 ISO/IEC JTC 1/SC 29/WG 04 N543

**ISO/IEC JTC 1/SC 29/WG 04  
MPEG Video Coding   
Convenorship: CN**

**Document type:** Output Document

**Title:** Common test conditions for video coding for machines

**Status:** Approved

**Date of document:** 2024-08-19

**Source:** ISO/IEC JTC 1/SC 29/WG 04

**Expected action:** None

**Action due date:** None

**No. of pages:** 18 (without cover page)

**Email of Convenor:** yul@zju.edu.cn

**Committee URL:** <https://sd.iso.org/documents/ui/#!/browse/iso/iso-iec-jtc-1/iso-iec-jtc-1-sc-29/iso-iec-jtc-1-sc-29-wg-4>

**INTERNATIONAL ORGANIZATION FOR STANDARDIZATION**

**ORGANISATION INTERNATIONALE DE NORMALISATION**

**ISO/IEC JTC 1/SC 29/WG 04 MPEG VIDEO CODING**

**ISO/IEC JTC 1/SC 29/WG 04 N543**

**July 2024, Sapporo, JP**

|  |  |
| --- | --- |
| **Title** | **Common test conditions for video coding for machines** |
| **Source** | **WG 04, MPEG Video Coding** |
| **Status** | **Approved** |
| **Serial Number** | **24281** |
| **Authors** | **Shan Liu, Chris Rosewarne, Olgierd Stankiewicz** |

**Abstract**

This document describes common test conditions (CTC), reference software and configurations, anchor generation, evaluation framework, reporting and other information to be used for experimenting and evaluating technologies for video coding for machines (VCM). These common test conditions are recommended for use in technical contributions to the 148th and following MPEG meetings, as applicable.

1. **Introduction**

Common test conditions (CTC) are desirable for conducting experiments in a well-defined environment followed by the comparison of the outcome of experiments. This document defines such common test conditions for experimenting and evaluating technologies for video coding for machines (VCM).

A set of test conditions (or configurations) are defined to reflect random-access, low-delay and intra-only settings for both inner and end-to-end codecs, as follows:

* Random access for inner (RA\_inner)
* Random access for end-to-end (RA\_e2e)
* Low delay for inner (LD\_inner)
* Low delay for end-to-end (LD\_e2e)
* All intra for inner (AI\_inner)
* All intra for end-to-end (AI\_e2e)

The reporting of results for each configuration is either mandatory or optional, dependent on the dataset, the machine task, etc. Detailed information about mandatory and optional configurations and results are provided in Section 2.

Evaluation scripts can be found at <https://git.mpeg.expert/MPEG/Video/VCM/vcm-ctc>. Access right to the repository is granted to registered experts in the MPEG WG 4.

The following sections define test sequences, anchors, test and training conditions, evaluation metrics and reporting. Anyone bringing input contributions for WG 4 VCM AHG to discuss shall provide a set of results that is as complete as possible and uses the CTC that apply to the proposal.

1. **Test datasets**

Three video datasets, i.e., SFU-HW-objects-v3.2 dataset (referred to as SFU-HW in this document), Tencent Video dataset (referred as TVD in this document) and PandaSet dataset (referred as Panda), are included in the CTC. Reporting results on SFU-HW and TVD, except class O of the SFU-HW dataset, for RA\_inner, RA\_e2e, LD\_inner, LD\_e2e, AI\_inner, AI\_e2e configurations are mandatory. Reporting results class O of the SFU-HW dataset and on PandaSet dataset for all configurations are optional. In addition, three image datasets are included in the CTC and can be tested using AI\_inner configuration. Reporting results on the image datasets is optional.

The relevant files and data corresponding to the test materials have been uploaded to the MPEG Expert area (<https://content.mpeg.expert/data/>), with access instructions as follows.

* Log in using username = mpeg
* The password is the same as for the MPEG Document Management System
* Subdirectories: MPEG-AI/VCM
* VCM path: <https://content.mpeg.expert/data/MPEG-AI/VCM/>

## *SFU-HW Dataset*

The SFU-HW-objects-v3.2 dataset (referred to as SFU-HW in this document) is a video dataset consisting of 14 sequences which are known from previous standardization efforts in JCT-VC and JVET. The sequences can be found on <ftp://hevc@mpeg.tnt.uni-hannover.de>. The annotations are available at https://data.mendeley.com/datasets/367kty6nw7/1. Specific license information is provided with each test sequence.

Table 1 provides the detailed information of each sequence from the SFU-HW dataset to be used in this CTC. Only a subset of each video sequence needs to be encoded. The first n frames of each sequence are skipped. The exact number n frames to skip and the number of frames to be coded can be found in Table 1. The n value in the table can be directly applied to VTM encoder [1] by using the FrameSkip parameter. The MD5 checksums for the 14 sequences in SFU-HW dataset are provided in Table 2.

Table 1. Test sequences in SFU-HW dataset

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Class** | **Sequence name** | **Frame count** | **Frame rate** | **Bit depth** | **Frames skipped** | **Frames coded** | **RA/LD/AI \_inner/e2e** |
| A | Traffic | 150 | 30 | 8 | 117 | 33 | M |
| B | ParkScene | 240 | 24 | 8 | 207 | 33 | M |
| B | BasketballDrive | 500 | 50 | 8 | 403 | 97 | M |
| B | BQTerrace | 600 | 60 | 8 | 471 | 129 | M |
| C | RaceHorsesC | 300 | 30 | 8 | 235 | 65 | M |
| C | BQMall | 600 | 60 | 8 | 471 | 129 | M |
| C | PartyScene | 500 | 50 | 8 | 403 | 97 | M |
| C | BasketballDrill | 500 | 50 | 8 | 403 | 97 | M |
| D | RaceHorsesD | 300 | 30 | 8 | 235 | 65 | M |
| D | BQSquare | 600 | 60 | 8 | 471 | 129 | M |
| D | BlowingBubbles | 500 | 50 | 8 | 403 | 97 | M |
| D | BasketballPass | 500 | 50 | 8 | 403 | 97 | M |
| O | Kimono | 240 | 24 | 8 | 207 | 33 | O |
| O | Cactus | 500 | 50 | 8 | 403 | 97 | O |

