Video compression dataset and benchmark of learning-based video-quality metrics

Anastasia Antsiferova[1,2*], **Sergey Lavrushkin**[1,2*], **Maksim Smirnov**[3*], **Alexander Gushchin**[3*], **Dmitriy Vatolin**[1,2,3*], **Dmitriy Kulikov**[2,3*]
ISP RAS Research Center for Trusted Artificial Intelligence[1]
MSU Institute for Artificial Intelligence[2]
Lomonosov Moscow State University[3]
{aantsiferova, sergey.lavrushkin, maxim.smirnov.2025, alexander.gushchin, dmitriy, dkulikov}@graphics.cs.msu.ru

Abstract

Video-quality measurement is a critical task in video processing. Nowadays, many implementations of new encoding standards — such as AV1, VVC, and LCEVC — use deep-learning-based decoding algorithms with perceptual metrics that serve as optimization objectives. But investigations of the performance of modern video- and image-quality metrics commonly employ videos compressed using older standards, such as AVC. In this paper, we present a new benchmark for video-quality metrics that evaluates video compression. It is based on a new dataset consisting of about 2,500 streams encoded using different standards, including AVC, HEVC, AV1, VP9, and VVC. Subjective scores were collected using crowdsourced pairwise comparisons. The list of evaluated metrics includes recent ones based on machine learning and neural networks. The results demonstrate that new no-reference metrics exhibit high correlation with subjective quality and approach the capability of top full-reference metrics.

1 Introduction

Video constitutes the largest part of the world's Internet traffic, and its volume has increased because of the Covid lockdowns. The network load has also increased, making efficient video compression extremely important. Development and comparison of new video encoders greatly relies on quality measurement, and many new compression standards implement machine-learning- and neural-network-based approaches. But traditional image- and video-quality metrics, such as PSNR and SSIM, emerged long before recent compression standards, and they did not account for neural-network-related artifacts. VMAF [26], a well-known video-quality metric from Netflix, was also trained using only H.264/AVC-compressed videos. Thus, quality measurement for new video-encoding standards is even more vital. The number of new image- and video-quality metrics has increased, and many recent algorithms employ learning-based approaches. Industry leaders have also created their own quality metrics: Apple's Advanced Video Quality Tool (AVQT) [2], Tencent's Deep Learning-Based Video Quality Assessment (DVQA) [1], and the aforementioned VMAF. Only a few of these metrics demonstrate high performance on independent benchmarks, however, and some new ones, including AVQT and DVQA, still await detailed analysis. A concern associated with

*A. Antsiferova designed methodology and performed results analysis, led the benchmark and the paper preparation, S. Lavrushkin performed results analysis and the paper preparation, M. Smirnov developed the benchmark of no-reference metrics, A. Gushchin developed the benchmark of full-reference metrics, both of them performed the dataset research, results analysis and contributed to the paper preparation, D. Vatolin and D. Kulikov coordinated the benchmark and the dataset creation

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

Figure 1: Crops from video sequences encoded using old and new standards relative to ground truth (GT). LCEVC employs super-resolution, which allows restoration of more details and creates new kinds of artifacts.

metric-result reproducibility and verification is the outdated datasets for measuring video-compression quality. Most datasets containing compressed videos and subjective scores only employ H.264/AVC compression. In-lab tests were the source of subjective scores for many such videos. Owing to the complexity and high cost of subjective comparisons, those tests involved a small number of viewers and garnered only a few scores per video.

Quality-metric development seldom takes into account artifacts produced by video encoders that implement contemporary standards. For example, super-resolution in LCEVC and in new neural-network-based encoders yields distortions that traditional metrics are unable to handle. Fig. 1 demonstrates the difference between frame crops of x265-encoded video and lcevc_x265-encoded video: the latter contains more detail despite its lower PSNR and SSIM scores. Existing benchmarks for image- and video-quality metrics do not consider artifacts produced by new compression standards. Our research therefore analyzed metric performance on videos with various compression artifacts.

The goal of our investigation was to evaluate new and state-of-the-art image- and video-quality metrics independently, using a large dataset representing diverse compression artifacts from different video encoders. We thus propose a new dataset of 2,486 compressed videos and subjective scores collected using a crowdsourced comparison with nearly 11,000 participants. We also present a new benchmark[2] based on that dataset, which we divide into open and hidden parts. This paper provides our assessment results for the open part as well as for the whole dataset.

2 Related Work

2.1 Video-Quality Datasets

Video-quality datasets with subjective scores break down into two types: legacy synthetically distorted (mainly through compression and transmission distortions, capture impairments, processing artifacts, and Gaussian blur), and authentic user-generated content (UGC). The former [43, 11, 42, 7, 32, 27, 37, 19] apply synthetic distortions to the original videos. The latter [14, 39, 36, 16, 44, 50] are gaining popularity, as videos produced today by amateurs often suffer from a wide variety of distortions. Many new video-quality metrics have undergone testing only for UGC videos. The latest studies employ a nearly identical pool of subjective video-quality datasets, summarized in Tab. 1.

2.2 Video-Quality Benchmarks

Most comparisons of IQA and VQA have appeared in papers that present new methods and a few benchmarks accept new methods for evaluation. Often, these comparisons either include an evaluation using open video datasets, for which existing metrics may have been tuned, or employ just a few methods. The authors of [39] published a comparison for a wide variety of datasets but evaluated only four methods. In [41] the authors compared no-reference VQA models using three UGC datasets

[2] https://videoprocessing.ai/benchmarks/video-quality-metrics.html

Dataset	Original videos	Average duration (s)	Distorted videos	Distortion	Subjective framework	Subjects	Answers
MCL-JCV (2016) [42]	30	5	1,560	Compression	In-lab	150	78K
VideoSet (2017) [43]	220	5	45,760	Compression	In-lab	800	-
UGC-VIDEO (2020) [25]	50	> 10	550	Compression	In-lab	30	16.5K
CVD-2014 (2014) [36]	5	10-25	234	In-capture	In-lab	210	-
LIVE-Qualcomm (2016) [14]	54	15	208	In-capture	In-lab	39	8.1K
GamingVideoSET (2018) [9]	24	30	576	Compression	In-lab	25	-
KUGVD (2019) [8]	6	30	144	Compression	In-lab	17	-
KoNViD-1k (2017) [16]	1,200	8	1,200	In-the-wild	Crowdsource	642	205K
LIVE-VQC (2018) [39]	585	10	585	In-the-wild	Crowdsource	4,776	205K
YouTube-UGC (2019) [44]	1,500	20	1,500	In-the-wild	Crowdsource	>8,000	600K
LSVQ (2020) [50]	39,075	5-12	39,075	In-the-wild	Crowdsource	6,284	5M
Our dataset: open part (2022)	36	10, 15	1,022	Compression (32 codecs)	Crowdsource	10,800	320K
Our dataset: hidden part (2022)	36	10, 15	1,464	Compression (51 codecs)	Crowdsource	10,800	446K
Our dataset (2022)	36	10, 15	2,486	Compression (83 codecs)	Crowdsource	10,800	766K

Table 1: Summary of subjective video-quality datasets and our new dataset.

Benchmark	Total number of videos	Total number of VQA methods	Total number of subjects	Distortion
Z. Sinno and A. Bovik (2018) [39]	585	4	4,776	In-the-wild videos, 80 mobile cameras, 18 resolutions
Y. Li *et al.* (2020) [24]	550	15	28	H.264, H.265 compression, QP: 22, 27, 32, 37, 42
UGC-VQA (2021) [41]	3,108 (LIVE-VQC, YouTube-UGC, KoNViD-1k)	13	>13,000	Compression, transmission
Our benchmark (2022)	2,486	26	10,800	Compression (H.264, H.265, AV1, VVC, etc.)

