

Supplementary Material for Deep Perceptual Preprocessing for Video Coding

Aaron Chadha

iSIZE, isize.co

Yiannis Andreopoulos

1. Ablation Study

In this section we present an ablation over the core hyperparameters λ and γ , which represent the rate loss coefficient and perceptual loss coefficient respectively. We centre our ablation on two sequences from XIPH, sunflower and crowd-run, which are representative of the extrema of content; relatively static with only camera motion versus fast and complex motion. We present results when varying λ on these sequences in Fig. 1 on H.264, for varying CRF, $C \in \{18, 22, 26, 30, 34, 38, 42\}$ (i.e. each point per curve represents a different CRF). As is evident, for the relatively static sunflower sequence, a lower λ is preferable, This is because the sequence is already easy to encode and a higher λ results in more content distortion - which is reflected in SSIM and its BD-rate. While the content may perceptually look more favorable (indicated by VMAF), there is too much deviation from the source. Conversely, for a more complex sequence such as crowd-run, which is more difficult to encode, there is more scope to change the content and remove motion/compression artifacts - hence BD-rate savings in both VMAF and SSIM for higher λ . We account for both types of sequences by using models trained with multiple $\lambda \in [0.001, 0.01]$ in our final implementation.

For the perceptual coefficient γ , the tradeoff is more straightforward - a higher γ will typically increase the perceptual quality of the content and thus increase VMAF (potentially at the cost of rate), but also result in a decrease in SSIM at higher bitrates when set too high; this is illustrated in Fig. 2 for the crowd-run sequence.

2. Additional Experimental Details

In this section, we provide additional experimental details for experimental results presented in the main paper. We report on tests with XIPH and CDVL sequences¹ (representing prime content). All sequences were at 1080p format and have been downloaded from their corresponding repositories in March 2018. All quality measurements are done

¹XIPH source material was downloaded from <https://media.xiph.org/video/derf/> and CDVL material from <https://www.cdvl.org/>

against each input video source in its native resolution and frame rate.

2.1. Encoding Recipes

AVC/H.264: The encoding recipe for H.264 is:

```
ffmpeg -y -i video_in.y4m -vf scale=WxH
  ↳ :flags=lanczos -c:v libx264 -
  ↳ profile:v high -threads 4 -preset
  ↳ veryslow -crf C -refs 5 -g 150 -
  ↳ tune ssim -x264opts ssim=1 -
  ↳ keyint_min 150 -sc_threshold 0 -f
  ↳ mp4 video_out.mp4
```

The utilized CRF values for H.264 encoding are: $C \in \{18, 22, 26, 30, 34, 38, 42\}$. The output resolution is set as $W \times H$, where W and H denote width and height.

AV1: The aomenc AV1 recipe (Lanczos downscaling takes place losslessly prior to encoding with aomenc):

```
aomenc --passes=2 --pass=1 --fpf=aoml.
  ↳ log --target-bitrate=B --cpu-used
  ↳ =5 --threads=8 --tile-columns=1
  ↳ --tile-rows=0 --tune=T --kf-max-
  ↳ dist=150 -o "video_out.webm" "
  ↳ video_in.y4m" 2>&1
```

```
aomenc --passes=2 --pass=2 --fpf=aoml.
  ↳ log --target-bitrate=B --cpu-used
  ↳ =5 --threads=8 --tile-columns=1
  ↳ --tile-rows=0 --tune=T --kf-max-
  ↳ dist=150 -o "video_out.webm" "
  ↳ video_in.y4m" 2>&1
```

The utilized bitrates are: $B \in \{138, 230, 385, 642, 1070, 1800, 3000, 5000, 8000\}$ kbps, i.e., increase by 40% per step. The perceptual tuning options were $T \in \{\text{vmaf_with_preprocessing}, \text{ssim}\}$ for tuning for VMAF_NEG/VMAF and SSIM, respectively.

