Domain-Specific Fusion Of Multiple Objective Quality Metrics

Presenter: Yiannis Andreopoulos

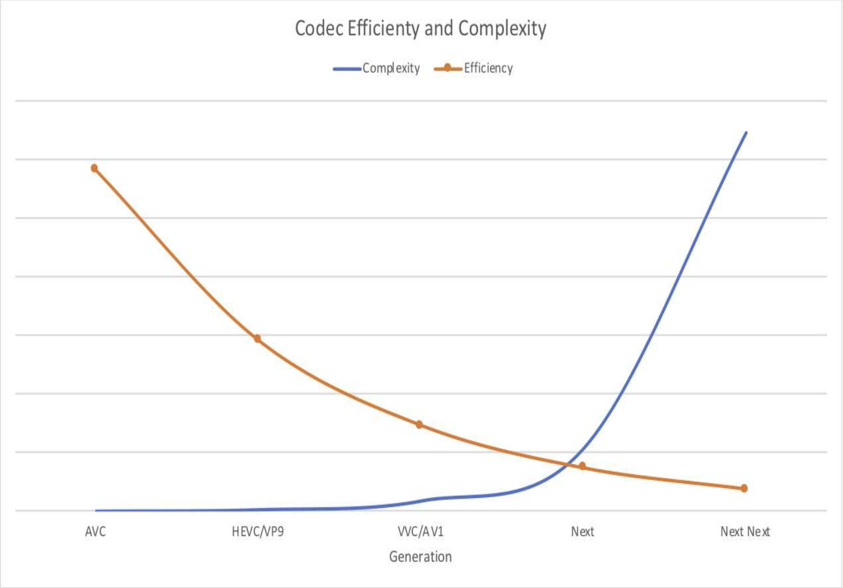
VQEG meeting, May 2022
Rennes, France

Joint work of iSIZE, industrial collaborators and the Innovate UK SEQUOIA consortium, project: 96984

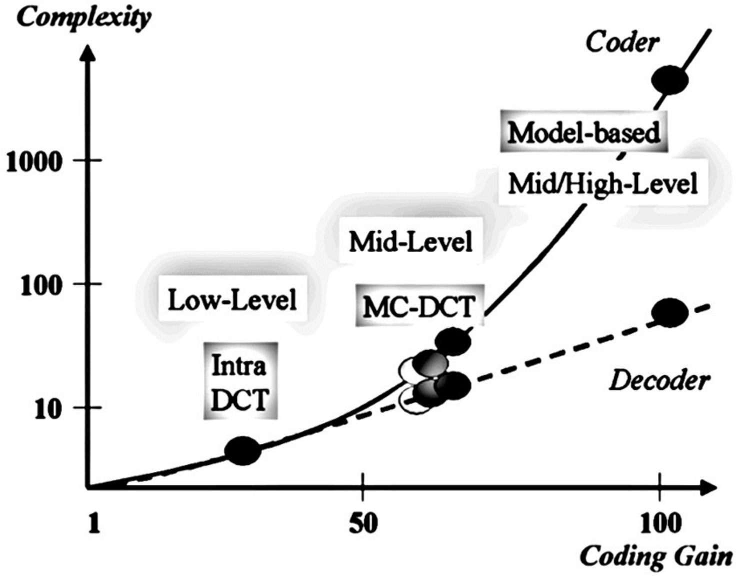
iSIZE: What we do

Problem we solve	Solution	Target market	Description
Perceptual Quality	Deep perceptual optimization BITSAVE demo.isize.co	• Entertainment / Media • Video streaming • Gaming • Social media	• Deep psychovisual preprocessing for maximum bitrate savings. Significantly advance the development of AI-based quality metrics and quality scoring • AI-based preprocessing that requires no change in encoding, delivery or decoding devices
Noisy Video Content	Deep perceptual denoising BITCLEAR http://bitclear.isize.co/	• Social media/user uploads • Post-decoder enhancement	• Remove compression noise from video content by addressing the problem across the quality-bitrate-complexity space • Can work both as a server and as a client component (post-decoder)
Low-Bitrate/Low-Latency Video Delivery	Domain-specific generative video representations BITGEN	• Conversational services • Virtual reality/telepresence • IoT/driverless technologies	• Extreme reduction in video bitrate, working in a compact latent space. • Enable telepresence with near-zero latency.

iSIZE: Why we do it



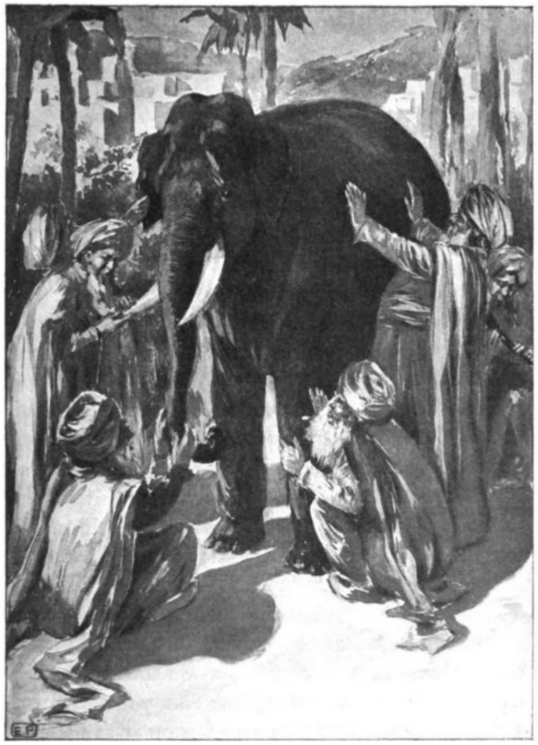
https://www.linkedin.com/pulse/encoder-complexity-hits-wall-david-ronca/ (D. Ronca, Meta, 2019)



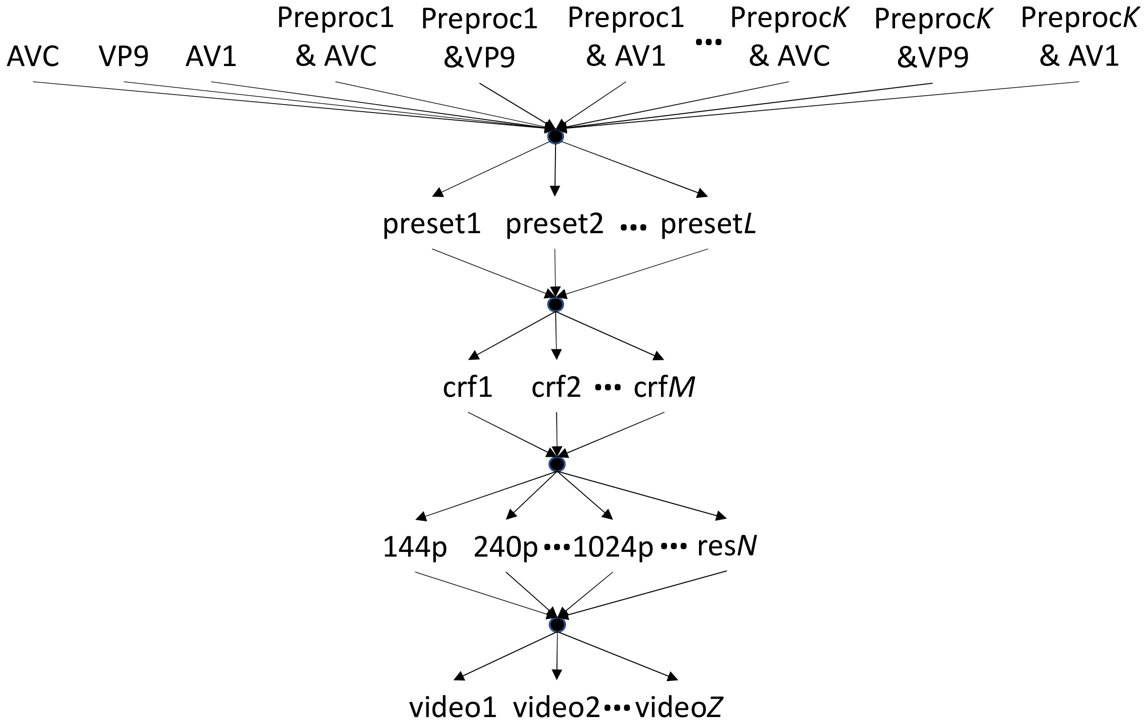
Sikora, Proc. of the IEEE, 2005, https://doi.org/10.1109/JPROC.2004.839601

