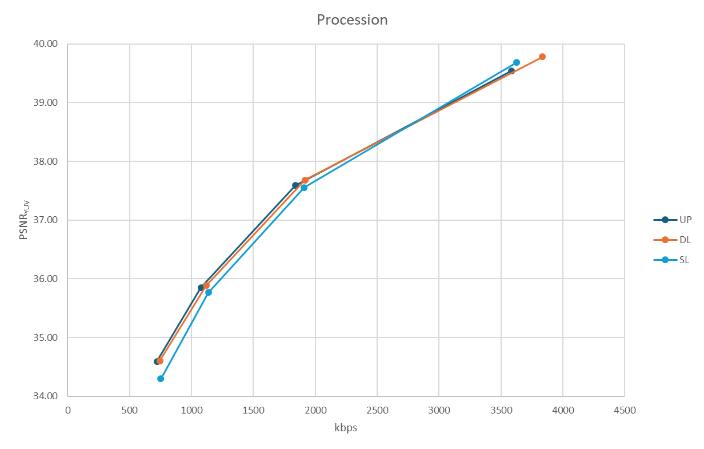
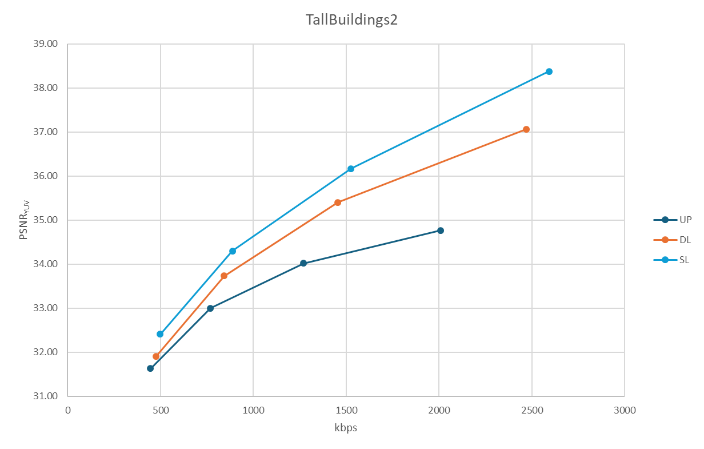
    

Figure 1: PSNR-over-rate plots for the PVSs of the experiment

UP = upscaled single layer coded at base resolution, DL = Dual layer, SL = Single layer at full resolution

The PSNR-over-rate plots in Figure 1 are very consistent with the RD-curves presented in JVET-AF0186. The BD-rate improvement of around 0.4% reported by JVET-AJ0255 is too small to be visible on the curves.

## Test method and test design

The assessment was performed with the reconstructed UHD output. Two DCR test session with 37 BTCs were presented to the viewers. The duration of each test session was about 19 min.

### Degradation category rating

#### Session setup

The test sequences were evaluated using an 11-grade scale shown in Figure 2 below.

Ein Bild, das Text, Screenshot, Schrift, Zahl enthält.

Automatisch generierte Beschreibung

Figure 2: Meaning of the used 11 grade numerical scale

The test sessions included 37 basic test cells (BTCs) showing the original uncompressed video sequence and one processed video sequence. Each BTC was structured as followed (PVS denotes the processed video sequence under evaluation):

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

The session included a stabilization phase of three BTCs, trapping BTCs where original full resolution test sequences were to be scored in the PVS, as well as the uncompressed upscaled base layer resolution test sequences.

#### Viewer information and training

The viewers were explained the viewing task, the organization of the session and the meaning of the grading scale. The viewers were organized into two groups of 11 and 12 participants, respectively. Training, viewing of session 1 and viewing of session 2 were done in an interleaved fashion, where one group performed the viewing tasks at the four displays while the other group had a break. During the viewing sessions, coring was done on paper sheets. The breaks were used to enter the acquired scores into the test database.

For training, the participants were shown a test session of 12 BTCs including representative examples of the impairments shown in the tests.

# Subjective results and analysis

## Data processing

As a first step, the complete set of votes experts was screened for the trapping sequences. No viewers were excluded on this basis. The Pearson correlation coefficient of the viewers scores relative to the MOS was assessed. The scores of two viewers were removed based on a PCC threshold of 0.75%. As a last step, the non-normalized -scores of the viewer scores were compared to a threshold of 2.5, where is the score of viewer for PVS , is the mean opinion score for that PVS for the viewers. Based on this threshold, 7.6% of the scores were removed from the dataset.

The tests might be influenced by the non-disabled post-processing setting on two of the displays, as reported above. To visualize the impact of the error, curves for MOS values using only unaffected datapoints are plotted in the Annex. The results and discussion below are based on the full set of scores.

## 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 uncompressed original was only assessed for CorsairJoy2, DrivingPOVLogo, and NxBoxeLogo. For the other two sequences, Procession and TallBuildings2, this data is not reported.