*Note: M – mandatory; O – optional.*

|  |  |  |
| --- | --- | --- |
| **Class** | **Sequence name** | **MD5 checksum** |
| A | Traffic | 4f03a86b03b47fc821acffb8baea56f6 |
| B | ParkScene | b7ada0912d693304165254177d08343d |
| B | BasketballDrive | d38951ad478b34cf988d55f9f1bf60ee |
| B | BQTerrace | efde9ce4197dd0b3e777ad32b24959cc |
| C | RaceHorsesC | 0a351df99f22d837bc528bd4901c6968 |
| C | BQMall | f889efea02b0c9a7d174b0f7a99cb51b |
| C | PartyScene | 4766c455665b6d228a6390e3d3ff2647 |
| C | BasketballDrill | bd215136fed04067d82c10b2e49b2c7c |
| D | RaceHorsesD | 290a63e86213abc4459fce1dbd39edbe |
| D | BQSquare | 713ef64958345859b9bae986c3a3f763 |
| D | BlowingBubbles | 50a520722f0e906b7884b6b9fea48699 |
| D | BasketballPass | bfd9abbdc677790130dc4023b4e409f0 |
| O | Kimono | 4a83005bc719012ac148dd3898e5e4ed |
| O | Cactus | 3fddb71486f209f1eb8020a0880ddf82 |

Overall BD-rate performance of SFU-HW video content is calculated and reported on a per-dataset basis and the reporting template contains assigned fields and auto-functions to facilitate this purpose. Note that class O (Cactus and Kimono) is not included in calculating the overall results.

## *Tencent Video Dataset (TVD)*

The Tencent Video Dataset (TVD) is a video dataset consisting of 7 sequences in 1920x1080 resolution used for object tracking. Detailed information about these sequences is provided in Table 3. The dataset with corresponding annotations is available at [2] <https://multimedia.tencent.com/resources/tvd>.

Table 3. Test sequences in TVD

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Sequence name** | **Frame count** | **Frame rate** | **Bit depth** | **Frames skipped** | **Frames coded** | **RA/LD/AI \_inner/e2e** |
| TVD-01-1 | 3000 | 50 | 8 | 1500 | 500 | M |
| TVD-01-2 | 3000 | 50 | 8 | 2000 | 500 | M |
| TVD-01-3 | 3000 | 50 | 8 | 2500 | 500 | M |
| TVD-02-1 | 636 | 50 | 10 | 0 | 636 | M |
| TVD-03-1 | 2334 | 50 | 10 | 0 | 500 | M |
| TVD-03-2 | 2334 | 50 | 10 | 500 | 500 | M |
| TVD-03-3 | 2334 | 50 | 10 | 1000 | 500 | M |

Table 4. MD5 checksum of the test sequences in TVD

|  |  |
| --- | --- |
| **Sequence name** | **MD5 checksum** |
| TVD-01 | 1dddac6c82e5c8e59f06d283458e2db7 |
| TVD-02 | aad63df298fa6401c16a36ede61e9798 |
| TVD-03 | 9aa26e98ac34e7da9712c3ed4677da4b |

As the original sequences are available in .mp4 format, they shall be converted to YUV420p using FFmpeg [3]:

ffmpeg -i {input.mp4} {output.yuv}

The MD5 checksums for TVD sequences in YUV420 format are provided in Table 4.

## *PandaSet Dataset (optional)*

The PandaSet dataset is a video dataset consisting of 106 sequences in 1920x1080 resolution used for semantic segmentation. Among the total 106 sequences, 73 are selected to be included in the CTC to provide optional test results. They are grouped into six classes, as shown in Table 5.

Table 5. Test sequences in Panda dataset

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Class** | **Sequence ID** | **Frame count** | **Frame rate** | **Bit depth** | **Frames skipped** | **Frames coded** | **RA/LD/AI \_inner/e2e** |
| A1 | 57, 58, 69, 70, 72, 73, 77 | 80 | 10 | 8 | 0 | 0 | O |
| B1 | 3, 11, 16, 17, 21, 23, 27, 29, 30, 33, 35, 37, 39, 43, 53, 56, 97 | 80 | 10 | 8 | 0 | 0 | O |
| C1 | 88, 89, 90, 95, 109, 112, 113, 115, 117, 119, 122, 124 | 80 | 10 | 8 | 0 | 0 | O |
| A2 | 64, 65, 66, 67, 71, 78, 149 | 80 | 10 | 8 | 0 | 0 | O |
| B2 | 1, 5, 13, 15, 19, 24, 28, 32, 34, 38, 40, 41, 42, 44, 46, 52, 54, 139 | 80 | 10 | 8 | 0 | 0 | O |
| C2 | 80, 84, 94, 101, 102, 103, 105, 106, 110, 116, 123, 158 | 80 | 10 | 8 | 0 | 0 | O |

More details about PandaSet dataset are available at <https://pandaset.org/>. The dataset with annotation can be downloaded from [https://content.mpeg.expert/data/MPEG-AI/VCM/Pandaset/](https://nam12.safelinks.protection.outlook.com/?url=https%3A%2F%2Fcontent.mpeg.expert%2Fdata%2FMPEG-AI%2FVCM%2FPandaset%2F&data=05%7C02%7Cshanl%40global.tencent.com%7Ccd449e2f1aa44159c57f08dc7316b773%7Ca32856f21731405cb53d480e26413adf%7C1%7C0%7C638511788605563761%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C0%7C%7C%7C&sdata=EgxlQAFwJIoxpnTD9ea6h77d7q%2B6CmHb2Y3JKIbtEZA%3D&reserved=0).

The MD5 checksums of the Panda sequences in YUV420 format are provided in Table 6.