- Device power+heat dissipation and cloud-based scaling have both hit the wall
- Inflection point: quality metrics and neural network hardware now allow for AI-based pre- and post-processing
- Codecs are amazing SNR/SSIM-vs.-bitrate machines, but these loss functions have significant limitations

iSIZE The three challenges with visual quality assessment



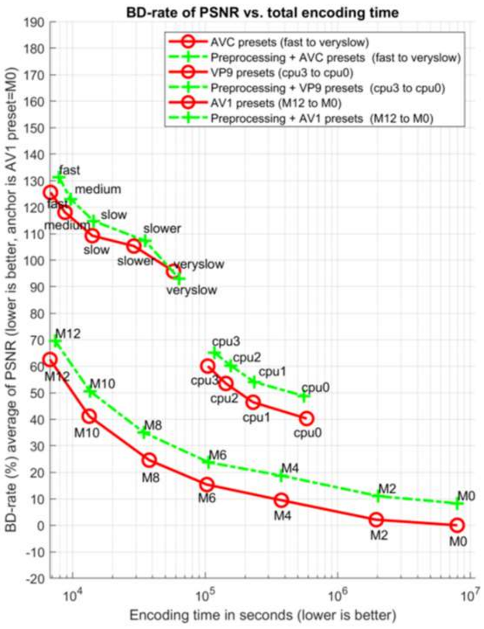
https://en.wikipedia.org/wiki/Blind_men_and_an_elephant

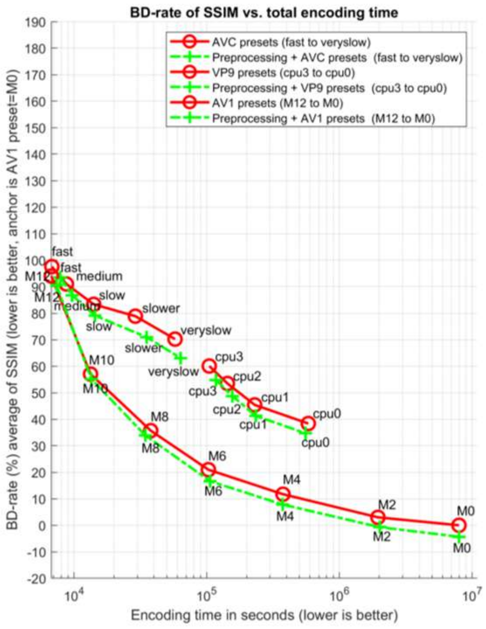


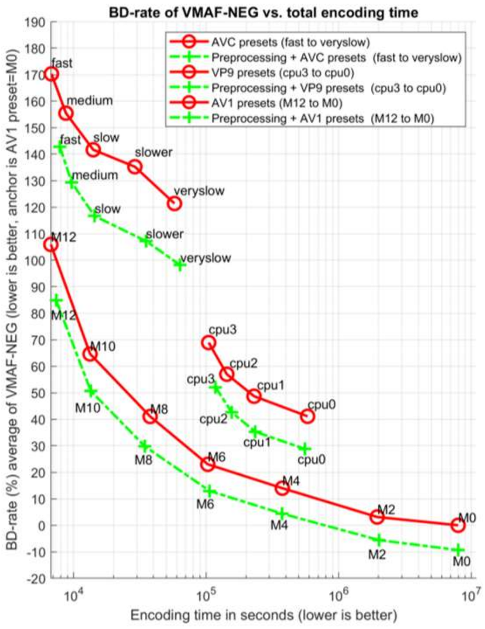
1. Objective metrics (and humans) are myopic

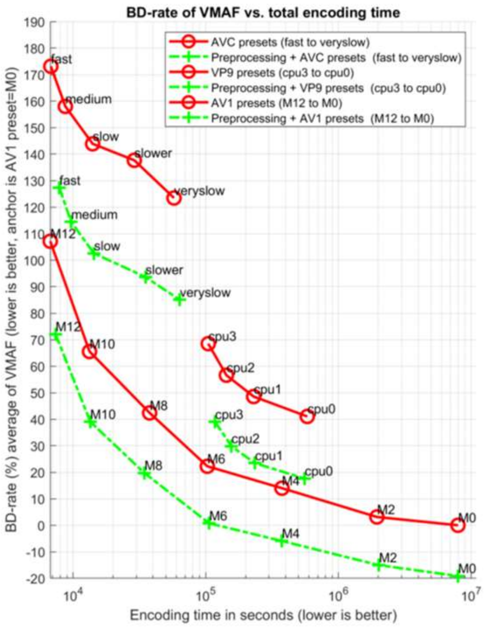
2. The exploration space can surpass 1m tests for a 100-video library

iSIZE The three challenges with visual quality assessment


BD-rate of PSNR vs. total encoding time


BD-rate of SSIM vs. total encoding time

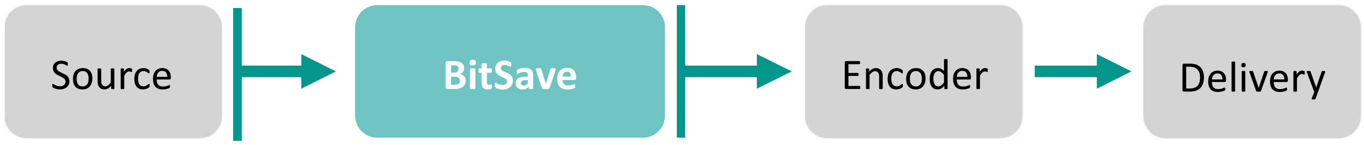

BD-rate of VMAF-NEG vs. total encoding time


BD-rate of VMAF vs. total encoding time

3. Video processing algorithms are now increasingly optimized for perceptual quality metrics instead of signal distortion

→ This means that they may score well for metrics like VMAF, but this may be because of metric overfitting.