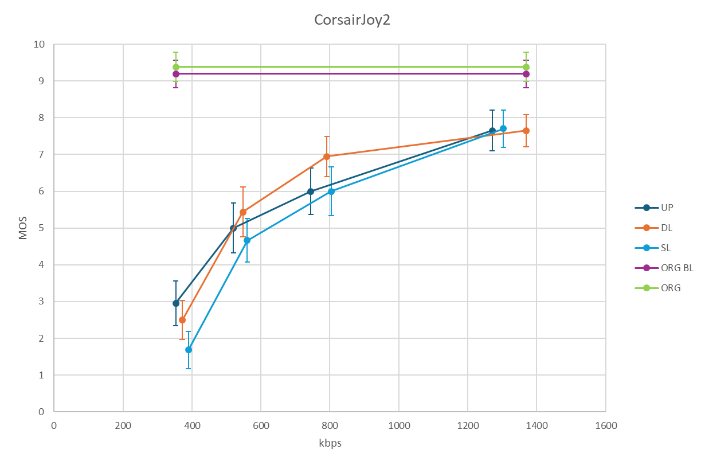
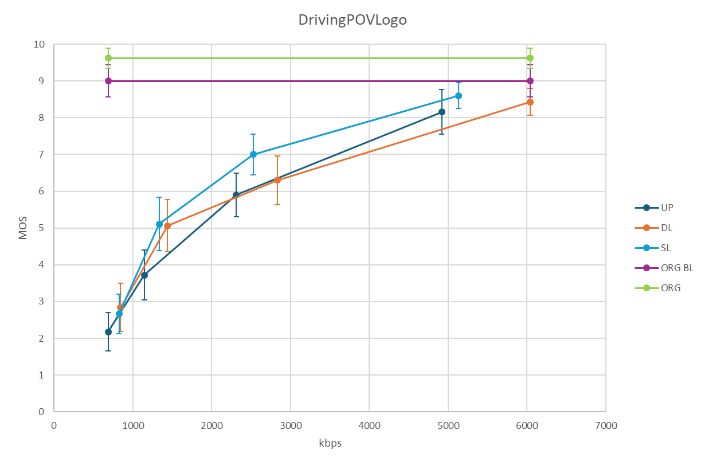
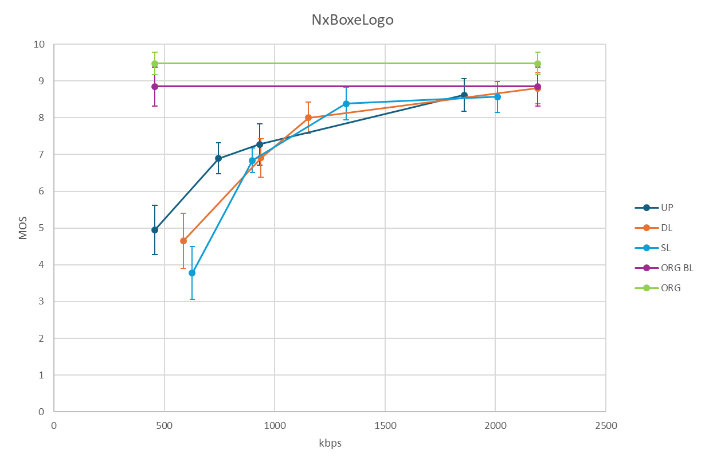
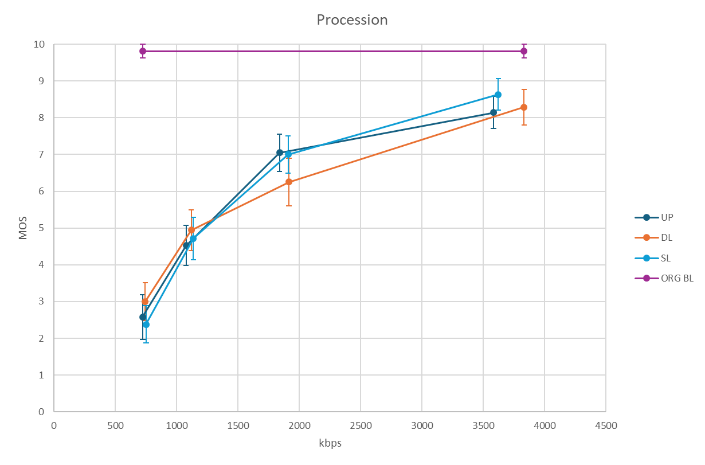
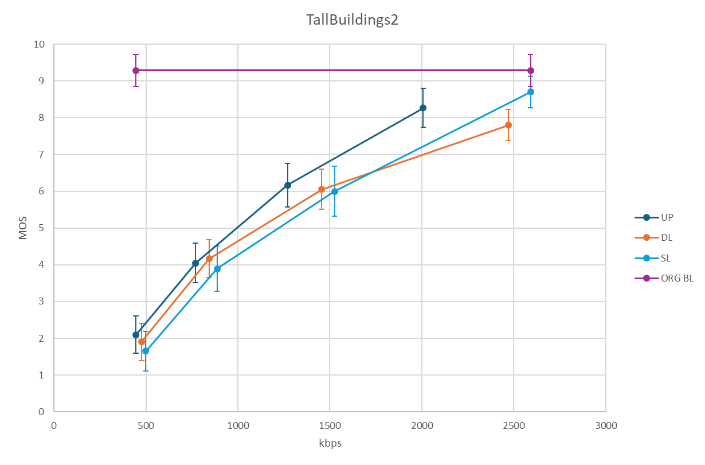
    

Figure 3: MOS-over-rate plots for the DCR experiment

UP = upscaled single layer coded at base resolution, DL = Dual layer, SL = Single layer at full resolution, ORG BL = upscaled half-resolution original, ORG = full resolution original

