ISO/IEC JTC 1/SC29/WG1 N100250
96th JPEG Meeting, Online, 25-29 July 2022

ISO/IEC JTC 1/SC 29/WG 1
(ITU-T SG16)

Coding of Still Pictures

JBIG
Joint Bi-level Image Experts Group

JPEG
Joint Photographic Experts Group

TITLE:	Report on the JPEG AI Call for Proposals Results
SOURCE:	REQ Group
STATUS:	Final
REQUESTED ACTION:	Distribute
DISTRIBUTION:	Public

Contact:
ISO/IEC JTC 1/SC 29/WG 1 Convener – Prof. Touradj Ebrahimi
EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland
Tel: +41 21 693 2606, Fax: +41 21 693 7600, E-mail: Touradj.Ebrahimi@epfl.ch

1 Purpose of This Document

This document reports the JPEG AI Call for Proposals (CfP) responses evaluation results for the standard reconstruction task. In total 10 valid codecs were evaluated in several aspects, namey the decoded quality provided for the bitrates under consideration and the decoding complexity. These deep learning-based image codecs employed different tools and methods but also training methodologies and loss functions. The submissions have covered all parts of learning-based image codecs, such as the use of attention layers in the non-linear transformation, RD optimized quantization and context-based probability models for entropy coding. All JPEG AI CfP proponents have submitted a detailed technical description of the entire image codec, as well as encoder and decoder implementations in software, and the decoded test images, according to the timeline defined in the JPEG AI CfP document [1]. The documents that describe all submissions are available as input documents in the 96th JPEG meeting [2]-[10].

The rate-quality tradeoff was measured using several objective quality metrics and a subjective assessment test. The subjective evaluation was performed using a crowdsourcing approach, using a web platform for presenting the images and collecting the votes using a double-stimulus methodology.

2 Standard Reconstruction Test Set

To avoid overfitting, the test set for JPEG AI CfP was kept hidden and was announced only after all decoders and models have been submitted; write access to the upload location was closed after submission date. The JPEG AI CfP test set was created by a hidden test set committee formed by Fernando Pereira, Thomas Richter and João Ascenso. The JPEG AI CfP test set consist of 21 different images of different resolutions, that are shown in Figure 1. The spatial resolutions of each image are shown in Table 1.

Figure 1. *JPEG AI test images.*

Table 1. *JPEG AI test images spatial resolution.*

Number	Width	Heigth	Megapixels
1	1192	832	0.99
2	1280	848	1.09
3	3032	1856	5.63
4	1920	1080	2.07
5	3680	2456	9.04
6	2192	1520	3.33
7	1248	832	1.04
8	2464	1640	4.04
9	1536	1024	1.57
10	1984	1320	2.62
11	1784	1296	2.31
12	3680	2456	9.04
13	800	1200	0.96
14	2000	2496	4.99
15	976	1472	1.44
16	560	888	**0.50**
17	1752	1856	3.25
18	7680	5120	**39.32**
19	2120	1608	3.41
20	1072	928	0.99
21	2048	1080	2.21

There are two images (#14 and #21) that belong to the screen content category. Those images require special coding tools, which according to the JPEG AI Common Training and Test Conditions (CTTC) [11] were enabled for HEVC, but not enabled for VVC. Since screen content images were not sufficiently represented in the JPEG AI training set, which all proponents are mandated to use, learning-based image coding solutions trained on the JPEG AI training set are not expected to work with high performance. Thus, natural captured and screen content results are reported independently. Regarding training, all the learning-based codecs should have been trained with a procedure that uses the JPEG AI training and validation sets, which was made previously available to all proponents with CC0 licensing. More information is available in the JPEG AI CTTC document [11].

3 Anchors and Bitrates

The encoding and decoding of the JPEG AI anchors are defined the JPEG AI CTTC, namely the software used, encoder configurations, color space conversions, etc. The JPEG AI anchors are the following:

- JPEG
- JPEG 2000
- HEVC
- VVC

The JPEG codec was not used for subjective evaluation due to its poor coding efficiency. All rate points listed in the JPEG AI CTTC, this means 0.03, 0.06, 0.12, 0.25, 0.50, 0.75, 1.0, 1.5, 2.0 bpp were included in the objective assessment that is reported in the next Section. However, it should be noted that only 5 rate points are mandatory for CfP response, namely 0.06, 0.12, 0.25, 0.50, 0.75 bpp. Moreover, it was defined not to exceed target rate more than 10% to make simple subjective assessment with stimuli at similar bitrates.

