Supplementary Material for Perceptual Video Compression with Neural Wrapping

This document provides

- additional visual comparisons (Sec. 1)
- additional objective comparisons (Sec. 2)
- additional results of pretraining (Sec. 3)
- description of the proposed inference architecture and its complexity (Sec. 4),
- codec recipes, methodology for BD-rate measurements and combined plots, and utilized content (Sec. 5-Sec. 9),
- details on the type of subjective testing used, and a summary of its key attributes. (Sec. 10),
- PSNR performance (Sec. 11)

1. Visual Comparison

Beyond the visual results on gaming sequences presented in the main paper, additional visual comparisons to DCVC-DC [7] are provided in Fig. 1 on gaming and natural sequences. When assessing visuals, the proposed method is found to better preserve text and fine details/textures in the content at lower bitrates than DCVC-DC. Visual comparison vs DCVC-FM [8] is shown in Fig. 2. It is again noticed that the proposed method provides better visual quality compared to DCVC-FM.

2. Additional Objective Results

Refer to Fig. 3 to see the rate distortion plots for all the methods. These plots show that the proposed approach outperforms competing methods across a variety of metrics.

3. Additional Results of Pretraining

A novel pretraining method for a proxy codec model has been proposed in the main paper. Fig. 4 shows scatter plots of different metrics for the actual codec vs the codec proxy before and after the pretraining. The plots clearly show better correlation between SVT-AV1 and the codec model after pretraining, indicating that the proposed approach allows the codec model to learn a meaningful representation of the rate and distortion behavior of an actual codec.

4. Network Architectures

A similar network architecture is used for both the pre and postprocessor. An initial convolutional layer processes an input and produces a 64-channel feature. This is followed by five ResBlocks [3], each with 64 channels. This is then followed by a final convolutional layer. Importantly, both pre and postprocessors have a single frame latency. The preprocessor processes the luma channel only, given that: *(i)* the human visual system has much higher sensitivity to luminance deviation than chromatic deviation; *(ii)* luminance contributes substantially more to rate; *(iii)* it was found empirically that high-frequency information embedded by the preprocessor in the chroma channels is likely to be removed by the target codec and is not propagated to the postprocessor. After an initial end-to-end training of the pre and the postprocessor for 100000 iterations, the trained models are pruned to 16 channels per ResBlock. Specifically, structured pruning is performed by ranking weights based on their channel-wise L1 norms and lower ranked channels are pruned. The pruned networks are trained end-to-end for 500000 iterations. Afterwards, static quantization is performed resulting in int8 quantized weights. Per output pixel, this results to 7.7KMACs and 9.5KMACs for the pre- and postprocessor, respectively. Given that preprocessing is applied once for all quality levels, these compute requirements are significantly lower than those of neural codecs like DCVC-DC [7], which tend to require 350KMACs/pixel or higher [8]. Putting this into context: modern standard codecs (e.g., AV1 and VVC) are in the range of 40KMACs/pixel for encoding and 4KMACS/pixel for decoding [14]; therefore, the proposed approach bundled with VVC or AV1 reaches: (7.7+40)KMACs/pixel for encoding and (4+9.5)KMACs/pixel for decoding. This totals 61.2KMACs/pixel, which is more than 5.7 times lower than the complexity of DCVC-FM [8]. The networks are shown graphically in Fig. 5.

Although ResNets [3] were proposed relatively early and are somewhat simple, preliminary experiments with other architectures demonstrated ResNets to be the optimal choice for the pre and postprocessors. Experiments with U-Nets [15] indicated a tendency to overfit to the training data, making them less suitable. Transformer architectures, such as Swin Transformer blocks [11], were not used for two reasons: their performance significantly degrades at lower computational complexity, and they struggle to gen-

