 ISO/IEC JTC 1/SC 29/AG 5 N170

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

**MPEG Visual Quality Assessment**

**Convenorship: DE**

**Document type:** General

**Title:** Report on **CVQM - Dataset of compressed video for study of quality metrics (version 2)**

**Status:** Approved

**Date of document:** 2025-8-15

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

**Expected action:** Information

**Action due date:** None

**No. of pages:** 16 (with this cover page)

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

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

**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**

**Daejeon, KR – July 2025**

|  |  |
| --- | --- |
| **Title** | Report on CVQM - Dataset of compressed video for study of quality metrics (version 2) |
| **Source** | AG 5 MPEG visual quality assessment |
| **Status** | Approved |
| **Serial Number** | 25454 |
| **Authors** | Yan Ye (Alibaba), Zhibo Chen (University of Science and Technology of China), Mathias Wien (RWTH Aachen University), Adam Wieckowski (Fraunhofer HHI), Vittorio Baroncini (VABTech) |

# Introduction

Exploring accurate and generalized objective video quality metrics is essential for the development of video coding standards, as well as for quality assurance in the multimedia content production and delivery ecosystem. The effectiveness of objective video quality metrics is usually evaluated by comparing them to the ground-truth subjective video quality scores, typically measured as mean opinion scores (MOS). Hence, it is essential to maintain a dataset of original video data and their corresponding coded versions for which ground truth subjective test results are available. There are already some published datasets to study compressed video quality assessment, including Waterloo 4K VQA [1], BVI-HD [2], MCL-V [3], CGVDS [4], SCVD [5], and PEA265 [6]. These datasets include MOS scores and objective metrics for video sequences with various characteristics (such as resolution, frame rate, scene representation, etc.) that are coded with different codecs at various coded quality.

However, these datasets are not suitable for AG 5’s work, because some of their copyright terms do not permit them to be used in standard development organizations such as MPEG. Further, AG 5 is interested in studying objective quality metrics suitable for future codecs that may rely on learning-based methods, whereas these existing datasets only use traditional video codecs.

To solve this problem, a dataset with diverse content and compression distortion types was proposed in [7]. Following discussions at the 12th meeting of AG 5 in July 2023, it was agreed to establish a dataset of compressed video for study of quality metrics, referred to as the CVQM dataset hereafter. This document provides detailed information about the original and compressed content in the CVQM dataset, including the sources and characteristics of the original content, coding parameters used to obtain the coded video data and objective quality metrics of the coded content.

In draft 1 of this document, coding results (coded bit rates, objective quality metrics) of a first round of encoding of the CVQM sequences are provided, with plans to further refine encoding parameters.

In draft 2 of this document, further analyses of the uncompressed video content in CVQM are provided, and more quantization parameters (QPs) are used to generate the compressed video data to cover a wider range of reconstructed qualities and coded bitrates. All bitstreams are re-encoded and cross-checked. Objective metrics are also re-calculated and cross-checked.

In version 1 of this document, complete results of the CVQM phase 1 study are included. The CVQM phase 1 study covers formal subjective viewing results and key objective quality metrics of the HD sequences in CVQM dataset coded using four traditional codecs and two learning-based codecs. Furthermore, editorial changes (e.g. re-arranging some sections) are made to streamline the presentation and make it more extensible for future CVQM phase 2 study. In this document, CVQM phase 1 and CVQM HD are used interchangeably, the same for CVQM phase 2 and CVQM UHD.

In version 2 of this document, results from further correlation study based on CVQM HD data are included.

# CVQM site content

## Uncompressed video content in CVQM

A set of 33 original video sequences from Youku, the JVET ftp site and Waterloo 4K datasets are included in the CVQM dataset, with further details provided in Annex A and in the accompanying excel (AG05N00150\_CVQM\_sequences.xlsx). These sequences are selected from a total of 108 original video sequences using the methodology described in Annex B. Copyright information of these sequences is provided in Annex C. Among these 33 CVQM sequences, 2 of them are from Alibaba (including 1 from Alibaba’s Youku), 8 of them are from Waterloo 4K and 23 of them are from JVET. Among the 23 JVET sequences, 6 of them are used in various JVET common test conditions (CTC) and 18 are not used in JVET CTC. All of these sequences can be obtained from the following site: <https://vqa.lfb.rwth-aachen.de/index.php/apps/files/?dir=/CVQM>.

Only accredited MPEG members can access the ftp site, and login information can be obtained by sending request to AG 5 Convenor Mathias Wien at [wien@lfb.rwth-aachen.de](mailto:wien@lfb.rwth-aachen.de).

### Characteristics of the uncompressed video content

The 33 original video sequences in CVQM span a wide variety of content category and spatial-temporal complexity. Figure 1 depicts the video categories and the range of spatial information (SI) and temporal information (TI) values covered by these sequences. More detailed information about the content categories can be found in Annex B.

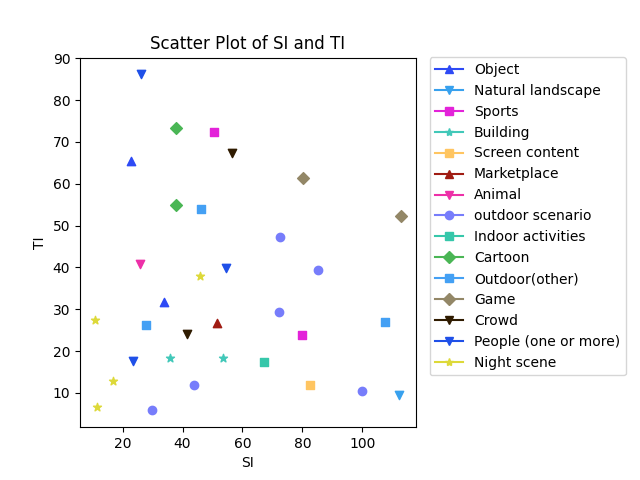


Figure 1. Annotation of the 33 original videos in the CVQM dataset, as well as their SI/TI distribution

Besides SI and TI analysis shown in Figure 1 , Video Complexity Analyzer (VCA) [9] is applied to the CVQM content. VCA is an open-source tool designed to assess spatial and temporal complexity by relying on the Discrete Cosine Transform (DCT) energy. Among the seven block-wise DCT-based energy features, visualization of VCA’s spatial complexity “E” and temporal complexity “H” is provided in Figure 2. It can be seen that the CVQM content has sufficient coverage in the VCA space.