# Discussion

## Results of the experiment

With respect to the goals of this experiment, the following observations are made:

* For all sequences except DrivingPOVLogo, UP is rated higher than SL at least in the lower quality range. This is particularly obvious for TallBuildings, CorsairJoy2, and NxBoxeLogo. This tends to verify that adaptive resolution coding makes sense for VVC.
* For CorsairJoy, the DL approach shows somewhat superior performance compared to SL and UP in the mid-high rate range. At the very low bitrates, UP seems to prevail. At the highest rate, SL starts to be competitive. It is noted that the original and the upscaled BL original show a very similar MOS value. It is concluded that the viewers had difficulties to see a difference between the two versions.
* For DrivingPOVLogo, a clearer difference between the two uncompressed versions of the sequence is observed. The single layer curve shows best performance for all but the lowest rate point. DL performs second in the mid-low rate range but is outperformed by UP at the mid-high and highest rate points. It seems that the visual improvements observed for these points come at a (too) high bitrate cost in the investigated configuration.
* The NxBoxeLogo sequence seems to have been difficult to assess, since all scores are relatively high. The sequence shows a typical behavior with UP being best at lowest rates, then DL taking over, followed by SL. It is noted that the curves converge to the saturation range at or after the second-highest rate point. It is further noted that the lower points of this sequence have been rated higher than those reported in JVET-AF311 in Hannover. This might be attributed to the different viewing conditions at the two meetings.
* For Procession, it is noted that the upscaled BL original appears to be very close to transparency. Accordingly, the UP and SL versions perform very similarly. The DL curve shows competitive performance at the lower rate range.
* TallBuildings2 shows an interesting behavior since here, the UP variant seems to be quite clearly superior to SL. DL slightly outperforms SL for all but the highest rate point. The performance of the UP configuration might indicate that still, there might be an issue with the overall configuration of the reference software.

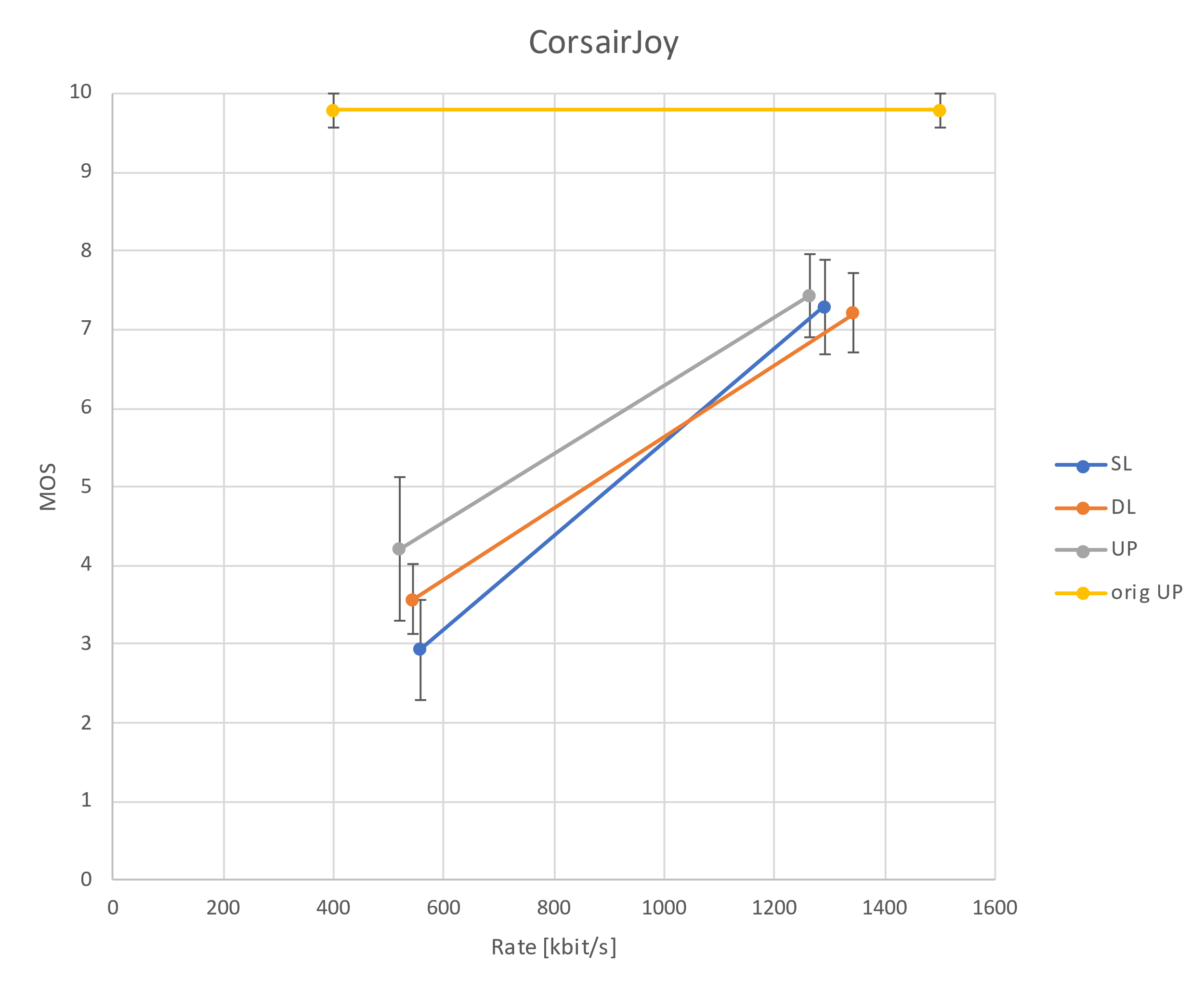
It is recommended to take these results into account by reconsidering the test setup with respect to bitrate ranges to be tested for certain resolutions and encoder configurations.

