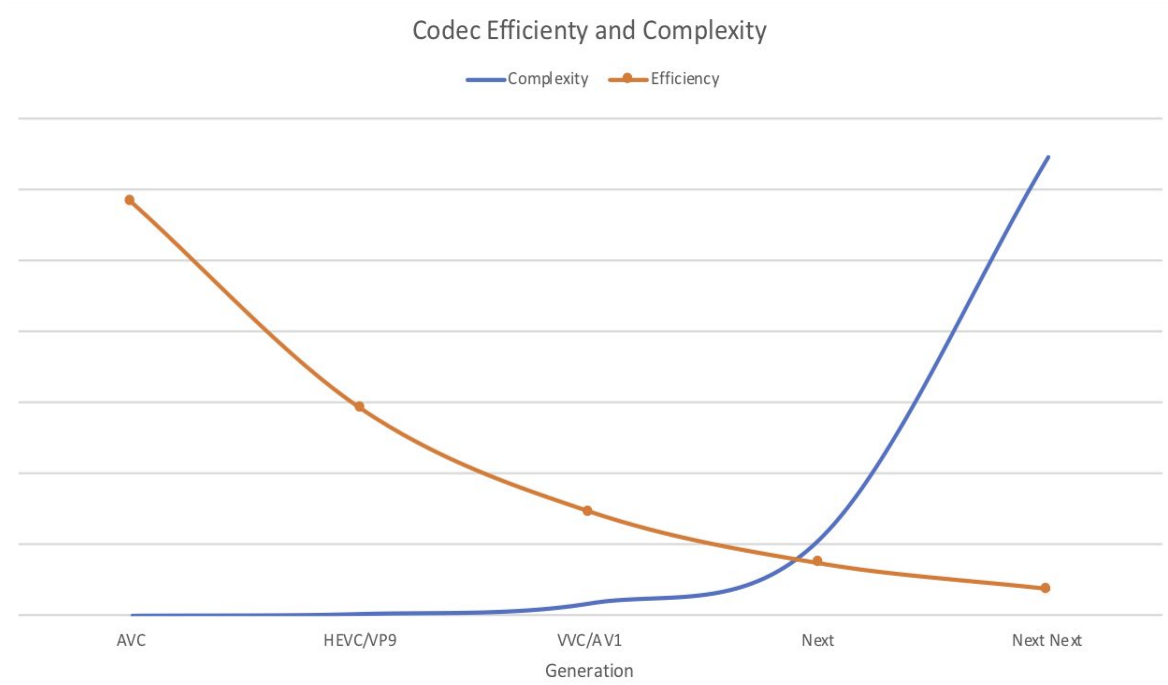
Toward Generalized Psychovisual Preprocessing For Video Encoding

Yiannis Andreopoulos

Joint work with: Aaron Chadha, Russell Anam, Matthias Treder, Ilya Fadeev
www.isize.co

November 2021

Motivation



[Encoder complexity hits wall](#) (D. Ronca)

Database / Metric	UHD (108) SROCC	LCC	OR	RMSE
PSNR	0.5517	0.6278	0.3056	8.7540
SSIM	0.5911	0.5853	0.3148	9.2195
MSSSIM	0.7426	0.7436	0.2130	7.4102
VIF	0.7464	0.7749	0.1852	6.9273
VSNR	0.5961	0.6580	0.2500	8.4062
VMAF	**0.8463**	**0.8375**	**0.1574**	**5.9972**

[Comparing VVC, HEVC and AV1](#) Zhang *et al.*

- Moore's law (mainly device power + heat dissipation) and cloud-based scaling **have both hit the wall**
- **We are at an inflection point:** perceptual metrics and ML are mature enough to allow for robust perceptual encoding
- **Codecs are amazing SNR/SSIM-to-bitrate machines, but these loss functions have significant limitations**

Related work: Machine-learning in video encoding

Server-side enhancement incudes **ML/neural-network based encoder tool optimization** within AVC, HEVC, AV1, VVC