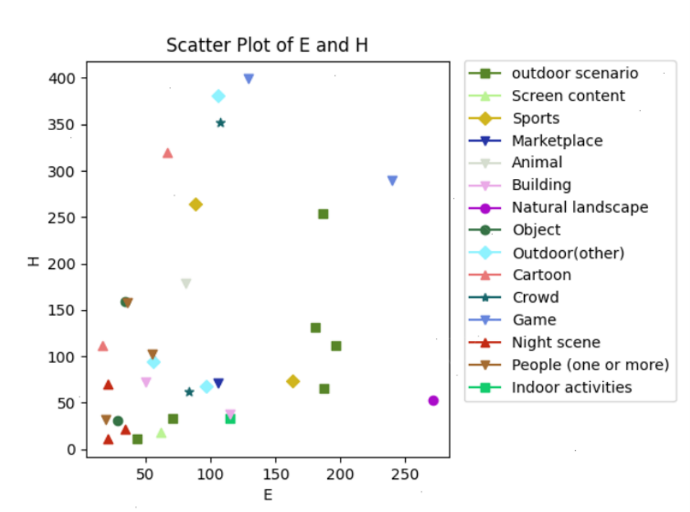
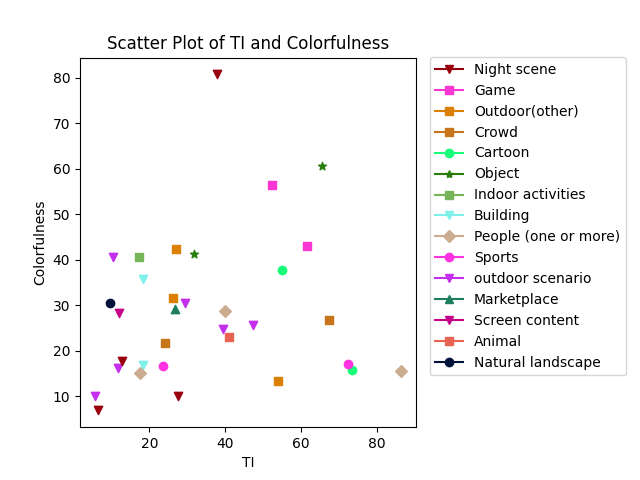
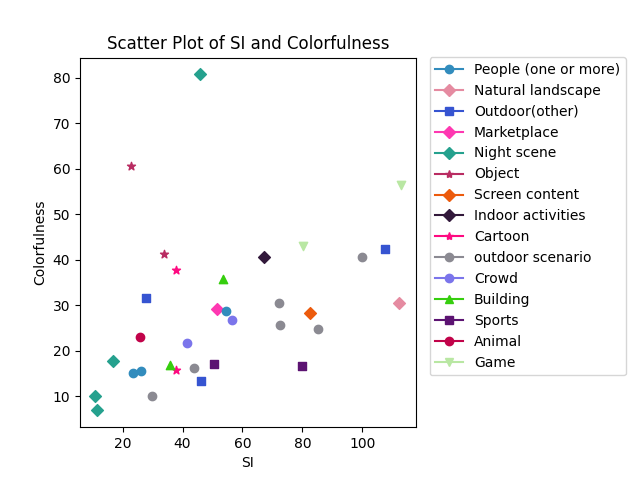


Figure 2. Distribution of VCA’s spatial and temporal complexity (“E” and “H”) of the CVQM sequences

Colourfulness (CF) [9][10] is commonly used to measure the variety and intensity of colour information in the images. The CF vs. SI and CF vs. TI plots for the CVQM content are provided in Figure 3 (a) and Figure 3 (b), respectively.



(a) (b)

Figure 3. CF vs SI distribution (a) and CF vs TI distribution (b) of the CVQM sequences

The presence of scene cut(s) in a video could affect the coded bitrates and perceived quality of the reconstructed video due to temporal motion discontinuity. Scene cuts could be the result of a sudden change between two scenes (e.g. due to sudden change in camera angle) or a more gradual change between two scenes (e.g. due to the use of fade-in, fade-out, or cross-fade techniques). Whether scene cuts are present, the number of scene cuts if present, and the type of scene cuts if present, are recorded in the accompanying spreadsheet AG05N00150\_CVQM\_sequences.xlsx.

## Compressed video data in CVQM

Both traditional video codecs and learning-based video codecs are used to encode the CVQM sequence and to obtain the compressed video data in CVQM. The same traditional codecs are used for HD and UHD sequences, whereas phase 1 learning-based codecs used to code HD content and phase 2 learning-based codecs used to code UHD content are different.

### Traditional video codecs

Two well-known ITU-T and MPEG codecs, namely H.265/HEVC and H.266/VVC, are selected to produce the coded video data in CVQM. Two implementations per codec are used, one from reference software and the other from open source. For H.265/HEVC, [HM-18.0](https://vcgit.hhi.fraunhofer.de/jvet/HM/-/tags/HM-18.0) and [x265-3.5+103-8f18e3a](https://bitbucket.org/multicoreware/x265_git/commits/8f18e3ad32684eee95e885e718655f93951128c3) are used. For H.266/VVC, [VTM-20.2](https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/tags/VTM-20.2) and [VVenC-1.9.0](https://github.com/fraunhoferhhi/vvenc/releases/tag/v1.9.0) are used.

All coded bitstreams can be obtained from the following site: <https://vqa.lfb.rwth-aachen.de/index.php/apps/files/?dir=/CVQM>.

The HM and VTM are coded with random access configuration. For x265 and VVenC, slower preset and medium preset with QPA enabled are used, respectively. And for all of these encoders, QP are set as {12, 17, 22, 27, 32, 37, 42, 47}. Furthermore, only the first 10 seconds for each sequence were used for encoding.

The same login information as those for the uncompressed data is needed to access and download these bitstreams.

For objective quality metrics, PSNR, SSIM, MS-SSIM, VMAF, and XPSNR [11] are provided along with the coded video data. PSNR is calculated according to the technical paper “Working practices using objective metrics for evaluation of video coding efficiency experiments” as specified in [HSTP-VID-WPOM](https://www.itu.int/dms_pub/itu-t/opb/tut/T-TUT-ASC-2020-HSTP1-PDF-E.pdf). PSNR, SSIM and MS-SSIM metrics are calculated using [HDRTools v0.24](https://gitlab.com/standards/HDRTools/-/commit/b6fde1487851ab417b5a317e03fee6d522d45f45), VMAF is calculated using [VMAF v2.3.1](https://github.com/Netflix/vmaf/releases/tag/v2.3.1), and XPSNR is calculated using [FFmpeg 7.1](https://ffmpeg.org/releases/ffmpeg-7.1.tar.xz).

All metrics for the coded data are reported in the accompanying excel (AG05N00150\_CVQM\_phase1\_data.xlsx). Without specific order, the HEVC codecs are referred to as H1 and H2, and the VVC codecs as V1 and V2.

In draft 1 of this document, the encoding used 4 QPs per sequence per encoder. In draft 2 of this document, to cover a wider range of reconstructed qualities and coded bitrates, the encoding was expanded to using 8 QPs for the HEVC and VVC encoders. All bitstreams were re-encoded and cross-checked. Objective metrics (PSNR, SSIM, MS-SSIM, and VMAF) were also re-calculated and cross-checked. In version 1 of this document, XPSNR was added as an additional objective metric.