4 Objective Assessment of JPEG AI CfP Submissions

According to the JPEG AI CTTC, the objective assessment procedure for the standard reconstruction task includes in total 7 metrics for objective performance evaluation: MS-SSIM, IW-SSIM, VIF, PSNR-HVS-M, NLDP, FSIM, VMAF. Other aspect that was also assessed regards the coding complexity which is measured by several complexity metrics, namely: 1) number of parameters (weights) for the size of the largest model and the total number of parameters for all models, including models for all mandatory rate points; 2) running time with CPU only (mandatory) and with GPU enabled (recommended), for both encoder and decoder; 3) MAC operations, number of Multiply

Accumulate operations per sample (kilo), for encoder (submitted bitstreams) and decoder (worst case) operations. The decoding run time should be relative to anchor using the same CPU. In case it is supported, the decoding time should also be reported for GPU platforms to assess the potential for this type of hardware. The hidden test was used to evaluate the objective performance of all the learning-based codecs submissions, namely the camera captured images, computed for the 5 mandatory target rates 0.06, 0.12, 0.25, 0.5, 0.75 bpp. To have compact results, the average BD-rate performance across all 7 quality metrics is used in the following results. The JPEG AI anchors comparison for their decoder run-time (CPU) and BD-rate is shown in Table 2.

Table 2. *JPEG AI anchors BD-rate performance and decoding complexity.*

	BD-rate Performance			CPU Dec. Time		
	J2K	HEVC	VVC	J2K	HEVC	VVC
J2K	**0.0%**	**50.0%**	**71.2%**	1.0	1.0	0.8
HEVC	**-28.8%**	**0.0%**	**13.1%**	1.0	1.0	0.8
VVC	**-37.6%**	**-11.3%**	**0.0%**	1.3	1.3	1.0

The BD-rate performance and decoder run-time relatively to JPEG AI anchors for all CfP submitted learning-based image codecs are summarized in Table 3. The experimental results were obtained for the 3rd phase submission of the CfP are used, since some submissions were updated relatively to the 2nd phase submission.

Table 3. *JPEG AI CfP BD-rate performance assessment and CPU/GPU decoding time.*

TEAMID	BD-rate performance			CPU dec. time			GPU dec. time
	J2K	HEVC	VVC	J2K	HEVC	VVC	HEVC
TEAM12	**-39.3%**	**-13.2%**	**-3.1%**	601	606	484	NA
TEAM13	**-31.5%**	**-2.1%**	**10.6%**	21	21	16	1.9
TEAM14	**-57.2%**	**-39.6%**	**-32.3%**	39	39	31	7.4
TEAM15	**-6.7%**	**33.6%**	**51.2%**	25	25	19	NA
TEAM16	**-47.7%**	**-26.6%**	**-17.9%**	44	44	34	0.7
TEAM17	**-21.5%**	**15.4%**	**32.0%**	98	98	75	25.0
TEAM19	**-34.2%**	**-4.4%**	**8.6%**	21	21	16	2.3
TEAM21	**-33.4%**	**1.6%**	**13.8%**	153	153	118	NA
TEAM22	**-32.6%**	**-4.9%**	**7.2%**	136	136	105	NA
TEAM24	**-56.5%**	**-37.4%**	**-29.9%**	44	44	34	0.7

Regarding other complexity metrics, namely the decoding kMAC/px and model size (for the largest model and for all the models) is reported in Table 4. The computational complexity measured in kMAC/pxl is much higher than can be supported by modern GPU (such as RTX 3080) to ensure 30 img/s of 4K resolution decoding (128 kMAC/pxl). The largest model size for single image decoding varies from 4 Million parameters (TEAM22) to 208 Million (TEAM13 and TEAM19). The total size of all models is the largest for TEAM16 (428 Million parameters) and for TEAM12 (479 Million parameters).

Table 4. *JPEG AI CfP complexity assessment (for decoding kMAC/pxl and model size) and training set used.*

TEAMID	Decoding complexity			Training set
	kMAC/pxl	Largest Model Size	Total Model Size	
TEAM12	no data	47	479	CLIC, LIU4K
TEAM13	419	208	344	JPEG AI
TEAM14	1266	38	152	JPEG AI

TEAM15	1262	13	39	vimeo-90k
TEAM16	576	20	428	JPEG AI
TEAM17	961	40	40	JPEG AI
TEAM19	478	208	344	JPEG AI
TEAM21	1348	25	59	JPEG AI
TEAM22	281	4	17	Flicker2W
TEAM24	593	20	326	JPEG AI

The cross-check of the results reported by each team was also performed, namely using different hardware platform, e.g. the CPU used for decoding may not be the same as the CPU for encoding and in case is supported, GPU could be used for decoding. The decoding of submitted bitstreams was made by each proponent in a cross-check fashion, this means that proponent A has decoded the bitstreams of proponent B and has measured the bitstream size and objective quality and vice-versa. When the performance assessment results reported by proponent and cross-checker are very similar (BD-rate less than 0.5%) and no decoder crash was reported by cross-checker, the proposal passed the cross-check. In total 5 teams have successfully passed cross-check: TEAM13, TEAM14, TEAM16, TEAM19 and TEAM24 as shown. In the CfP submissions, it was also observed that at least three teams used different training data from the one specified in the CTTC, which is the JPEG AI training set. Table 5 shows a summary of the objective evaluation of all the proposals, which are sorted in descending order using the BD-rate performance relatively to the VVC Intra anchor.