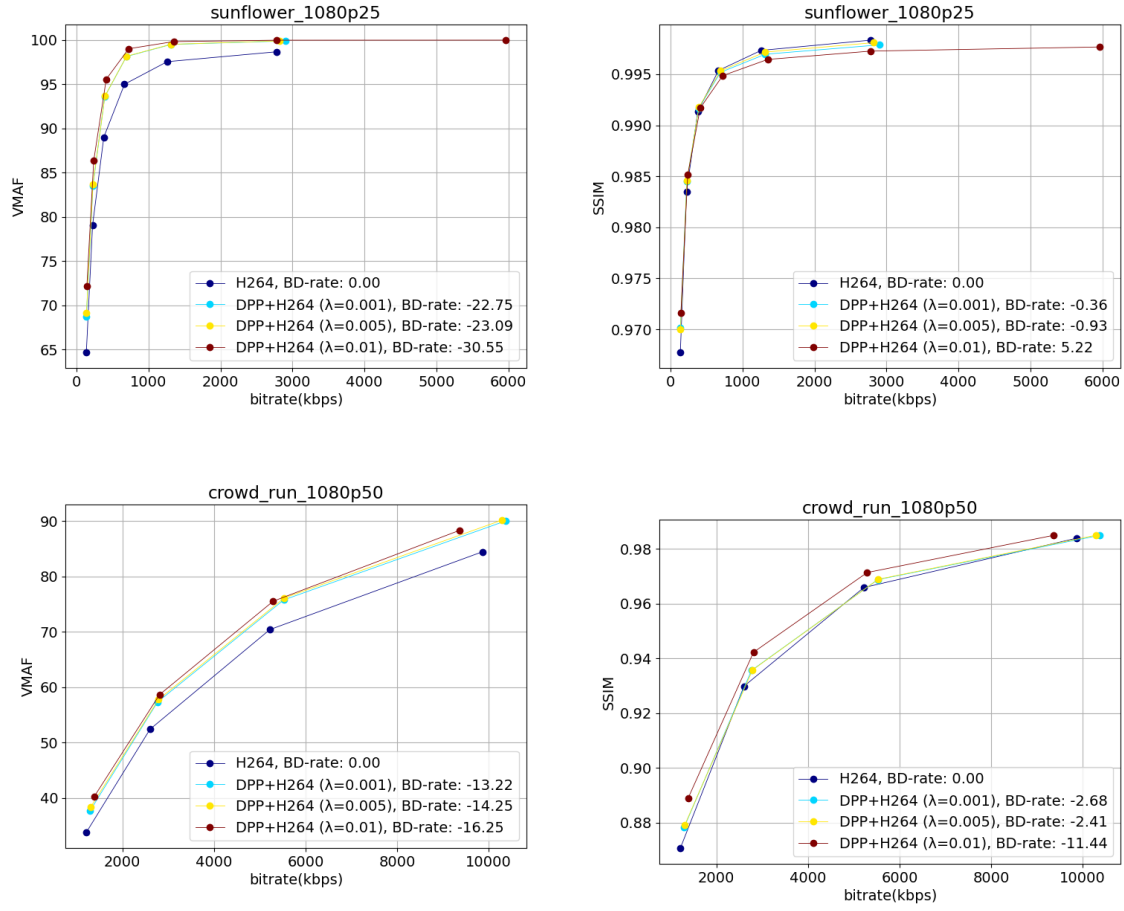


Figure 1: VMAF and SSIM for sunflower (top row) and crowd-run (bottom row) when varying rate coefficient λ