### Phase 1 learning-based video codecs

ISO/IEC SC 29/AG 5 issued the first version of Call for learning-based video codecs for study of quality assessment in October 2023 [12], and received two learning-based codecs [13][14] in response to the Call in January 2024. These two learning-based video codecs are briefly summarized as follows:

* MaskCRT by NYCU and LUH [13]: MaskCRT is a Transformer-based P-frame codec that features a pixel-adaptive hybrid approach combining conditional motion coding and conditional residual coding. It includes an I-frame codec and a P-frame codec, with the P-frame codec comprising a motion estimation network, motion extrapolation network, conditional motion codec, motion compensation network, mask generator, and masked conditional inter-frame codec. Both the conditional motion codec and masked conditional inter-frame codec utilize a Transformer-based conditional autoencoder. To support a wide bit-rate range with high visual quality, MaskCRT adopts three variable-rate models, all optimized for MS-SSIM in the RGB domain using Vimeo-90k dataset [15]. For encoding, MaskCRT adopts IPPP coding structure with intra period of 32, and its coding protocol involves converting YUV420 sequences to RGB444 using BT.709 for encoding, then reverting back to YUV420 after decoding, with all decoded videos stored in 10-bit. The MaskCRT codec shows coding performance comparable to that of HM-18.0 in Random Access in terms of MS-SSIM-RGB.
* NVC by USTC [14]: Neural Video Codec (NVC) is an end-to-end video coding neural network model with a dedicated P-frame structure. NVC applies separate models for I-frame and P-frame, with the P-frame model relying solely on unidirectional reference information. Each intra period follows an IPPP reference structure. The P-frame model comprises several modules: motion estimation, motion coding, temporal context mining, contextual coding, and frame generation. The codec model is trained in the RGB domain on the Vimeo-90k dataset [15]. Given the CVQM video content is in YUV420 format, forward and backward colour space conversion between YUV420 and RGB is applied to the coding pipeline before encoding and after decoding, respectively. The NVC codec demonstrates objective performance comparable to that of HM-18.0 in Random Access configuration.

Given these two learning-based codecs only had the capability to encode video at HD resolution at the time of submission (in January 2024), CVQM phase 1 study mainly focused on the subjective quality evaluation of the 8 HD-resolution sequences in the CVQM dataset. In April 2024, all bitstreams from these two learning-based video codecs were submitted [16][17] and determined to satisfy the phase 1 subjective testing requirements. The coded bitrates of these two learning-based codecs approximated the coded bitrates of HM-18.0 within a reasonable range to ensure meaningful bitrate and quality coverage.

All metrics for the coded data are reported in the accompanying excel (AG05N00150\_CVQM\_phase1\_data.xlsx). The learning-based codecs are referred to as N1 and N2 without specific order.

### Phase 2 learning-based video codecs

To be completed in the future.

## Other information provided on the CVQM site

The following information can be found on the CVQM site <https://vqa.lfb.rwth-aachen.de/index.php/apps/files/?dir=/CVQM>:

* Readme file with relevant administrative information
* cvqm\_mos-metrics\_v01.csv, containing information about the objective metrics and MOS scores
* Folder original\_sequences/: Uncompressed YUV source files
* Folder yuvfiles/{N1,N2}/: reconstructed YUV files for the N1, N2 rate points packed in zip files
* Folders bitstreams/{H1,H2,N1,N2,V1,V2}/: Bitstreams for the codecs under test. The H1/H2 bitstreams can be decoded using an HEVC decoder, the V1/V2 bitstreams can be decoded using a VVC decoder.
* Folder codecs/{N1,N2}/: software packages for the N1/N2 codecs. In order to correctly decode the bitstreams, the usage of the graphics cards indicated in the software packages may be necessary.

# Phase 1 CVQM (CVQM HD) subjective quality study

A total of four laboratories participated in phase 1 CVQM study: University of Science and Technology of China (USTC), RWTH Aachen University, Fraunhofer HHI, and VABTech (see Annex D for information about each of these testing laboratories regarding their setup and naïve subjects). The testing process and results are detailed in this section.

## QP selection for subjective viewing

For each of the video sequences, four bitrates were selected for subjective evaluation. Small-scale subjective evaluations of the encoded sequences in the CVQM phase 1 study were conducted to select four suitable rate points (out of eight total rate points) for formal subjective viewing. The small-scale subjective experiments were conducted on VTM- and HM-encoded sequences across eight QP values (i.e., 12, 17, 22, 27, 32, 37, 42, 47) and involved six viewers experienced in video processing and subjective quality assessment.

Based on the small-scale viewing results, the mean opinion scores were plotted against bit rates on the rate-MOS curve. To ensure the selected rate points cover a sufficient quality range, and to maintain distinguishable quality between adjacent rate points, four evenly distributed quality scores were identified on the rate-MOS curve as follows. Denote the min and max quality scores on the rate-MOS curve as Q1 to Q2, the four quality scores are selected as , , and.

Given these quality scores, the closest bit rate points (and their corresponding QP values) are selected. Further fine-tuning is conducted to ensure uniform distribution across the whole bit rate range as much as possible. Further details about the process to select the rate points (and corresponding QPs) for HM and VTM used for formal subjective viewing can be found in [18].

For other encoders, bit rate points closest to the rate points chosen for HM and VTM are chosen for formal subjective viewing. Specifically, rate points for x265 and two learning-based encoders (MaskCRT and NVC) are aligned with the HM rate points, while rate points for VVenC are aligned with those of VTM.

## Preparation for phase 1 subjective testing

### Calibration experiment

Before conducting the formal experiment, pre-experiments were carried out in all participating laboratories to check for the alignment of the experimental setups. The calibration test included 16 PVSs (Processed Video Sequences) with different frame rate and different content. This exercise ensures that the viewing is properly set up when playing a test session with variable frame rates at all participating laboratories, and that scores collected across the different sites are comparable. The calibration tests were conducted successfully without any technical difficulties.

### Design of the test sessions

Each PVS was evaluated in a basic test cell (BTC) of the following structure:

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

The PVS under test were randomized and assigned to test sessions for each laboratory. Each laboratory was assigned three test sessions with 29 PVSs each which were to be presented in a defined order. The randomization, session design, and allocation to the laboratories was performed by RWTH. Each session included a stabilization phase of three BTCs and two BTCs with original vs. original comparison for sanity check purposes. In order to allow for cross-checking between the laboratories, the first session to be shown included overlapping PVS across the sessions.

### Training and instructions to the viewers

An identical set of instructions (see attachment [AG05N00170\_Instructions\_CVQM\_DCR\_test.pdf](file:///D:\Sciebo\MwiMeetings\20250626_MPEG151_JVET-AM_AG5-20_Daejeon\AG5\AG05N00170_CVQM\AG05N00170_Instructions_CVQM_DCR_test.pdf)) was provided to all laboratories, which were given to the participants. Furthermore, a common training session was used at all the laboratories. The training session included 18 BTCs with examples of different levels of impairments across all codecs being tested.

## Subjective testing results

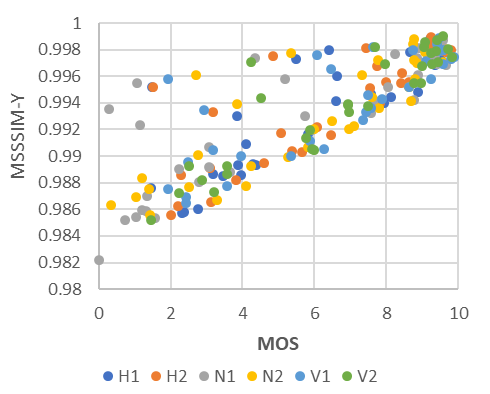
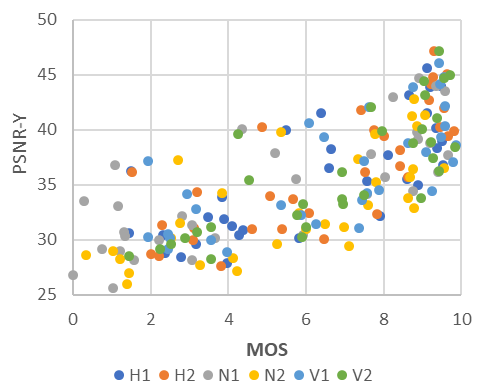
### Data processing