## Comparison of the results to the Hannover test

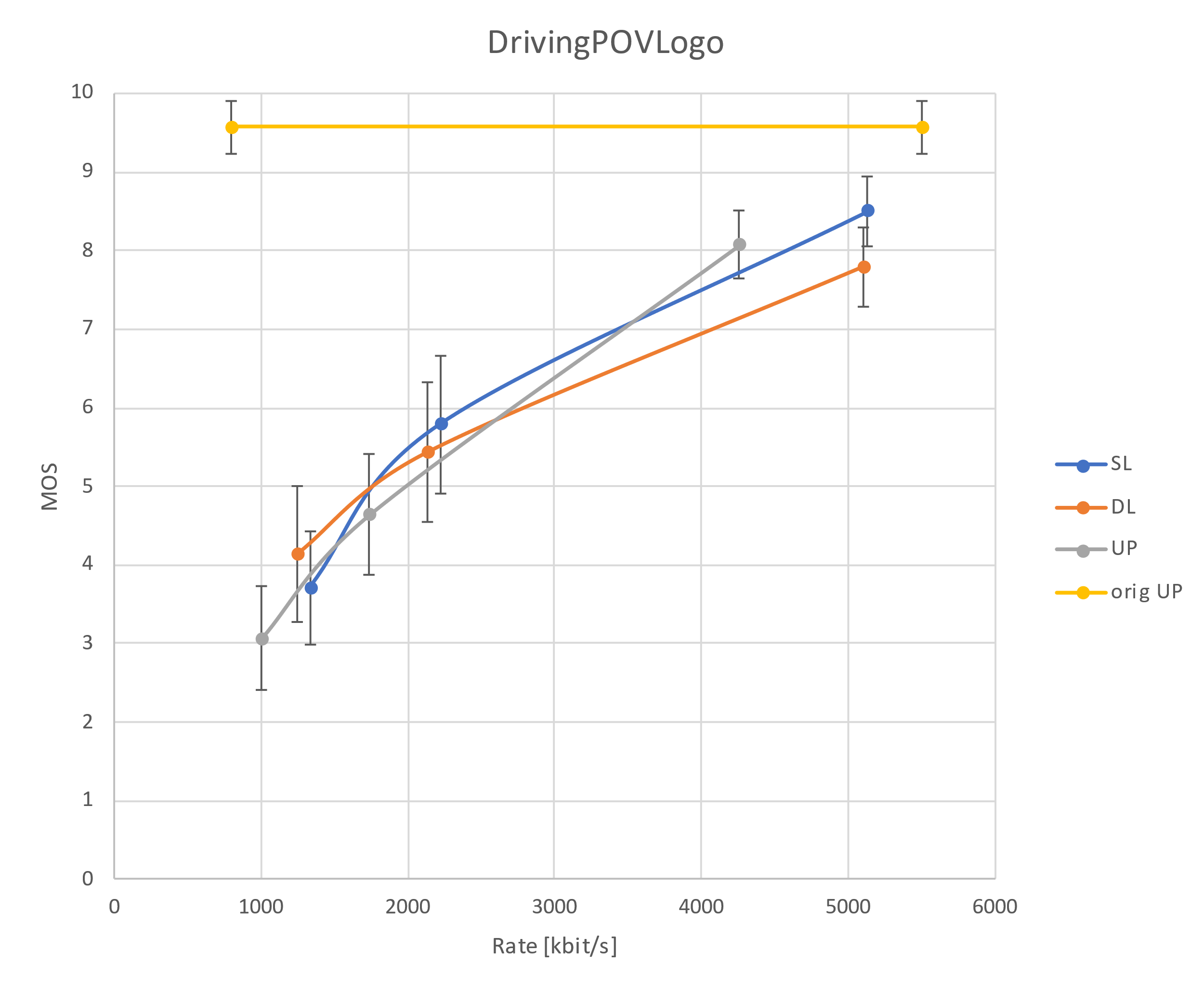
The impact of the changes from AJ0225 and AJ0067 may be assessed by comparing the scores reported above to those reported in JVET-AF0311. The following observations are made:

* For CorsairJoy: a clear improvement on the low rate point tested in Hannover (QP 34+41). DL is now above UP (so as the next higher point) whereas it was clearly below in JVET-AF0311. The order (from better to worst) is now DL/UP/SL instead of UP/DL/SL.
* For DrivingPOVLogo: the QPs are slightly different. A slight improvement for DL (e.g. the 2rd lowest rate point (lowest in JVET-AF0311)).
* For NxBoxeLogo: the 3rd rate point is improved for DL
* For TallBuildings: The curves show a similar tendency, otherwise difficult to compare