Methods: iSIZE BitSave preprocessing (mk3)



AI-based pre-processing prior to encoding (AVC, HEVC, VP9, AV1)

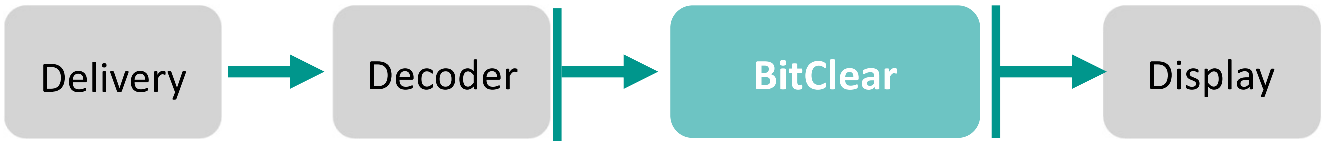
One frame latency

Single pass processing per content for an entire ABR ladder

Improves encoding quality as measured by standard perceptual quality metrics (VMAF, SSIM, VIF), can also work in tandem with BitClear

Integrated within Intel OpenVINO, ONNX and Dolby Vision, easy to plug&play within any existing workflow

Methods: iSIZE BitClear post-processing (mk3.5)



AI-based post-processing after decoding (AVC, HEVC, VP9, AV1)

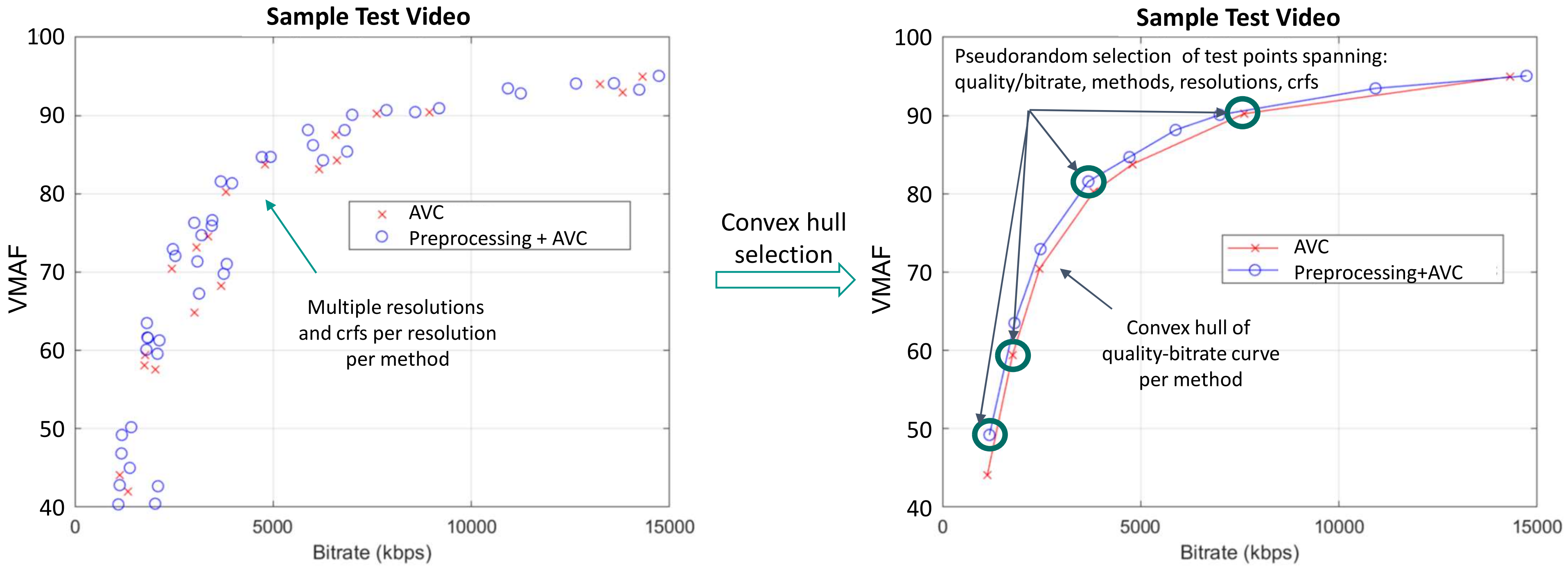
One frame latency

Single pass processing per content for an entire ABR ladder

Improves decoding quality as measured by standard perceptual quality metrics (VMAF, SSIM, VIF), can also work in tandem with BitSave

Integrated within Intel OpenVINO and ONNX, easy to plug&play within any existing workflow

Domain-specific fusion of multiple quality metrics



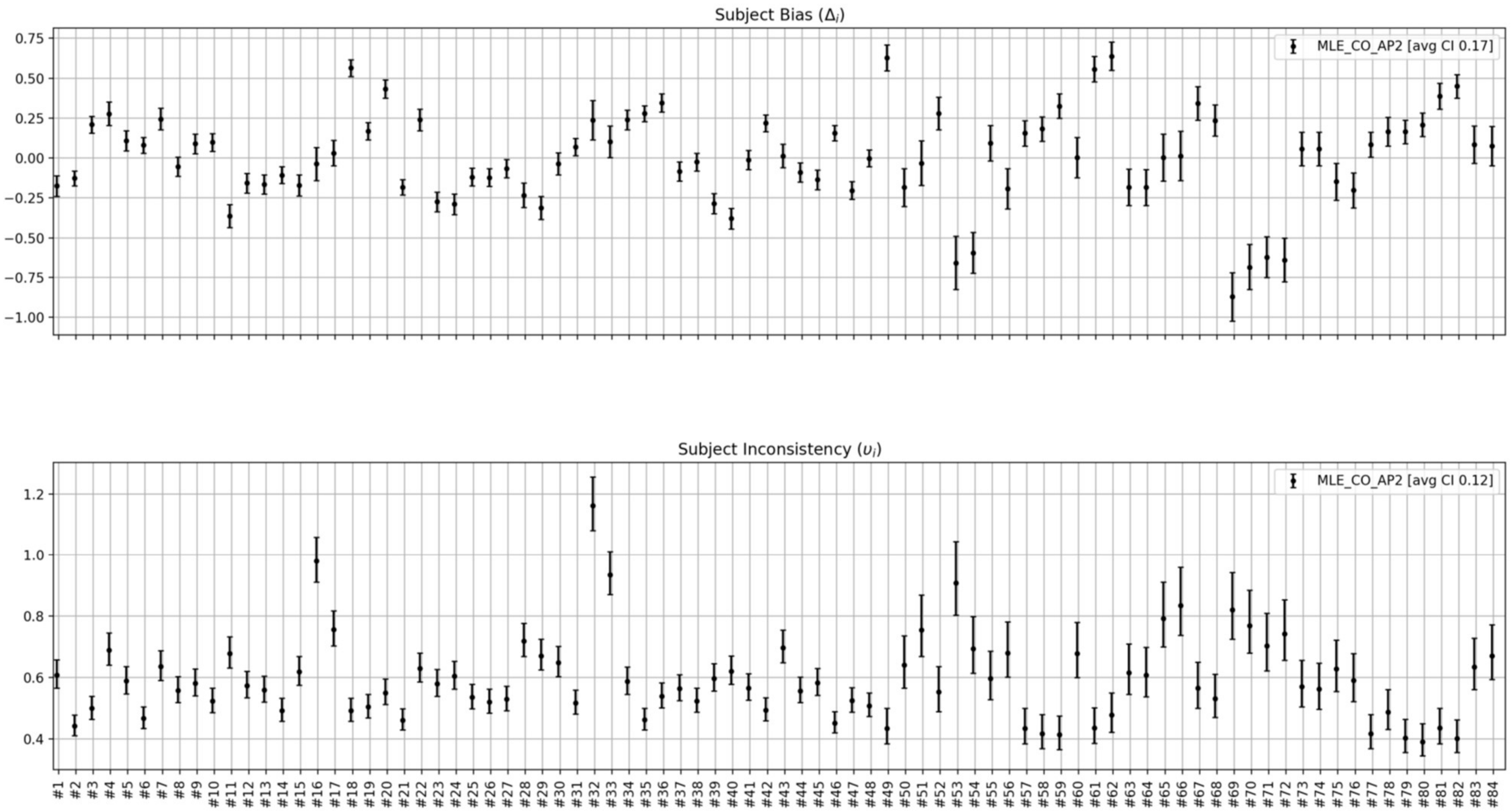
Three steps:

1. Convex-hull selection based on VMAF
2. Carry out P.910 ACR and post-processing
3. Fuse metrics to recovered quality scores using support vector regression

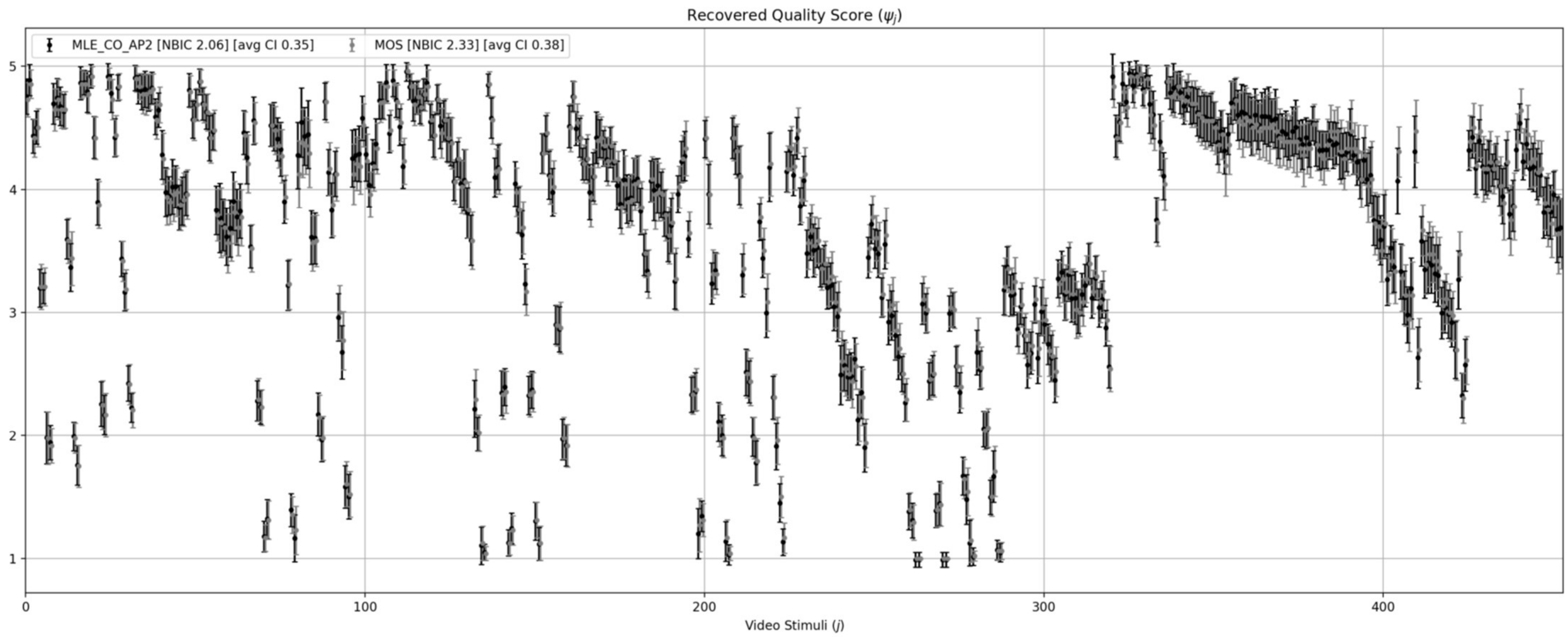
P.910 ACR test setup and conditions

Setup Component	What was Used	Further Details on Settings	Comments
Encoders	AVC x264 (Lavc58.134.100 libx264) WEBM VP9 (v1.10.0-48-g4ec84326c)	• 1080p, 720p, 540p, 360p, 216p (only underlined done for post-processing) • Per resolution: AVC preset=veryslow, CRF={22,30,38,46} (medium used for post-processing) • Per resolution: VP9 preset=0, CRF={32,38,44,46,48,50,52,54,56,58,60} (underlined CRFs done for 720p & 540p, preset=5 used for post-processing)	• The slowest preset of each encoder was used for preprocessing, faster presets for post-processing • Constant-CRF encoding ensures quality remains consistent, no effects from rate control algorithms • The range of CRFs ensures the full quality range of relevance to each resolution & application is sampled • All lower resolutions were upscaled to 1080p for viewing using FFmpeg Lanczos-5
Content and test conditions	AV2 CTC content https://media.xiph.org/video/aomctc/test_set/ P.910 ACR standard test conditions applied	• 3H distance, controlled lighting, same screen conditions for all tests • Ratings from 1-5 • Raters were briefed on task and how to use the quality scaling	• All content replayed at 25fps, 1080p@50Hz TV screen, all TV filters were off • 21 sequences at 1080p resolution (8bit) used, comprising a mixture of entertainment, sports, UGC, gaming, web browsing, and artistic content (16 sequences for post-processing)
Raters and data processing	• 48 raters for preprocessing (the underlined VP9 CRFs had 36 additional raters) • 24 raters for post-processing • The SUREAL package was used for post-processing	• All raters were screened for color blindness and good eyesight • All 16368 ratings were used	• SUREAL: https://github.com/Netflix/sureal • The full maximum likelihood estimation (MLE) model of SUREAL was used, which takes into account both subjects and contents • For quality-bitrate plots per resolution and cross-resolution combined quality-bitrate plots, an MLE fit per codec was carried out and the recovered quality scores were used