Table 2: Summary of video-quality-measurement benchmarks and our new benchmark.

and various experiments. They analyzed metrics applied to videos with different content types, resolution and quality subsets, temporal pooling, and computational-complexity-evaluation methods. Compression artifacts, however, played a minor role in that study. The main idea of [24] was to compare full- and no-reference metrics through subjective evaluation of UGC videos transcoded using different compression standards and levels, but this work only tested a few no-reference methods and codecs.

3 Benchmark

3.1 List of Metrics

This study aimed to evaluate new and state-of-the-art neural-network-based video- and image-quality metrics on a compression-oriented video dataset. We excluded several well-known metrics such as BRISQUE [33] and VIIDEO [34] because of their low correlations in many other studies [50, 22, 24].

3.1.1 No-Reference Video-Quality Metrics

The no-reference video-quality metric **VIDEVAL** (2021) [41] chooses 60 features (related to motion, certain distortions, and aesthetics) from previously developed quality models. It performs well on existing UGC datasets, but it may suffer from overfitting, as users must set many of its parameters.

Most recent quality-assessment papers emphasize deep-learning-based approaches. **MEON** (2017) [29] is a model consisting of two sub-networks: a distortion-identification one and a quality-prediction one. It can also determine the distortion type.

VSFA (2019) [20] employs a pretrained ResNet-50 [15] as well as a deep content-aware feature extractor followed by a temporal-pooling layer for temporal memory. It performed poorly in the cross-dataset evaluation, so the authors proposed an enhanced version, **MDTVSFA** (2021) [22]. This enhanced version follows a mixed-dataset training strategy and may have high computational complexity owing to recurrent layers and full-size-image inputs.

PaQ-2-PiQ (2020) [51] uses a deep region-based architecture trained on a large subjective image-quality dataset of 40,000 pictures. **KonCept512** (2020) [17] is based on InceptionResNetV2 and was trained on the proposed KonIQ-10k dataset. **SPAQ** (2020) [13] implements three extra modifications of its baseline model: EXIF-data processing (MT-E), image-attribute observation (MT-A), and observation of a scene's high-level semantics (MT-S). The creators of **Linearity** (2020) [21] introduced their own loss function, "norm-in-norm", which converges 10 times faster than the MAE and MSE loss functions. **NIMA** (2018) [40] was trained on the large-scale Aesthetic Visual Analysis (AVA) dataset and predicts a quality-rating distribution.

3.1.2 Full-Reference Video-Quality Metrics

PSNR and **SSIM** [45] are among the most popular image- and video-quality metrics. We compared variations of SSIM and **MS-SSIM** [46] in our benchmark; the latter is an advanced version of the former calculated over multiple scales using subsampling.

LPIPS (2019) [52] is based on AlexNet and VGG. We chose a VGG-based version for testing because it serves as a generalization of "perceptual loss" [18]. **DISTS** (2020) [12] was designed to tolerate texture resampling and to be sensitive to structural differences. It combines structure- and texture-similarity measurements for corresponding image embedments and is based on a pretrained VGG network.

Tencent's **DVQA** (2020) [1] is based on the C3DVQA network [49]. It uses 3D convolutional layers to learn spatiotemporal features and 2D convolutional layers to extract spatial information.

The main feature of **FovVideoVDP** (2021) [31] is consideration of peripheral visual acuity. This method models the human visual system's response to temporal changes across the visual field. It can estimate flickering, juddering, and other temporal distortions, as well as spatiotemporal artifacts such as those appearing at different degrees of peripheral vision.

ST-GREED (2021) [30] can quantify reference and distorted videos of different frame rates without temporal preprocessing. It offers two primary features: SGreed and TGreed. The latter quantifies the statistics of temporal bandpass responses to both spatial and temporal distortions. The former obtains spatial bandpass responses using a local filtering scheme. Calculation of the final ST- GREED value employs the support-vector regressor.

Nowadays **VMAF** (2018) [26] is one of the most popular VQA metrics. It computes three base features—the detail-loss metric (DLM) [23], visual-information fidelity (VIF) [38], and temporal information (TI)—and combines them with a support-vector regressor. We also evaluated **AVQT** (2021) [2], developed by Apple, but the company has yet to publish any technical information.

3.2 Video Dataset

To analyze the relevance of quality metrics to video compression, we collected a special dataset of videos exhibiting various compression artifacts. For video-compression-quality measurement, the original videos should have a high bitrate or, ideally, be uncompressed to avoid recompression artifacts. We chose from a pool of more than 18,000 high-bitrate open-source videos from www.vimeo.com. Our search included a variety of minor keywords to provide maximum coverage of potential results—for example "a," "the," "of," "in," "be," and "to." We downloaded only videos that were available under CC BY and CC0 licenses and that had a minimum bitrate of 20 Mbps. The average bitrate of the entire collection was 130 Mbps. We converted all videos to a YUV 4:2:0 chroma subsampling. Our choice employed space-time-complexity clustering to obtain a representative complexity distribution. For spatial complexity, we calculated the average size of x264-encoded I-frames normalized to the uncompressed frame size. For temporal complexity, we calculated the average P-frame size divided by the average I-frame size. We divided the whole collection into 36 clusters using the K-means algorithm [28] and, for each cluster, randomly selected up to 10 candidate videos close to the cluster center. From each cluster's candidates we manually chose one video, attempting to include different

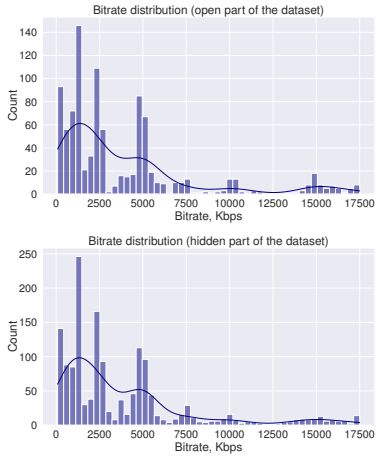

Figure 2: Bitrate distribution of videos in our dataset.

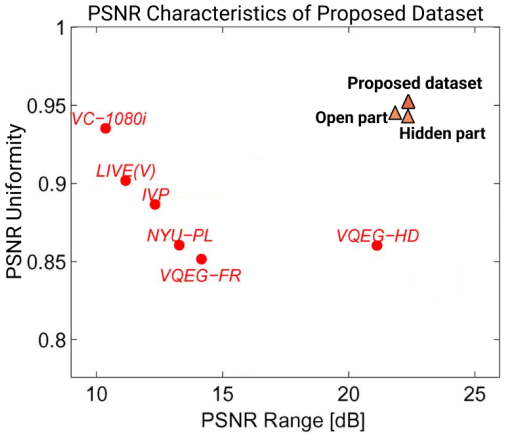

Figure 3: PSNR (range and uniformity) comparison for our dataset versus other video datasets.

genres in the final dataset (sports, gaming, nature, interviews, UGC, etc.). The result was 36 FullHD videos for further compression.

We obtained numerous coding artifacts by compressing videos through several encoders: 11 H.265/HEVC encoders, 5 AV1 encoders, 2 H.264/AVC encoders, and 4 encoders based on other standards. To increase the diversity of coding artifacts, we also used two different presets for many encoders: one that provides a 30 FPS encoding speed and the other that provides a 1 FPS speed and higher quality. The list of settings for each encoder is presented in the supplementary materials. Not all videos underwent compression using all encoders. We compressed each video at three target bitrates — 1,000 kbps, 2,000 kbps, and 4,000 kbps — using a VBR mode (for encoders that support it) or with corresponding QP/CRF values that produce these bitrates. Major streaming-video services recommend at most 4,500–8,000 kbps for FullHD encoding [3, 4, 5]. We avoided higher target bitrates because visible compression artifacts become almost unnoticeable, hindering subjective comparisons. Fig. 2 shows the distribution of video bitrates for our dataset. The distribution differs from the target encoding rates because we used the VBR encoding mode, but it complies with the typical recommendations.