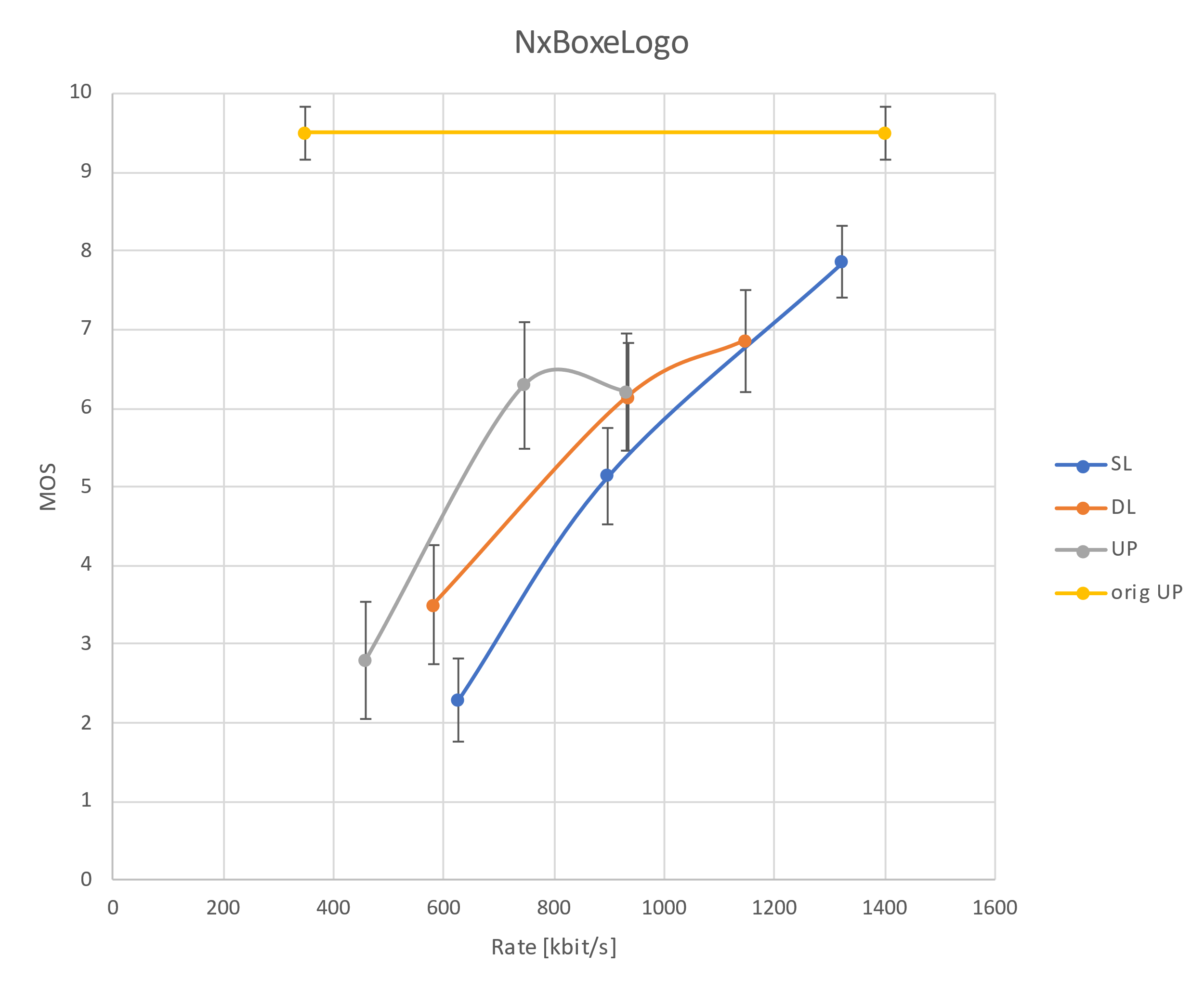
A graph with colored lines and numbers

Description automatically generated 

A graph with colored lines and numbers

Description automatically generated

A graph with colored lines and numbers

Description automatically generated

A graph of a graph with colored lines

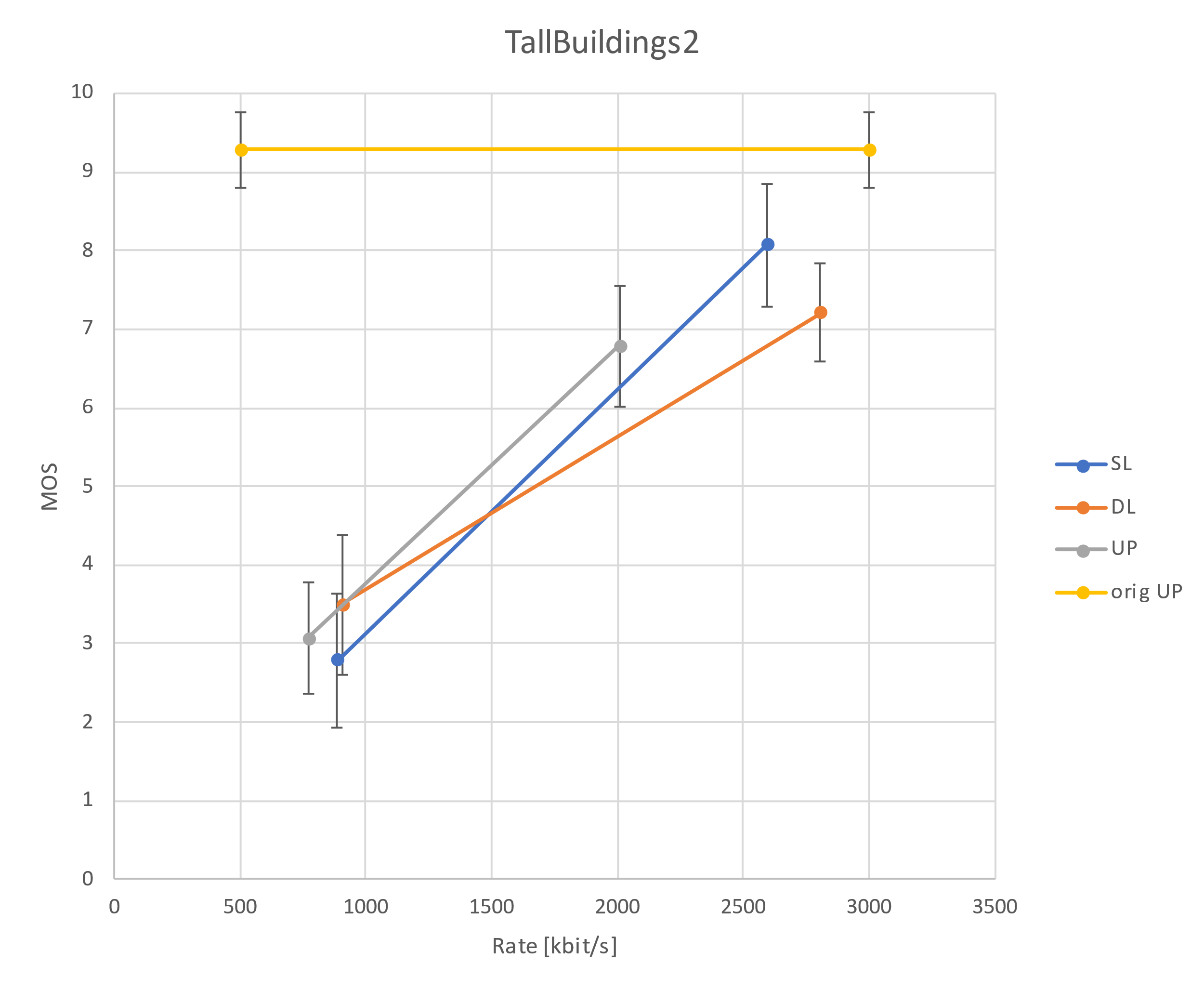
Description automatically generated with medium confidence

Figure 4: MOS-over-rate plots comparison of new results (left) with those from JVET-AF0311 (right)

## Visual examples

The impact of the enhancement layer on top of UP in terms of detail enhancement may not have been reflected in visual scores, as e.g. in the second lower quality point of NxBoxeLogo (QP 32+39, where the same score is obtained for UP and DL despite a significant bitrate in EL), or the two highest quality point of TallBuildings.

It could be studied whether this is related to other artefacts that justify the score given to UP, and are still present in DL, with detail enhancement being much less visually prominent and not a sufficient reason to change the score, despite being real as shown is Figure 5 and Figure 6. Or if there are still artefacts created by EL that would be reflected in the score. Or if consistency of quality across the whole picture is key.

A close up of a black object

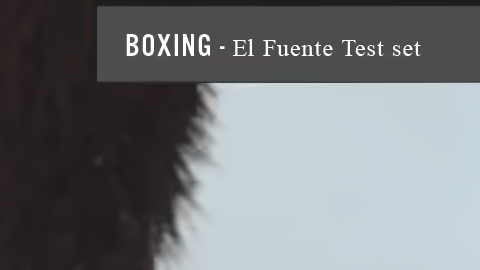
Description automatically generated 

Figure 5: Crops of NxBoxeLogo frame #97, at QP 32+39. Left=BL, right=EL  
(static fine text banners)

Figure 6: Crops of TallBuildings frame #97, at QP 32+42. Left=BL, right=EL  
A(slow camera motion = moving details)