Preprocessing results: Subject bias & inconsistency

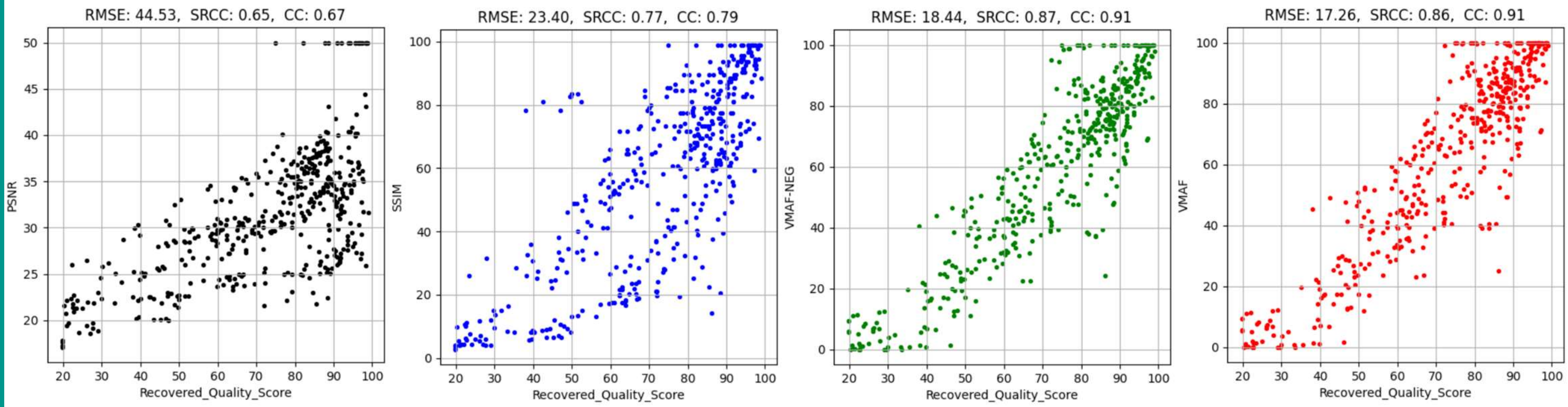


Preprocessing results: Recovered quality scores



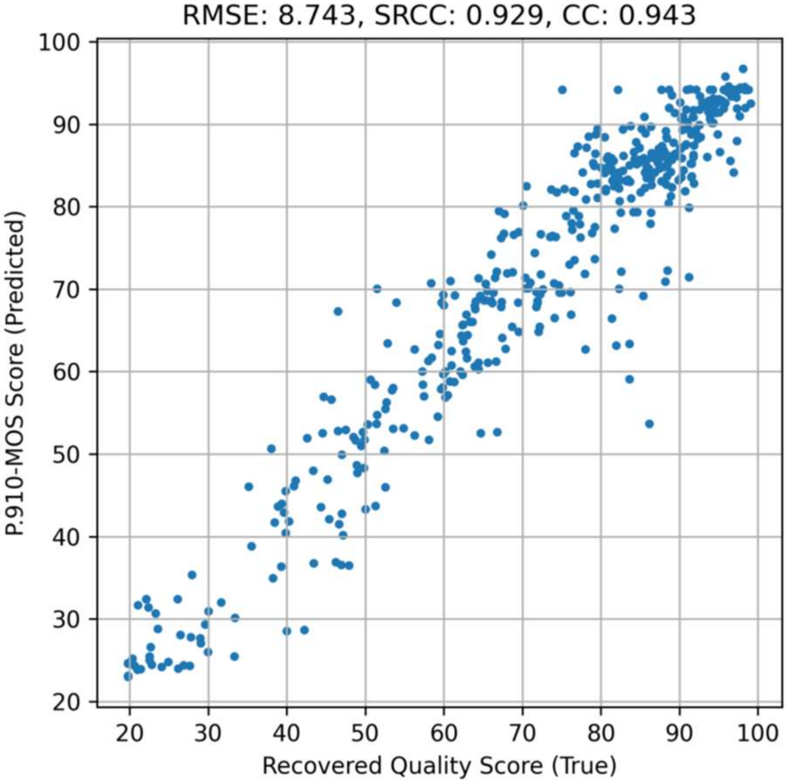
- The Recovered Quality Scores (RQS) span the entire quality range and are adjusted according to bias, uncertainty and inconsistency based on SUREAL's methodology

Preprocessing results: Metrics vs. RQS scatter plot



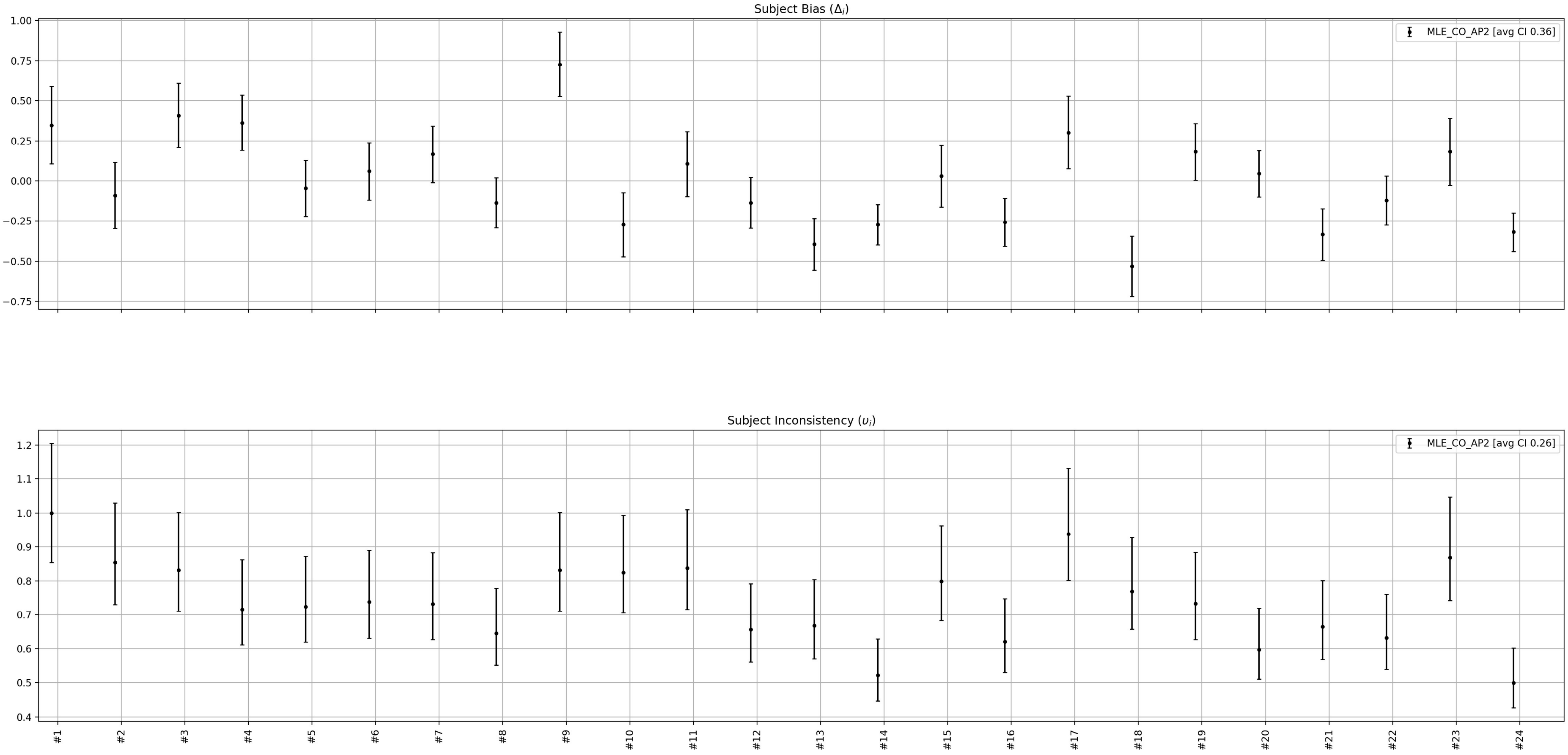
- VMAF-NEG and VMAF are well aligned to Recovered Quality Scores, with correlation of 91%