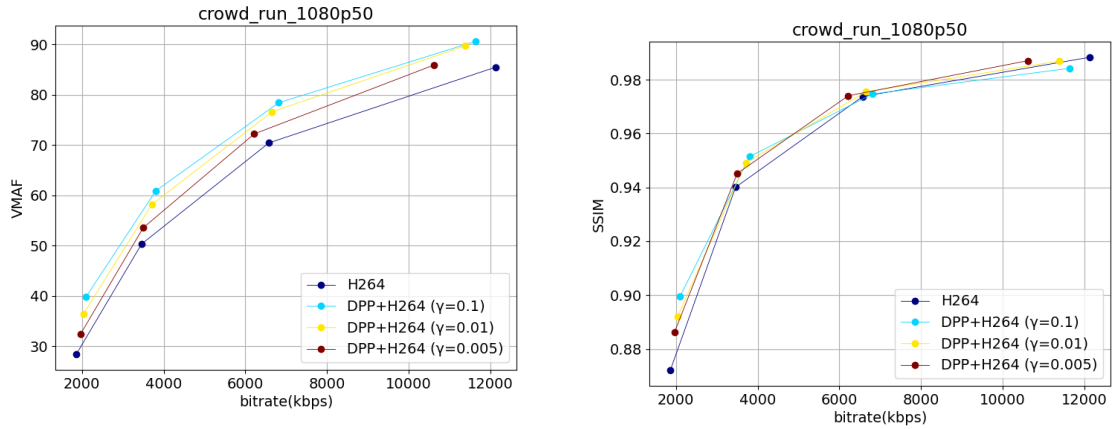
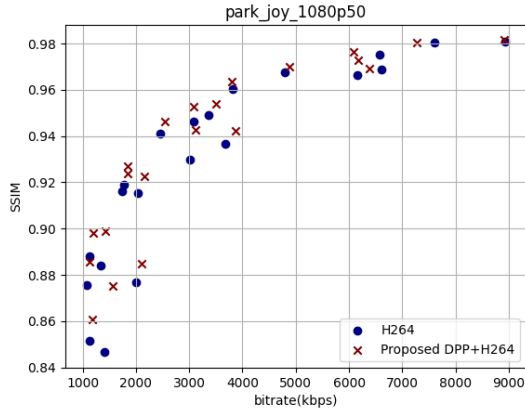
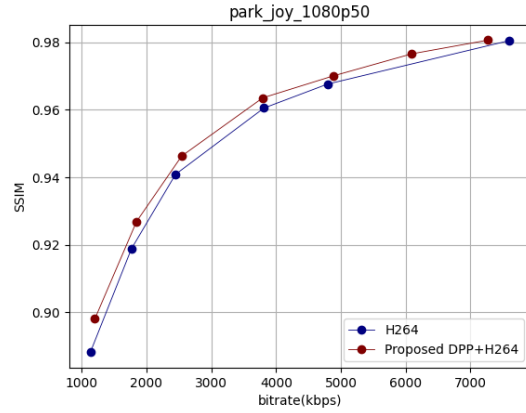


Figure 2: VMAF and SSIM for crowd-run when varying perceptual coefficient γ



(a) Bitrate-quality points



(b) Convex hull

Figure 3: a) Bitrate-quality points for parkjoy XIPH sequence plotted for multiple resolutions, crfs/bitrates and multiple $\lambda \in [0.001, 0.001]$. b) Convex hull over monotonically increasing points for H264 and proposed DPP+H264.

VVC/H.266: The encoding recipe for VVC is:

```
vvecapp -y -i video_in.yuv -s WxH -c
  ↳ yuv420 -r F --preset slow --qp C
  ↳ -qpa 0 -ip 128 -t 4 -o video_out
  ↳ .266
```

The utilized QP values correspond to those used for the CRF values for H.264 encoding. The input/output resolution is set as $W \times H$, where W and H denote width and height and the encoding frame rate F , corresponds to the input source frame rate. The GOP size was set to 128 frames.

2.2. Convex Hull Implementation

For all codec recipes, utilized resolutions are produced by FFmpeg Lanczos used on the source video or on the DPP output. They are: 1080p, 720p, 540p, 432p, 360p, 288p, 216, 144p. All decoded results are upsampled with FFmpeg bicubic² to 1080p prior to quality measurement at the native video resolution via libvmaf. All Bjontegaard delta-rates (BD-rates) [1] are produced by first finding the subset of monotonically-increasing bitrate-quality points that are in the convex hull of the quality-bitrate curve, and then using standard BD-rate measurements to find average rate saving per sequence. The convex hull is computed over bitrate-quality points of all resolutions, CRFs/bitrates/QPs and multiple rate coefficients $\lambda \in [0.001, 0.01]$, thus resulting in a single bitrate-quality curve for both the codec and our proposed DPP+codec. An example of the convex hull computation is illustrated in Fig. 3 for the ‘park joy’ sequence.