The dataset falls into two parts: open and hidden (40% and 60% of the entire dataset, respectively). We employ hidden part only for testing through our benchmark to ensure a more objective comparison of future applications. This approach may prevent learning-based methods from training on the entire dataset, thereby avoiding overfitting and incorrect results. To divide our dataset, we split the codec list in two; the encoded videos each reside in the part corresponding to their respective codec. We also performed x265-lossless encoding of all compressed streams to simplify further evaluations and avoid issues with nonstandard decoders.

Tab. 1 shows the characteristics of the final parts of the dataset. Links to source videos and additional details about the collection process are in the supplementary materials. We also compared the statistics of PSNR uniformity and range for our dataset using the approach in [47]. As Fig. 3 shows, this dataset provides wide quality and compression-rate ranges.

3.3 Subjective-Score Collection

We collected subjective scores for our video dataset through the Subjectify.us crowdsourcing platform. Subjectify.us is a service for pairwise comparisons; it employs a Bradley-Terry model to transform the results of pairwise voting into a score for each video. A more detailed description of the method is at `www.subjectify.us`.

Because the number of pairwise comparisons grows exponentially with the number of source videos, we divided the dataset into five subsets by source videos and performed five comparisons. Each subset contained a group of source videos and their compressed versions. Every comparison produced and

evaluated all possible pairs of compressed videos for one source video. Thus, only videos from the same source were in each pair. The comparison set also included source videos. Participants viewed videos from each pair sequentially in full-screen mode. They were asked to choose the video with the best visual quality or indicate that the two are of the same quality. They also had an option to replay the videos. Each participant had to compare a total of 12 pairs, two of which had an obviously higher-quality option and served as verification questions. All responses from those who failed to correctly answer the verification questions were discarded.

To increase the relevance of the results, we solicited at least 10 responses for each pair. In total, we collected 766,362 valid answers from nearly 11,000 individuals. After applying the Bradley-Terry model to a table of pairwise ranks, we received subjective scores that are consistent within each group of videos compressed from one reference video. A detailed description of the subjective-comparison process, as well as collected statistics, is in the supplementary materials. Tab. 1 summarizes the parameters of our dataset.

3.4 Methodology

We used public source code for all metrics without additional pretraining, and we selected the default parameters to avoid overfitting. To get a video's quality score using the IQA method, we compared the given distorted sequence and the reference video frame by frame, then averaged the resulting per-frame quality scores for each video. VQA methods generate a score for the whole distorted sequence and require no additional averaging.

Because the subjective scores are based on pairwise comparisons of videos produced from the same original sequence, they are comparable only within their respective groups. Each group size is three (the number of encoding bitrates) times the number of codecs applied to the reference video. For each reference-video/preset pair (resulting in one distorted-video group), we calculated Spearman and Kendall correlation coefficients (SROCC and KROCC, respectively) between the metrics and subjective scores. We then selected only those values calculated for groups whose number of samples exceed a threshold (15 for SROCC and 6 for KROCC) to provide more-statistically-reliable results. Our next step was to use the Fisher Z-transform [10] (inverse hyperbolic tangent) and average the results, weighted proportionally to group size. The inverse Fisher Z-transform yielded a single correlation for the entire dataset. We provide the link to code example in Sec. 4.

To analyze metric performance in more detail, we added a few mutually nonexclusive categories with videos from the dataset: User-Generated Content, Shaking, Sports, Nature, Gaming / Animation, Low Bitrate (up to 1,000 Kbps), and High Bitrate (above 6,000 Kbps). To assign each video to one of these categories, we conducted a subjective survey of five people from our laboratory.

3.5 Results

We examined the results for the open part of the dataset as well as for the whole dataset, including the hidden part. Tab. 3 shows the Spearman and Kendall correlation coefficients for the metrics we analyzed, along with subjective quality scores. For the whole dataset, VMAF and its variations calculated using different chroma-component ratios exhibited the highest correlations. VMAF was originally to be calculated only using the luma component, but here we proved that YUV-VMAF performs better. Also, VMAF NEG (a no-enhancement-gain version [6]) correlated less well with subjective quality than the original version did. MDTVSFA and Linearity had the highest correlations among no-reference methods: about 0.93, nearly matching the top results of full-reference metrics (VMAF at 0.94). For the open dataset, SSIM and PSNR showed the highest correlations in addition to VMAF, followed by a recently-released AVQT by Apple.

We compared metrics using different video subsets: videos with low and high bitrates; videos encoded using HEVC/H.265, AV1, and VVC/H.266 (Tab. 4); and videos with different content types — UGC, Shaking, Sports, Nature, and Gaming/Animation (Tab. 5).

"High bitrate" and "Low bitrate" encoding. All metrics showed their lowest correlations for videos encoded at 6,000 Kbps or higher. The reason may be the low confidence of the subjective scores for this category. As we described in Sec. 3.2, viewers apparently have difficulty spotting compression artifacts in videos that employ high-quality encoding. The no-reference MDTVSFA, VSFA, and

Dataset	All Dataset (Open+Hidden) 2486 videos		Open Dataset 1022 videos	
Metric	SROC	KROC	SROC	KROC
No-Reference				
MEON [29]	0.507 (0.495, 0.518)	0.376 (0.367, 0.384)	0.554 (0.534, 0.574)	0.350 (0.333, 0.366)
Y-NIQE [35]	0.599 (0.586, 0.611)	0.421 (0.411, 0.431)	0.701 (0.679, 0.721)	0.557 (0.541, 0.573)
VIDEVAL [41]	0.729 (0.719, 0.738)	0.541 (0.532, 0.551)	0.719 (0.700, 0.737)	0.558 (0.540, 0.575)
KonCept512 [17]	0.836 (0.831, 0.841)	0.661 (0.655, 0.666)	0.861 (0.853, 0.868)	0.696 (0.688, 0.703)
NIMA [40]	0.849 (0.844, 0.854)	0.675 (0.668, 0.681)	0.868 (0.860, 0.875)	0.729 (0.719, 0.738)
PaQ-2-PiQ [51]	0.871 (0.866, 0.875)	0.708 (0.702, 0.714)	0.901 (0.894, 0.908)	0.752 (0.743, 0.761)
SPAQ MT-A [13]	0.879 (0.875, 0.884)	0.715 (0.709, 0.720)	**0.912** (0.905, 0.919)	**0.796** (0.786, 0.805)
SPAQ BL [13]	0.880 (0.875, 0.884)	0.711 (0.704, 0.717)	**0.912** (0.905, 0.918)	**0.789** (0.780, 0.798)
SPAQ MT-S [13]	0.882 (0.878, 0.886)	0.719 (0.713, 0.724)	**0.912** (0.906, 0.918)	0.787 (0.778, 0.796)
VSFA [20]	**0.905** (0.901, 0.908)	**0.748** (0.743, 0.753)	0.891 (0.886, 0.897)	0.758 (0.750, 0.766)
Linearity [21]	**0.910** (0.907, 0.913)	**0.759** (0.754, 0.763)	0.905 (0.899, 0.911)	0.783 (0.774, 0.791)
MDTVSFA [22]	**0.929** (0.927, 0.931)	**0.788** (0.784, 0.792)	**0.930** (0.927, 0.934)	**0.813** (0.806, 0.819)
Full-Reference				
FOV VIDEO [31]	0.527 (0.521, 0.534)	0.375 (0.370, 0.380)	0.565 (0.551, 0.579)	0.492 (0.477, 0.507)
LPIPS [52]	0.749 (0.742, 0.756)	0.567 (0.561, 0.573)	0.787 (0.774, 0.799)	0.667 (0.655, 0.679)
DVQA [1]	0.763 (0.756, 0.770)	0.579 (0.572, 0.585)	0.774 (0.762, 0.786)	0.683 (0.670, 0.695)
GREED [30]	0.764 (0.759, 0.769)	0.588 (0.582, 0.593)	0.790 (0.782, 0.797)	0.644 (0.634, 0.654)
Y-VQM [48]	0.821 (0.815, 0.827)	0.644 (0.637, 0.651)	0.881 (0.870, 0.890)	0.767 (0.756, 0.777)
DISTS [12]	0.847 (0.842, 0.851)	0.671 (0.667, 0.676)	0.873 (0.866, 0.879)	0.753 (0.744, 0.761)
AVQT [2]	0.876 (0.872, 0.879)	0.720 (0.716, 0.725)	0.889 (0.882, 0.896)	0.792 (0.784, 0.800)
YUV-PSNR	0.883 (0.879, 0.887)	0.728 (0.722, 0.733)	**0.949** (0.946, 0.952)	0.849 (0.844, 0.854)
YUV-SSIM	0.906 (0.902, 0.909)	0.756 (0.750, 0.761)	**0.948** (0.945, 0.951)	**0.897** (0.872, 0.917)
Y-MS-SSIM [46]	0.909 (0.905, 0.912)	0.756 (0.751, 0.760)	0.946 (0.943, 0.949)	0.841 (0.835, 0.847)
Y-VMAF NEG [26]	0.914 (0.911, 0.917)	0.765 (0.760, 0.769)	0.945 (0.942, 0.948)	0.841 (0.836, 0.846)
YUV-VMAF NEG [26]	**0.917** (0.914, 0.920)	**0.769** (0.765, 0.774)	0.947 (0.944, 0.950)	**0.894** (0.869, 0.915)
Y-VMAF (v061) [26]	**0.942** (0.940, 0.944)	**0.809** (0.805, 0.813)	0.945 (0.942, 0.948)	0.888 (0.861, 0.910)
YUV-VMAF (v061) [26]	**0.943** (0.941, 0.945)	**0.810** (0.806, 0.814)	**0.948** (0.945, 0.951)	**0.895** (0.870, 0.916)