Only minimal data processing was applied to the recorded scores. The z-score for each vote was computed and the vote was discarded if the z-score exceeded the threshold of 2.5. The z-score was computed as follows:

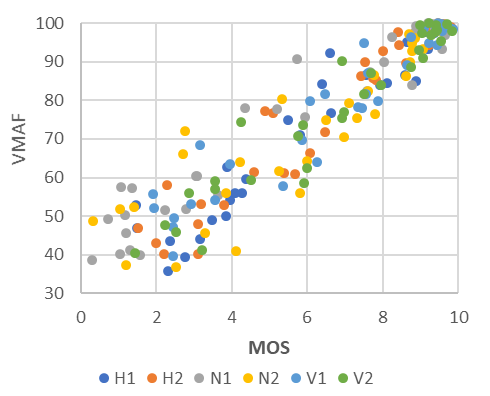
where is the score of viewer for PVS , and and are the MOS and the standard deviation for this PVS. Out a total of 9810 scores, 180 scores were removed (1.83%). No viewers were excluded while generating the scores presented in CVQM\_phase1\_data.xls.

### Objective metrics vs. MOS for all phase 1 test points

Figure 4 depicts the objective metrics vs. MOS scatter plots for all phase 1 test points.



(a) (b)

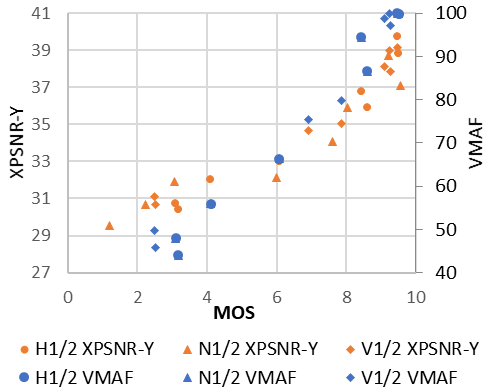
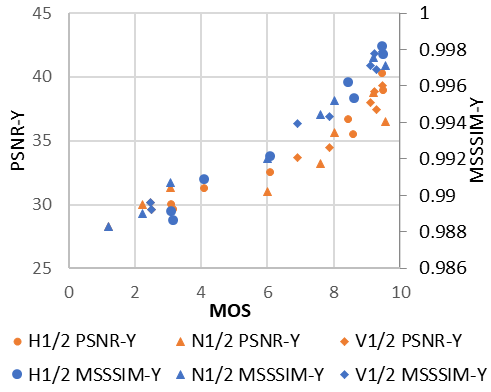


(c) (d)

Figure 4. Scatter plots of objective metrics vs. MOS for all phase 1 test points: (a) PSNR-Y vs. MOS, (b) MS-SSIM-Y vs. MOS, (c) XPSNR-Y vs. MOS and (d) VMAF vs. MOS

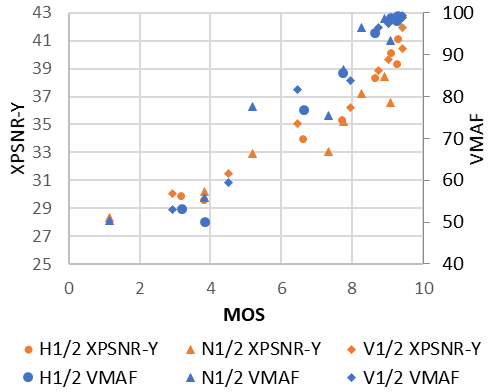
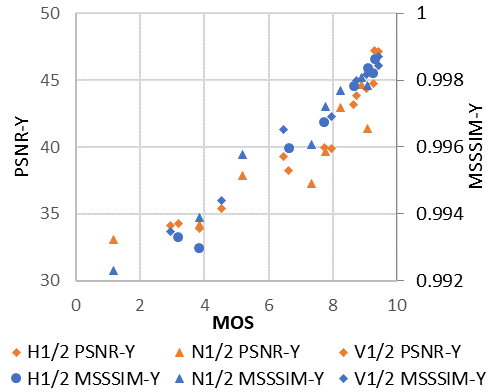
### Per-sequence plots of objective metrics vs. MOS

Figure 5 to Figure 12 provide the scatter plots of objective metrics over MOS for each of the phase 1 test sequences, with subfigure (a) plotting PSNR-Y (orange) and MS-SSIM-Y (blue) over MOS, and subfigure (b) plotting XPSNR-Y (orange) and VMAF (blue) over MOS.



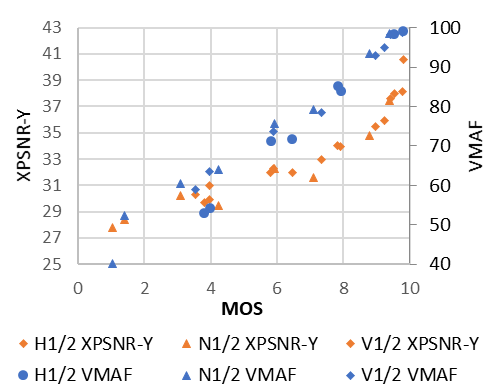
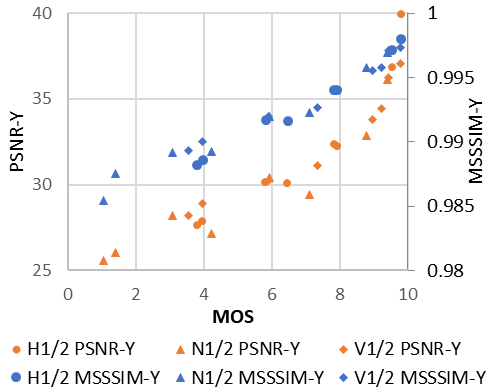
(a) (b)

Figure 5. Sequence “BasketballDrive”



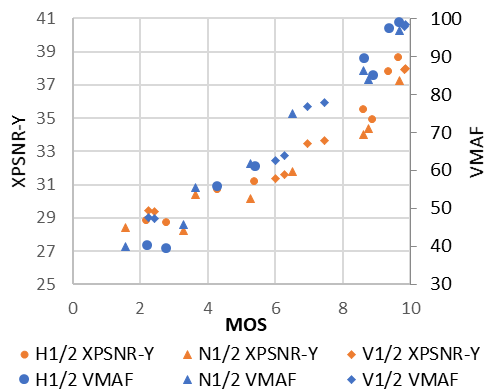
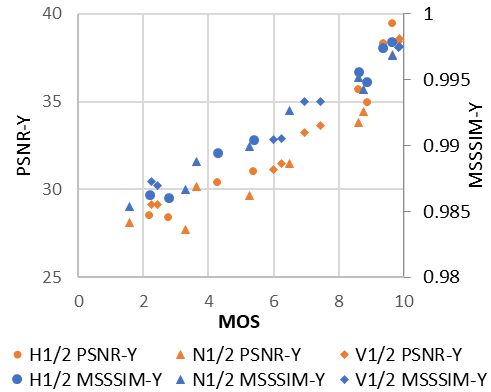
(a) (b)

Figure 6. Sequence “BGStatic2MovingOnePeople”