²as this represents the typical upscaler quality expected to be available in commodity client devices [2]

The BD-rate of each dataset is the numerical average of the BD-rates per sequence of the dataset. To avoid skewing the BD-rate averages from the use of very-low or very-high bitrate points, we remove points with excessively-high bitrates that would not be used in practice (i.e., greater than 12mbps). This type of measurement corresponds to a dynamic-optimizer type of BD-rate measurement per clip [2] for the quality regimes usable in video delivery. Importantly, for all experimental points per sequence (e.g., for the example of Fig. 3), DPP requires only a single-pass per frame (and per model setting λ , with the latency of each model being only 1 frame), and all encodings are then generated from the produced results. The DPP+codec curves contain only points produced from the DPP output and not from the codec-only output.

Finally, for the bitrate-quality figures of the paper (Fig. 5 of the paper), we simply average the corresponding resolution-CRF/bitrate/QP- λ points over sequences and the plot the convex hull of the averaged points as a single curve for the dataset.

2.3. Metrics

VMAF and SSIM were computed using the Netflix libvmaf library. With regards to VMAF, we used the standard 0.6.1 library with and without the latest update that includes parameters: ADM_ENHN_GAIN_LIMIT and VIF_ENHN_GAIN_LIMIT. By setting the gain limits to 1.0, we enable metric VMAF_NEG, which is designed to constrain metric sensitivity to linear enhancements.

3. BD-Rate Tables

We present BD-rates on individual sequences of XIPH and CDVL datasets for AVC/H.264, AV1 and VVC/H.266

in Tables 1-3 respectively. We also include results for the two presets of AVC/H.264 (slow and veryslow), in order to demonstrate that our results are consistent across different encoding presets. Similarly, for AV1 we include results with two perceptual tuning options (tunessim and tunevmf) in order to show that our framework offers consistent improvement on top of any perceptual tuning available within the encoder itself. The sequences range from very static (e.g. sunflower_1080p25) to fast and complex motion (e.g. crowd_run_1080p50). Our proposed DPP offers consistent gains over all encoders and all types of sequences, whilst being single pass and not requiring any change on the codec or the client device.

4. Visual Examples

In Figures 4-7, we present more visual examples for DPP+codec versus codec only, when encoding at the same bitrate or CRF value (for AV1 and AVC, respectively). We focused our comparison to 1mbps and CRF=30 to illustrate the effect of improvement on lower-bitrates where visual quality degradation is more apparent to viewers.

References

- [1] Gisle Bjontegaard. Improvements of the BD-PSNR model. In *ITU-T SG16/Q6, 35th VCEG Meeting, Berlin, Germany, July, 2008*, 2008.
- [2] Ioannis Katsavounidis and Liwei Guo. Video codec comparison using the dynamic optimizer framework. In *Applications of Digital Image Processing XLI*, volume 10752, page 107520Q. International Society for Optics and Photonics, 2018.



(a) Original

(b) AV1 tunevmf at 1070kbps, VMAF= 32.5, VMAF_NEG=31.5, SSIM=0.875

(c) Proposed DPP+AV1 tunevmf at 1070kbps, VMAF= 35.9, VMAF_NEG=34.0, SSIM=0.879



(d) AVC/H.264 at CRF30, VMAF= 37.5, VMAF_NEG=36.5, SSIM=0.918

(e) Proposed DPP+AVC/H.264 at CRF30, VMAF=43.1, VMAF_NEG=40.9, SSIM=0.926

Figure 4: Example frame segment of ‘duckscrowd run’ at the same bitrate or CRF value.