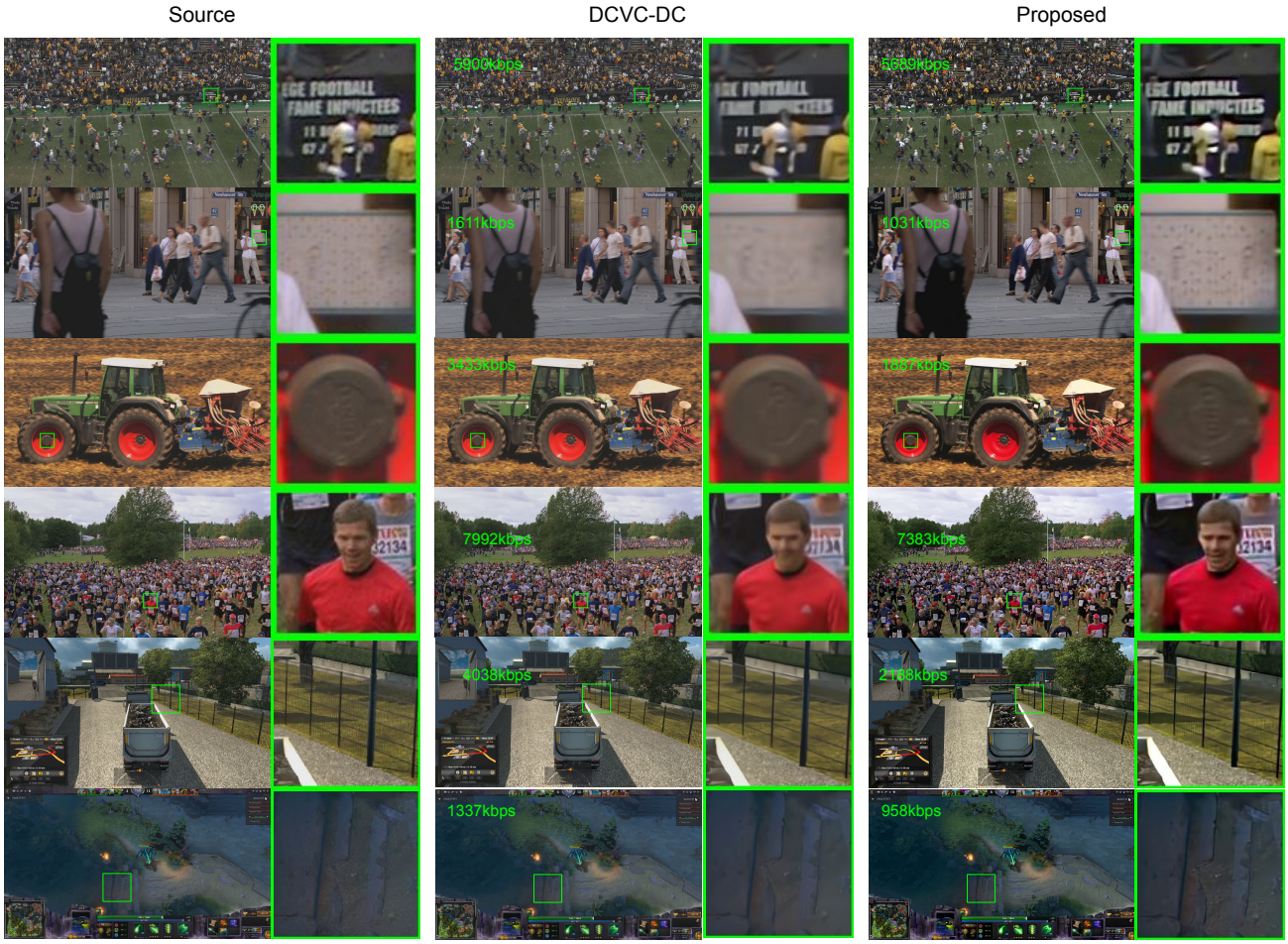


Figure 1. Visual comparison of the proposed method with DCVC-DC [7]. Zoom for better view.

erate high-frequency details without being cascaded with CNNs. Multiframe temporal networks, which are expected to perform better than single-frame models, were not used to avoid latency issues in the video streaming framework.

The input frames during training are $I_{in} \in \mathbb{R}^{b\times T\times 3\times M\times M}$ where b is the batch-size, T is the GOP, 3 is the number of channels (YUV) and M is the patch size. The first two dimensions of the input are flattened at the input of the pre and postprocessor to $I'_{in} \in \mathbb{R}^{bT\times 3\times M\times M}$ and unflattened at the output. Evaluation is performed with a single frame input to the pre and postprocessor. Inputs to the networks are scaled to the $[0, 1]$ and outputs are scaled based on the target bitwidth, e.g., $[0, 255]$. Note that the same procedure and training losses are used when training with the proxy codec model proposed in [4]. More specifically, a differentiable implementation of JPEG is used with varying block sizes. The bitrate during training is estimated as the scaled sum of the absolute values of the DCT co-efficients. Since the authors do not mention the quality factor of JPEG used in their training, we use a quality factor of 85% based on empirical evaluation.

The input to the proposed methods is a single frame. Unlike other single frame methods which produce temporal flicker when used with videos, the proposed method produces temporally coherent results because the codec in the middle processes multiple frames at once. In other words, the training pipeline is temporal. Temporal coherence of the proposed method is evident by the improved MOS results. Temporally incoherent videos perform extremely poor in MOS tests.

5. Codecs

The AV1 binary was built from the repository of SVT-AV1 https://gitlab.com/AOMediaCodec/SVT-AV1 with version v1.8.0. The VVC binary was built from the VVenC repository https://github.com/fraunhoferhhi/vvenc with version v1.10.0. Default build options were used for both the codecs.