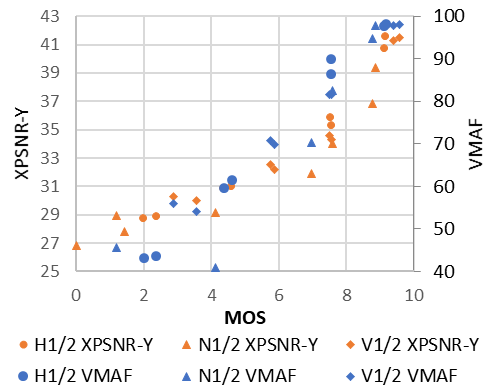
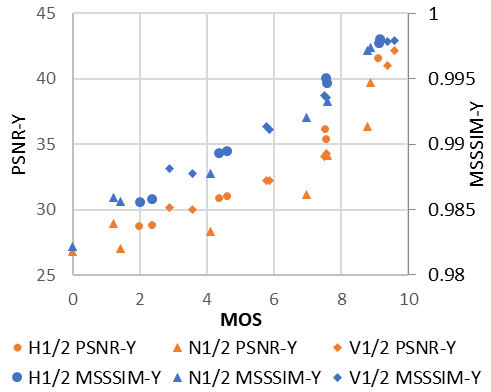
(a) (b)

Figure 7. Sequence “BQTerrace”



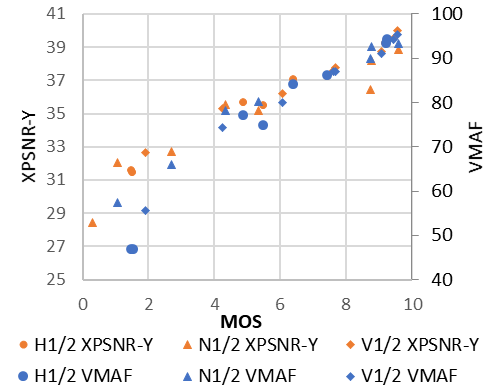
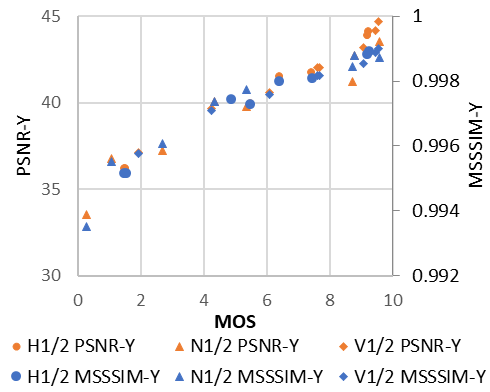
(a) (b)

Figure 8. Sequence “Cactus”



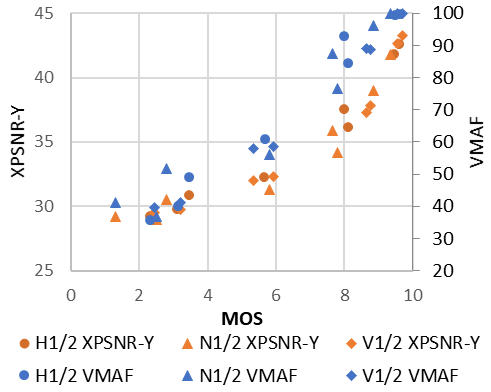
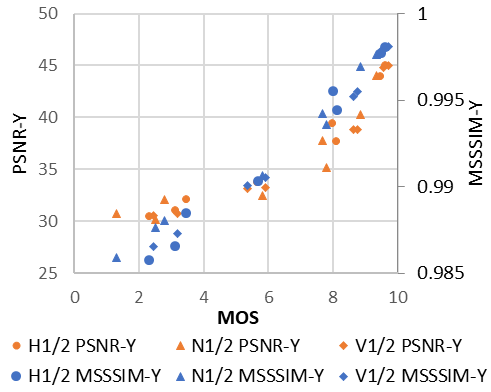
(a) (b)

Figure 9. Sequence “Fallout4”



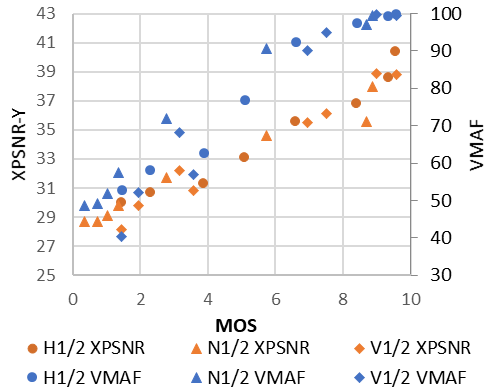
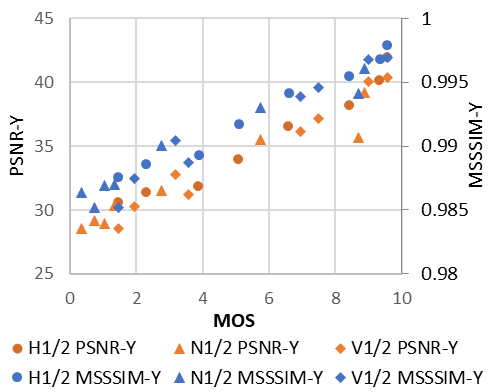
(a) (b)

Figure 10. Sequence “Meridian2”



(a) (b)

Figure 11. Sequence “RitualDance”



(a) (b)

Figure 12. Sequence “Stem”

### Correlation study

Correlation study between the various objective quality metrics and subjective MOS in CVQM HD has been conducted in [22]. For PSNR, XPSNR and MS-SSIM, the values of (Y, U, V) components are averaged together using weighting factor of (6,1, 1), as in the following equation; for VMAF, only the Y component is used.

Table 1 presents the overall correlations of each of the four objective metrics with the MOS for the CVQM HD sequences. As indicated in the table, VMAF demonstrates the strongest correlation with the MOS based on both the Pearson product-moment correlation coefficient and Spearman’s rank correlation coefficient. Further, it was observed that several high bit rate points in CVQM HD surpass the bit rates for HD resolution video in practical applications, causing saturation effects for both MOS and objective metric scores. Therefore, Table 1 also include Pearlson and Spearman correlation scores for which a 15 Mbit/s upper limit is applied, resulting in fourteen rate points being excluded from the calculation of correlation scores. While the correlation weakens for all four objective metrics, the general ranking remains, with VMAF ranked the highest, followed closely by XPSNR, and PSNR and MS-SSIM ranked lower.

Table 1. Pearson and Spearman correlation between objective metrics and MOS for all CVQM HD data points (left) and for only CVQM HD data points with bitrates not exceeding 15Mbit/s (right)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **no bitrate limit** | | **15 Mbit/s bitrate limit** | |
|  | **Pearson** | **Spearman** | **Pearson** | **Spearman** |
| **VMAF** | 0,943 | 0,941 | 0,937 | 0,937 |
| **PSNR(AVG)** | 0,735 | 0,756 | 0,727 | 0,754 |
| **XPSNR(AVG)** | 0,895 | 0,918 | 0,889 | 0,912 |
| **MS-SSIM(AVG)** | 0,720 | 0,733 | 0,708 | 0,728 |

Table 2 shows the Pearson correlation of the different objective metrics with the MOS per codec. It is noticeable that while on average each metric performs similarly well for different types of distortions, introduced by different codecs, each distortion metric has an outlier for which it either over- or under-performs. VMAF and XPSNR seem to underperform for the N2 codec, while still outperforming other metrics. SSIM and MS-SSIM seem to perform noticeably better for the V2 codec than for other codecs. Also, PSNR shows large correlation differences between the codecs V1 and V2, even though both implement the VVC standard.