(a) Original

(b) AV1 tunevmf at 1070kbps, VMAF= 38.6, VMAF_NEG=37.5, SSIM=0.877

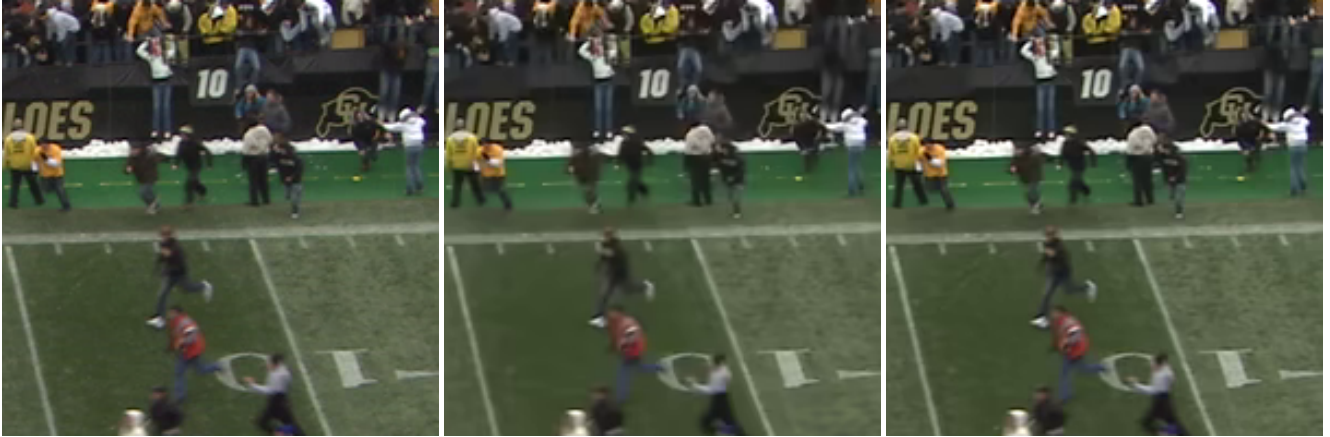
(c) Proposed DPP+AV1 tunevmf at 1070kbps, VMAF=42.3, VMAF_NEG= 39.9, SSIM=0.882



(d) AVC/H.264 at CRF30, VMAF= 41.5, VMAF_NEG=40.3, SSIM=0.919

(e) Proposed DPP+AVC/H.264 at CRF30, VMAF=47.0, VMAF_NEG=44.4, SSIM=0.927

Figure 5: Example frame segment of ‘park joy’ at the same bitrate or CRF value.



(a) Original

(b) AV1 tunevmf at 1070kbps, VMAF= 58.5, VMAF_NEG=56.8, SSIM=0.942

(c) Proposed DPP+AV1 tunevmf at 1070kbps, VMAF=63.2, VMAF_NEG= 59.1, SSIM=0.944



(d) AVC/H.264 at CRF30, VMAF= 53.8, VMAF_NEG=52.2, SSIM=0.947

(e) Proposed DPP+AVC/H.264 at CRF30, VMAF=59.8, VMAF_NEG=56.5, SSIM=0.952

Figure 6: Example frame segment of 'rush field cuts' at the same bitrate or CRF value.



(a) Original

(b) AV1 tunevmf at 1070kbps, VMAF= 80.6, VMAF_NEG=78.1, SSIM=0.978

(c) Proposed DPP+AV1 tunevmf at 1070kbps, VMAF=86.3, VMAF_NEG= 78.9, SSIM=0.979



(d) AVC/H.264 at CRF30, VMAF= 55.6, VMAF_NEG=53.9, SSIM=0.947

(e) Proposed DPP+AVC/H.264 at CRF30, VMAF=61.6, VMAF_NEG=58.2, SSIM=0.952

Figure 7: Example frame segment of ‘touchdown pass’ at the same bitrate or CRF value.

Table 1: Bjontegaard delta-rate results for DPP+H264 in terms of metrics that are increasingly perceptual in nature: SSIM, VMAF_NEG (ADM_ENHN_GAIN_LIMIT = VIF_ENHN_GAIN_LIMIT = 1.0) and VMAF. The first 16 sequences are from the XIPH website; the remaining sequences are from CDVL.