- These approaches are standard compliant, but operate under SNR/SSIM optimization of encoding
- They also need reimplementation and validation per encoder and vendor, and that can be complex

Neural encoders have been proposed for the last 8 years

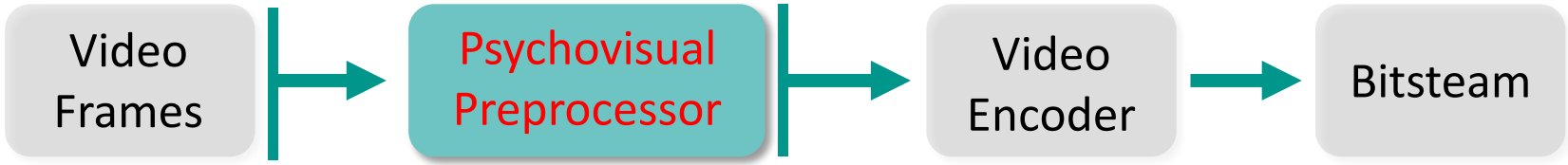
- However, they still fail to outperform HEVC under advanced configurations that use all encoding tools
- GAN-based approaches have been proposed, but they can cause deviations from the source content

Other approaches like **AV1 tune-VMAF** or **encoder-specific perceptual tuning** can offer visual quality improvement

- They need to be applied for each encoding
- Improvement on one metric (e.g., VMAF), but damage on all the others (SSIM, MS-SSIM, VIF)
- Improvement only in certain regimes of bitrate/quality, where it is not always quantifiable or visible
- Recent VMAF proposals (VMAF_NEG) limit the effect of uncontrolled preprocessing

Finding the best rate-perception-distortion-complexity trade-off is a hard problem!

Our proposal: Toward a generalized psychovisual preprocessor



Five principles:

1) *Psychovisual tuning*
   - Encapsulating known principles of human vision in a data-driven (i.e., learnable) manner

2) *Multi-metric gains*
   - Allow for gains over multiple quality metrics, and/or lead to measured quality improvement in P.910 tests

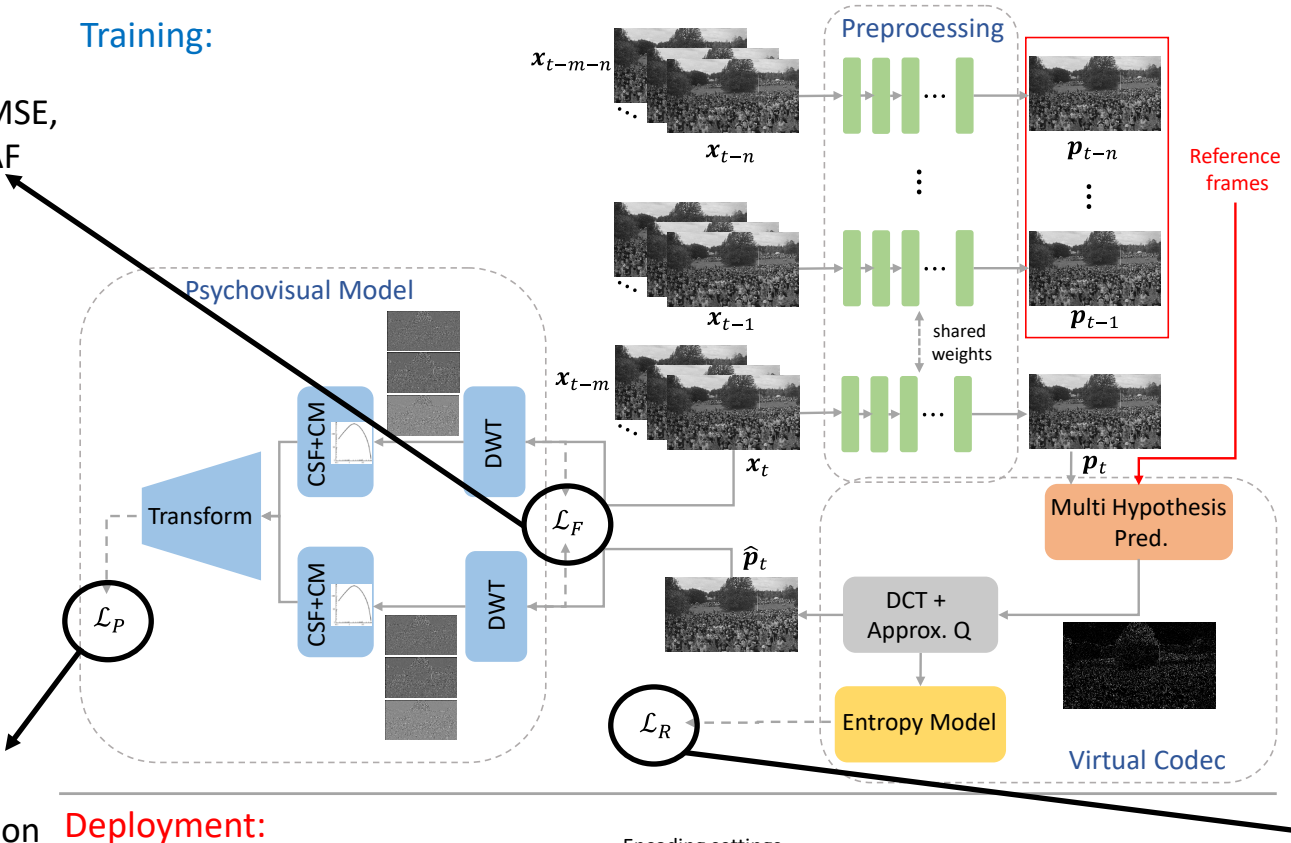
3) *Cross-content and cross-codec applicable*
   - Offering compounded gains over other optimization frameworks like content-adaptive or convex-hull encoding