Table 2: Pearson correlation filtered by codec

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **H1** | **H2** | **V1** | **V2** | **N1** | **N2** |
| **VMAF** | 0,959 | 0,962 | 0,953 | 0,960 | 0,962 | 0,887 |
| **PSNR(AVG)** | 0,727 | 0,755 | 0,690 | 0,782 | 0,754 | 0,717 |
| **XPSNR(AVG)** | 0,912 | 0,917 | 0,887 | 0,914 | 0,920 | 0,863 |
| **MS-SSIM(AVG)** | 0,673 | 0,686 | 0,662 | 0,801 | 0,715 | 0,772 |

# CVQM phase 2 (CVQM UHD) study

To be filled in the future.

# References

1. Z. Li, Z. Duanmu, W. Liu, Z. Wang, “Waterloo 4K Video Quality Assessment Database: A Comparative Study of Modern Video Encoders,” https://ece. uwaterloo. ca/~ z777li/4kvqa/.
2. F. Zhang, F. M. Moss, R. Baddeley, and D. R. Bull, “BVI-HD: A video quality database for HEVC compressed and texture synthesized content,” IEEE Transactions on Multimedia, vol. 20, no. 10, pp. 2620-2630, 2018.
3. J. Y. Lin, R. Song, C.-H. Wu, T. J. Liu, H. Wang and C.-C. J. Kuo, “MCL-V: A streaming video quality assessment database,” Journal of Visual Communication and Image Representation, Vol. 30, pp. 1-9, July 2015.
4. S. Zadtootaghaj, S. Schmidt, S. Shafiee Sabet, S. Moeller, and C. Griwodz, “Quality Estimation Models for Gaming Video Streaming Services Using Perceptual Video Quality Dimensions,” Proceedings of the 11th International Conference on Multimedia Systems. ACM. 2020.
5. S. Cheng, H. Zeng, J. Chen, J. Hou, J. Zhu, K.K. Ma, “Screen content video quality assessment: Subjective and objective study,” IEEE Transactions on Image Processing, Vol. 29, pp. 8636-8651, 2020.
6. L. Lin, S. Yu, L. Zhou, W. Chen, T. Zhao, Z. Wang, “PEA265: Perceptual assessment of video compression artifacts,” IEEE Transactions on Circuits and Systems for Video Technology, 30(11), pp. 3898-3910, 2020.
7. Y. Lu, F. Guan, Y. Qi, Y. Ren, X. Li, Z. Yu, Z. Chen, R.-L. Liao, Y. Ye, “Video dataset for study of learning-based quality metrics,” ISO/IEC JTC 1/SC 29/AG 5, document no. m64283, July 2023.
8. S. Wang, M. Wien and Y. Ye, "Preliminary databases for assessment of quality metrics for 2D video using neural-network (NN) based methodologies," ISO/IEC JTC 1/SC 29/AG 5, document no. m57121, July 2021.
9. Winkler S. Analysis of public image and video databases for quality assessment[J]. IEEE Journal of Selected Topics in Signal Processing, 2012, 6(6): 616-625.
10. D. Hasler and S. E. Suesstrunk, “Measuring colorfulness in natural images,” in Proc. SPIE Human Vision and Electronic Imaging, vol. 5007, Jun. 2003, pp. 87–95.
11. C. R. Helmrich, S. Bosse, H. Schwarz, D. Marpe, and T. Wiegand, “A study of the extended perceptually weighted peak signal-to-noise ratio (XPSNR) for video compression with different resolutions and bit depths,” ITU Journal: ICT Discoveries, Volume 3, Issue 1.
12. Call for learning-based video codecs for study of quality assessment, version 1, AG 5 MPEG visual quality assessment output doc. no. N00104, October 2023.
13. Y.-H. Chen, et. al., “Response to Call for Learning-Based Video Codecs for Study of Quality Assessment by NYCU and LUH”, MPEG doc. no. m66163, January 2024.
14. X. Pan, et. al., “Response to the call for learning-based video codecs for study of quality assessment by University of Science and Technology of China and Alibaba”, MPEG doc. no. m66296, January 2024.
15. T. Xue, et al. "Video enhancement with task-oriented flow." International Journal of Computer Vision 127 (2019): 1106-1125.
16. Y.-H. Chen, et. al., “Bitstream Generation and Bit Rate Fitting Results of MaskCRT for CVQM HD Sequences”, MPEG doc. no. m68079, April 2024.
17. X. Pan, et. al., “Coded bitstream generation using learning-based video codec for CVQM”, MPEG doc. no. m66975, April 2024.
18. Z. Yu, et. al., “Sequence-dependent QP Selection for the CVQM Subjective Experiments”, MPEG doc. no. m66270, January 2024.
19. D. Müllner, “Modern hierarchical, agglomerative clustering algorithms,” arXiv preprint arXiv:1109.2378, 2011.
20. Z. Li, Z. Duanmu, W. Liu, Z. Wang, “AVC, HEVC, VP9, AVS2 or AV1? — A Comparative Study of State-of-the-art Video Encoders on 4K Videos”, 16th International Conference on Image Analysis and Recognition, Waterloo, Ontario, Canada, August 27-29, 2019.
21. J. Chen, Y. Ye, R. Li, W. Jiang, “AHG4: New test sequences for JVET exploration,” the joint video experts team (JVET) of ITU-T SG16 and ISO/IEC JTC 1/SC 29, document no. JVET-Z0156, April 2022.
22. C. Lehmann, A. Wieckowski, B. Bross, “Detailed Correlation between Objective Metrics and MOS for CVQM HD Sequences”, MPEG doc. no. m73191, July 2025.