Table 5. *BD-rate performance relatively to VVC Intra anchor and cross-checking results.*

TEAMID	BD-rate vs VVC	Passed cross-check
TEAM14	**-32.3%**	**YES**
TEAM24	**-29.9%**	**YES**
TEAM16	**-17.9%**	**YES**
TEAM12	**-3.1%**	NO
TEAM22	**7.2%**	NO
TEAM19	**8.6%**	**YES**
TEAM13	**10.6%**	**YES**
TEAM21	**13.8%**	NO
TEAM17	**32.0%**	NO
TEAM15	**51.2%**	NO

5 Subjective Assessment of the JPEG AI CfP Submissions

To evaluate the performance of the JPEG AI CfP submissions, a crowdsourcing-based subjective test was performed, which was designed according to the JPEG AI CTTC, respecting the decisions made on the 95th JPEG meeting about the subjective evaluation procedure of JPEG AI proposals, as defined in [12].

5.1 Subjective Evaluation Methodology

The subjective assessment of all the learning-based image codec submissions was made according to a Double Stimulus Continuous Quality Scale (DSCQS) protocol. The subjects were presented with the original image and the impaired decoded image displayed side by side, where both images are rated with a continuous scale, as shown in Figure 2. This scale is divided into five equal lengths and is compliant with the ITU-R five-point quality scale,

namely: “Excellent”, “Good”, “Fair”, “Poor”, and “Bad”. The DSCQS method requires assessment of both the original and impaired versions for each test image. The observers are not aware which one is the reference image, and the position of the reference image (left of right) is changed in pseudo-random order. The subjects evaluate the overall quality of the original and decoded images by inserting a mark on the vertical scale. The two vertical scales are presented in pairs to reflect the double stimuli nature of the subjective test. The methodology follows BT500-14 [13] and a randomized presentation order of the stimuli, as described in ITU-T P.910 [14] was used; the same content is never displayed consecutively. There was no presentation or voting time limit per comparison.

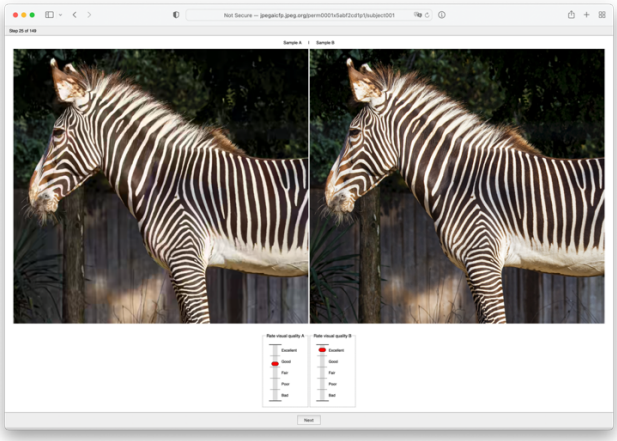

Figure 2. *Stimuli presentation and voting for the JPEG AI test image #11.*

Before the stimuli are scored by the subject, the display resolution was validated using the information obtained from the subject’s web-browser. If the display resolution was smaller than 1920×1080 or HiDPI/Retina mode was enabled subjects could not proceed. Figure 3 shows the web page that is shown to check that the display resolution meets the requirements.

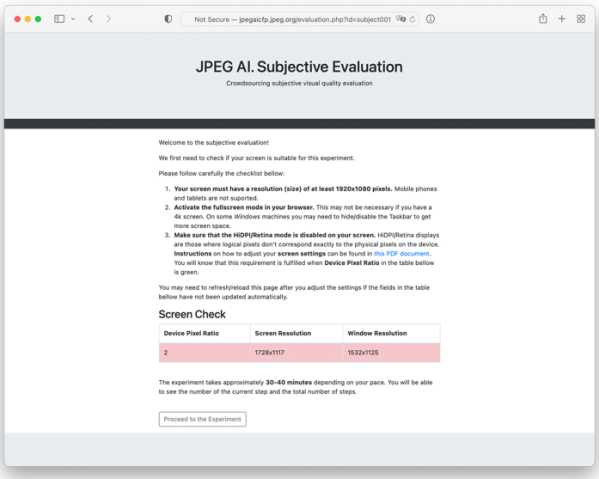

Figure 3. *Display resolution validation page.*

A training session took place at the beginning of the subjective test to familiarize participants with artefacts and distortions in the test images and the voting interface. Three training pairs of images were shown before the evaluation, representing “Bad”, “Excellent”, and “Fair” quality (in this order). Dishonest users are usually present on commercial crowdsourcing platforms such as Amazon Mechanical Turk which rate users by their success rate of finished jobs and thus must be detected. To detect these users, honeypot (or dummy) comparisons, where both images were high quality originals, were inserted in each session. The total number of the images to assess in the experiment was 416, a rather high number because of the number of codecs (13, 10 submissions and 3 anchors), rates (4) and

images (8). The evaluation of this high number of stimuli may take longer than two hours and thus cause fatigue. In this case, the stimuli were divided in three parts which resulted in three sessions of 142, 142, and 141 comparisons each, after including three honeypots per session. Each subject only did one session. After, the data from the three sessions was aggregated.