4) *Low delay*
   - Allow for single-pass per encoding resolution/bitrate, or comprise a single-pass model for multiple resolutions/bitrates

5) *Low complexity*
   - Inference complexity must be analogous to low-complex encoding, e.g., AVC x264 medium-preset encoding

Our proposal: Generalized psychovisual preprocessor

Fidelity loss:
Regularization of MSE, SSIM, DLM of VMAF



Perceptual loss:
A measure over a learned transform on top of the filtered source and reconstructed subbands

Rate loss: log2 output of entropy model of Virtual Codec

Experimental results: Codecs, recipes, measurement

Content: XIPH sequences for VoD/Live, YouTube UGC dataset for UGC Music/Sports/Gaming, all at 1080p resolution

Encoders: AVC, HEVC, VVC via: libx264 (veryslow), libx265 (veryslow), vvenc1.0.0 (slow), CRF={18,22,26,30,34,38,42} (similar for VVC); for all encoders, encoding at 1080p and quality measurement is done at 1080p, extending to ABR gives similar results

Measurement: libvmaf for VMAF/VMAF-NEG/SSIM, p.1204 github for p.1204, Apple AVQT binary for AVQT, BD-rates based on the libvmaf BD-rate calculator

BD rates (%)	XIPH VoD/Live Premium Content			UGC Music/Sports	UGC Gaming
Metric/Encoder	AVC	HEVC	VVC	AVC	HEVC
SSIM	-8.1	-8.4	-5.7	-16.6	-16.8
AVQT	-6.2	-7.9	-10.2	-16.5	-17.8
p.1204	-6.1	-8.9	N/A	-0.8	-17.5
VMAF-NEG	-11.4	-10.1	-6.8	-17.4	-1.1
VMAF	-20.2	-21.6	-19.6	-34.0	-26.8
Numerical average:	-10.4	-11.4	-10.6	-17.1	-16.0

Key takeaways:

- cross-metric improvement with single-pass processing for all encoders and encoding recipes/quality levels (CRFs)
- cross-content and cross-codec applicable with psychovisual tuning from the psychovisual model
- the average bitrate saving is not consistent across metrics, but not all metrics are equally important