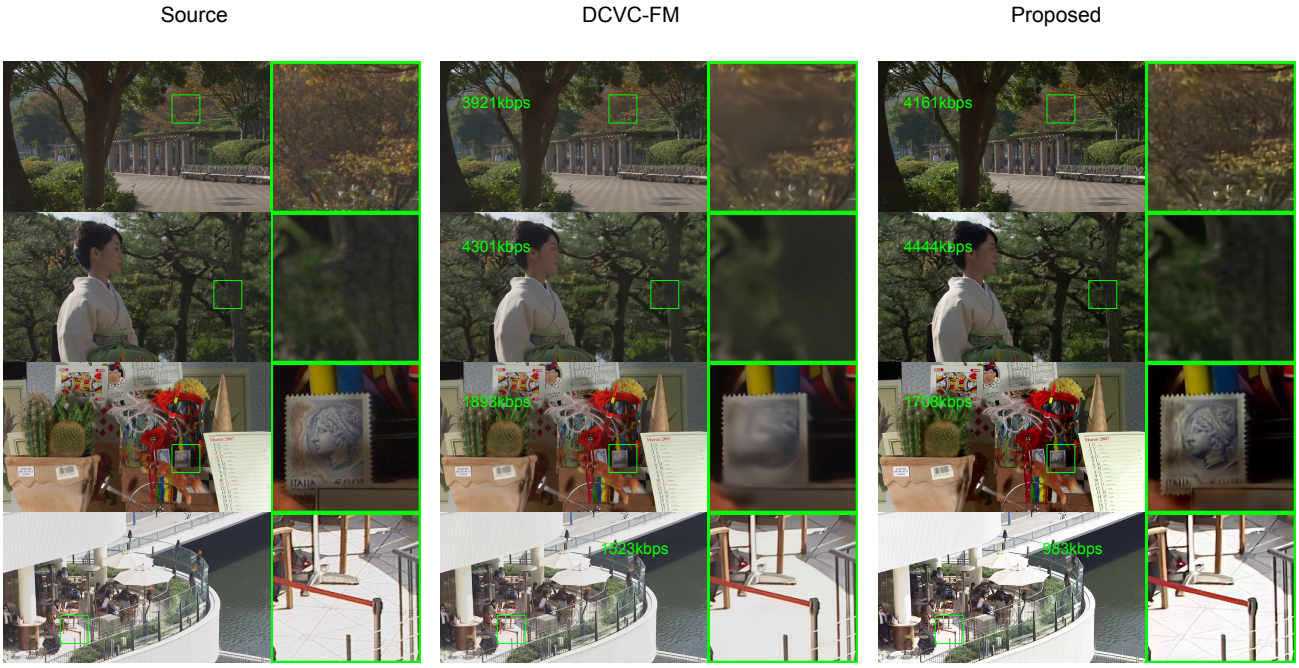


Figure 2. Visual comparison of the proposed method with DCVC-FM [8]. Zoom for better view.

6. Encoding Recipes

AV1: `SvtAv1EncApp -i <invideo> --preset 4 --keyint -1 --crf <crf> -b <outvideo>;`
AV1+SSIM: `SvtAv1EncApp -i <invideo> --preset 4 --keyint -1 --tune 2 --crf <crf> -b <outvideo>;`

The following low-delay recipe of AV1 was used in training and ablation:
AV1 Low-delay: `SvtAv1EncApp -i <invideo> --preset 8 --keyint -1 --fast-decode 1 --pred-struct 1 --crf <crf> -b <outvideo>;`

VVC: `vvencapp -i <invideo.yuv> -s <width>x<height> -c yuv420 -r <fps> --preset slow --qp <crf> --qpa 0 -ip 256 -t 4 -o <outstream.266>;`
VVC+QPA: `vvencapp -i <invideo.yuv> -s <width>x<height> -c yuv420 -r <fps> --preset slow --qp <crf> --qpa 1 -ip 256 -t 4 -o <outstream.266>;`

DCVC-DC:
```
python test_video.py
--i_frame_model_path
./checkpoints/cvpr2023_image_ssim.pth.tar
--p_frame_model_path
./checkpoints/cvpr2023_video_ssim.pth.tar
--rate_num 4 --test_config
<path_to_rgb_config_json> --cuda 1
--worker 1 --write_stream 1
--output_path output.json
--save_decoded_frame 1
--decoded_frame_path
<path_to_decoded_frames> --verbose 2
```

DCVC-FM:
```
python test_video.py
--model_path_i
./checkpoints/cvpr2024_image.pth.tar
--model_path_p
./checkpoints/cvpr2024_video.pth.tar
--rate_num 4 --test_config
<path_to_rgb_config_json> --cuda 1
--worker 1 --write_stream 1
--output_path output.json
--save_decoded_frame 1 --verbose 2
```

7. CRFs/QPs Used in Evaluation

AV1: {22, 27, 32, 37, 42, 47, 50, 52, 55, 57, 59, 61, 63};
VVC: {14, 16, 18, 20, 22, 27, 32, 37, 42, 47, 50, 52, 55, 57, 59, 61, 63}.