Table 3: Results for SROCC and KROCC on the full dataset and on the open part.

Linearity metrics performed better than the full-reference alternatives. The leaders for "Low Bitrate" remain the same as for the whole dataset supplemented by no-reference NIMA.

HEVC, AV1, and VVC encoding. Metric correlation for "H.265 encoding" is higher than for other standards. Because H.265 is older and more popular than newer standards, quality-assessment models may have been tuned to it. For the new VVC standard, the leaders differ relative to other encoding standards: the best no-reference metric is Linearity and the best full-reference one is SSIM. This result is unexpected, but SSIM's good performance may owe to its versatility. Also, the sample for this category is small, so further analysis of the best metrics for estimating VVC encoding quality would likely require a larger dataset.

Leaders by video content category: "UGC", "Shaking", "Sports", "Nature", and "Gaming/Animation". VMAF retained its lead in all categories, but its original luma-only (Y-VMAF) version performed better on "UGC" and "Shaking" content; its modified version using chroma components (YUV-VMAF) is the best for other categories. The no-reference MDTVSFA metric leads in "UGC", "Shaking" and "Nature"; Linearity is ahead in "Sports"; and VSFA is best for "Gaming/Animation". PaQ-2-PiQ also achieves to precisely estimate gaming-content quality.

We performed a one-sided Wilcoxon rank-sum test on SROCC, which we computed for most of the methods in Tab. 4 and Tab. 5 using different groups of videos from our dataset to get the average correlation value. A table of results appears in the supplementary materials. Different versions of SPAQ (BL, MT-S, and MT-A) behaved in a statistically equal manner—except in the "Sports" and "Low Bitrate" categories, where SPAQ BL was superior. In addition, VMAF, the top full-reference metric in average SROCC for the full dataset, yielded to MDTVSFA, the leading no-reference metric, only on videos encoded using AV1 and videos with a high bitrate. A rarely used encoding standard for training quality-assessment methods, VVC was difficult for these methods to handle, and videos encoded using it form the only part of our dataset where MDTVSFA fell short of Linearity. Among related metrics, the test revealed that VMAF NEG and VSFA were statistically worse than or equal to VMAF (v061) and MDTVSFA, respectively, depending on the subset. AVQT was superior to most metrics for the "Low Bitrate" and "Shaking" categories, making it valuable when predicting subjective quality.

Tab. 6 shows the computational complexity of the metrics we studied.

Fig. 4 shows the distribution of normalized metric scores for our dataset. Many metrics have a nonuniform "real-life" distribution of values resulting from compression artifacts. For example, the average SSIM is about 0.85, which corresponds with common statistics (an SSIM of 0.5 does not mean average quality, and values below 0.5 seldom appear in real situations).