Table 6. MD5 checksum of the test sequences in Panda dataset

|  |  |  |
| --- | --- | --- |
| **Class** | **Sequence name** | **MD5 checksum** |
| A1 | 57 | 9130490082689713a69e575a80fb4ad6 |
| A1 | 58 | 1440c40a45549a2f445e7f77d68ade95 |
| A1 | 69 | 3910301cc88baf0784a857ab7311dfc7 |
| A1 | 70 | f2010059d5c7cc6994f6fe57f6c0b3b8 |
| A1 | 72 | 680260f1c9b02008e274ce95f9b695bb |
| A1 | 73 | 17901de39c20e19997428387797a9161 |
| A1 | 77 | 15fb8f97f221ad0d09ef712ba98f3b32 |
| B1 | 3 | b9adefa4b91950dff73d68e3084f0087 |
| B1 | 11 | 30d36645cd4f91392cfa6b2af591efe5 |
| B1 | 16 | 59deb96e6d0384d10afaa5ce10a90e26 |
| B1 | 17 | 206ad92bdfaf0cfa78825500b239908f |
| B1 | 21 | ec53bde219de6dc4561829654365cc42 |
| B1 | 23 | 0212ceac5619c8fd5ea83d39072fcd2f |
| B1 | 27 | 63b21eb8e620e19e61397321a7cc8023 |
| B1 | 29 | b97e9951a61340441d1ee7e69d420b9e |
| B1 | 30 | f0df4c2f6bb9101a34c74a738ad6d2bb |
| B1 | 33 | 5854fb8637c1712376722531ba3e44ee |
| B1 | 35 | 1ffc71a4adbad0c6404f9feb07703943 |
| B1 | 37 | 8dccf94cab1f9a1540e8a58dd1b6a3b9 |
| B1 | 39 | 2c17b499c276424ff661531e4f624208 |
| B1 | 43 | 622b88fc748ad35775c9dd11b912e1de |
| B1 | 53 | ec850c3929eb14e93819211a1a4bec25 |
| B1 | 56 | 4f90b8bcb8201e9c15d3913c078f125b |
| B1 | 97 | 3d2e46d5522ea4f15ff630435177a30d |
| C1 | 88 | 64ac63da2e4ad4e2138385ffe79eceac |
| C1 | 89 | 2c157de05be51e95a534b711f80f6ef5 |
| C1 | 90 | d10c40118ac9cb5aab853788f8bfefef |
| C1 | 95 | 534cb091fa6355c02239bbea7e055851 |
| C1 | 109 | 51bbec10cdbf61e9ae0136d4c3d76103 |
| C1 | 112 | 5f267a5086c69cd8580f5cccb7475ca4 |
| C1 | 113 | a09b649c860fc3f905e1ecead478af27 |
| C1 | 115 | 3fb89c43967c5a1faf1952e2b88f734c |
| C1 | 117 | 0adf67cd7d0c77961f57aae4dd528856 |
| C1 | 119 | a1ba6f0ef28ab8ad80e05676be07224b |
| C1 | 122 | b2b614dead35d3fbfa50b97530deee1a |
| C1 | 124 | 3ae37079ec6e32deb138dae6d3ad7832 |
| A2 | 64 | 840e927178cf14eff1a6b1a4eaf3a466 |
| A2 | 65 | 12075d8f069d017d7dd4c4358d8ef0fd |
| A2 | 66 | c582ba5619715acc9aead8487ee13ba7 |
| A2 | 67 | 404f8281d8b41a9eb454e54bf0491c51 |
| A2 | 71 | 87a48d966c8ba0a4e9ac24b647c0e9be |
| A2 | 78 | 11d3e7d488907424bc4cdd5bd2f306a5 |
| A2 | 149 | aac60318588523e07778107c8700bc3c |
| B2 | 1 | 5fc6b2c96a16d37a4094856e2f84a83b |
| B2 | 5 | b472fd9af8cec51457cddc3c233c00b0 |
| B2 | 13 | 0e74b2133e0cec1d79f75e6473fdaa9f |
| B2 | 15 | f22bcb4cca87f4efc11ff2227cf89790 |
| B2 | 19 | 9d7726b0f7ab826627b6dfccc4179212 |
| B2 | 24 | 5400c8c7e4be2d9d6d6bfe22dde300fa |
| B2 | 28 | f89cb49f08f27db2df2cd7fc9018eb71 |
| B2 | 32 | 7ced67a4ca3d801866d19d71443d3331 |
| B2 | 34 | a871e52e0038da9c7abd4a92214b504d |
| B2 | 38 | ac28704f6f74063f9e0a79591d4e636f |
| B2 | 40 | 20936234433ab631bed08ed5a6909c6a |
| B2 | 41 | 889c2719182256259fd67e49adb63e04 |
| B2 | 42 | 13965016f866c14bc3992acc8fcd69e8 |
| B2 | 44 | 932aa68b07ca51bcfdca7918a5da6267 |
| B2 | 46 | 176236bc25abc2333f9223f6cf36973a |
| B2 | 52 | a4ce077c7782b6a0dd74d036624868b4 |
| B2 | 54 | 41b24a74ac7444b8d7bd4b9ec4c7b6c4 |
| B2 | 139 | 7e802b6f216cd872e80d40bb25086b96 |
| C2 | 80 | cfb25bb1b0e21e58e8ed279cfd3b6f2a |
| C2 | 84 | f806095ed1dea3fe04c4cae9f4c9b4fc |
| C2 | 94 | 763a5f1218685d9ad3b1a55f44b6ee75 |
| C2 | 101 | b984600ebf0877b43a36266614b0967e |
| C2 | 102 | 775135212b887e48cb25e08bb4261a30 |
| C2 | 103 | e9aa3f4b3624152dff19451964e70d29 |
| C2 | 105 | 7804195c442963b721ea7e7ece7dcdcb |
| C2 | 106 | 38696f8e3a440f926742f8d1b84543c9 |
| C2 | 110 | 964c7860cc1a495c80c85a1cba0262bc |
| C2 | 116 | a2f5af8f1f67d01603f2775da64e2958 |
| C2 | 123 | 2b97973e20f1a77b00d87e78bed3316d |
| C2 | 158 | d26de8bfbb63aeed3529abf6e80fd138 |

## *Image Datasets (optional)*

In addition, three image datasets are included to provide optional information using all-intra configuration.