Codec	H264 (preset=slow, tune_ssim)			H264 (preset=veryslow, tune_ssim)		
Sequence	SSIM (%)	VMAF_NEG (%)	VMAF (%)	SSIM (%)	VMAF_NEG (%)	VMAF (%)
red_kayak_1080p30	-2.67	-11.32	-17.24	-2.82	-12.25	-16.67
ducks_take_off_1080p50	-4.88	-13.28	-26.92	-5.18	-20.86	-26.46
controlled_burn_1080p30	-3.10	-13.18	-24.02	-3.79	-11.31	-22.46
tractor_1080p25	-3.57	-14.59	-22.16	-3.14	-13.95	-21.65
old_town_cross_1080p50	-2.94	-15.73	-24.42	-2.26	-15.29	-23.90
riverbed_1080p25	-2.69	-12.72	-17.64	-1.77	-6.10	-10.59
west_wind_easy_1080p30	-2.30	-10.00	-18.90	-2.09	-7.09	-16.66
sunflower_1080p25	-5.08	-14.67	-22.83	-3.00	-6.98	-15.65
rush_field_cuts_1080p30	-2.49	-12.21	-18.90	-2.35	-11.76	-18.20
blue_sky_1080p25	-2.57	-5.63	-15.11	-2.43	-10.10	-13.93
pedestrian_area_1080p25	-2.91	-10.10	-18.51	-2.91	-10.09	-18.56
crowd_run_1080p50	-3.65	-20.46	-23.39	-3.18	-15.23	-18.44
rush_hour_1080p25	-1.74	-11.53	-21.39	-1.82	-12.73	-22.53
park_joy_1080p50	-3.86	-7.28	-11.13	-3.99	-7.52	-11.48
aspen_1080p30	-4.02	-11.85	-20.68	-3.81	-11.32	-18.66
touchdown_pass_1080p30	-2.88	-10.11	-17.43	-2.34	-9.34	-16.69
vqeghd5_src06_original	-1.98	-14.50	-18.19	-1.90	-11.11	-15.02
vqeghd4_src02_original	-2.37	-11.75	-16.82	-2.59	-10.80	-17.02
vqeghd5_csrc14_original	-3.54	-11.55	-20.39	-3.68	-11.73	-20.49
vqeghd4_src01_original	-2.58	-11.71	-20.02	-1.36	-11.92	-20.08
vqeghd5_src01_original	-2.93	-11.96	-20.60	-2.53	-12.42	-20.34
vqeghd4_csrc14_original	-3.54	-11.55	-20.39	-3.68	-11.73	-20.49
vqeghd4_src04_original	-1.00	-9.26	-17.76	-1.00	-6.78	-18.76
vqeghd4_src05_original	-2.39	-11.94	-20.68	-3.65	-12.72	-20.92
vqeghd4_csrc12_original	-3.74	-12.21	-21.66	-3.32	-11.52	-20.50
vqeghd4_src07_original	-1.39	-12.86	-19.58	-1.97	-12.68	-19.32
vqeghd4_src08_original	-3.61	-16.54	-22.77	-2.80	-14.00	-20.58
vqeghd5_src08_original	-2.97	-14.54	-23.17	-1.86	-13.02	-21.80
vqeghd4_src03_original	-4.63	-14.62	-26.09	-4.27	-13.09	-24.57
vqeghd5_src09_original	-3.25	-11.88	-19.98	-2.80	-12.28	-20.31
vqeghd5_csrc13_original	-4.19	-9.42	-16.67	-4.33	-8.91	-16.15
vqeghd5_src04_original	-3.47	-13.35	-19.36	-3.21	-17.62	-23.77
vqeghd4_src06_original	-2.24	-12.64	-22.08	-1.85	-12.50	-25.18
vqeghd5_csrc11_original	-2.18	-11.02	-18.58	-2.78	-13.84	-18.67
vqeghd5_csrc12_original	-3.74	-12.21	-21.66	-3.33	-11.52	-20.50
vqeghd4_src09_original	-3.00	-1.88	-16.61	-2.18	-1.14	-15.63
vqeghd4_csrc13_original	-4.19	-9.42	-16.67	-4.33	-8.93	-16.16
vqeghd5_src05_original	-3.46	-11.23	-19.74	-1.81	-12.32	-20.57
vqeghd5_src02_original	-3.09	-9.93	-26.04	-1.53	-7.40	-23.37
vqeghd4_csrc11_original	-2.18	-11.03	-18.59	-2.78	-13.83	-18.66
Average	-3.07	-11.84	-20.12	-2.81	-11.39	-19.28
Av. AH-VMAF & VMAF	–	-15.98		–	-15.34	
Average of all three	-11.68			-11.16		

Table 2: Bjontegaard delta-rate results for DPP+AV1 in terms of metrics that are increasingly perceptual in nature: SSIM, VMAF_NEG (ADM_ENHN_GAIN_LIMIT = VIF_ENHN_GAIN_LIMIT = 1.0) and VMAF. The first 16 sequences are from the XIPH website; the remaining sequences are from CDVL.