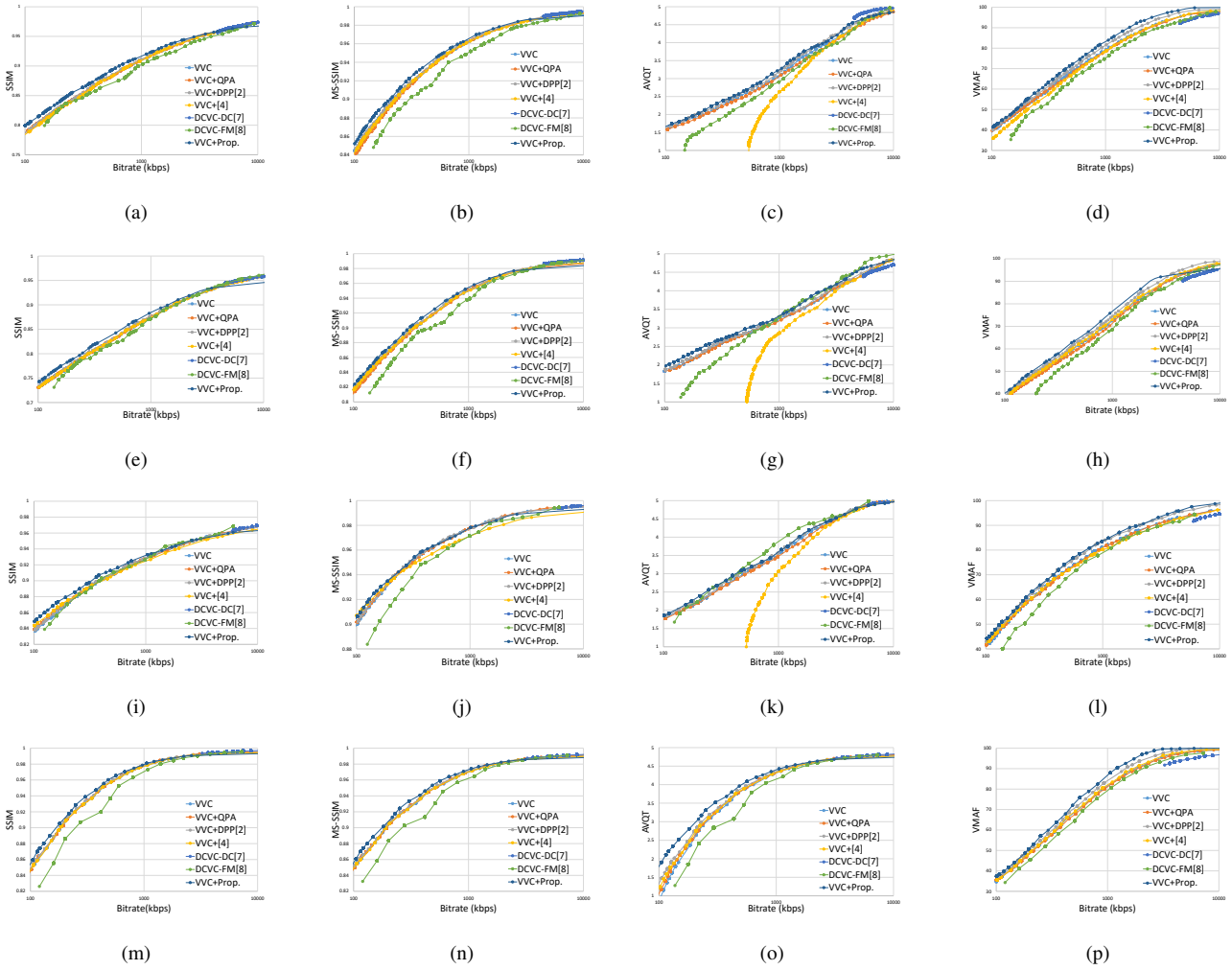


Figure 3. Combined rate-quality plots over Gaming ((a), (b), (c), (d)), XIPH ((e), (f), (g), (h)), UVG ((i), (j), (k), (l)) and HEVC-B ((m), (n), (o), (p)) datasets. Unlike DPP [2] and the rest of the methods, current version of DCVC-DC [7] has limited coverage of the entire bitrate-quality regime of VVC.

8. BD-rate Measurement and Slope-based Integration for Combined Plots

All BD-rate measurements are done with the code from the libvmaf library of Netflix [9]. The entire range of each quality metric is used and BD-rates are averaged over all sequences of each dataset. AVQT is measured using the Linux library provided by Apple.