## Tencent Video Dataset (Image)

The Tencent Video Dataset (Image), or in short TVD-I, is an image dataset of 166 images of 1920x1080 resolution that have annotations for object detection and instance segmentation. The dataset with corresponding annotations is available at [2] <https://multimedia.tencent.com/resources/tvd>.

## OpenImages v6

The OpenImages dataset consists of around 9 million images. A subset of the validation set of its version 6 containing 5000 images are selected for testing object detection in this activity. The dataset with corresponding annotations is available at <https://storage.googleapis.com/openimages/web/index.html>. The list of images that are selected can be found in the script package included in [4].

## FLIR

The FLIR dataset used in the VCM group is a dataset consisting of 300 infrared images. The images, annotations and the fine-tuned model for thermal images can be found on the MPEG content repository (<https://content.mpeg.expert/data/>).

1. **Anchor**

## *Anchor software*

VCM-RS release v0.10 is used for the compression anchors for the VCM experiments. The VCM-RS is available at

<https://git.mpeg.expert/MPEG/Video/VCM/VCM-RS>.

Access right to the repository is granted to registered experts in the MPEG WG 4.

## *Anchor configuration*

The VCM-RS contains placeholders for some pre-inner codec and post-inner codec modules, as well as a neural-network based inner codec as an alternative for the inner codec module. These placeholders are by-passed when generating anchors. VTM 20.0 [1] is included in VCM-RS v0.10 and used as the basis for inner codec in VCM-RS.

Six configurations are defined for encoding video sequences with different GOP and latency strucures using the VCM-RS as follows:

* Random access for inner (RA\_inner)
* Random access for end-to-end (RA\_e2e)
* Low delay for inner (LD\_inner)
* Low delay for end-to-end (LD\_e2e)
* All intra for inner (AI\_inner)
* All intra for end-to-end (AI\_e2e)

Configuration files corresponding to the above configurations are provided together with VCM-RS for experiments and evaluations. The following parameters may be changed for generating compressed data for each test point:

* **input directory** to specify the directory where input image files are stored or where the frames of an input video are stored
* **output\_dir** to specify the output directory where the bitstream files and reconstructed images or frames are stored
* **directory\_as\_video** to specify the directory containing frames of an input video
* **FrameRate** to reflect the frame rate of a given video sequence
* **FrameSkip** to reflect the number of frames to be skipped
* **FramesToBeEncoded** to reflect the frame count of a given video sequence to be encoded
* **IntraPeriod** to reflect the intra refresh period in the random-access test cases. The intra refresh period is dependent on the frame rate of the source and the GOP size in use: a value 32 shall be used for sequences with a frame rate equal to 20fps, 24fps, 25fps and 30fps, 64 for 50fps, and 60fps, and 96 for 100fps.
* **quality** to reflect the quantization parameter value, corresponding to the quantization parameter (QP) value used in VVC
* **Configuration** to reflect the GOP and latency structure setting used for encoding

Note that in this CTC, GOP size for RA configuration is determined by the value of IntraPeriod. When IntraPeriod is greater than or equal to 16, GOP is set to 16, otherwise it is set to 8.

Some coding tools are implemented or setup differently for inner and end-to-end configurations, as illustrated in Table 7 and Table 8.

Table 7 Tool settings for inner configurations

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **RA\_inner** | **LD\_inner** | **AI\_inner** |
| Temporal Resampling | On | On | On |
| ROI | On | On | On |
| RoIAccumulationPeriod | 0 (Intra Period) | 32 (8 in inner codec) | 1 |
| RoIAccumulationWindowExtension | 0 | 0 | 0 |
| RoIRetargetingMode | sequence | sequence | sequence |
| RoIRetargetingMaxNumRoIs | 11 | 11 | 11 |
| Spatial Resampling | On | On | On |
| Bit Truncation | On | On | On |
| Neural Network-based In-loop Filter | On | On | On |
| GOP structure | “Random Access” | “Low delay” | “All Intra” |

Table 8 Tool settings for end-to-end configurations

|  |  |  |  |
| --- | --- | --- | --- |
| **Tool** | **RA\_e2e** | **LD\_e2e** | **AI\_e2e** |
| Temporal Resampling | On | Off | Off |
| ROI | On | On | On |
| RoIAccumulationPeriod | 0 (Intra Period) | 1 | 1 |
| RoIAccumulationWindowExtension | 0 | 32 | 0 |
| RoIRetargetingMode | dynamic+pad | dynamic+pad | dynamic+fit |
| RoIRetargetingMaxNumRoIs | 3 | 2 | 11 |
| Spatial Resampling | On | On | On |
| Bit Truncation | On | On | On |
| Neural Network-based In-loop Filter | On | On | On |
| GOP structure | “Random Access” | “Low delay” | “All Intra” |

## *Anchor generation*

The following input arguments shall be specified according to the CTC when using VTM 20.0 as the inner codec of VCM-RS v0.10.

--FrameRate <frame rate> \

--IntraPeriod <intra period> \

--FrameSkip <frame skip> \

--FramesToBeEncoded <number of frames to be encoded> \

--quality <quality> \

--output\_dir <output directory> \

<input yuv file>

Six equally spaced QP values with an interval of 4 are used to generate the anchor results for SFU-HW dataset and TVD. For example, QP1=QP0+4, QP2=QP1+4=QP0+8, …, QP5=QP4+4=QP0+20. The starting QPs (QP0) for SFU-HW dataset and TVD are provided in Table 9 and Table 10, respectively. For Panda dataset and Image dataset, the interval is 5. The QP values for generating the anchor results for Panda dataset can be found in Table 11, and the QP values for generating the anchor results for image datasets can be found in Table 12.