Codec	AV1 (cpu=5, tune_ssim)			AV1 (cpu=5, tune_vmaf)		
Sequence	SSIM (%)	VMAF_NEG (%)	VMAF (%)	SSIM (%)	VMAF_NEG (%)	VMAF (%)
red_kayak_1080p30	-2.13	-10.58	-18.90	-1.83	-7.11	-15.98
ducks_take_off_1080p50	-4.73	-16.82	-22.42	-5.74	-17.09	-22.49
controlled_burn_1080p30	-3.40	-10.20	-33.90	-0.70	-5.53	-31.52
tractor_1080p25	-3.42	-13.31	-25.94	-3.31	-9.03	-19.58
old_town_cross_1080p50	-1.26	-7.99	-21.66	0.07	-3.90	-15.47
riverbed_1080p25	-2.19	-11.62	-16.07	-1.80	-10.55	-15.53
west_wind_easy_1080p30	-1.43	-3.74	-21.61	-4.90	1.50	-30.34
sunflower_1080p25	2.23	-6.80	10.61	-0.17	-10.39	-30.53
rush_field_cuts_1080p30	-2.46	-11.41	-18.50	-2.45	-9.52	-17.32
blue_sky_1080p25	-1.70	-7.68	-16.74	-0.97	-4.92	-6.60
pedestrian_area_1080p25	-1.49	-9.70	-18.42	-1.31	-7.39	-17.41
crowd_run_1080p50	-3.93	-10.44	-15.12	-3.68	-10.22	-14.92
rush_hour_1080p25	-2.50	-13.02	-31.11	-1.12	-7.46	-20.47
park_joy_1080p50	-4.32	-11.62	-17.78	-3.77	-10.85	-18.65
aspen_1080p30	-3.03	-10.20	-25.14	-1.70	-7.75	-24.24
touchdown_pass_1080p30	-3.49	-15.11	-26.95	-2.35	-7.55	-18.13
vqeghd5_src06_original	-4.30	-13.08	-17.14	-3.21	-11.02	-16.30
vqeghd4_src02_original	-1.59	-9.35	-17.57	-2.06	-9.19	-19.06
vqeghd5_csrc14_original	-3.93	-9.16	-21.60	-2.67	-7.61	-21.05
vqeghd4_src01_original	-2.28	-11.46	-22.24	-1.69	-8.80	-20.96
vqeghd5_src01_original	-1.04	-8.09	-16.25	-0.40	-5.54	-15.19
vqeghd4_csrc14_original	-3.93	-9.16	-21.60	-2.67	-7.61	-21.05
vqeghd4_src04_original	-2.22	-12.91	-29.39	-0.62	-6.19	-14.97
vqeghd4_src05_original	-2.02	-19.64	-41.95	0.30	-13.45	-44.34
vqeghd4_csrc12_original	1.63	-6.93	-32.06	6.56	-3.72	-31.69
vqeghd4_src07_original	-3.27	-15.22	-22.90	-2.92	-11.91	-19.84
vqeghd4_src08_original	-2.88	-20.18	-27.59	-2.54	-17.89	-25.92
vqeghd5_src08_original	-3.85	-10.70	-28.83	-0.97	-1.53	-30.45
vqeghd4_src03_original	-1.82	-11.66	-30.73	-3.43	-8.01	-29.72
vqeghd5_src09_original	-1.72	-10.44	-25.76	0.49	-5.93	-21.66
vqeghd5_csrc13_original	-4.21	-16.31	-23.92	-2.92	-14.40	-23.84
vqeghd5_src04_original	-2.34	-14.18	-21.75	-1.61	-11.53	-18.54
vqeghd4_src06_original	-5.80	-9.89	-31.81	-7.59	-6.35	-30.87
vqeghd5_csrc11_original	-2.10	-11.26	-22.58	-2.07	-9.42	-22.07
vqeghd5_csrc12_original	1.63	-6.93	-32.06	6.56	-3.72	-31.69
vqeghd4_src09_original	-1.66	-9.63	-23.89	-1.20	-11.09	-35.26
vqeghd4_csrc13_original	-4.21	-16.31	-23.92	-2.92	-14.40	-23.84
vqeghd5_src05_original	-0.20	-3.23	-14.01	-0.24	0.32	-12.95
vqeghd5_src02_original	-5.03	-6.77	-26.75	1.09	1.57	-23.53
vqeghd4_csrc11_original	-2.10	-11.26	-22.58	-2.07	-9.42	-22.07
Average	-2.46	-11.10	-22.96	-1.61	-8.12	-20.25
Av. AH-VMAF & VMAF	-	-17.03	-	-	-14.19	-
Average of all three	-12.17	-	-	-9.99	-	-