Complexity, latency and P.910: Completing the 5 points

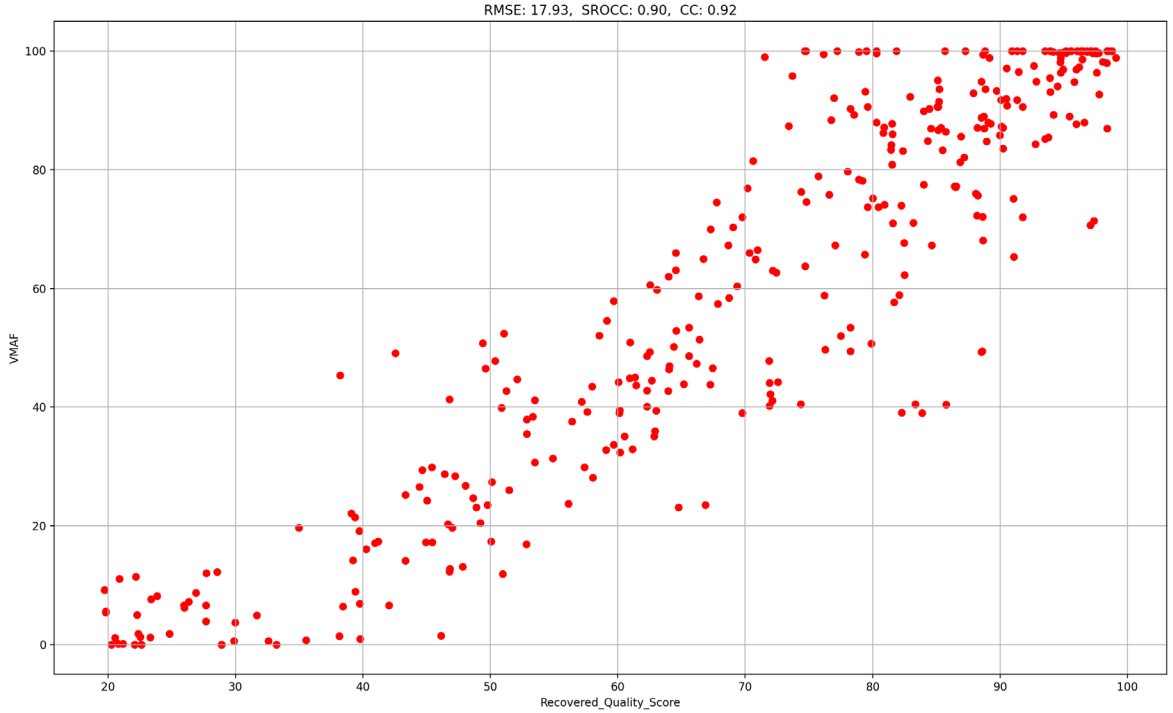
Runtime performance on CPU and GPU for 1080p: 218fps on Intel Xeon (x264 medium is 114fps), 714fps on NVIDIA T4

Latency: Single-frame processing, i.e., less than 5ms on CPU, less than 1.5ms on GPU