Table 9. Starting QPs (QP0) for sequences in SFU-HW dataset to generate anchor

|  |  |  |  |
| --- | --- | --- | --- |
| **Class** | **Sequence name** | **RA/LD\_inner RA/LD\_e2e** | **AI\_inner/e2e** |
| QP0 | QP0 |
| A | Traffic | 32 | 28 |
| B | ParkScene | 20 | 18 |
| B | BasketballDrive | 28 | 18 |
| B | BQTerrace | 28 | 24 |
| C | BasketballDrill | 18 | 16 |
| C | BQMall | 26 | 26 |
| C | PartyScene | 22 | 16 |
| C | RaceHorsesC | 20 | 14 |
| D | BasketballPass | 16 | 14 |
| D | BQSquare | 16 | 14 |
| D | BlowingBubbles | 18 | 16 |
| D | RaceHorsesD | 18 | 14 |
| O | Kimono | 26 | 24 |
| O | Cactus | 32 | 20 |

Table 10. Starting QPs (QP0) for sequences in TVD to generate anchor

|  |  |  |  |
| --- | --- | --- | --- |
| **Sequence name** | **RA/LD/AI\_inner** | **RA/LD\_e2e** | **AI\_e2e** |
| QP0 | QP0 | QP0 |
| TVD-01-1 | 18 | 14 | 16 |
| TVD-01-2 | 18 | 18 | 16 |
| TVD-01-3 | 22 | 20 | 16 |
| TVD-02-1 | 20 | 20 | 22 |
| TVD-03-1 | 26 | 26 | 24 |
| TVD-03-2 | 26 | 26 | 24 |
| TVD-03-3 | 24 | 24 | 24 |

Table 11. QPs for Panda datasets to generate anchor

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Class** | **All Configuration** | | | | | |
| QP0 | QP1 | QP2 | QP3 | QP4 | QP5 |
| A1 | 22 | 27 | 32 | 37 | 42 | 47 |
| B1 | 22 | 27 | 32 | 37 | 42 | 47 |
| C1 | 22 | 27 | 32 | 37 | 42 | 47 |
| A2 | 22 | 27 | 32 | 37 | 42 | 47 |
| B2 | 22 | 27 | 32 | 37 | 42 | 47 |
| C2 | 22 | 27 | 32 | 37 | 42 | 47 |

Table 12. QPs for image datasets to generate anchor

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Dataset** | **AI Configuration** | | | | | |
| QP0 | QP1 | QP2 | QP3 | QP4 | QP5 |
| OpenImageV6 | 22 | 27 | 32 | 37 | 42 | 47 |
| FLIR (IR) | 22 | 27 | 32 | 37 | 42 | 47 |
| TVD-I | 22 | 27 | 32 | 37 | 42 | 47 |







1. **Evaluation**

## *Evaluation pipeline*

Figure 1 and Figure 2 show the pipeline for evaluation of proposed VCM technologies for image and video datasets, respectively. These pipelines consist of the following stages:

1. For video datasets, the original video (in YUV420 format) and for image datasets, images (in PNG format) are encoded by the VCM-RS encoder into bitstreams. Encoding can be conducted on one GPU, or more than one GPUs simultaneously. The scripts for enabling multi-GPU encoding are available at https://git.mpeg.expert/MPEG/Video/VCM/VCM-RS/-/blob/main/Scripts/VTM\_InnerCodec/run\_all.sh.
2. Bitstreams are decoded by the VCM-RS decoder to produce decoded video (in YUV420 format) and images (in PNG format).
3. For video datasets, the task network is performed using the decoded video after conversion from YUV420 to PNG format. For image datasets decoded PNGs are directly supplied to the task network to produce a task network output.
4. A task evaluator produces a task result using the task network output and the ground truth.
5. An objective visual evaluator produces Y-PSNR, U-PSNR, V-PSNR, and RGB-PSNR results using the decoded video (image) and the original video (image).

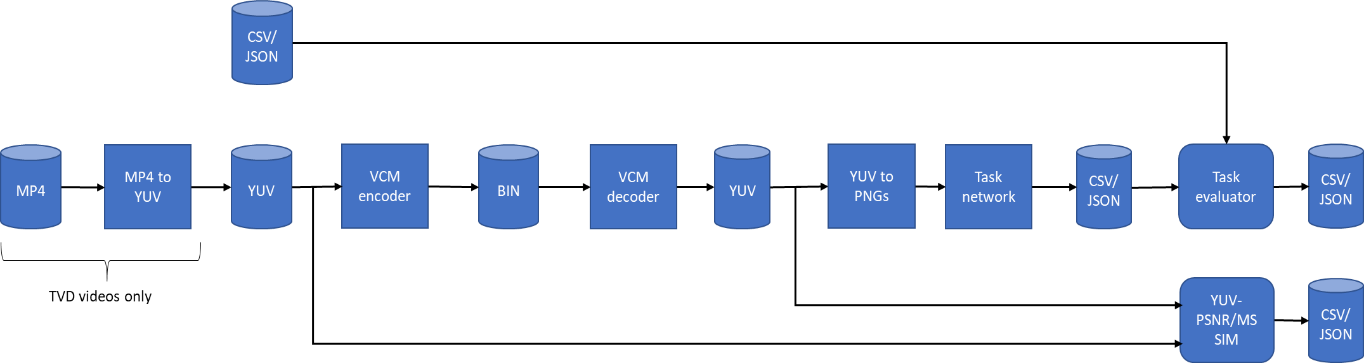


Figure 1. Pipeline of VCM evaluation framework for video datasets.