The combined plots of Fig. 3 and Fig. 2 of the main paper are produced based on the slope-based integration method of Wu et al. [17]. This approach generalizes individual rate-quality convex-hulls towards convex-hull curves produced over the entire dataset. As shown by Fig. 3 and Fig. 2 of the main paper, this provides for [17]: *(i)* significantly higher number of points per method across the entire bitrate-quality range; *(ii)* a way to compare the performance of multiple methods over entire datasets, instead of focusing on individual sequences. Note that the appearance of non convexity on some curves is due to the use of log-scale in the bitrate axis. For completeness and reproducibility, the operation of slope-based integration is summarized here based on the description of Wu *et al.* [17]. The combined curves generalize the "constant-slope" method that is the core of the dynamic optimizer framework [6]. The following steps are applied:

- All test sequences of each dataset are grouped together into one class.
- For each such class, each sequence is treated as a "shot", i.e., part of a longer sequence referred to as the combined sequence that is the virtual "collage" of all sequences in the class. For example, for the seven 1080p sequences of the UVG dataset, with a total of 7× 96 frames, the

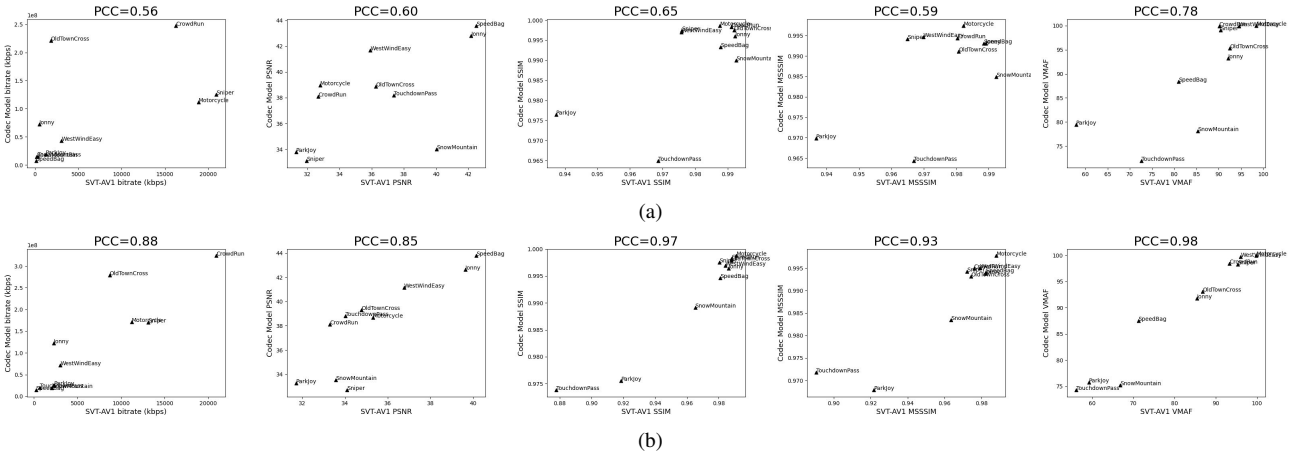


Figure 4. Alignment between the codec model and actual codec (a) before and (b) after the proposed pretraining method.

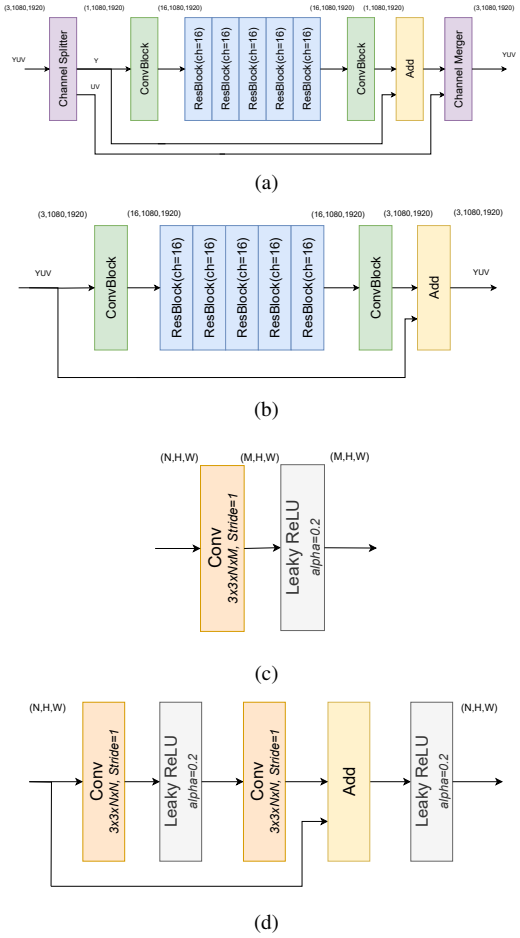


Figure 5. The (a) preprocesor, (b) postprocessor, (c) ConvBlock and (d) ResBlock.

ensemble of these 672 frames is considered as a single combined video sequence.