Dataset	Low Bitrate (up to 1,000 kbps) 477 videos		High Bitrate (above 6,000 kbps) 384 videos		H.265 Encoding 1139 videos		AV1 Encoding 482 videos		VVC Encoding 251 videos	
Metric	SROC	KROC	SROC	KROC	SROC	KROC	SROC	KROC	SROC	KROC
No-Reference										
MEON [29]	0.039 (0.000, 0.093)	0.069 (0.032, 0.106)	0.127 (0.097, 0.158)	0.107 (0.089, 0.125)	0.834 (0.825, 0.842)	0.703 (0.661, 0.740)	0.800 (0.768, 0.828)	0.734 (0.625, 0.815)	0.709 (0.662, 0.750)	0.535 (0.492, 0.576)
Y-NIQE [35]	0.313 (0.263, 0.361)	0.214 (0.181, 0.246)	0.027 (0.000, 0.063)	0.013 (0.000, 0.040)	0.722 (0.706, 0.736)	0.519 (0.431, 0.597)	0.629 (0.574, 0.678)	0.640 (0.501, 0.747)	0.389 (0.311, 0.462)	0.271 (0.218, 0.323)
VIDEVAL [41]	0.615 (0.580, 0.647)	0.415 (0.387, 0.441)	0.290 (0.256, 0.322)	0.209 (0.189, 0.229)	0.804 (0.791, 0.817)	0.661 (0.576, 0.731)	0.754 (0.707, 0.794)	0.625 (0.596, 0.653)	0.635 (0.576, 0.688)	0.490 (0.444, 0.533)
KonCept512 [17]	0.891 (0.883, 0.899)	0.719 (0.708, 0.729)	0.204 (0.149, 0.259)	0.164 (0.136, 0.192)	0.904 (0.898, 0.909)	**0.876** (0.837, 0.907)	0.819 (0.793, 0.842)	**0.990** (0.972, 0.996)	0.849 (0.819, 0.874)	0.693 (0.658, 0.725)
NIMA [40]	**0.904** (0.895, 0.913)	**0.764** (0.751, 0.776)	0.361 (0.315, 0.405)	0.256 (0.232, 0.280)	0.879 (0.872, 0.886)	0.791 (0.748, 0.828)	0.807 (0.774, 0.835)	0.953 (0.901, 0.978)	0.712 (0.648, 0.765)	0.540 (0.486, 0.589)
PaQ-2-PiQ [51]	0.880 (0.871, 0.888)	0.689 (0.675, 0.703)	0.402 (0.343, 0.457)	0.291 (0.261, 0.321)	0.911 (0.906, 0.915)	0.861 (0.823, 0.891)	0.870 (0.851, 0.886)	0.978 (0.949, 0.991)	0.888 (0.865, 0.908)	**0.749** (0.717, 0.777)
SPAQ MT-A [13]	0.842 (0.835, 0.849)	0.689 (0.677, 0.702)	0.393 (0.356, 0.430)	0.308 (0.286, 0.331)	0.898 (0.892, 0.904)	0.816 (0.777, 0.849)	0.870 (0.853, 0.886)	0.957 (0.909, 0.980)	**0.894** (0.869, 0.915)	0.727 (0.689, 0.760)
SPAQ BL [13]	0.844 (0.835, 0.852)	0.672 (0.659, 0.685)	0.401 (0.358, 0.442)	0.332 (0.307, 0.356)	0.901 (0.894, 0.907)	0.820 (0.781, 0.852)	0.875 (0.858, 0.890)	0.888 (0.812, 0.935)	0.887 (0.862, 0.908)	0.729 (0.694, 0.760)
SPAQ MT-S [13]	0.810 (0.804, 0.816)	0.648 (0.640, 0.656)	0.417 (0.382, 0.450)	**0.344** (0.319, 0.368)	0.891 (0.883, 0.897)	0.808 (0.767, 0.842)	0.882 (0.867, 0.896)	0.959 (0.914, 0.981)	**0.908** (0.886, 0.926)	**0.764** (0.731, 0.793)
VSFA [20]	0.894 (0.890, 0.898)	**0.757** (0.748, 0.764)	**0.517** (0.467, 0.565)	**0.370** (0.339, 0.401)	**0.927** (0.922, 0.932)	0.790 (0.782, 0.799)	**0.914** (0.900, 0.927)	0.989 (0.973, 0.996)	0.846 (0.818, 0.870)	0.694 (0.662, 0.723)
Linearity [21]	**0.900** (0.894, 0.906)	0.731 (0.721, 0.740)	**0.470** (0.424, 0.514)	0.336 (0.311, 0.361)	**0.932** (0.928, 0.936)	**0.902** (0.870, 0.926)	**0.906** (0.892, 0.918)	**0.993** (0.981, 0.997)	**0.919** (0.903, 0.933)	**0.791** (0.768, 0.812)
MDTVSFA [22]	**0.943** (0.940, 0.946)	**0.818** (0.811, 0.824)	**0.560** (0.511, 0.606)	**0.363** (0.331, 0.394)	**0.945** (0.941, 0.948)	**0.871** (0.843, 0.895)	**0.932** (0.919, 0.943)	**0.997** (0.991, 0.999)	0.882 (0.859, 0.902)	0.746 (0.718, 0.772)
Full-Reference										
FOV VIDEO [31]	0.526 (0.512, 0.539)	0.372 (0.361, 0.384)	0.158 (0.135, 0.182)	0.116 (0.100, 0.133)	0.558 (0.544, 0.571)	0.403 (0.393, 0.413)	0.381 (0.349, 0.414)	0.539 (0.377, 0.670)	0.281 (0.243, 0.319)	0.211 (0.185, 0.238)
LPIPS [52]	0.774 (0.761, 0.786)	0.577 (0.565, 0.589)	0.270 (0.246, 0.293)	0.179 (0.160, 0.198)	0.814 (0.803, 0.824)	0.815 (0.757, 0.860)	0.532 (0.499, 0.563)	0.477 (0.452, 0.502)	0.464 (0.422, 0.504)	0.356 (0.325, 0.387)
DVQA [1]	0.781 (0.766, 0.794)	0.584 (0.572, 0.596)	0.103 (0.075, 0.130)	0.100 (0.088, 0.113)	0.786 (0.775, 0.797)	0.828 (0.767, 0.874)	0.458 (0.434, 0.482)	0.466 (0.435, 0.495)	0.503 (0.446, 0.555)	0.366 (0.327, 0.403)
GREED [30]	0.823 (0.811, 0.834)	0.642 (0.628, 0.656)	0.210 (0.189, 0.231)	0.145 (0.128, 0.162)	0.805 (0.796, 0.815)	0.808 (0.748, 0.854)	0.593 (0.562, 0.621)	0.682 (0.557, 0.777)	0.593 (0.554, 0.630)	0.448 (0.417, 0.478)
Y-VQM [48]	0.752 (0.721, 0.779)	0.586 (0.559, 0.611)	0.265 (0.213, 0.315)	0.149 (0.117, 0.181)	0.842 (0.833, 0.851)	0.833 (0.780, 0.873)	0.787 (0.754, 0.816)	0.589 (0.565, 0.613)	0.843 (0.804, 0.874)	0.700 (0.656, 0.740)
DISTS [12]	0.901 (0.896, 0.906)	0.731 (0.723, 0.739)	**0.417** (0.397, 0.436)	**0.245** (0.229, 0.260)	0.866 (0.860, 0.873)	0.873 (0.828, 0.908)	0.711 (0.695, 0.726)	0.731 (0.621, 0.812)	0.626 (0.601, 0.650)	0.460 (0.440, 0.478)
AVQT [2]	0.923 (0.918, 0.927)	0.784 (0.777, 0.791)	0.176 (0.129, 0.222)	0.075 (0.042, 0.107)	0.894 (0.889, 0.899)	0.872 (0.831, 0.903)	0.857 (0.833, 0.877)	0.926 (0.858, 0.962)	0.842 (0.812, 0.867)	0.698 (0.665, 0.728)
YUV-PSNR	0.907 (0.901, 0.912)	**0.869** (0.814, 0.908)	0.239 (0.184, 0.293)	0.119 (0.085, 0.152)	0.893 (0.886, 0.898)	0.911 (0.874, 0.937)	0.813 (0.786, 0.837)	0.641 (0.616, 0.664)	**0.900** (0.878, 0.919)	**0.773** (0.743, 0.800)
YUV-SSIM	0.937 (0.935, 0.940)	0.820 (0.813, 0.826)	**0.302** (0.256, 0.347)	0.177 (0.144, 0.209)	0.915 (0.910, 0.920)	0.921 (0.888, 0.944)	**0.869** (0.848, 0.887)	0.874 (0.788, 0.926)	**0.912** (0.891, 0.929)	**0.806** (0.776, 0.832)
Y-MS-SSIM [46]	**0.952** (0.950, 0.955)	**0.897** (0.854, 0.928)	0.259 (0.211, 0.306)	0.134 (0.102, 0.166)	0.910 (0.905, 0.914)	0.901 (0.864, 0.928)	0.860 (0.839, 0.879)	0.819 (0.741, 0.876)	**0.905** (0.882, 0.923)	**0.778** (0.747, 0.805)
Y-VMAF NEG [26]	**0.946** (0.943, 0.948)	0.823 (0.818, 0.827)	0.268 (0.215, 0.320)	0.163 (0.128, 0.197)	0.910 (0.905, 0.915)	0.902 (0.865, 0.928)	0.863 (0.842, 0.881)	**0.957** (0.909, 0.980)	0.880 (0.855, 0.900)	0.731 (0.700, 0.759)
YUV-VMAF NEG [26]	0.945 (0.942, 0.947)	0.836 (0.831, 0.841)	0.245 (0.196, 0.293)	0.146 (0.113, 0.179)	**0.920** (0.915, 0.925)	**0.925** (0.894, 0.947)	0.861 (0.840, 0.880)	0.884 (0.806, 0.933)	0.896 (0.873, 0.915)	0.766 (0.736, 0.793)
Y-VMAF (v061) [26]	0.932 (0.928, 0.936)	0.803 (0.798, 0.809)	**0.453** (0.405, 0.499)	**0.366** (0.337, 0.394)	**0.940** (0.937, 0.944)	**0.922** (0.893, 0.943)	**0.905** (0.890, 0.919)	**0.997** (0.992, 0.999)	0.874 (0.849, 0.895)	0.732 (0.701, 0.760)
YUV-VMAF (v061) [26]	**0.952** (0.950, 0.954)	**0.846** (0.841, 0.851)	0.274 (0.225, 0.322)	**0.216** (0.186, 0.246)	**0.946** (0.942, 0.949)	**0.939** (0.914, 0.957)	**0.910** (0.894, 0.924)	**0.996** (0.988, 0.999)	0.897 (0.874, 0.916)	0.767 (0.737, 0.794)

Table 4: Results for SROCC and KROCC on five subsets of our dataset (by encoding category).