# Appreciation

TCL and VESTEL are gratefully thanked for providing the displays for viewing. Fraunhofer HHI is thanked for providing the playout PC. The participating experts are gratefully thanked for their evaluation efforts.

# References

1. O. Chubach, P. de Lagrange, M. Wien, “Verification test plan for VVC multilayer coding (update 4),” Doc. JVET-AI2021, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, 35rd meeting, Sapporo, JP, Jul. 2024.
2. M. Wien, “Results of expert viewing at AF meeting for the spatial scalability category of the VVC multilayer VT,” Doc. JVET-AF0311, Joint Video Experts Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 5, 32nd meeting, Hannover, DE, Oct. 2023.
3. M. Le Pendu, P. de Lagrange, F. Urban, “AHG3/AHG10: RDO adjustments for inter-layer predictions,” Doc. JVET-AJ0225, 36th meeting, Kemer, Nov. 2024.
4. K. Andresson, “AHG10: Suggested updates of settings for scalable VVC coding,” Doc. JVET-AJ0067, 36th meeting, Kemer, Nov. 2024.
5. P. de Lagrange, F. Urban, “AHG4: experiments related to VVC spatial scalability visual testing,” Doc. JVET-AF0186, 32nd meeting, Hannover, Oct. 2023.
6. P. de Lagrange, “AHG10: Teleconference on encoder optimization for multi-layer coding,” Doc. JVET-AJ0042, 36th meeting, Kemer, Nov. 2024.
7. M. Le Pendu, P. de Lagrange, F. Urban, “AHG3/AHG10: RDO adjustments for inter-layer predictions,” Doc. JVET-AJ0225, 36th meeting, Kemer, Nov. 2024.
8. Recommendation ITU-T P.910 (2008), *Subjective video quality assessment methods for multimedia applications*.
9. Recommendation ITU-R BT.500-14 (2019), Methodologies for the subjective assessment of the quality of television images.