- After the convex hull is obtained for each "shot", the multiple convex hulls (seven for the previous example) are combined as described in the dynamic optimizer framework [6], i.e., using the constant slope principle.

This results in a single rate-quality curve describing the coding performance over the entire combined sequence [17] (i.e., over the entire dataset).

9. Test Sequences

The following sequences are used for evaluation:
Gaming: {*apex_legends, doom, dota2, euro_truck_simulator2, fallout4, flatout3_a, fortnite_b, grid2a, gtav, injusticegodsauv2, metro_exodus_c, quake_champions, warframe*};
XIPH: {*aspen, blue_sky, controlled_burn, crowd_run, ducks_take_off, old_town_cross, park_joy, pedestrian_area, red_kayak, riverbed, rush_field_cuts, rush_hour, sunflower, touchdown_pass, tractor, west_wind_easy*};
UVG: {*Beauty, Bosphorus, HoneyBee, Jockey, ReadySteadyGo, ShakeNDry, YachtRide*}.
HEVC-B: {*BQTerrace, BasketballDrive, Cactus, Kimono, ParkScene*}.

All sequences are 8-bit and 1080p. For DCVC-DC, we use GOP of 32 as suggested by the authors [7]. DCVC-FM [8] uses infinite GOP as per authors recommendation. All encodings with the standard codecs use an infinite GOP. The total number of frames per sequence is limited to 96 frames, as per DCVC-DC [7] and DCVC-FM [8] testing methodology.

The gaming sequences comprise first-person, third-person, platform and simulator games, active and static scenes, and varying levels of texture detail. All sequences

are in 1080p resolution and are captured via a professional live gameplay capture, at framerates of 30fps or 60fps. The XIPH sequences are all the synonymous 1080p sequences available at https://media.xiph.org/video/derf/. Finally, as in prior work [7], the UVG dataset is using the seven 1080p sequences [13] available at https://github.com/ultravideo/UVG-4K-Dataset.

10. Subjective Evaluation Protocol

For all subjective tests, the ITU-T P.910 subjective quality scoring [5] is used. As per ITU-T Visual Quality Experts Group (VQEG [1]) recommendations, for a fixed amount of testing time, the 5-scale absolute categorical rating with hidden reference provides for the best accuracy (i.e., smallest confidence intervals).[1] A total of 72 sequences are used in each test, taken from the encodings of the Gaming dataset. Specifically, there are 8 hidden-reference (source) sequences, selected as a representative subset, with 4 encodings per sequence at varying QP levels. The raw scores of the test with AV1 and VVC are shown in Fig. 7a and Fig. 7b, respectively, in the grayscale mapping produced by SUREAL [10]. The strong presence of vertical lines indicates agreement between raters, while the varying shades of gray per row indicate that the utilized test content ensures wide coverage of the rating scale. The raters' biases and inconsistencies are shown in Fig. 8 and Fig. 9, and are within expectations (from the guidance provided by the authors of SUREAL). The inconsistencies before and after SUREAL processing are shown in Fig. 10 and Fig. 11, showing how maximum-likelihood estimation carried out by SUREAL adjusts the recovered MOS score per clip according to the subject bias and inconsistency. Finally, the combined plots of bitrate-MOS of Fig. 4 of the main paper are established by using these recovered MOS scores in conjunction with the methodology for the combined plots of Sec. 8.

11. PSNR Results

The PSNR BD-rates of the proposed method are shown in Tab. 1. Since the method is optimized for perceptual quality, we see a decreased PSNR performance. However, the proposed approach shows improvement on other fidelity metrics such as SSIM and MS-SSIM, as seen from the results in the paper. It is widely known that PSNR is not a valid indicator of perceptual quality of a video. Refer to Fig. 6 as an example which compares DCVC-FM with the proposed approach. DCVC-FM shows excellent performance on PSNR, surpassing numerous recent codecs. This is also indicated in the figure as DCVC-FM has better PSNR compared to the proposed method. However, the visual quality of the proposed method is superior compared to DCVC-FM. The proposed method does a significantly better job of preserving geometric and textural details. In line with the recent remarkable performance of neural codecs, especially DCVC-FM, future research on neural codecs will benefit from an increased focus on perceptual quality of videos.

[1] Differential Categorical Rating (DCR) may offer smaller confidence intervals per rater than Absolute Category Rating (ACR) as subjects rate the difference between source and decoded result, but takes twice the time, allowing only half the scores. As confidence intervals decrease polynomially with more scores [12, 16], ACR-HR is preferred for superior accuracy. Moreover, increasing granularity in the testing scale does not improve accuracy significantly, and tends to prolong the rating time per clip [16].

Table 1. PSNR BD-rates of the proposed method with VVC as baseline.