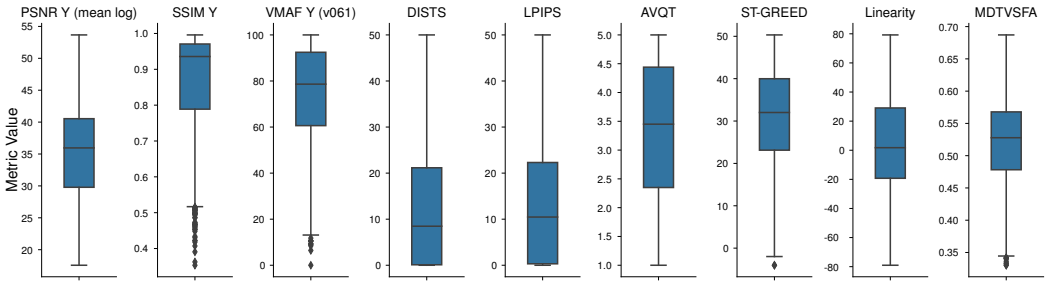

Figure 4: Distribution of metric scores. Each metric appears on a separate axis.

4 Conclusion

We created a new diverse dataset containing 2,486 videos compressed by various encoding standards, including AVC, HEVC, AV1, and VVC. We used it to analyze the correlation between new learning-based objective-quality metrics and subjective-quality scores. Our analysis revealed that some new no-reference metrics, such as MDTVSFA, have already caught up with full-reference metrics. At the same time, VMAF showed the highest correlation with subjective scores, making it the best full-reference option for assessing video-compression quality. The open part of the dataset is available publicly [3]. The code, with an example metric launch running on that part of the dataset, is also available [4].

[3] https://videoprocessing.ai/datasets/vqa.html

[4] https://github.com/msu-video-group/MSU_VQM_Compression_Benchmark

Dataset	User-Generated Content 203 videos		Shaking 594 videos		Sports 582 videos		Nature 1216 videos		Gaming / Animation 204 videos	
Metric	SROC	KROC	SROC	KROC	SROC	KROC	SROC	KROC	SROC	KROC
No-Reference										
MEON [29]	0.072 (0.021, 0.122)	0.102 (0.061, 0.143)	0.221 (0.201, 0.241)	0.176 (0.162, 0.190)	0.645 (0.625, 0.665)	0.490 (0.473, 0.507)	0.524 (0.508, 0.539)	0.405 (0.393, 0.416)	0.707 (0.677, 0.734)	0.531 (0.503, 0.557)
Y-NIQE [35]	0.232 (0.148, 0.312)	0.164 (0.102, 0.223)	0.473 (0.437, 0.507)	0.332 (0.305, 0.358)	0.672 (0.658, 0.685)	0.485 (0.474, 0.496)	0.621 (0.604, 0.638)	0.443 (0.429, 0.457)	0.810 (0.794, 0.826)	0.623 (0.602, 0.644)
VIDEVAL [41]	0.422 (0.365, 0.475)	0.304 (0.260, 0.346)	0.535 (0.511, 0.559)	0.386 (0.367, 0.406)	0.836 (0.827, 0.844)	0.645 (0.634, 0.655)	0.739 (0.726, 0.751)	0.548 (0.536, 0.560)	0.923 (0.912, 0.933)	0.773 (0.751, 0.792)
KonCept512 [17]	0.789 (0.771, 0.805)	0.629 (0.612, 0.646)	0.833 (0.819, 0.847)	0.651 (0.634, 0.666)	0.843 (0.836, 0.849)	0.669 (0.661, 0.676)	0.809 (0.802, 0.816)	0.642 (0.634, 0.649)	0.755 (0.732, 0.777)	0.565 (0.543, 0.587)
NIMA [40]	0.826 (0.809, 0.842)	0.674 (0.654, 0.693)	**0.849** (0.838, 0.859)	**0.677** (0.664, 0.689)	0.837 (0.831, 0.843)	0.656 (0.649, 0.664)	0.792 (0.783, 0.801)	0.609 (0.599, 0.618)	0.814 (0.804, 0.823)	0.637 (0.628, 0.645)
PaQ-2-PiQ [51]	0.801 (0.778, 0.822)	0.659 (0.637, 0.679)	0.807 (0.792, 0.821)	0.643 (0.627, 0.659)	0.908 (0.903, 0.913)	0.753 (0.745, 0.761)	0.827 (0.818, 0.835)	0.658 (0.648, 0.667)	0.963 (0.961, 0.965)	0.855 (0.851, 0.860)
SPAQ MT-A [13]	0.786 (0.733, 0.829)	0.624 (0.570, 0.672)	0.797 (0.776, 0.816)	0.607 (0.585, 0.629)	0.848 (0.841, 0.854)	0.678 (0.671, 0.684)	0.857 (0.848, 0.864)	0.690 (0.680, 0.700)	0.942 (0.937, 0.947)	0.801 (0.793, 0.810)
SPAQ BL [13]	0.764 (0.709, 0.810)	0.608 (0.553, 0.657)	0.797 (0.773, 0.818)	0.603 (0.578, 0.627)	0.869 (0.864, 0.875)	0.699 (0.693, 0.706)	0.865 (0.856, 0.872)	0.693 (0.683, 0.704)	0.924 (0.921, 0.926)	0.773 (0.769, 0.778)
SPAQ MT-S [13]	0.780 (0.739, 0.815)	0.619 (0.575, 0.661)	0.756 (0.735, 0.776)	0.568 (0.546, 0.589)	0.890 (0.884, 0.895)	0.727 (0.720, 0.734)	0.863 (0.855, 0.871)	0.693 (0.683, 0.703)	0.954 (0.949, 0.958)	0.830 (0.819, 0.839)
VSFA [20]	**0.852** (0.843, 0.861)	**0.690** (0.678, 0.702)	0.830 (0.818, 0.841)	0.654 (0.640, 0.667)	**0.911** (0.905, 0.916)	**0.758** (0.750, 0.765)	**0.873** (0.867, 0.878)	**0.708** (0.701, 0.716)	**0.975** (0.972, 0.978)	**0.881** (0.873, 0.889)
Linearity [21]	**0.893** (0.882, 0.903)	**0.753** (0.738, 0.767)	**0.864** (0.852, 0.874)	**0.688** (0.674, 0.702)	**0.931** (0.927, 0.936)	**0.787** (0.780, 0.795)	**0.904** (0.899, 0.909)	**0.754** (0.747, 0.761)	**0.964** (0.961, 0.967)	**0.856** (0.849, 0.862)
MDTVSFA [22]	**0.912** (0.904, 0.918)	**0.776** (0.763, 0.788)	**0.897** (0.888, 0.905)	**0.742** (0.729, 0.754)	**0.924** (0.920, 0.928)	**0.780** (0.773, 0.787)	**0.917** (0.913, 0.921)	**0.772** (0.766, 0.778)	**0.971** (0.968, 0.973)	**0.866** (0.860, 0.872)
Full-Reference										
FOV VIDEO [31]	0.625 (0.596, 0.651)	0.464 (0.441, 0.486)	0.540 (0.528, 0.553)	0.387 (0.378, 0.397)	0.560 (0.546, 0.574)	0.401 (0.389, 0.412)	0.516 (0.506, 0.527)	0.373 (0.365, 0.380)	0.566 (0.550, 0.581)	0.396 (0.384, 0.408)
LPIPS [52]	0.724 (0.692, 0.753)	0.587 (0.565, 0.609)	0.639 (0.617, 0.660)	0.473 (0.455, 0.490)	0.854 (0.846, 0.861)	0.674 (0.665, 0.683)	0.746 (0.734, 0.758)	0.565 (0.554, 0.576)	0.785 (0.772, 0.798)	0.597 (0.586, 0.608)
DVQA [1]	0.847 (0.824, 0.867)	0.689 (0.664, 0.711)	0.708 (0.687, 0.727)	0.528 (0.510, 0.545)	0.841 (0.832, 0.849)	0.653 (0.642, 0.663)	0.783 (0.772, 0.792)	0.600 (0.590, 0.610)	0.883 (0.872, 0.894)	0.709 (0.694, 0.723)
GREED [30]	0.719 (0.685, 0.749)	0.572 (0.543, 0.600)	0.726 (0.712, 0.740)	0.551 (0.538, 0.564)	0.805 (0.800, 0.811)	0.640 (0.634, 0.645)	0.748 (0.739, 0.755)	0.580 (0.573, 0.588)	0.828 (0.812, 0.842)	0.643 (0.624, 0.662)
Y-VQM [48]	0.809 (0.787, 0.830)	0.656 (0.637, 0.675)	0.810 (0.798, 0.822)	0.634 (0.620, 0.647)	0.867 (0.857, 0.877)	0.689 (0.675, 0.702)	0.848 (0.841, 0.855)	0.677 (0.669, 0.685)	0.930 (0.922, 0.936)	0.779 (0.766, 0.791)
DISTS [12]	0.854 (0.831, 0.874)	0.694 (0.668, 0.718)	0.794 (0.780, 0.807)	0.613 (0.599, 0.627)	0.893 (0.888, 0.898)	0.726 (0.719, 0.733)	0.834 (0.826, 0.841)	0.657 (0.649, 0.665)	0.896 (0.888, 0.902)	0.727 (0.717, 0.737)
AVQT [2]	**0.919** (0.908, 0.928)	**0.791** (0.774, 0.807)	0.849 (0.838, 0.860)	0.684 (0.669, 0.698)	0.902 (0.896, 0.908)	0.757 (0.748, 0.765)	0.877 (0.871, 0.883)	0.719 (0.712, 0.727)	0.913 (0.910, 0.915)	0.772 (0.768, 0.776)
YUV-PSNR	0.770 (0.740, 0.797)	0.638 (0.616, 0.658)	0.810 (0.797, 0.822)	0.645 (0.632, 0.658)	0.933 (0.928, 0.937)	0.793 (0.785, 0.801)	0.868 (0.860, 0.876)	0.710 (0.701, 0.719)	0.944 (0.938, 0.950)	0.807 (0.795, 0.818)
YUV-SSIM	0.779 (0.750, 0.805)	0.642 (0.618, 0.665)	0.811 (0.797, 0.824)	0.648 (0.633, 0.663)	**0.952** (0.948, 0.957)	**0.828** (0.818, 0.837)	0.900 (0.893, 0.907)	0.750 (0.740, 0.759)	0.958 (0.951, 0.964)	0.837 (0.823, 0.850)
Y-MS-SSIM [46]	0.895 (0.882, 0.907)	0.746 (0.729, 0.762)	0.851 (0.841, 0.862)	0.680 (0.666, 0.693)	0.942 (0.938, 0.947)	0.808 (0.800, 0.817)	0.901 (0.895, 0.907)	0.746 (0.738, 0.754)	0.955 (0.950, 0.960)	0.832 (0.821, 0.841)
Y-VMAF NEG [26]	0.916 (0.907, 0.925)	0.779 (0.765, 0.791)	**0.861** (0.851, 0.870)	0.688 (0.675, 0.700)	0.945 (0.940, 0.949)	0.810 (0.802, 0.818)	0.909 (0.903, 0.914)	0.757 (0.749, 0.765)	0.960 (0.956, 0.964)	0.841 (0.832, 0.849)
YUV-VMAF NEG [26]	0.916 (0.907, 0.924)	0.773 (0.761, 0.785)	**0.861** (0.851, 0.870)	**0.691** (0.678, 0.703)	0.947 (0.942, 0.951)	0.815 (0.807, 0.823)	**0.913** (0.908, 0.918)	**0.763** (0.756, 0.771)	**0.968** (0.965, 0.970)	**0.860** (0.854, 0.866)
Y-VMAF (v061) [26]	**0.946** (0.940, 0.952)	**0.835** (0.822, 0.846)	**0.891** (0.882, 0.900)	**0.730** (0.717, 0.743)	**0.959** (0.955, 0.962)	**0.836** (0.828, 0.843)	**0.942** (0.939, 0.946)	**0.810** (0.803, 0.816)	**0.967** (0.964, 0.969)	**0.860** (0.854, 0.866)
YUV-VMAF (v061) [26]	**0.942** (0.935, 0.948)	**0.820** (0.807, 0.832)	**0.879** (0.870, 0.888)	**0.716** (0.703, 0.729)	**0.961** (0.958, 0.964)	**0.843** (0.836, 0.850)	**0.944** (0.941, 0.947)	**0.813** (0.806, 0.819)	**0.972** (0.971, 0.974)	**0.872** (0.868, 0.876)