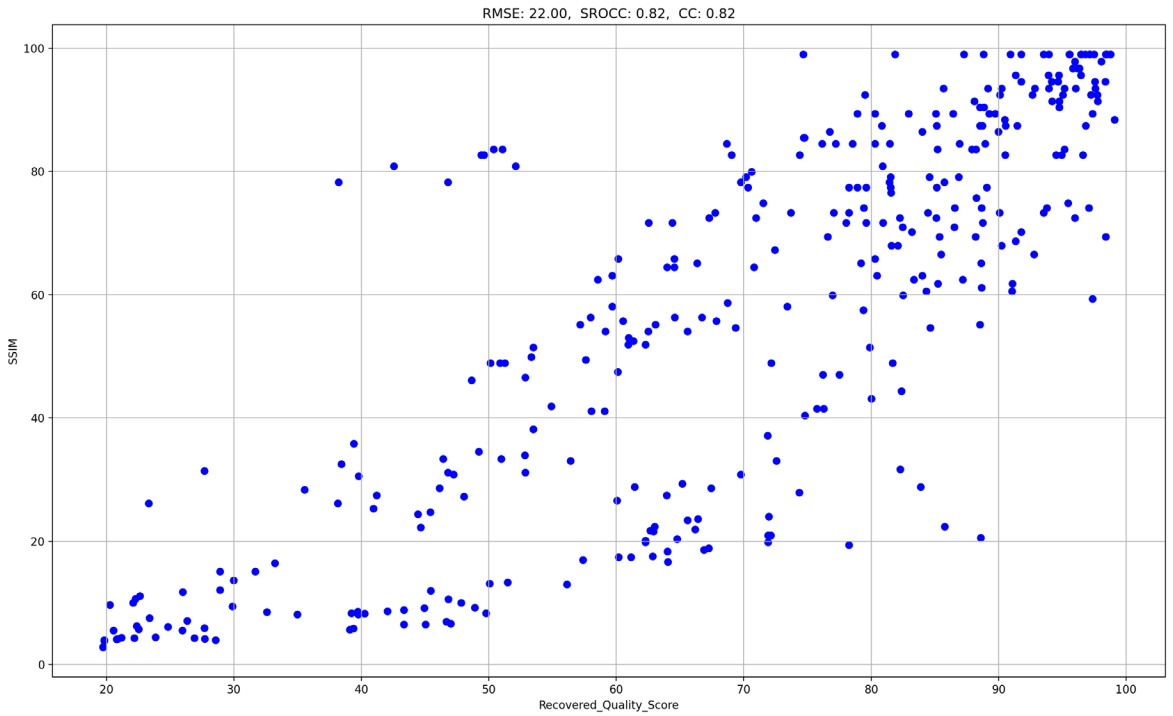
Recovered Quality (RQ) Scoring: We have carried out a P.910 test validating that VMAF is the best fit to RQ scores

- *This type of framework meets the 5 points postulated for a generalized psychovisual preprocessing framework*