Table 3: Bjontegaard delta-rate results for DPP+VVC in terms of metrics that are increasingly perceptual in nature: SSIM, VMAF_NEG (ADM_ENHN_GAIN_LIMIT = VIF_ENHN_GAIN_LIMIT = 1.0) and VMAF. The first 16 sequences are from the XIPH website; the remaining sequences are from CDVL.

Codec	VVC/H.266 (slow)		
Sequence	SSIM (%)	VMAF_NEG (%)	VMAF (%)
red_kayak_1080p30	-1.99	-7.80	-16.56
ducks_take_off_1080p50	-9.40	-8.23	-24.99
controlled_burn_1080p30	-9.75	-1.88	-18.01
tractor_1080p25	2.78	-7.72	-17.59
old_town_cross_1080p50	-9.13	2.54	-16.35
riverbed_1080p25	-2.88	-6.22	-13.30
west_wind_easy_1080p30	-0.47	-1.40	-23.23
sunflower_1080p25	-8.13	-3.86	-13.42
rush_field_cuts_1080p30	-1.00	-4.49	-16.88
blue_sky_1080p25	-16.47	-0.69	-11.96
pedestrian_area_1080p25	-2.84	-6.81	-15.25
crowd_run_1080p50	3.31	-5.24	-12.34
rush_hour_1080p25	-2.31	-9.48	-22.31
park_joy_1080p50	-4.58	-4.97	-17.96
aspen_1080p30	-6.83	-3.24	-16.01
touchdown_pass_1080p30	-3.18	-5.89	-17.10
vqeghd5_src06_original	-2.82	-7.06	-13.17
vqeghd4_src02_original	-4.70	-2.04	-14.73
vqeghd5_csrc14_original	-2.48	-4.13	-14.30
vqeghd4_src01_original	8.49	-7.74	-18.59
vqeghd5_src01_original	3.61	-4.25	-11.35
vqeghd4_csrc14_original	-2.48	-4.13	-14.30
vqeghd4_src04_original	-0.99	-1.31	-11.31
vqeghd4_src05_original	4.65	-3.37	-26.27
vqeghd4_csrc12_original	-7.76	0.36	-18.06
vqeghd4_src07_original	-2.64	-3.57	-12.76
vqeghd4_src08_original	-4.59	-13.21	-25.38
vqeghd5_src08_original	-8.21	-4.81	-18.71
vqeghd4_src03_original	-1.65	-4.70	-21.02
vqeghd5_src09_original	0.29	-7.82	-21.09
vqeghd5_csrc13_original	-6.46	-11.40	-25.27
vqeghd5_src04_original	-1.38	-9.63	-17.63
vqeghd4_src06_original	2.10	-6.14	-23.62
vqeghd5_csrc11_original	2.02	-4.62	-14.77
vqeghd5_csrc12_original	-7.76	0.36	-18.06
vqeghd4_src09_original	-9.53	-4.75	-21.09
vqeghd4_csrc13_original	-6.46	-11.40	-25.27
vqeghd5_src05_original	-6.17	2.38	-25.48
vqeghd5_src02_original	-8.13	-0.63	-18.51
vqeghd4_csrc11_original	2.02	-4.62	-14.77
Average	-3.35	-4.84	-17.97
Av. AH-VMAF & VMAF	–	-11.40	
Average of all three	-8.72		