Table 5: Results for SROCC and KROCC on five subsets of our dataset (by content type).

No-Reference Metric	VIDEVAL[1] (CPU)	MEON[1] (CPU)	Linearity[1]	KonCept512[1]	SPAQ MT-S[1]	SPAQ BL[1]	SPAQ MT-A[1]	NIMA[1]	MDTVSFA[1]	VSF[1]A	PaQ-2-PiQ[1]	NIQE[1]
Computation Complexity (FPS)	0.62	2.27	3.41	4.18	6.48	6.49	6.66	7.24	9.12	9.26	11.10	80.00

Full-Reference Metric	LPIPS[1]	DVQA[1]	GREED[1]	DISTS[1]	FOV VIDEO[1]	AVQT (CPU)[2]	VMAF[1]	MS-SSIM[1]	SSIM[1]	VQM[1]	PSNR[1]
Computation Complexity (FPS)	3.20	5.75	7.10	8.65	37.57	37.66	52.62	99.36	160.58	280.00	371.96

Table 6: FPS evaluation for videos from the dataset. The metric testing used a configuration with two Intel Xeon Silver 4216 processors running Ubuntu 20.04 at 2.10 GHz with a Titan RTX GPU, and another configuration with an Intel Core i9 processor running at 2.3 GHz with 16 GB of RAM and AMD Radeon Pro 5500M 4 GB graphics card.

Our proposed dataset will be useful for researchers and developers of image- and video-quality metrics that evaluate video-compression artifacts. It can serve in training models that assess video-compression quality to achieve more-precise results and higher correlation with subjective scores. Our benchmark will remain an unbiased test of compression quality for new image- and video-quality metrics.

We are accepting new methods for evaluation using our benchmark[5]. During the few months since its publication, we have already received several submissions, as well as good reviews and requests for further development. Our plan is to further increase the number of original videos and add new encoders. Because the subjective tests are expensive, we estimate our current dataset cost about $15,000. We are open to collaboration and sponsorship to improve the dataset more quickly and to provide more-reliable and more-valuable results.

[5] https://videoprocessing.ai/benchmarks/video-quality-metrics.html

4.1 Limitations

We did not retrain the tested metrics on the open part of our dataset. We used already trained models without tuning their parameters. This approach allowed us to prevent metrics from overfitting on our dataset. Nevertheless, some methods are not fitted for data that was absent from the training set (for instance, compression artifacts) or simply underwent training on small datasets. As a result, these metrics may show weak performance on our dataset. Future work will therefore include metric retraining on open part of the dataset and assessment of their quality on the hidden part.

4.2 Acknowledgments

The work received support through a grant for research centers in the field of artificial intelligence (agreement identifier 000000D730321P5Q0002, dated November 2, 2021, no. 70-2021-00142 with the Ivannikov Institute for System Programming of the Russian Academy of Sciences).