VMAF



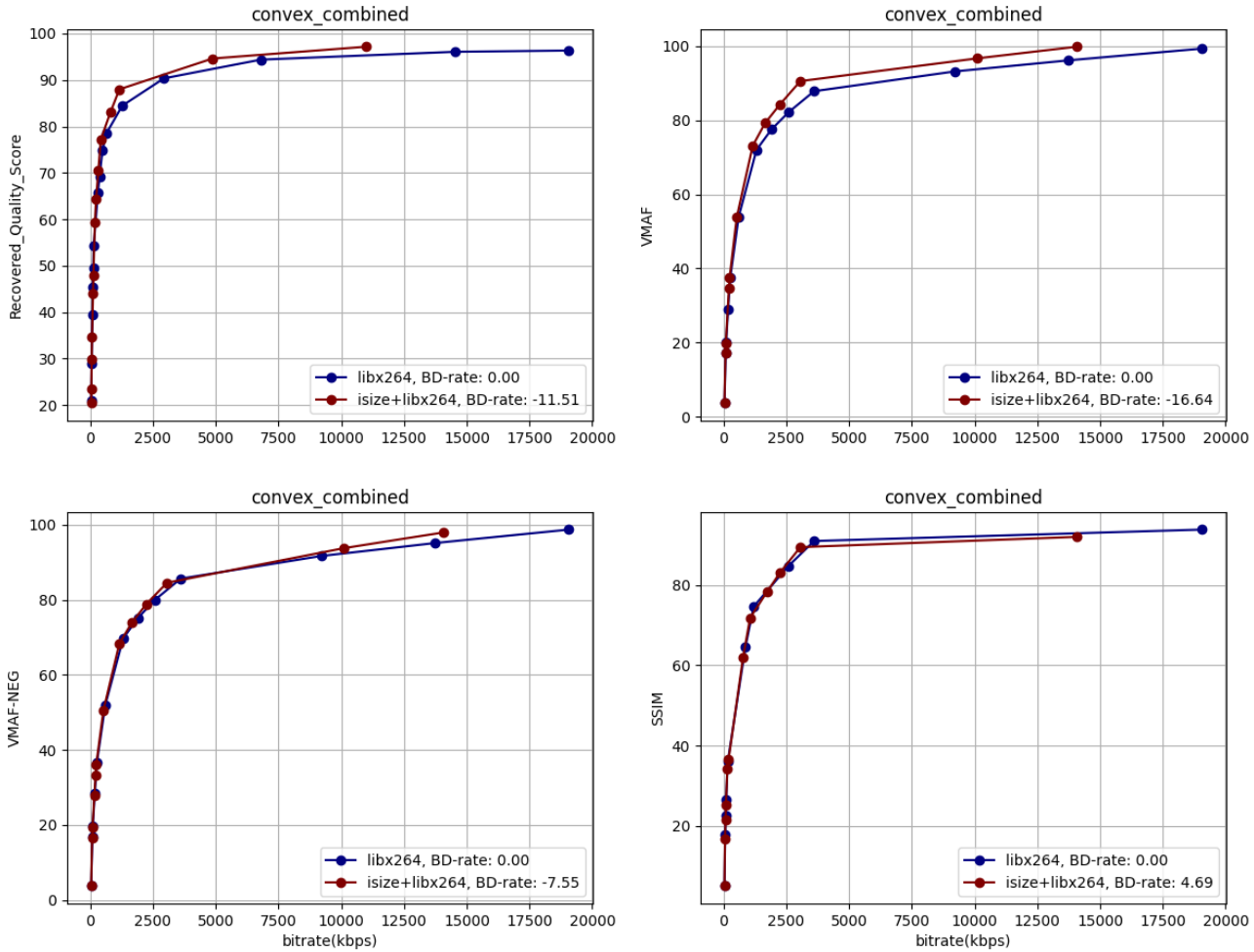
SSIM



[P.910 testing in collaboration with BBC R&D and Queen Mary University of London, Innovate UK SEQUOIA project]