The QualityCrowd3 crowdsourcing framework [15] was used in the evaluation. This is a web-based framework that allows presenting the stimuli and collecting the votes. The QualityCrowd3 satisfies the requirements of the procedure that was defined in the JPEG AI CTTC and was previously used in the CfE proposal evaluation. All the subjects were recruited using AWS Mechanical Turk platform. All subjects were rewarded with 8 USD for completing the experiment. A virtual machine running Linux operating system (Ubuntu 22.04 LTS) was created to host the QualityCrowd3 platform. In addition, a third-level domain name was registered for the host: jpegaicfp.jpeg.org. QualityCrowd3 was deployed on jpegaicfp.jpeg.org and configured for receiving external subjects from AWS Mechanical Turk. Moreover, to use AWS Mechanical Turk service, an account named WG1JPEG was registered on aws.amazon.com and requester.mturk.com.

5.2 Test Image Selection and Cropping

Some selected images from the JPEG AI CfP Test Set (defined in Section 2) were selected and cropped to be used in the subjective evaluation to fit into the screen, in a side-by-side comparison. First, 10 images were selected by the JPEG AI hidden test set committee and some cropping area selection was proposed by the same committee. Then, based on the results of the subjective quality assessment of the JPEG AI anchors [16], which was performed for these 10 images, 8 images were selected. The images from the JPEG AI CfP test set were cropped according to the parameters presented in Table 5. Figure 4 illustrates the cropped regions in the selected images. The full reconstructed decoded images were provided by the proponents in a lossless PNG format, which were used for cropping purposes.

Table 5. *Definition of the cropped image for each JPEG AI test image that was used in the subjective evaluation.*

Test image file name	Crop top left coordinate (X,Y)	Crop width	Crop height
00001_TE_1192x832_8bit_sRGB.png	(0,0)	945	832
00007_TE_1248x832_8bit_sRGB.png	(152,0)	945	832
00009_TE_1536x1024_8bit_sRGB.png	(154,71)	945	880
00011_TE_1784x1296_8bit_sRGB.png	(180,31)	945	880
00012_TE_3680x2456_8bit_sRGB.png	(1679,350)	945	880
00015_TE_976x1472_8bit_sRGB.png	(15,592)	945	880
00019_TE_2120x1608_8bit_sRGB.png	(1052,30)	945	880
00020_TE_1072x928_8bit_sRGB.png	(60,30)	945	880

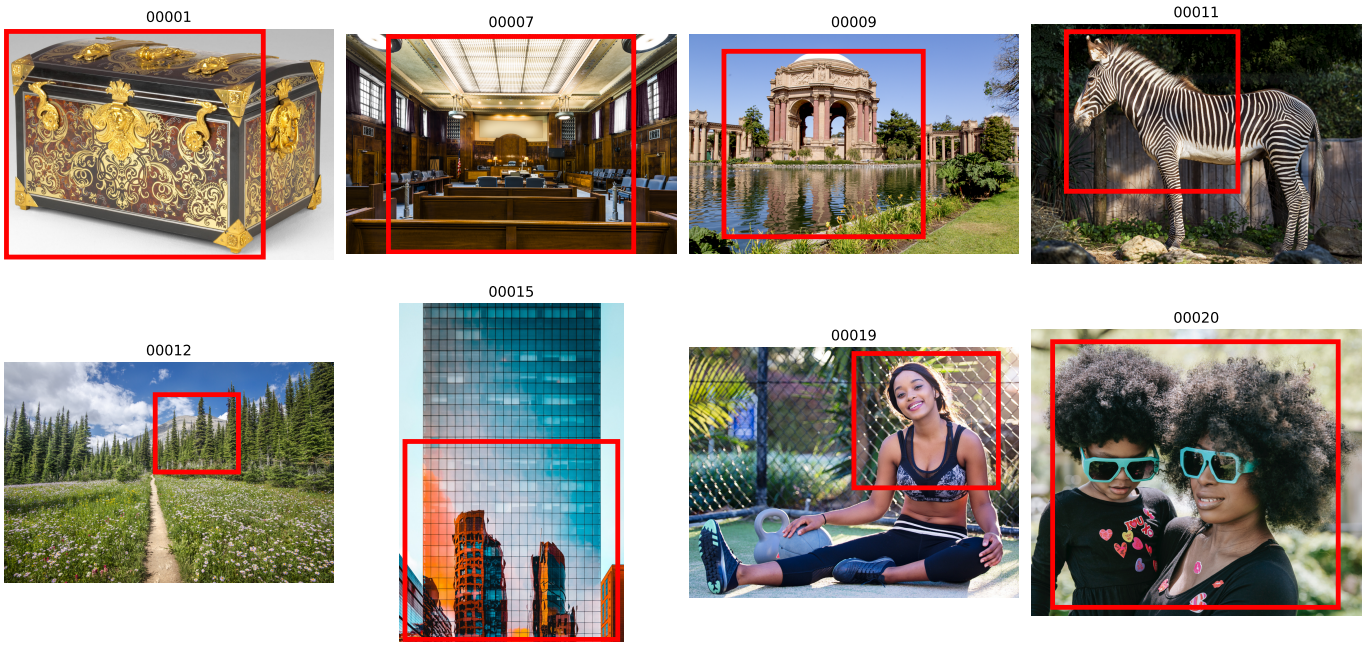

Figure 4. *Cropping areas of the images selected for the subjective evaluation from the JPEG AI CfP Test Set. Due to the screen size restriction all images had to be not bigger than 945x880.*