Method	Gaming	Xiph	UVG	HEVC-B
VVC+Prop.	15	13	22	11

References

[1] Jochen Antkowiak, T Jamal Baina, France Vittorio Baroncini, Noel Chateau, France FranceTelecom, Antonio Claudio França Pessoa, F Stephanie Colonnese, Italy Laura Contin, Jorge Caviedes, and France Philips. Final report from the video quality experts group on the validation of objective models of video quality assessment march 2000. *Final report from the video quality experts group on the validation of objective models of video quality assessment*, 2022. 6

[2] Aaron Chadha and Yiannis Andreopoulos. Deep perceptual preprocessing for video coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14852–14861, 2021. 4

[3] 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. 1

[4] Yueyu Hu, Chenhao Zhang, Onur G Guleryuz, Debargha Mukherjee, and Yao Wang. Standard compliant video coding using low complexity, switchable neural wrappers. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 1922–1928. IEEE, 2024. 2

[5] ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications. *ITU*, 2022. 6

[6] Ioannis Katsavounidis and Liwei Guo. Video codec comparison using the dynamic optimizer framework. In *Applications of Digital Image Processing XLI*, page 107520Q. International Society for Optics and Photonics, 2018. 4, 5

[7] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22616–22626, 2023. 1, 2, 4, 5, 6

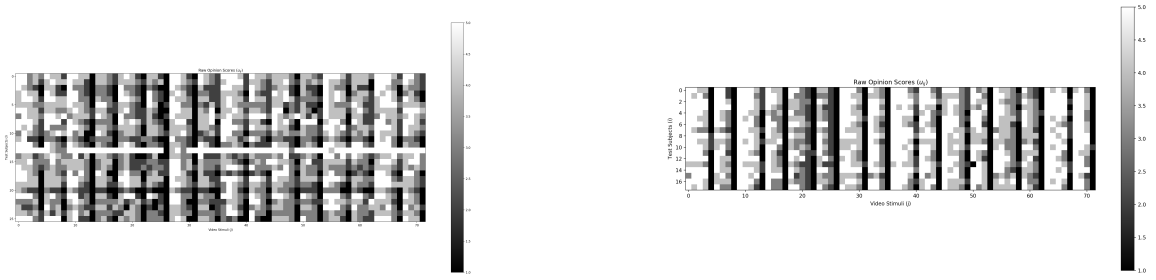
[8] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26099–26108, 2024. 1, 3, 5

Figure 6. Visual comparison of DCVC-FM and the proposed method.

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

[10] Zhi Li, Christos G Bampis, Lukáš Krasula, Lucjan Janowski, and Ioannis Katsavounidis. A simple model for subject behavior in subjective experiments. *arXiv preprint arXiv:2004.02067*, 2020. 6

[11] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1



(a) AV1. (b) VVC.

Figure 7. Raw MOS scores.

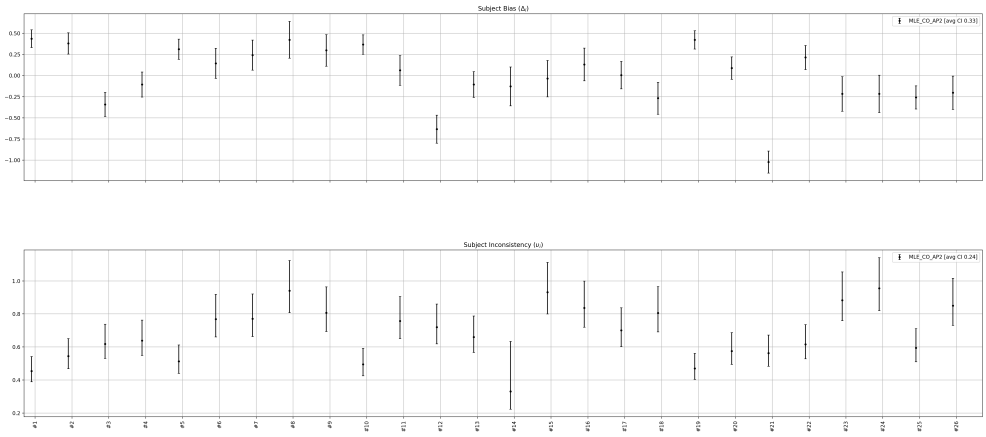


Figure 8. Subjects bias and inconsistency with SVT-AV1.