# Annex: MOS data

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Single-layer** |  |  |  | **Dual-layer** |  |  |  | **BL-upscaled** |  |  |  |
| **Seq** | **QP** | **Rate** | **MOS** | **CI** | **QP** | **Rate** | **MOS** | **CI** | **QP** | **Rate** | **MOS** | **CI** |
| CorsairJoy2 | qp33 | 1303.58 | 7.70 | 0.51 | qp27 | 1369.46 | 7.65 | 0.44 | qp27 | 1270.70 | 7.65 | 0.55 |
| CorsairJoy2 | qp37 | 804.13 | 6.00 | 0.66 | qp31 | 791.45 | 6.94 | 0.55 | qp31 | 745.22 | 6.00 | 0.63 |
| CorsairJoy2 | qp40 | 560.54 | 4.67 | 0.59 | qp34 | 548.09 | 5.44 | 0.67 | qp34 | 520.43 | 5.00 | 0.67 |
| CorsairJoy2 | qp43 | 389.51 | 1.68 | 0.51 | qp37 | 372.07 | 2.50 | 0.54 | qp37 | 353.29 | 2.95 | 0.60 |
| DrivingPOVLogo | qp32 | 5133.62 | 8.60 | 0.35 | qp26 | 6042.96 | 8.43 | 0.37 | qp26 | 4920.79 | 8.16 | 0.61 |
| DrivingPOVLogo | qp37 | 2527.17 | 7.00 | 0.55 | qp31 | 2834.07 | 6.30 | 0.66 | qp31 | 2313.65 | 5.90 | 0.59 |
| DrivingPOVLogo | qp42 | 1333.51 | 5.11 | 0.72 | qp36 | 1441.61 | 5.06 | 0.71 | qp36 | 1147.35 | 3.72 | 0.68 |
| DrivingPOVLogo | qp46 | 824.47 | 2.67 | 0.54 | qp40 | 843.88 | 2.84 | 0.65 | qp40 | 687.17 | 2.18 | 0.52 |
| NxBoxeLogo | qp30 | 2009.27 | 8.57 | 0.42 | qp24 | 2192.88 | 8.80 | 0.42 | qp24 | 1858.63 | 8.62 | 0.44 |
| NxBoxeLogo | qp34 | 1323.84 | 8.38 | 0.44 | qp30 | 1152.75 | 8.00 | 0.43 | qp30 | 931.65 | 7.28 | 0.56 |
| NxBoxeLogo | qp38 | 899.03 | 6.84 | 0.33 | qp32 | 937.41 | 6.90 | 0.52 | qp32 | 745.55 | 6.89 | 0.42 |
| NxBoxeLogo | qp42 | 626.43 | 3.78 | 0.71 | qp37 | 585.22 | 4.65 | 0.75 | qp37 | 456.66 | 4.95 | 0.67 |
| Procession | qp35 | 3623.72 | 8.63 | 0.43 | qp29 | 3834.45 | 8.29 | 0.48 | qp29 | 3585.65 | 8.14 | 0.44 |
| Procession | qp40 | 1909.60 | 7.00 | 0.50 | qp34 | 1916.26 | 6.25 | 0.64 | qp34 | 1839.38 | 7.05 | 0.51 |
| Procession | qp44 | 1136.71 | 4.71 | 0.58 | qp38 | 1119.73 | 4.94 | 0.55 | qp38 | 1077.40 | 4.52 | 0.55 |
| Procession | qp47 | 751.52 | 2.38 | 0.51 | qp41 | 744.43 | 3.00 | 0.51 | qp41 | 722.86 | 2.58 | 0.61 |
| TallBuildings2 | qp34 | 2591.27 | 8.70 | 0.43 | qp28 | 2470.21 | 7.80 | 0.42 | qp28 | 2008.17 | 8.26 | 0.53 |
| TallBuildings2 | qp38 | 1525.45 | 6.00 | 0.68 | qp32 | 1455.50 | 6.05 | 0.54 | qp32 | 1270.29 | 6.17 | 0.60 |
| TallBuildings2 | qp42 | 886.94 | 3.89 | 0.62 | qp36 | 842.99 | 4.17 | 0.52 | qp36 | 769.42 | 4.05 | 0.54 |
| TallBuildings2 | qp46 | 499.01 | 1.65 | 0.53 | qp40 | 476.16 | 1.90 | 0.50 | qp40 | 445.49 | 2.10 | 0.50 |

# Annex: Results for correctly calibrated displays only

Figure 7 shows the MOS scores from test, analogue to Figure 3, but only for the subjects participating in from of correctly configured displays.