5.3 Bitrate Selection

The results obtained for the subjective quality assessment of the JPEG AI anchors [16] were used for the selection of the bitrates, and thus decoded images, that will be evaluated. Therefore, bitrate selection was performed for each image in such a way that the quality measured by DMOS would span from 60 to 90 approximately (scale of 1 to 100) for the VVC Intra anchor. The VVC Intra anchor was selected since it is nowadays the most powerfull standard based codec and the DMOS limits correspond to medium-low quality to transparent quality, i.e. almost no difference between the original and the decoded images. Considering this procedure, the following four target bitrate points were used:

- 0.06, 0.12, 0.25 and 0.5 bpp for image #11 and #15
- 0.06, 0.12, 0.25 and 0.75 bpp for image #19
- 0.12, 0.25, 0.5 and 0.75 bpp for the remaining images

5.4 Data Processing

To obtain subjective scores using a DSCQS methodology, it is first computed the mean opinion score (MOS) for each source reference and impaired stimuli according to:

$$\mathrm{MOS}_i = \frac{1}{N}\sum_{j=1}^{N} s_{ij} \tag{1}$$

where N is the number of valid subjects and s_{ij} is the score by subject j for the impaired image i. Then, the differential score DMOS is computed according to:

$$DMOS(IMP) = MOS(IMP) - MOS(SRC) + MAX\_SCALE \tag{1}$$

where SRC is the source reference and IM is impaired stimuli. The maximum of the rating scale is 100. The individual scores of each stimuli are represented in the scale of [0,1000] to represent a continuous scale; during the processing all the scores are re-scaled to [0,100] through dividing by 10 in float precision.

5.5 Experimental Results

In this Section the experimental results obtained according to the subjective assessment procedure described in the Section 5.1, with the selected test images and bitrates (Section 5.2 and 5.3) using the data processing procedure described in Section 5.4.

5.5.1 Subject Statistics

There were 288 subjects recruited from AWS Mechanical Turk which completed the experiment. Since the stimuli set was divided in three parts, the actual distribution of subjects among the parts was as follows:

- Part 1: 95 subjects
- Part 2: 97 subjects
- Part 3: 96 subjects

The age and gender for the subject population is distributed as follows:

- Females: 88, Males: 200
- Age from 20 to 72
- Age Mean: 38.11, Age Median: 36.00

Moreover, Figure 5 shows the histogram of the age of the subjects in 5-year bins plotted separately for females and for males. As shown, male subjects with an age between 30 and 40 were often selected to perform the subjective test.

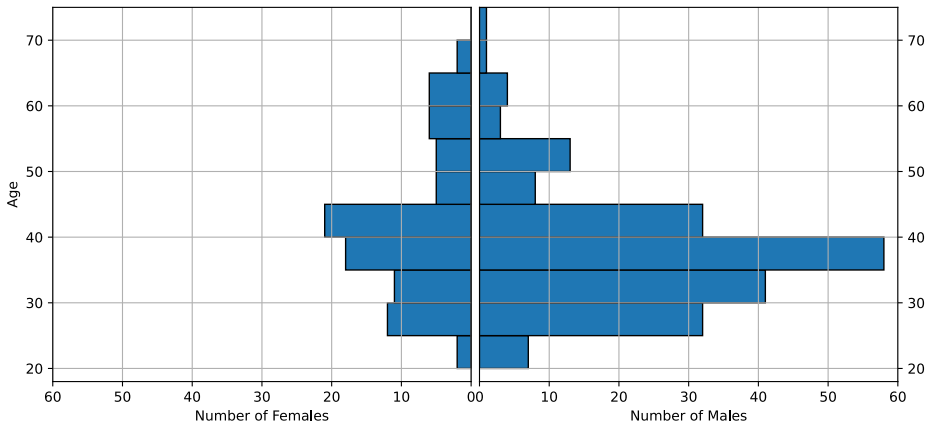

Figure 5. *Subjects age distribution.*

Table 6 presents the statistics on the screen size and the location of the participants. As expected, most of the subjects were from the USA and the most common screen size was 1920×1080.

Table 5. *Display resolution and subjects' country.*

Display resolution	**Subjects**
1920x1080	204
2560x1440	21
1950x1050	20

Country	**Subjects**
United States	215
India	21
Brazil	20

1920x1200	9
2560x1600	6
3440x1440	5
3840x2160	4
2880x1800	3
2560x1080	3
2048x1280	2
2048x1152	2
1950x1100	2
2256x1504	1
2000x1100	1
2896x1629	1
3840x1600	1
1920x1280	1
5120x1440	1
5120x2880	1

Italy	8
Not found	4
Latvia	3
Spain	2
Bangladesh	2
Honduras	1
Austria	1
Greece	1
Kenya	1
France	1
Estonia	1
Romania	1
Colombia	1
Turkey	1
United Kingdom	1
Australia	1

5.5.2 Sanity Check and Outlier Detection

Before the computation of the subjective scores, dishonest, unreliable, and inattentive subjects must be identified; these subjects often give random scores. First, a sanity check was done, for each subject, using the answers given to the honeypot questions to detect dishonest subjects. The honeypot comparisons correspond to the presentation of two identical original images. Second, outlier detection according to the ITU-R BT.500-14 [13] was performed to detect unreliable subjects. For the sanity check, the distributions of individual differential opinion scores were plotted for each honeypot comparison, which are shown in Figure 1. If a subject voted outside of the two standard deviations interval $[\mu - 2\sigma, \mu + 2\sigma]$ for two or more honeypot comparisons (out of three), all the data for this subject was removed. By using this simple sanity check procedure, 14 subjects were identified.