# Annex A: List of CVQM sequences

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Video Name** | **source** | **width** | **height** | **frame rate** | **Bit depth** | **# of frames** | **scene cut** | **MD5** |
| Waterloo\_01 | Waterloo | 3840 | 2160 | 24 | 8 | 240 | No | 015de9df9398279ec2988e67ef4d7ad8 |
| Waterloo\_02 | Waterloo | 3840 | 2160 | 25 | 8 | 250 | No | 3dc18fd4da4584cb9db067b0aafa5559 |
| Waterloo\_04 | Waterloo | 3840 | 2160 | 24 | 8 | 240 | No | ec40855a65b104e0529c9edf937b5eb6 |
| Waterloo\_05 | Waterloo | 3840 | 2160 | 30 | 8 | 300 | No | 017b0d2cc3c5cd03b3c90c7a72074271 |
| Waterloo\_06 | Waterloo | 3840 | 2160 | 30 | 8 | 300 | No | d798d32ff68ff24045cd1e5d5578b410 |
| Waterloo\_11 | Waterloo | 3840 | 2160 | 30 | 8 | 300 | Yes | 5b09e998e8a1059af962d7db23e4ba94 |
| Waterloo\_12 | Waterloo | 3840 | 2160 | 30 | 8 | 300 | Yes | c16169fe2f1593c56007e48dc1bae6cd |
| Waterloo\_18 | Waterloo | 3840 | 2160 | 25 | 8 | 250 | No | 99a20338575a84b9b84694fb5dad2052 |
| 3DGameKit4K | JVET | 3840 | 2160 | 60 | 10 | 601 | No | f92dde781b7213821e89b03ef77eaa83 |
| BasketballDrive | JVET | 1920 | 1080 | 50 | 8 | 501 | No | d38951ad478b34cf988d55f9f1bf60ee |
| BQTerrace | JVET | 1920 | 1080 | 60 | 8 | 601 | No | efde9ce4197dd0b3e777ad32b24959cc |
| Cactus | JVET | 1920 | 1080 | 50 | 8 | 500 | No | 3fddb71486f209f1eb8020a0880ddf82 |
| Campfire | JVET | 3840 | 2160 | 30 | 10 | 300 | No | 63d3d9f9e4e8b5c344e89840e84e6428 |
| DroneTakeOff | JVET | 3840 | 2160 | 30 | 10 | 300 | No | b787e95fb3d3e066f4717ccf8875f972 |
| Fallout4 | JVET | 1920 | 1080 | 60 | 8 | 3602 | Yes | 09980ed191d285449d591fb0cf3dbe43 |
| FoodMarket4 | JVET | 3840 | 2160 | 60 | 10 | 720 | No | a378b34190f54f688d048a9a8b46a8ac |
| Fountains | JVET | 3840 | 2160 | 30 | 8 | 300 | No | 662ce078ab5bcfa068b6b5f45f5cd930 |
| NeptuneFountain3 | JVET | 3840 | 2160 | 60 | 10 | 600 | No | 88fd87ea57df4a36200946025e8618aa |
| Netflix\_Dancers | JVET | 4096 | 2160 | 60 | 10 | 1199 | No | e440efb517c01d27ecf072c29e31197b |
| Netflix\_DinnerScene | JVET | 4096 | 2160 | 60 | 10 | 1199 | No | 4cf8bb5eb4610ab20266332a638c3c55 |
| Netflix\_Meridian2 | JVET | 1920 | 1080 | 60 | 10 | 600 | Yes | 473757104c47837efe3fc3b85aef642a |
| Netflix\_RollerCoaster | JVET | 4096 | 2160 | 60 | 10 | 1199 | No | 8e99b7f486025ce8fccda4f3e818a186 |
| Netflix\_SquareAndTimelapse | JVET | 4096 | 2160 | 60 | 10 | 600 | No | a7ffe67fde0a41b313c7cc9317195d44 |
| Netflix\_ToddlerFountain | JVET | 4096 | 2160 | 60 | 10 | 1199 | Yes | 31eb61104464922d91f42e5d5b49097e |
| NightRoad | JVET | 3840 | 2160 | 60 | 10 | 600 | No | e13f46ff51607fd5eda429cd6f9b3d2d |
| OberbaumSpree\_ | JVET | 3840 | 2160 | 60 | 10 | 600 | No | 6975b81c9e63c92b3bf4223796102da1 |
| ResidentialBuilding | JVET | 3840 | 2160 | 30 | 8 | 300 | No | e1b3a496b20a88fa88f7cd2e0261c68a |
| RitualDance | JVET | 1920 | 1080 | 60 | 10 | 600 | Yes | a3cb399a7b92eb9c5ee0db340abc43e4 |
| Stem | JVET | 1920 | 1080 | 25 | 8 | 250 | No | 9e17f565a8c1beb46a5bc02e680c64ff |
| TiergartenParkway | JVET | 3840 | 2160 | 60 | 10 | 600 | No | cc17b64e1fb93879c0873444ca13e290 |
| Wood | JVET | 3840 | 2160 | 30 | 8 | 300 | No | d7c6e0954971b4d08a4de724f06f5e44 |
| BGStatic2Moving\_OnePeople | Alibaba1 | 1920 | 1080 | 30 | 8 | 450 | No | fe3b72f9f4fd3474a93b66e4db10c4be |
| PostalService | Alibaba1 | 3840 | 2160 | 25 | 8 | 375 | No | 203db979f31f431c6b9f7afe0918892d |

1. The sequence “BGStatic2Moving\_OnePeople” is from JVET-T0060 and the sequence “PostalService” is from JVET-Z0156.

# Annex B: sequence selection methodology

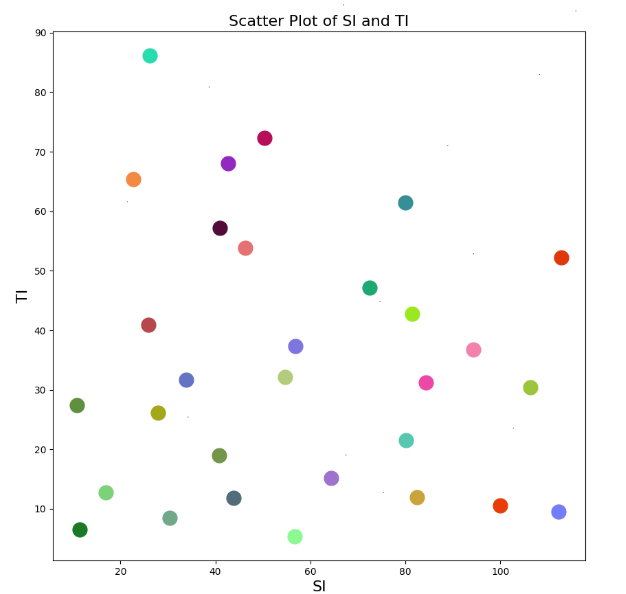
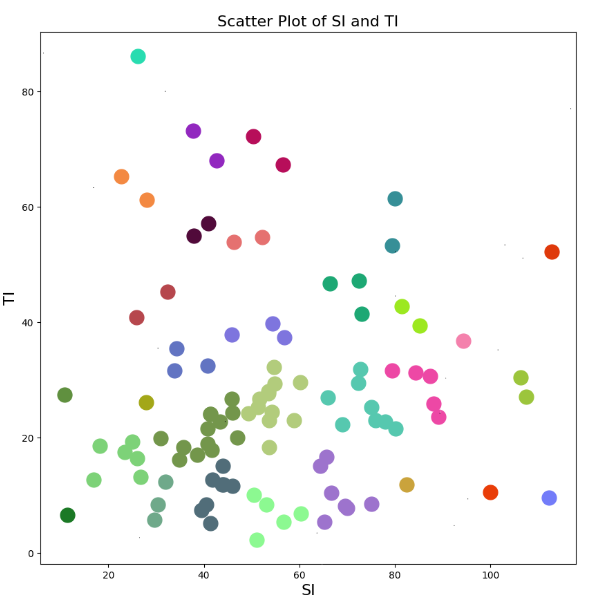
The sequence selection process started from a total of 108 candidate sequences from Youku, JVET ftp site and Waterloo 4K datasets. All these sequences are in YUV420 format, and have a duration of at least 10 seconds. If a sequence is longer than 10 seconds, then a 10-second segment is cropped from it to become the candidate sequence.

All candidate sequences are annotated with scene categories, encompassing a wide range of commonly encountered video compression scenes. There are 15 scene categories in total, including animal, building, cartoon, crowd, game, indoor activities, marketplace, natural landscape, night scene, object, outdoor scene, outdoor (other), people (one or more), screen content, and sports.

The degree of distortion in the compressed video is related to the spatial-temporal complexity of the content, which can be measured by two commonly used attributes, spatial information (SI) and temporal information (TI). Therefore, sequences in the CVQM dataset should have wide coverage of SI and TI values.