Average bitrate saving (BD-rate) results from P.910: AVC

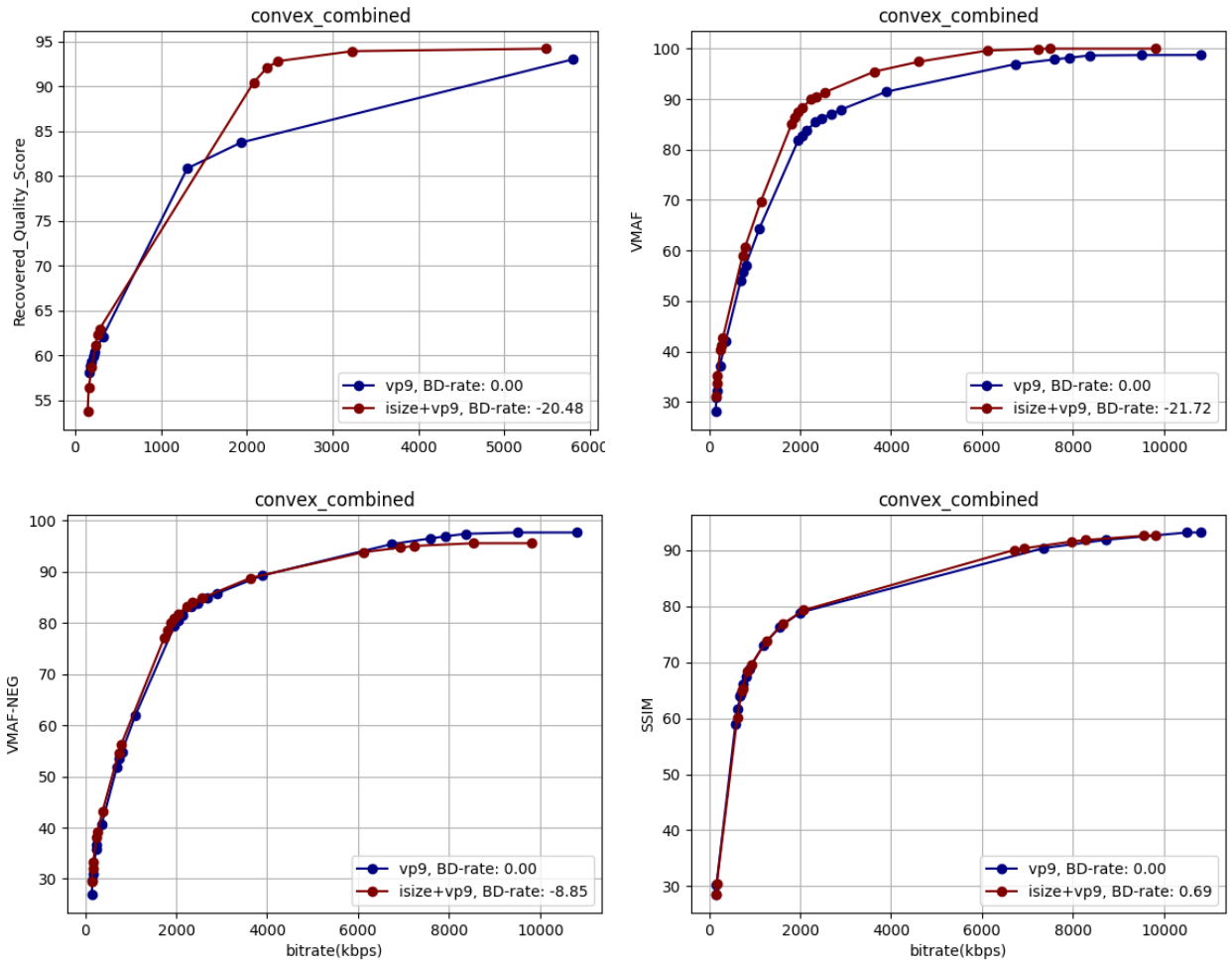
- BD-rate of Recovered Quality (RQ) Score is -11.5%
- Non-negative least squares fit shows that VMAF is the dominant feature that is correlated to RQ, followed by SSIM (coeffs of 0.74 and 0.26, resp.)
- Weighted BD-rate of VMAF and SSIM is -11.1%
- We did not yet measure fits of AVQT and p.1204 to RQ



[P.910 testing in collaboration with BBC R&D and Queen Mary University of London, Innovate UK SEQUOIA project]