References

[1] Dvqa - deep learning-based video quality assessment. https://github.com/Tencent/DVQA. Accessed: 2022-08-06. 1, 4, 7, 8, 9

[2] Evaluate videos with the advanced video quality tool. https://developer.apple.com/videos/play/wwdc2021/10145. Accessed: 2022-08-06. 1, 4, 7, 8, 9

[3] Recommended encoding settings for ibm watson media. https://support.video.ibm.com/hc/en-us/articles/207852117-Internet-connection-and-recommended-encoding-settings. Accessed: 2022-08-12. 5

[4] Recommended encoding settings for twitch streaming. https://stream.twitch.tv/encoding/. Accessed: 2022-08-12. 5

[5] Recommended encoding settings for youtube. https://support.google.com/youtube/answer/2853702. Accessed: 2022-08-12. 5

[6] Toward a better quality metric for the video community, netflix technology blog. https://netflixtechblog.com/toward-a-better-quality-metric-for-the-video-community\-7ed94e752a30. Accessed: 2022-08-17. 6

[7] Christos George Bampis, Zhi Li, Anush Krishna Moorthy, Ioannis Katsavounidis, Anne Aaron, and Alan Conrad Bovik. Study of temporal effects on subjective video quality of experience. *IEEE Transactions on Image Processing*, 26(11):5217–5231, 2017. 2

[8] Nabajeet Barman, Emmanuel Jammeh, Seyed Ali Ghorashi, and Maria G Martini. No-reference video quality estimation based on machine learning for passive gaming video streaming applications. *IEEE Access*, 7:74511–74527, 2019. 3

[9] Nabajeet Barman, Saman Zadtootaghaj, Steven Schmidt, Maria G Martini, and Sebastian Möller. Gamingvideoset: a dataset for gaming video streaming applications. In *2018 16th Annual Workshop on Network and Systems Support for Games (NetGames)*, pages 1–6. IEEE, 2018. 3

[10] David Corey, William Dunlap, and Michael Burke. Averaging correlations: Expected values and bias in combined pearson rs and fisher's z transformations. *Journal of General Psychology - J GEN PSYCHOL*, 125:245–261, 07 1998. 6

[11] Francesca De Simone, Marco Tagliasacchi, Matteo Naccari, Stefano Tubaro, and Touradj Ebrahimi. A h. 264/avc video database for the evaluation of quality metrics. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2430–2433. IEEE, 2010. 2

[12] Keyan Ding, Kede Ma, Shiqi Wang, and Eero Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 12 2020. 4, 7, 8, 9

[13] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3677–3686, 2020. 4, 7, 8, 9

[14] Deepti Ghadiyaram, Janice Pan, Alan C Bovik, Anush Krishna Moorthy, Prasanjit Panda, and Kai-Chieh Yang. In-capture mobile video distortions: A study of subjective behavior and objective algorithms. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9):2061–2077, 2017. 2, 3

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[16] Vlad Hosu, Franz Hahn, Mohsen Jenadeleh, Hanhe Lin, Hui Men, Tamás Szirányi, Shujun Li, and Dietmar Saupe. The konstanz natural video database (konvid-1k). In *2017 Ninth international conference on quality of multimedia experience (QoMEX)*, pages 1–6. IEEE, 2017. 2, 3

[17] Vlad Hosu, Hanhe Lin, Tamas Sziranyi, and Dietmar Saupe. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing*, 29:4041–4056, 2020. 4, 7, 8, 9

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[19] Christian Keimel, Arne Redl, and Klaus Diepold. The tum high definition video datasets. In *2012 Fourth international workshop on quality of multimedia experience*, pages 97–102. IEEE, 2012. 2

[20] Dingquan Li, Tingting Jiang, and Ming Jiang. Quality assessment of in-the-wild videos. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2351–2359, 2019. 4, 7, 8, 9

[21] Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 789–797, 2020. 4, 7, 8, 9

[22] Dingquan Li, Tingting Jiang, and Ming Jiang. Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4):1238–1257, 2021. 3, 4, 7, 8, 9

[23] Songnan Li, Fan Zhang, Lin Ma, and King Ngan. Image quality assessment by separately evaluating detail losses and additive impairments. *IEEE Transactions on Multimedia*, 13:935–949, 10 2011. 4

[24] Yang Li, Shengbin Meng, Xinfeng Zhang, Meng Wang, Shiqi Wang, Yue Wang, and Siwei Ma. User-generated video quality assessment: A subjective and objective study. *IEEE Transactions on Multimedia*, 2021. 3

[25] Yang Li, Shengbin Meng, Xinfeng Zhang, Shiqi Wang, Yue Wang, and Siwei Ma. Ugc-video: Perceptual quality assessment of user-generated videos. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 35–38, 2020. 3

[26] Zhi Li, Christos Bampis, Julie Novak, Anne Aaron, Kyle Swanson, Anush Moorthy, and JD Cock. Vmaf: The journey continues. *Netflix Technology Blog*, 25, 2018. 1, 4, 7, 8, 9

[27] Joe Yuchieh Lin, Rui Song, Chi-Hao Wu, TsungJung Liu, Haiqiang Wang, and C-C Jay Kuo. Mcl-v: A streaming video quality assessment database. *Journal of Visual Communication and Image Representation*, 30:1–9, 2015. 2

[28] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982. 4

[29] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2017. 3, 7, 8, 9

[30] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. ST-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction. *IEEE Trans. Image Process.*, 2021. 4, 7, 8, 9

[31] Rafał K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021. 4, 7, 8, 9

[32] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Mylene CQ Farias, and Alan Conrad Bovik. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Transactions on Image Processing*, 29:6054–6068, 2020. 2

[33] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012. 3

[34] Anish Mittal, Michele A Saad, and Alan C Bovik. A completely blind video integrity oracle. *IEEE Transactions on Image Processing*, 25(1):289–300, 2015. 3

[35] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012. 7, 8, 9

[36] Mikko Nuutinen, Toni Virtanen, Mikko Vaahteranoksa, Tero Vuori, Pirkko Oittinen, and Jukka Häkkinen. Cvd2014—a database for evaluating no-reference video quality assessment algorithms. *IEEE Transactions on Image Processing*, 25(7):3073–3086, 2016. 2, 3

[37] Pradip Paudyal, Federica Battisti, and Marco Carli. A study on the effects of quality of service parameters on perceived video quality. In *2014 5th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2014. 2

[38] Hamid Sheikh, Alan Bovik, and Gustavo Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *Image Processing, IEEE Transactions on*, 14:2117–2128, 01 2006. 4

[39] Zeina Sinno and Alan Conrad Bovik. Large-scale study of perceptual video quality. *IEEE Transactions on Image Processing*, 28(2):612–627, 2018. 2, 3

[40] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018. 4, 7, 8, 9

[41] Zhengzhong Tu, Chia-Ju Chen, Yilin Wang, Neil Birkbeck, Balu Adsumilli, and Alan C Bovik. Video quality assessment of user generated content: A benchmark study and a new model. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1409–1413. IEEE, 2021. 2, 3, 7, 8, 9

[42] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1509–1513. IEEE, 2016. 2, 3

[43] Haiqiang Wang, Ioannis Katsavounidis, Jiantong Zhou, Jeonghoon Park, Shawmin Lei, Xin Zhou, Man-On Pun, Xin Jin, Ronggang Wang, Xu Wang, et al. Videoset: A large-scale compressed video quality dataset based on jnd measurement. *Journal of Visual Communication and Image Representation*, 46:292–302, 2017. 2, 3

[44] Yilin Wang, Sasi Inguva, and Balu Adsumilli. Youtube ugc dataset for video compression research. In *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–5. IEEE, 2019. 2, 3

[45] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13:600 – 612, 05 2004. 4

[46] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 4, 7, 8, 9

[47] Stefan Winkler. Analysis of public image and video databases for quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):616–625, 2012. 5

[48] Feng Xiao et al. Dct-based video quality evaluation. *Final Project for EE392J*, 769, 2000. 7, 8, 9

[49] Munan Xu, Junming Chen, Haiqiang Wang, Shan Liu, Ge Li, and Zhiqiang Bai. C3dvqa: Full-reference video quality assessment with 3d convolutional neural network. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4447–4451. IEEE, 2020. 4

[50] Zhenqiang Ying, Maniratnam Mandal, Deepti Ghadiyaram, and Alan Bovik. Patch-vq:'patching up'the video quality problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14019–14029, 2021. 2, 3

[51] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepti Ghadiyaram, and Alan Bovik. From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3575–3585, 2020. 4, 7, 8, 9

[52] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595. IEEE, 2018. 4, 7, 8, 9