[12] Rafał K Mantiuk, Anna Tomaszewska, and Radosław Mantiuk. Comparison of four subjective methods for image quality assessment. In *Computer graphics forum*. Wiley Online Library, 2012. 6

[13] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. Uvg dataset: 50/120fps 4k sequences for video codec analysis and development. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 297–302, 2020. 6

[14] Debargha Mukherjee. Challenges in incorporating ml in a mainstream nextgen video codec. *CVPR, CLIC 2022 competition*, 2022. 1

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 1

[16] Toshiko Tominaga, Takanori Hayashi, Jun Okamoto, and Akira Takahashi. Performance comparisons of subjective quality assessment methods for mobile video. In *QoMEX*, 2010. 6

[17] Ping-Hao Wu, Ioannis Katsavounidis, Zhijun Lei, David Ronca, Hassene Tmar, Omran Abdelkafi, Colton Cheung, Foued Ben Amara, and Faouzi Kossentini. Towards much better svt-av1 quality-cycles tradeoffs for vod applications. In *Applications of Digital Image Processing XLIV*, pages 236–256. SPIE, 2021. 4, 5

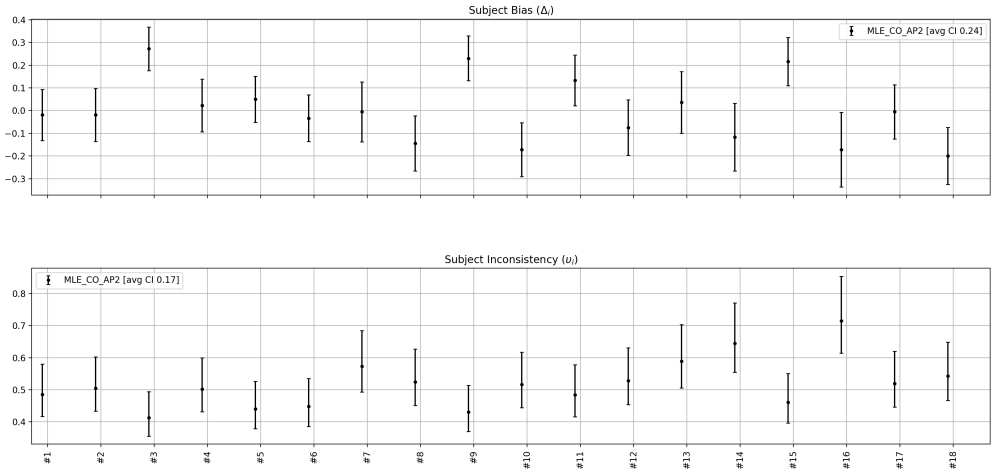


Figure 9. Subjects bias and inconsistency with VVC.

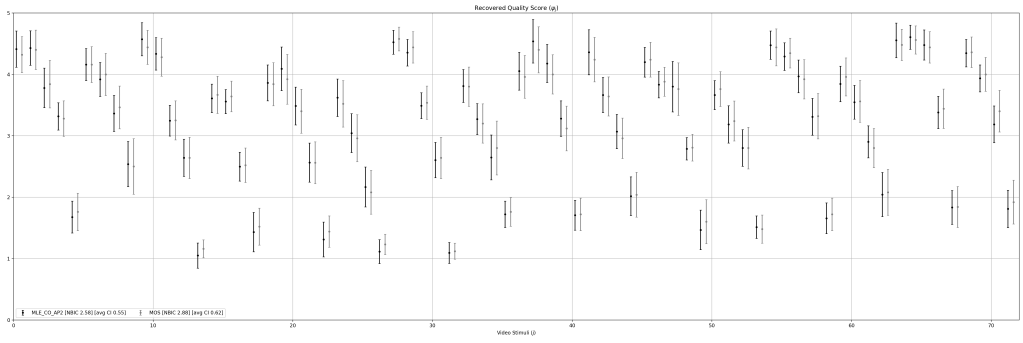


Figure 10. Recovered quality scores with SVT-AV1.

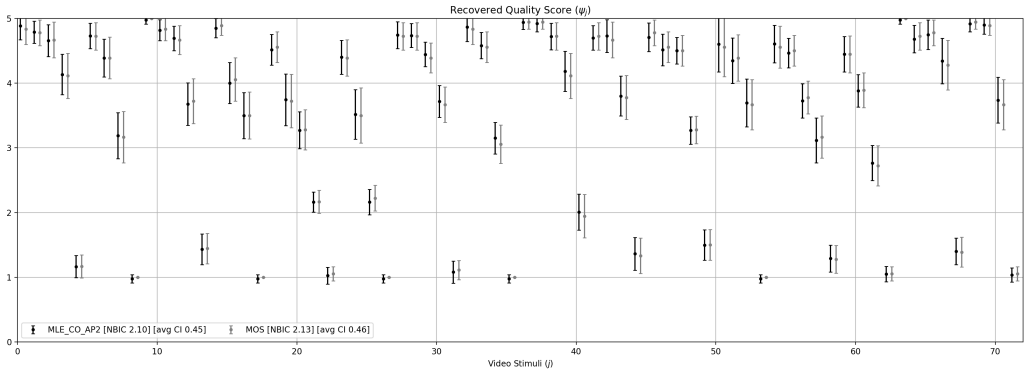


Figure 11. Recovered quality scores with VVC.