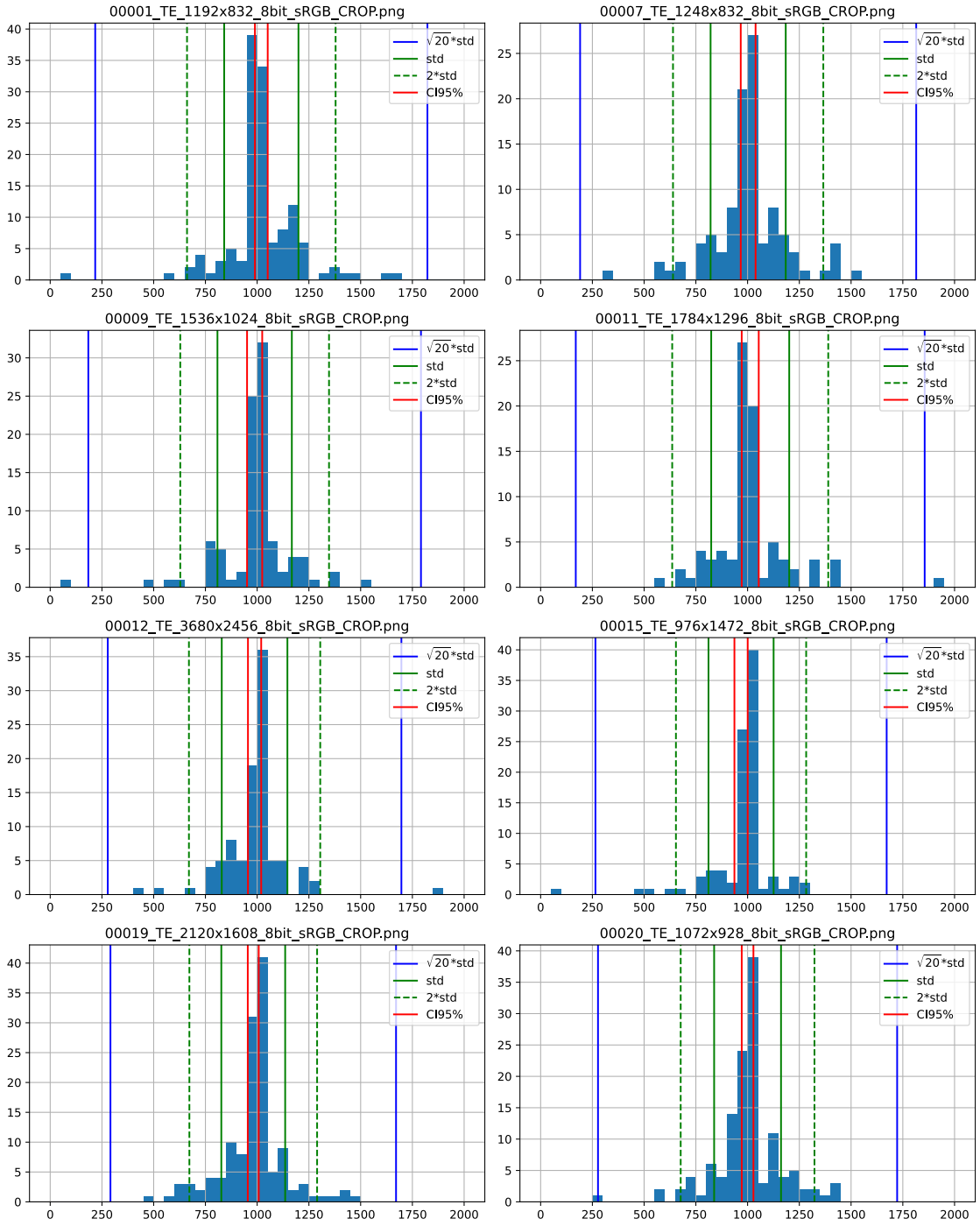

Figure 1. *Distribution of individual votes for hidden reference images. The horizontal axis is the value of an individual differential opinion score, i.e. {impaired image score} – {reference image score} + 1000. In this figure, the impaired image and reference image scores are in the scale of [0,1000]; and the differential scores are in the scale of [0,2000]. The vertical axis is the number of individual differential opinion scores that fall into the respective histogram bin.*

Following this sanity check, the outliers detection method as described in ITU-R BT.500-14, Section A1-2.3.1 was used. This procedure counts the number of instances that a subject's opinion score deviates by a few sigmas (i.e. standard deviation). The first step is to check if the distribution of scores is Normal or not by computing the Kurtosis; if the Kurtosis is between 2 and 4, the distribution is Normal. The intervals used for rejection are $[\mu - 2\sigma, \mu + 2\sigma]$ when the scores are Normally distributed and $[\mu - 4.47\sigma, \mu + 4.47\sigma]$ for the other case. If the number of votes outside the corresponding interval, for a subject, is > 5% and the difference between number of votes lying above the upper and bellow the lower interval limits is 30% (i.e. outlying votes are rather balanced) the subject is considered as an outlier. After the sanity check, 0 outliers were detected.