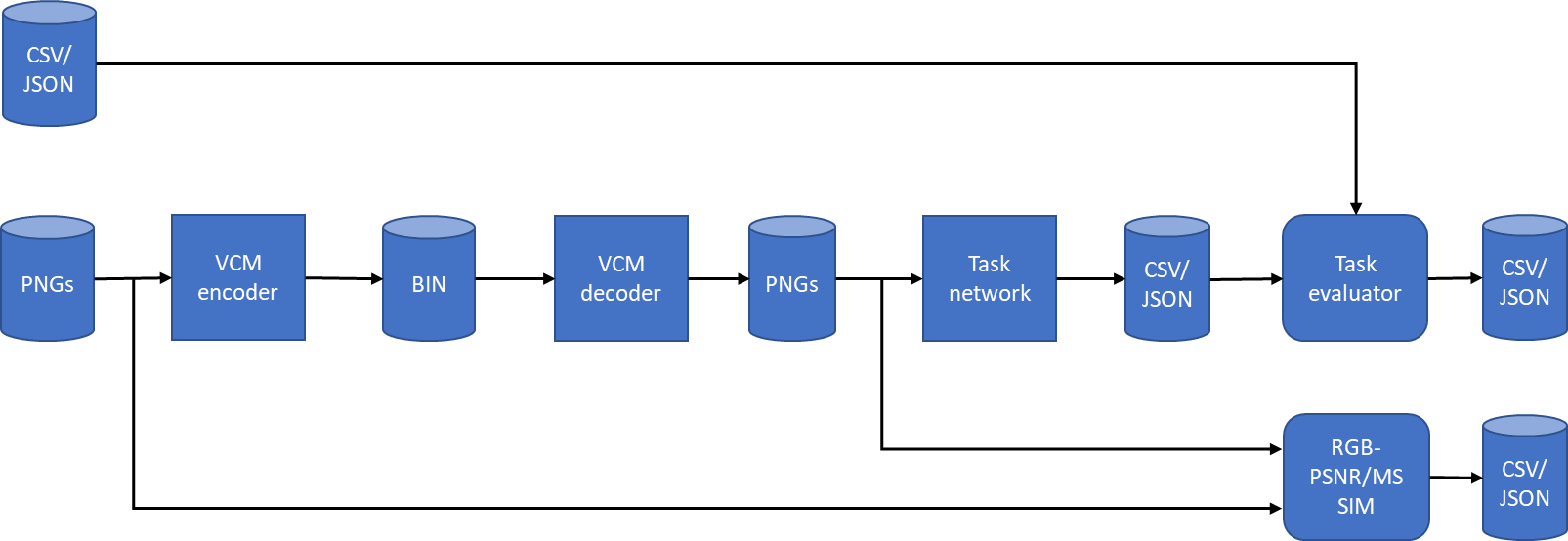


Figure 2. Pipeline of VCM evaluation framework for image datasets.

The video dataset pipeline of Fig. 1 is applied to the following datasets:

* SFU-HW Dataset
* Tencent Video Dataset (TVD)

The image dataset pipeline of Fig. 2 is applied to the following datasets:

* Tencent Video Dataset (Image)
* OpenImages v6
* FLIR

Note that the SFU-HW videos are provided in YUV420 format, while the TVD videos are provided in MP4 format and require conversion to YUV420 before encoding by VCM-RS.

FFmpeg [3] should be used when format conversion is needed for coding and evaluation. Some example command lines are as follows.

* png to yuv: ffmpeg -i {input.png} -f rawvideo -pix\_fmt yuv420p -dst\_range 1 {output.yuv}
* yuv to png (per frame): ffmpeg -f rawvideo -s {width}x{height} -pix\_fmt yuv420p10le -i {input.yuv} -vsync 1 -y -pix\_fmt rgb24 {outputPngfolder}/%06d.png

## *Task networks*

## Object detection

The following network architecture is used for the evaluation of object detection task.

* Faster R-CNN [3] X101-FPN (part of Facebook AI Research’s Detectron2 [4])

Model parameters file is ‘model\_final\_68b088.pkl’ available [here](https://dl.fbaipublicfiles.com/detectron2/COCO-Detection/faster_rcnn_X_101_32x8d_FPN_3x/139173657/model_final_68b088.pkl). Note that FLIR anchor is generated using retrained model parameters that are available in the MPEG Content Repository.

## Object tracking

The following network architecture is used for the evaluation of object tracking task.

* JDE-1088x608 [5]

Model parameters file is ‘jde.1088x608.uncertainty.pt’ available [here](https://drive.google.com/open?id=1nlnuYfGNuHWZztQHXwVZSL_FvfE551pA) or [here](https://pan.baidu.com/s/1Ifgn0Y_JZE65_qSrQM2l-Q).

## Semantic segmentation

The following network architecture is used for the evaluation of semantic segmentation task.

* Panoptic FPN R-101-FPN 3x (part of Facebook AI Research’s Detectron2 [4])

Model parameter file is available [here](https://dl.fbaipublicfiles.com/detectron2/COCO-PanopticSegmentation/panoptic_fpn_R_101_3x/139514519/model_final_cafdb1.pkl).

1. **Test conditions**

## *Fixed QP*

For the proposals with a quantization concept, results shall be provided using the six quantization parameter values that match to the anchor’s QP values. These values define the initial QP values that are specified as the input QP of the proposals. Adaptation of QP during compression and processing shall be described in the proposals. If the task R-D performance of a proposed method does not overlap with that of the anchor, and thus impacting the BD-rate calculation, extra QP points shall be used to generate R-D task performance results within a range that overlaps sufficiently with the anchor.

## *Bit-rate Target*

For proposals that do not have a quantization concept substantially similar to the anchor, such as so-called super-resolution and/or end-to-end (E2E) architectures, results shall be provided for six rate-distortion points for each sequence corresponding to the anchor quantization values. Each of the six provided points is requested to be ±10% of the rate of the corresponding quantization parameter of the anchor. This rate comparison shall be performed on a sequence basis.