Preprocessing results: P.910 SVR model

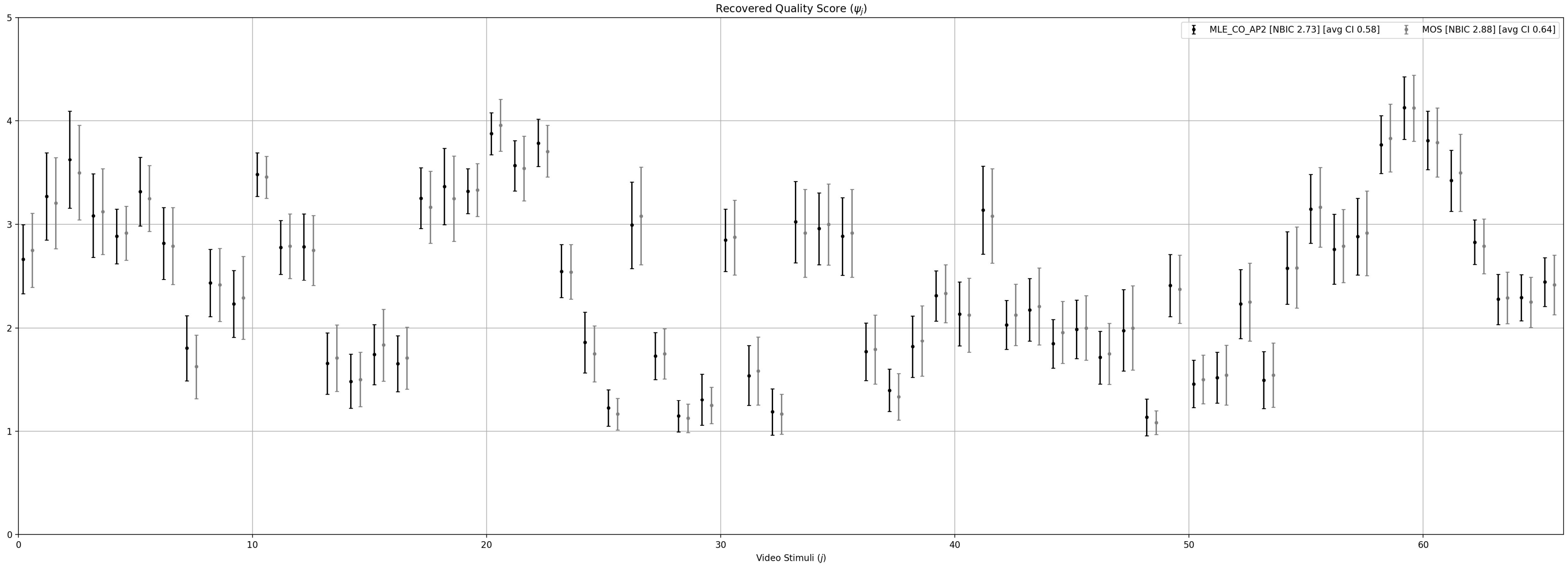


- Scatter plot of SVR with *ν*=0.5 (proportion of support vectors vs. total samples), *γ*=0.85 (radius of RDF), *C*=1 (regularization term) predicted scores vs recovered quality scores

Post-processing results: Subject bias & inconsistency

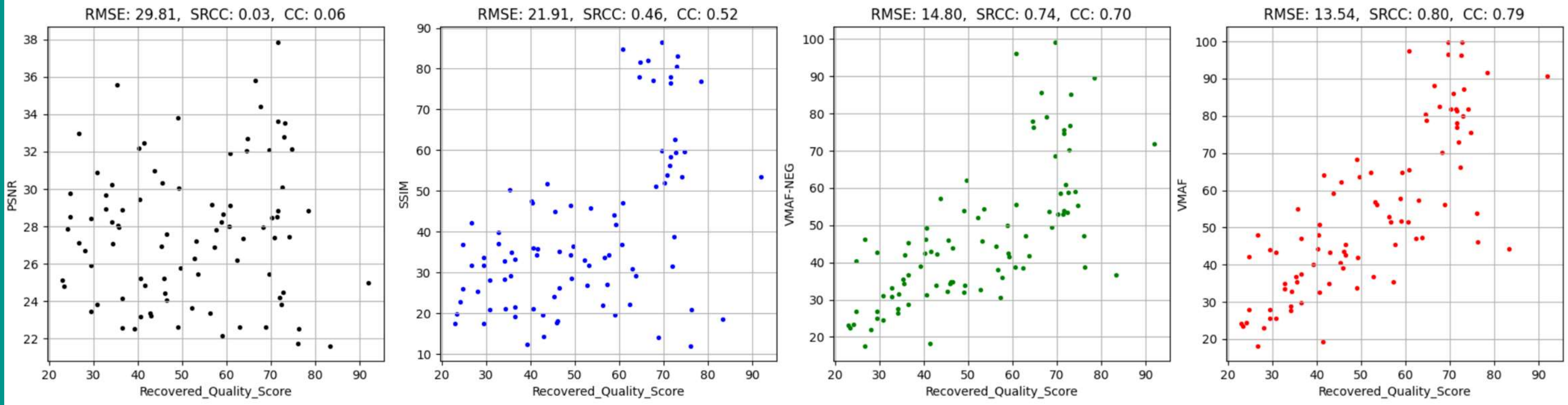


Post-processing results: Recovered quality scores



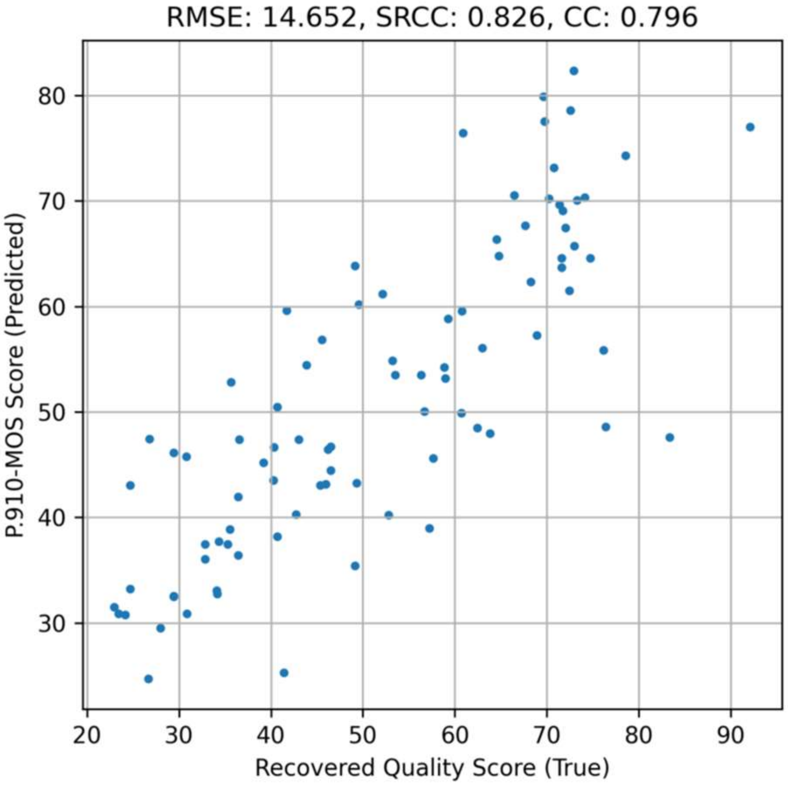
- The Recovered Quality Scores (RQS) span the entire quality range and are adjusted according to bias, uncertainty and inconsistency based on SUREAL's methodology