Average bitrate saving (BD-rate) results from P.910: VP9

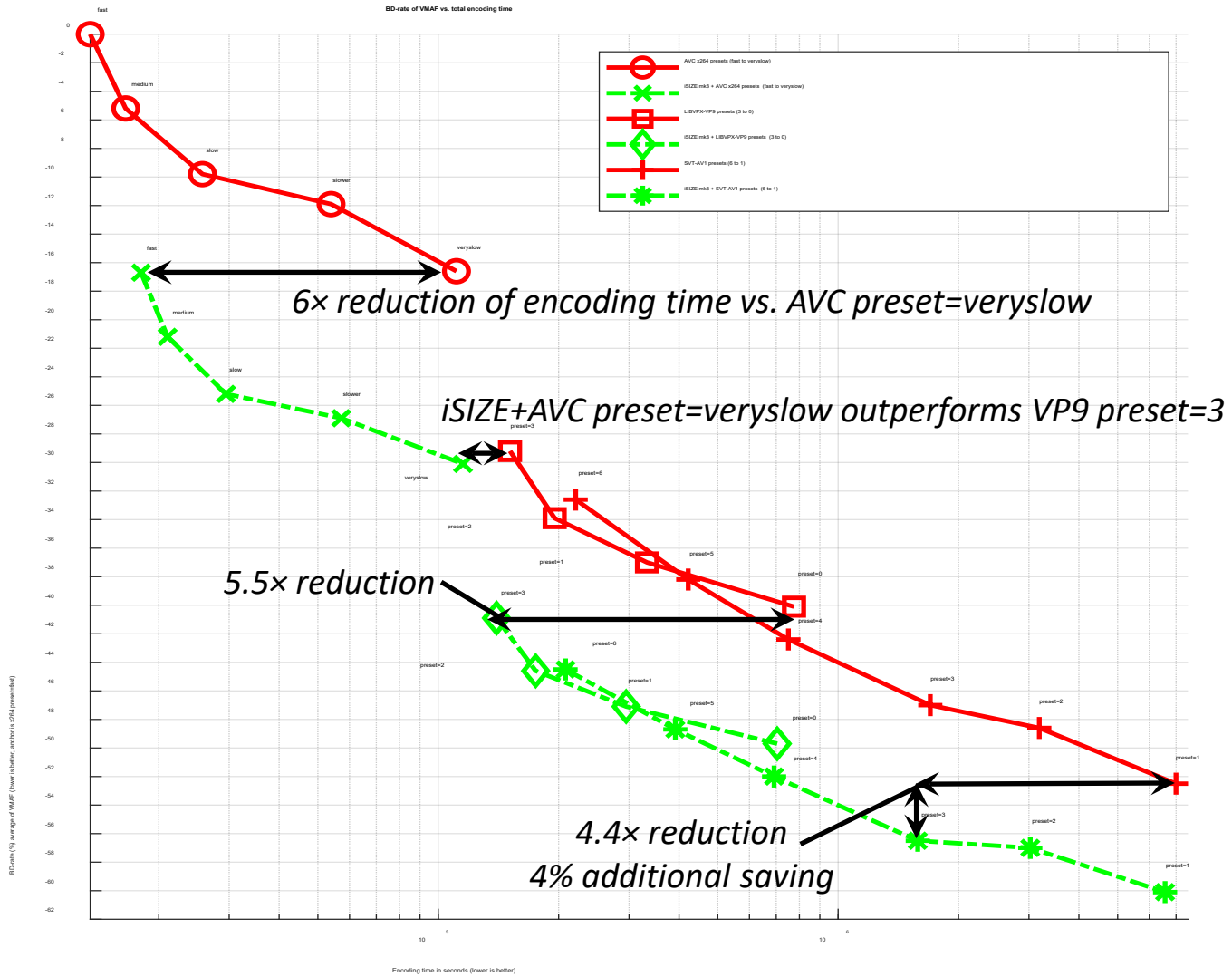
- BD-rate of Recovered Quality (RQ) Score is -20.5%
- Non-negative least squares fit shows that VMAF is by far the dominant feature that is correlated to RQ (weight of 1.0)
- BD-rate of VMAF is -21.7%
- We did not yet measure fits of AVQT and p.1204 to RQ
- On-going further testing for the middle bitrate region of VP9



[P.910 testing in collaboration with BBC R&D and Queen Mary University of London, Innovate UK SEQUOIA project]

Bitrate savings vs. encoding time: AVC, VP9 and AV1

- Complexity-vs.-BD-rate graph corresponding to very extensive testing on the entirety of AV2 CTC test content, 8 ABR points and 6 CRFs/resolution, all encoding presets
- 4-to-6 fold complexity reduction
- Average bitrate saving 8%-15%
- Enables high-spec VP9 performance with just AVC
- Enables high-spec AV1 with VP9
- Interesting trade-offs to explore further



Final remarks

- We believe we can go further in VMAF improvements across multiple video coding standards
- Our approach offers compounded gains to any encoder-specific perceptual quality optimization, e.g., within AV1 & VVC
- The single-pass nature and decoupling from specific encoder standards and vendor implementations allows for easy deployment on custom hardware or high-performance CPU/GPU clusters
- The current implementation complexity already allows for real-time operation under GPU or multi-CPU environment, and we plan to showcase further optimizations
- The current framework meets the requirements for a generalized psychovisual preprocessor for video encoding