1. **Training conditions**

A proposal can train the compressor with various training materials. The training materials shall be publicly available with license terms that allow commercial usage or can be used by standards committees or standardization activities. No images or videos from the VCM validation dataset shall be used in training. It is required that the usage of any training materials be described in the contribution for reproducing training results.

Table 11 describes a list of recommended training sets. It is desirable that all proposals use the specific training sets. Results using sequences not in the list of specific sequences may also be provided as supplemental information.

Table 11 Recommended source for training datasets

|  |  |
| --- | --- |
| Database | Location |
| OpenImageV6-train+ | <https://storage.googleapis.com/openimages/web/index.html> |
| TVD-train\* | <https://multimedia.tencent.com/resources/tvd> |
| BVI-DVC dataset | [https://vcgit.hhi.fraunhofer.de/jvet-ahg-nnvc/nnvc-ctc](https://linkprotect.cudasvc.com/url?a=https%3a%2f%2fvcgit.hhi.fraunhofer.de%2fjvet-ahg-nnvc%2fnnvc-ctc&c=E,1,0XHkUYpmU7-yMa8rpRq1nhSLpJozbvEcO_0pARY1q7bt3BjfM8kuKR0XKj8LdXiXgGQKaUjshaZoRhbQIccDHXv7ETDJpY52qahVgN_spUtLhNT7yw,,&typo=1) |

+OpenImageV6-train: the “train” subset of OpenImageV6.

\*TVD-train: TVD video dataset excluding sequences listed in Appendix.

In general, if an image or video, or part of it is used for testing, it shall not be used for training.

1. **Reporting**

## *Machine task performance*

* ***Bitrate measurement***

For image datasets, bits per pixel (BPP) shall be used. For video sequences, the bitrate shall be measured in kilobits per second (kbps).

* ***PSNR***

The Peak Signal to Noise Ratio.

* ***SSIM***

The Structural Similarity.

* ***MS-SSIM***

The Multi-scale Structural Similarity.

* ***mAP for detection and segmentation tasks***

For both object detection and object segmentation, mean Average Precision (mAP) shall be used to measure the performance of the detection results. Please refer to [6] for more information about mAP.

* ***MOTA for tracking tasks***

For the object tracking task, Multiple Object Tracking Accuracy (MOTA) shall be used to measure the tracking performance. Please refer to [6] for more information about MOTA.

* ***MIoU for semantic segmentation tasks***

For semantic segmentation task, Mean Intersection over Union (MIoU) shall be used to measure the performance.

Assuming *k* is the number of categories, MIoU is calculated as

where refers to the number of pixels inferred from category as category , refers to the true positive. and refer to the false positive and false negative, respectively.

* ***BD-rate machine task calculation***

A cubic polynomial function-based curve fitting method is applied to the mAP or MOTA score values before they are used to calculate the machine task BD-rate performance in the reporting template. This curve fitting process is automated using the scripts provided in <http://mpegx.int-evry.fr/software/MPEG/Video/VCM/vcm-ctc/scripts> directory. After curve fitting, the middle four QP points are used for calculating BD-rate, i.e., the lowest and the highest QPs are excluded from the BD-rate calculation.

## *Runtime*

Runtime including Encoding time (EncT) and Decoding time (DecT) shall be reported for complexity measurement. The proposed runtime methods for a VCM solution are:

* **EncT:** Time needed to convert input video or image to bitstream.
* **DecT:** Time needed to convert bitstream to reconstructed output video or image.

For the purpose of reporting encoding and decoding runtimes, the anchor and proposal shall be simulated on the same platform, e.g. similar CPU and GPU configuration, to have reliable time comparison. It is encouraged to report runtimes of modules, e.g., Temporal Resample, Spatial Resample, RoI, Bit-depth truncation and Inner codec, etc.

## *Inference information*

In additional to encoding and decoding time, the information described below is required to be provided for the inference process for both encoding and decoding process.

* **Network Visualization:** Graphical representation of the neural network
* **Param. Number**: Total numbers of parameters in the neural network.
* **Param. Precision**: Bits for storing one parameter. Besides, using “I” for indicating the integer, and using “F” for indicating the floating number. For example, if the proposed method uses 16-bit integer to represent a parameter, you can report this information as “16 (I)”.
* **MAC (Kilo)**: Number of multiply–accumulate operations in inference stage per pixel, where the multiply–accumulate operation is a common step that computes the product of two numbers and adds that product to an accumulator.
* **Mem.T (MB)**: Temporary memory. It denotes the memory used to store the output feature map in each intermediate layer (forward pass). Since different size of input may influence the value, it is suggested to use 3840x2160 as the input size for unification, if there is no parallel operation. Or, if block level parallel operation is used in the proposed method, the block size can be used as the input for calculation, while the input size should be reported. For reporting Mem.T (MB) the calculation process is also suggested to be provided for crosschecking.
* **Patch Size**: size of input to the neural networks during inference (patchW×patchH×patchT, e.g. 64x64x3)

## *Training information*

It is required to report and discuss the following information for the training process.

* **Epoch**: The number of complete passes through the training data (e.g. 100)
* **Batch Size**: The number of samples processed before the model is updated. (e.g. 4Kx16frames)
* **Training Time**: CPU and/or GPU (e.g. 48h)
* **Learning Curve:** Plot of the training loss and validation loss (or similar) versus the number of epochs
* **Training Sets**: Training sets used. If a pre-trained model is used, the source of the pre-trained model and its training sets should be reported in detail.
* **Training Configuration per Rate-Distortion Point**: Any changes in the requested information used to generate different rate-distortion points. Additional training information could also help to better understand the proposed neural network-based method and thus encouraged to be included in the contribution.
* **Number of Iterations:** number of gradient updates within an epoch
* **Patch Size**: size of input to the neural networks (patchW×patchH×patchT, e.g. 64x64x3)
* **Learning Rate**: The amount that the weights are updated during training (e.g. 5e-4)
* **Optimizer**: The algorithm used to change the attributes of proposed neural networks (e.g. ADAM)
* **Loss Function**: The function to calculate the model error during training and optimization (e.g. L1, L2, etc.)
* **Preprocessing**: (e.g. preprocessing procedure, normalization, cropping method, rotation, zoom etc.)