Post-processing results: Metrics vs. RQS scatter plot



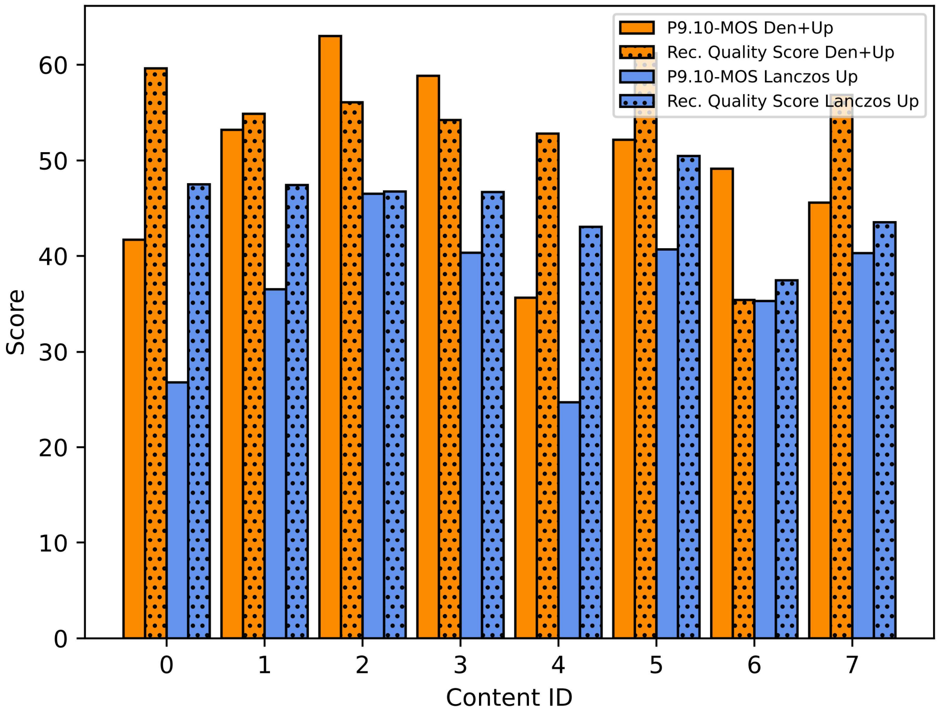
- VMAF-NEG and VMAF are better aligned to Recovered Quality Scores, with correlation of 70% to 79%

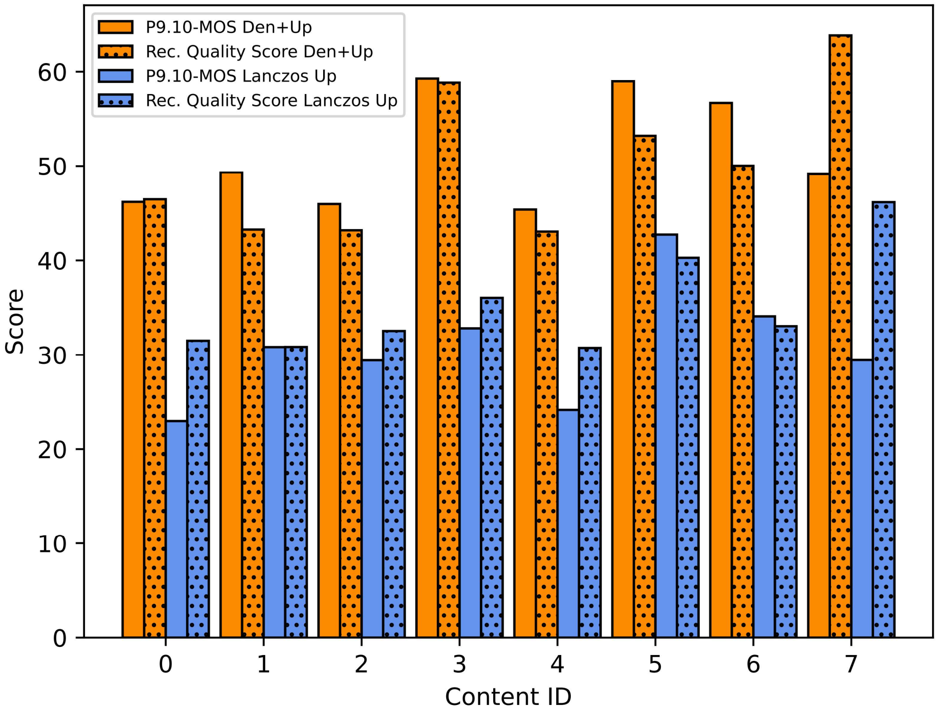
Post-processing results: P.910 SVR model



- In this case, it is mainly VMAF-NEG and VMAF that contribute to the SVR fit

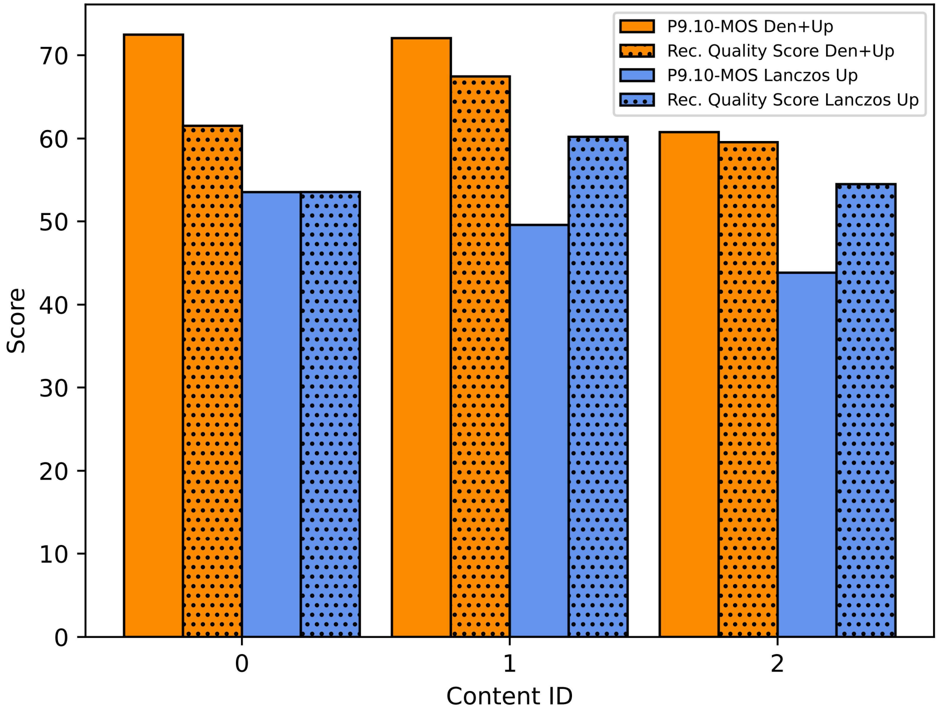
Post-processing results: Bar plots for AVC medium+high CRF