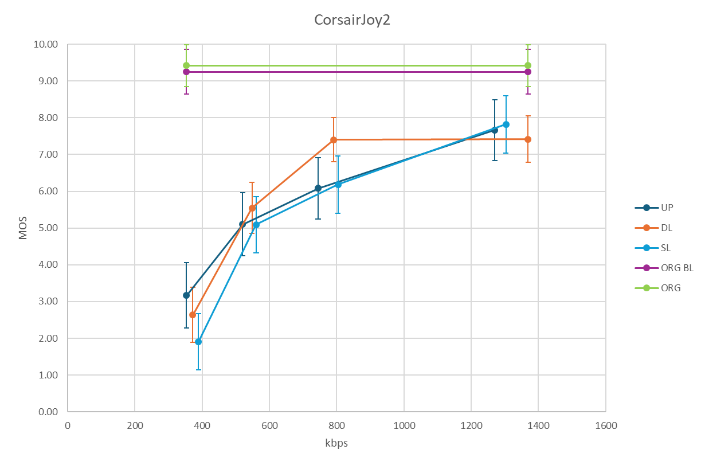
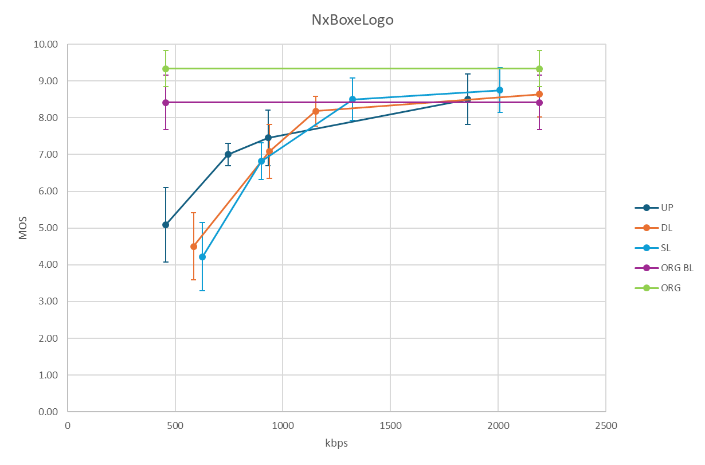
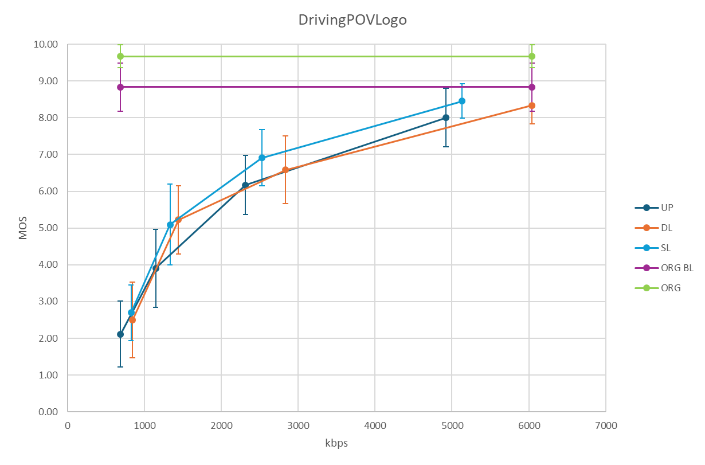
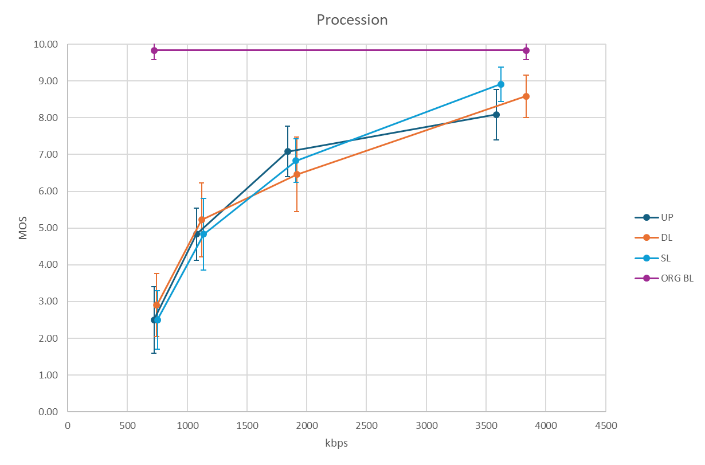
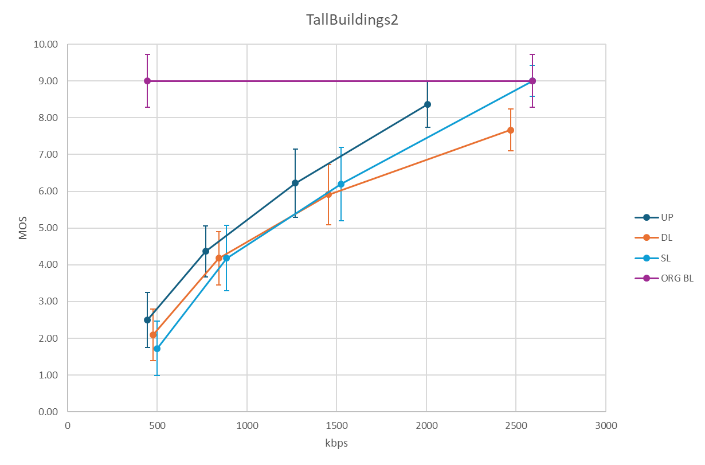
   

Figure 7: MOS-over-rate plots for the DCR experiment excluding the 9 viewers sitting at the miss-configured displays (see Section 2.1).