To select appropriate sequences from these 108 candidate sequences, the following steps are followed:

1. The SI and TI values of the 108 candidate sequences are calculated using the VQEG SITI-Tools (available at <https://github.com/VQEG/siti-tools>). Figure 5 (a) shows the SI and TI distribution of all 108 candidate sequences;
2. Based on the SI/TI scatter plot, a "cluster then select" paradigm is applied to down select a smaller set of representative sequences from the 108 candidates. The Agglomerative Clustering method [17] is used for clustering. The sequences belonging to the same cluster have similar SI/TI values, therefore one is selected and other redundant samples are removed. This is performed in an iterative manner, and finally a smaller set of intermediate sequences are selected. As shown in Figure 5 (b), these intermediate sequences cover a wide SI/TI range and the distribution exhibits good uniformity;



1. （b）

Figure 5. SI/TI distribution of (a) all 108 candidate sequences, and (b) intermediate selected sequences

1. Further minor adjustments are performed to ensure that all 15 scene categories are represented by the dataset. Finally, 33 video sequences in total are selected. Figure 1 depicts the final selected sequences in the CVQM dataset, their respective category labels, and the range of SI/TI values they cover. Within this set, night scene and outdoor scene contain multiple samples. The samples from night scene are concentrated in the low SI/TI region, which is consistent with the observation that these night scene videos contain lower color richness and lower temporal fluctuations. For outdoor scene, the samples cover a wider SI/TI range, which reflects the richness of the spatial-temporal complexity of the video content in this category.

# Annex C: Copyright statements of the original content

The video sequences in the CVQM dataset are from one of three sources: Waterloo 4K, Youku, and JVET. Their respective copyright statements are provided in this Annex.

## C.1 Waterloo 4K content copyright

The copyright statement of Waterloo 4K dataset can be found at [https://ivc.uwaterloo.ca/database/4KVQA.html](https://ivc.uwaterloo.ca/database/4KVQA.htmlf), and is copied and pasted below. It provides reasonable terms of use by standard development organization such as MPEG, and upon our communication with the original authors, they have given explicit permission for AG 5 to use their video data.

*“Permission is hereby granted, without written agreement and without license or royalty fees, to use, copy, modify, and distribute this database (the images, the results and the source files) and its documentation for any purpose, provided that the copyright notice in its entirity appear in all copies of this database, and the original source of this database, Image and Vision Computing Laboratory (IVC, https://ivc.uwaterloo.ca/) at the University of Waterloo (UW, http://www.uwaterloo.ca), is acknowledged in any publication that reports research using this database.”*

It is hereby acknowledged that Waterloo 4K content is provided with the following publication [18]:

* Zhuoran Li, Zhengfang Duanmu, Wentao Liu, Zhou Wang, “AVC, HEVC, VP9, AVS2 or AV1? — A Comparative Study of State-of-the-art Video Encoders on 4K Videos”, 16th International Conference on Image Analysis and Recognition, Waterloo, Ontario, Canada, August 27-29, 2019.

## C.2 Youku content copyright

Youku sequence copyright is provided in an earlier JVET contribution JVET-Z0156 [19], copied and pasted below:

*“The proposed sequences are produced by Alibaba Group（licensed by one of its subsidiary Youku） and all intellectual property rights remain with Alibaba Group.*

*The following uses are allowed for the contributed sequences:*

*1. Sequences may be published in technical papers, played at technology research and development events.*

*2. Sequences may be used by Standards committees. (e.g., ITU, MPEG, VQEG).*

*The following uses are NOT allowed for the contributed sequences:*

*1. Do not publish snapshots in product brochures.*

*2. Do not use video for marketing purposes.*

*3. Do not redistribute video with a commercial product.*

*4. Do not use in television shows, commercials, or movies.*

*5. Do not sublicense or transfer the sequences to anyone who has not received this license.”*

## C.3 JVET content copyright

Throughout the years, various organizations have provided content for the use of standards development to the joint teams of ITU-T SG16 and ISO/IEC JTC 1/SC 29. These sequences are stored on the JVET ftp site with accompanying copyright statements that permit the use by MPEG.

# Annex D: Setup at each of the CVQM phase 1 testing laboratories

At the USTC laboratory, the subjective experiment setup included a single 55-inch SONY screen with a resolution of 3840×2160, connected via HDMI to a DeckLink 4K Extreme 12G video board, enabling stable playback of raw 10-bit YUV data from the SSD drives at the required frame rate. The HD video signal was centrally displayed on the UHD screen, surrounded by mid-grey padding in the unused areas, and the viewing angle was maintained at 90 degrees. Participants were seated at a distance of 1.5H from the UHD display. The experiment involved 30 naive viewers (3 females and 27 males, all under 25 years old).

HHI performed the subjective experiments using two 65-inch LG OLED screens (LG OLED65E97LA) with UHD resolution (3840x2160). The displays were connected to an HDMI-splitter connected to DecLink 4G Extreme 4G video board using HDMI. The HD content was displayed centered in the display surrounded by mid-gray padding. The smooth 3840p 60fps 10bit video playback capability of the overall setup was verified on both displays (UHD playback required to also display padding). The participants were sitting at 1.5H distance to the display (i.e. 3H distance relative to the displayed content without padding). The test was performed with 29 subjects, 19 of which were male, and 10 were female. Eight of the subjects didn’t provide additional personal data. The 21 subjects providing personal information were aged between 21 and 33 (average 27.1). No visual acuity screening was performed other than subject self-assessment. The tests were performed in 5 sessions of maximal 6 viewers sitting at 90° or 60° angle to the display (3 per display). For the sessions with less than 6 subjects, the center viewing position (90°) was always used.

At RWTH Aachen, four displays were used for the viewing task: 1× Sony 55” PVM X550 (3840×2160), connected via Quad-SDI, 1× LG OLED65CX (3840×2160), connected via HDMI, and 2× LG OLED55G19LA (3840×2160), connected via HDMI as well. The X550 was driven by a PC with a DeckLink Extreme 4G video board via Quad-SDI, the signal further converted to HDMI by an AJA Hi5-4K-Plus converter and sent in parallel to the three LG displays via an HDMI splitter. The viewing distance of the viewers was 1.5H per UHD display, with two viewers placed at each display. The HD video signal was displayed centered to the UHD screen with a mid-grey padding for the unused area. A total of 31 naïve viewers participated in the tests. The viewers were checked for acuity and color blindness. The group consisted of 7 females and 24 males from 14 different nationalities.

The VABTech test setup used a 65” display (TV set LG OLED 65E97LA) with an UHD resolution of 3840x2150 driven through an HDMI 2.1 connection by a high speed PC operating under Windows 11. The HD content was displayed centered in the display surrounded by mid-gray padding. The smooth 3840p 60fps 10bit video playback capability of the overall setup was verified on both displays (UHD playback required to also display padding). The participants were seated at 1.5H distance to the display considering the actual active part of the screen. The test was performed with 16 subjects, checked for visual acuity and color blindness. Two viewers were sitting in front of the screen. Due to logistics constraints and limited subjects availability Vabtech run only 16 verified subjects. VATech run all the 9 test sessions, to assess in an additional way the cross correlation of results among the three labs. The results of this additional experiment show the same level of correlation provided running session 0.