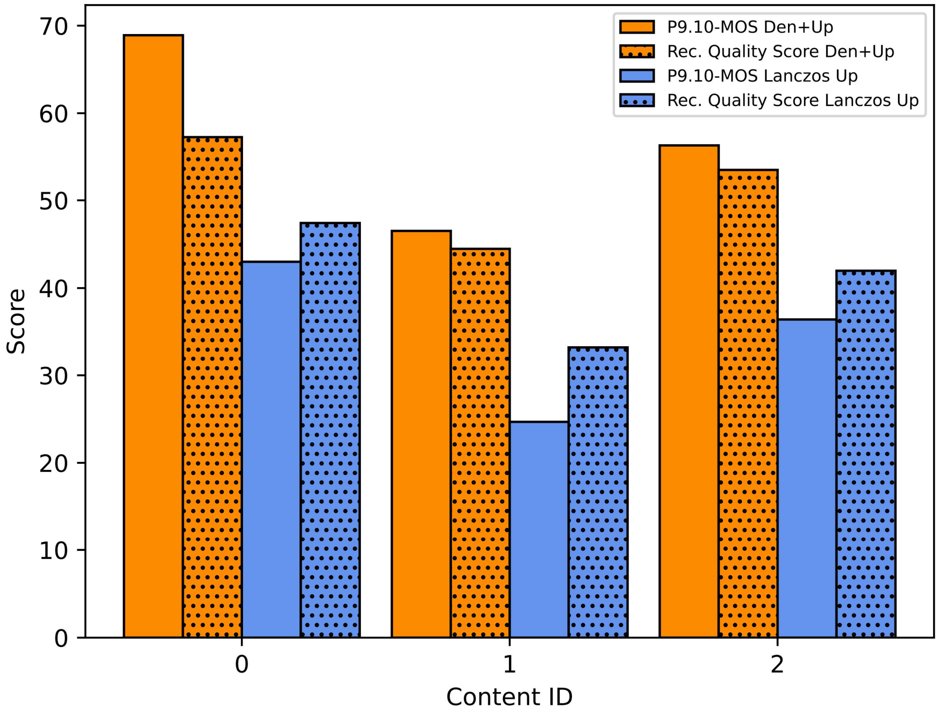




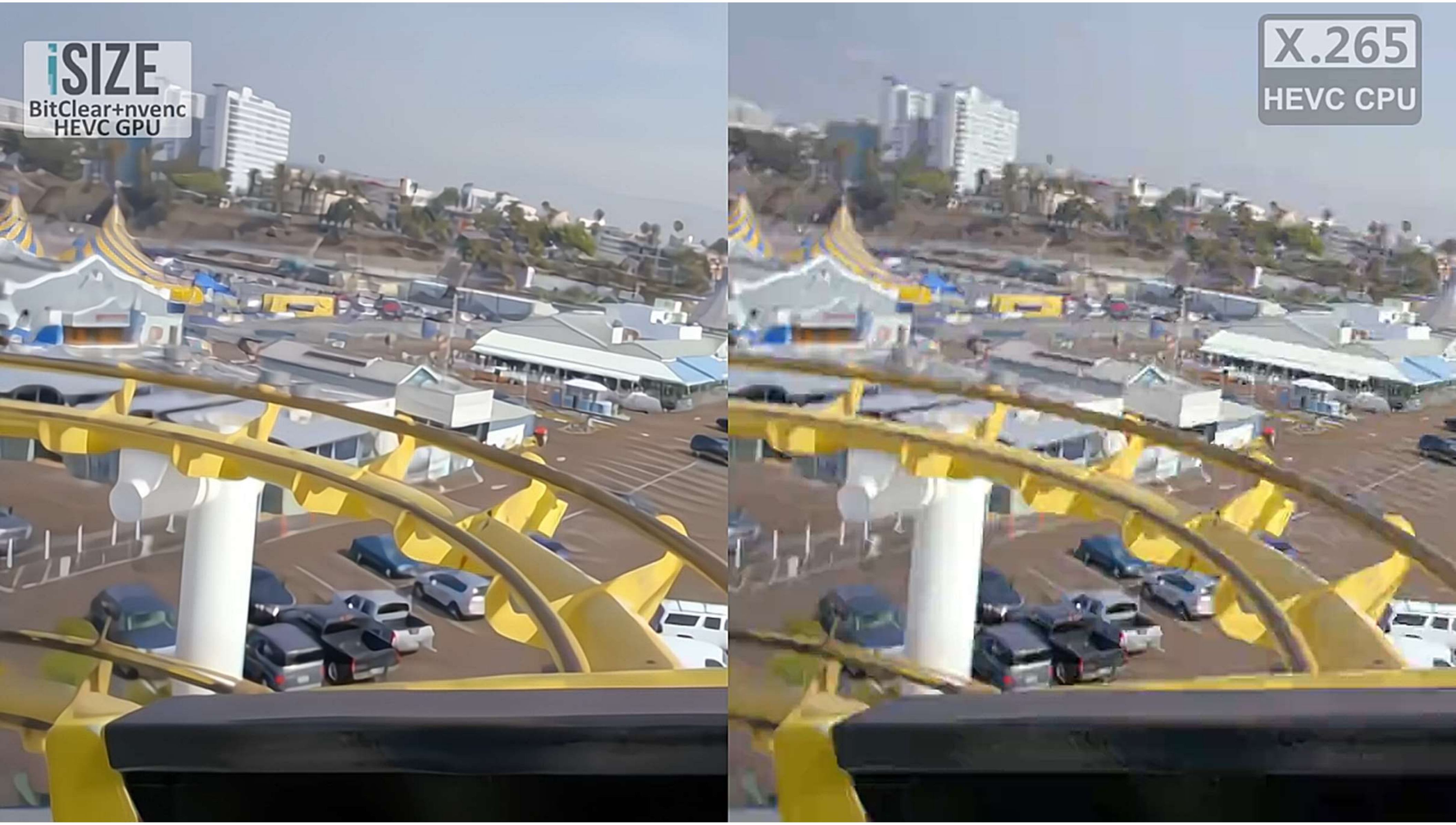
- In this case, P.910-MOS of post-processing (“Den+Up”) and Lanczos comes close to the results of VMAF

Post-processing results: Bar plots for VP9 medium+high CRF





- In this case, P.910-MOS of post-processing (“Den+Up”) and Lanczos comes close to the results of VMAF


iSIZE
BitClear+nvenc
HEVC GPU
X.265
HEVC CPU

Conclusion: Some key take-aways

- Domain-specific fusion of metrics can help get closer to the true ACR recovered quality scores if a single metric does not dominate
- The presented methodology is easy to apply and allows for quick testing (and re-testing!) as versions improve
- In the case of iSIZE preprocessing, we found that P.910 ACR results come between VMAF-NEG and VMAF (i.e., VMAF-NEG with some allowance for gain limit may suffice (e.g., 2%-5%)
- In the case of iSIZE post-processing, due to the use of GAN losses, only VMAF-NEG and VMAF remain relevant; the overall average gains of post-processing were 1 point in the 5-point ACR scale or 14 VMAF points
- Pseudo-random sampling of the convex hull of points helps (100-fold reduction in sampling), there are probably further ways of optimizing the distribution sampling that we have not considered
- It would probably be interesting to add other metrics (LPIPS, no-reference metrics) to our tests