5.5.3 Rate-DMOS Performance

Differential Mean Opinion Scores (DMOS) were computed for each of 416 stimuli. Figure 2 depicts the resulting DMOS plotted separately for each image. The confidence intervals were computed assuming Student's t-distribution and the designated confidence level of 95%. From the experimental results obtained, the following conclusions can be taken:

- JPEG 2000 has the worst coding efficiency, which was expected, especially considering that more recent but also more complex standard-compliant codecs are available nowadays. Nevertheless, it was able to achieve higher coding efficiency for image #20 when compared to TEAM22 codec and to image #12 for low bitrates when compared to HEVC Intra.
- VVC Intra has higher or comparable coding efficiency compared to HEVC Intra, which is also expected due to the introduction of new Intra coding tools in the VVC standard. The only exception is image #15 for which HEVC Intra has higher performance due to the usage of screen content tools (not used in VVC Intra) which clearly bring advantages for this type of content.
- Many learning-based image codecs have better performance compared to VVC Intra, sometimes very significantly. Approximately, 4 learning-based image codecs have managed to have better coding efficiency compared to VVC Intra in a consistent way, namely TEAM 16, TEAM 14, TEAM 24 and TEAM 12. For some images, such as image #20 which is very textured with high frequency details, 8 learning-based image codecs have higher coding efficiency compared to VVC Intra. For a DMOS of 80, a minimum of 33% of rate reduction can be observed between VVC Intra and the best performing learning-based image codec for image #01. However, much higher performance gains can be achieved for other images, e.g., between 65%-70% of rate reduction can be observed for image #7, #15 and #19. These are very encouraging results which hold the promise of a very successful JPEG AI standard.
- TEAM16 learning-based image codec has better coding efficiency compared to other learning-based codec submissions for several images and a wide range of bitrates, especially low bitrates. The cases where the performance gains over other codecs are not clear are for image #20, image #7, image #11 and image #12. After, TEAM14 learning-based image codec also has very high performance, especially for images #11, #12, #20. TEAM24 learning-based image codec has also reached very high performance, close to TEAM14 for many cases and even overcoming it for medium and high bitrates for image #7.