1. **Cross-check criteria**

When a technology is proposed in VCM, it shall pass a successful cross-check by an independent expert not affiliated with the proponent for consideration of adoption. The cross checker shall report bitrate, machine task performance, encoding decoding and runtimes.

The successful cross-check criteria are as follows:

* The bitrates reported by the proposal and cross-check report shall match exactly.
* The difference in machine task performance between results provided by the proponent and the cross-checker shall be less than 1%. For example, any machine task performance value between 9.9% and 10.1% in the cross-check report is considered successful when the reported machine task performance value in the proposal is 10%.

1. **Reference**
2. ISO/IEC JTC1/SC29/WG5, "VVC Reference Model (VTM)," <https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM.git>.
3. X. Xu, S. Liu and Z. Li, "A Video Dataset for Learning-based Visual Data Compression and Analysis," in 2021 International Conference on Visual Communications and Image Processing (VCIP), Dec. 2021.
4. S. Ren, K. He, R. Girshick, et al. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.
5. Y. Wu, A. Kirillov, F. Massa, et al. "Detectron2," <https://github.com/facebookresearch/detectron2>
6. Z. Wang, L. Zheng, Y. Liu, et al. "Towards real-time multi-object tracking," in European Conference on Computer Vision (ECCV). 2020: 107-122.
7. ISO/IEC JTC1/SC29/WG4, “Evaluation framework and methodologies for video coding for machines,” N00277, Mainz, October 2022.
8. <https://git.mpeg.expert/MPEG/Video/VCM/VCM-RS>.

# Appendix : TVD-train dataset

TVD-train is a training dataset consisting video sequences from TVD excluding sequences listed below:

\*MovingBikes\_3840x2160\_50fps\_8bit\_420

FallingLeaves\_3840x2160\_50fps\_8bit\_420

FallenLeaves\_3840x2160\_50fps\_8bit\_420

GirlThrowingLeaves\_3840x2160\_50fps\_8bit\_420

GirlWalkingOnStreet\_3840x2160\_50fps\_8bit\_420

GirlWatchingPhone\_3840x2160\_50fps\_8bit\_420

StaticWaterAndBikes2\_3840x2160\_50fps\_10bit\_420

GirlRunningOnGrass\_3840x2160\_50fps\_10bit\_420

GirlsOnGrass1\_3840x2160\_50fps\_10bit\_420

GirlsOnGrass2\_3840x2160\_50fps\_10bit\_420

\*PeopleOnGrass\_3840x2160\_50fps\_10bit\_420

\*MovingBikesAndPedestrian4\_3840x2160\_50fps\_10bit\_420

BoyMakingUp1\_3840x2160\_50fps\_10bit\_420

BoyMakingUp2\_3840x2160\_50fps\_10bit\_420

BoyDressing1\_3840x2160\_50fps\_10bit\_420

BoyDressing2\_3840x2160\_50fps\_10bit\_420

BoyWithCostume\_3840x2160\_50fps\_10bit\_420

MountainsAndStairs1\_3840x2160\_24fps\_10bit\_420

MountainsAndStairs4\_3840x2160\_25fps\_10bit\_420

GirlWithTeaSet1\_3840x2160\_25fps\_10bit\_420

GirlWithTeaSet2\_3840x2160\_25fps\_10bit\_420

GirlWithTeaSet3\_3840x2160\_25fps\_10bit\_420

CableCar\_3840x2160\_25fps\_10bit\_420

HotelClerks\_3840x2160\_25fps\_10bit\_420

RestaurantWaitress1\_3840x2160\_25fps\_10bit\_420

RestaurantWaitress2\_3840x2160\_25fps\_10bit\_420

ChefCuttingUp1\_3840x2160\_25fps\_10bit\_420

ChefCuttingUp2\_3840x2160\_25fps\_10bit\_420

ChefCooking2\_3840x2160\_25fps\_10bit\_420

ChefCooking3\_3840x2160\_25fps\_10bit\_420

ChefCooking4\_3840x2160\_25fps\_10bit\_420

ChefCooking5\_3840x2160\_25fps\_10bit\_420

RawDucks\_3840x2160\_25fps\_10bit\_420

PeopleNearDesk\_3840x2160\_25fps\_10bit\_420

HotPot\_3840x2160\_25fps\_10bit\_420

RiverAndTrees\_3840x2160\_25fps\_10bit\_420

BuildingTouristAttraction2\_3840x2160\_25fps\_10bit\_420

BuildingTouristAttraction3\_3840x2160\_25fps\_10bit\_420

RoomTouristAttraction1\_3840x2160\_25fps\_10bit\_420

RoomTouristAttraction2\_3840x2160\_25fps\_10bit\_420

RoomTouristAttraction3\_3840x2160\_25fps\_10bit\_420

RoomTouristAttraction4\_3840x2160\_25fps\_10bit\_420

RoomTouristAttraction5\_3840x2160\_25fps\_10bit\_420

RoomTouristAttraction6\_3840x2160\_25fps\_10bit\_420

RoomTouristAttraction7\_3840x2160\_25fps\_10bit\_420

BlackBird\_3840x2160\_25fps\_10bit\_420

StampCarving1\_3840x2160\_25fps\_10bit\_420

StampCarving2\_3840x2160\_25fps\_10bit\_420

Weave\_3840x2160\_25fps\_10bit\_420

ManWithFilmMachine\_3840x2160\_25fps\_10bit\_420

LyingDog\_3840x2160\_25fps\_10bit\_420

OilPainting1\_3840x2160\_25fps\_10bit\_420

OilPainting2\_3840x2160\_25fps\_10bit\_420

# \*These sequences are used in video task testing; while other sequences contain extracted frames used for image task testing.