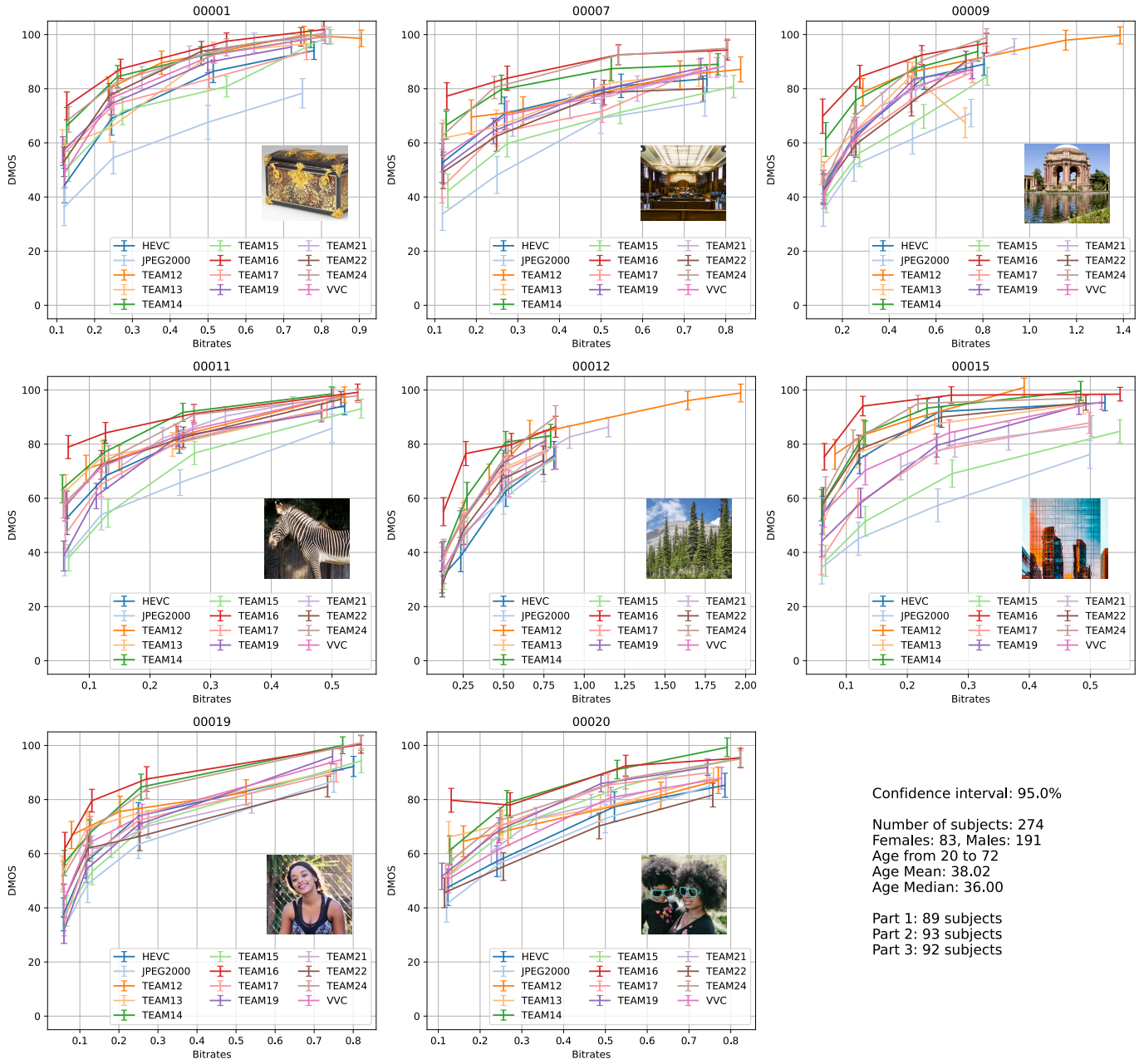

Figure 2. *Differential Mean Opinion Scores (DMOS) with the corresponding 95% confidence intervals (CI) plotted with respect to the bitrate.*

References

[1] ISO/IEC JTC 1/SC29/WG1 N100095, “Final Call for Proposals for JPEG AI", 94th JPEG Meeting, Online, January 2022.

[2] ISO/IEC JTC 1/SC29/WG1 M96016, “Presentation of the Huawei response to the JPEG AI Call for Proposals”, 96th JPEG Meeting, Online, July 2022.

[3] ISO/IEC JTC 1/SC29/WG1 M96050, “Tencent response to the JPEG AI Call for Proposals: Task-driven end-to-end image compression”, 96th JPEG Meeting, Online, July 2022.

[4] ISO/IEC JTC 1/SC29/WG1 M96051, “Tencent response to the JPEG AI Call for Proposals: Learning-based image compression”, 96th JPEG Meeting, Online, July 2022.

[5] ISO/IEC JTC 1/SC29/WG1 M96053, “Bytedance's response to the JPEG AI Call for Proposals”, 96th JPEG Meeting, Online, July 2022.

[6] ISO/IEC JTC 1/SC29/WG1 M96054, “NYCU-PUT Response to the JPEG AI Call for Proposals”, 96th JPEG Meeting, Online, July 2022.

[7] ISO/IEC JTC 1/SC29/WG1 M96056, “XidianUniversity-OPPO response to the JPEG AI Call for Proposals”, 96th JPEG Meeting, Online, July 2022.

[8] ISO/IEC JTC 1/SC29/WG1 M96066, “NJUVISION Response to the JPEG AI Call for Proposals”, 96th JPEG Meeting, Online, July 2022.

[9] ISO/IEC JTC 1/SC29/WG1 M96067, “USTC Response to the JPEG AI Call for Proposals”, 96th JPEG Meeting, Online, July 2022.

[10] ISO/IEC JTC 1/SC29/WG1 M96068, “XJU-SFU-Google Response to the JPEG AI Call for Proposals”, 96th JPEG Meeting, Online, July 2022.

[11] ISO/IEC JTC 1/SC29/WG1 N100106, “JPEG AI Common Training and Test Conditions,” 94th JPEG Meeting, Online, January 2022.

[12] ISO/IEC JTC 1/SC29/WG1 N100203, “Subjective Evaluation Procedure of JPEG AI Proposals,” 95th JPEG Meeting, Online, April 2022.

[13] ITU-R Recommendation BT.500-14, “Methodology for the Subjective Assessment of the Quality of Television Pictures,” International Telecommunications Union, Geneva, Switzerland, 2012.

[14] ITU-T Recommendation P. 910, “Subjective Video Quality Assessment Methods for Multimedia Applications,” International Telecommunication Union, Geneva, Switzerland, 2008.

[15] C. Keimel, J. Habigt, C. Horch, and K. Diepold, “Qualitycrowd: a Framework for Crowd-based Quality Evaluation,” Picture Coding Symposium, Krakow, Poland, May 2012.

[16] ISO/IEC JTC 1/SC29/WG1 N100205, “Report on the Objective and Subjective Quality Assessment of the JPEG AI Anchors”, 95th Meeting, Online, April 2022.

Acknowledgment

We would like to thank to all proponents that volunteered for the subjective evaluation and for the creation and management of the crowdsourcing platform, especially Evgeniy Upenik. Also, we would like to acknowledge the test set creation committee for its valuable work and Huawei, Bytedance and Google for the anchor generation.