Supplementary Material for: Perceptual Neural Video Compression with Color Separation and Rank Chain

Xiongzhuang Liang Chuanbo Tang Zhuoyuan Li Li Li Dong Liu
MOE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition,
University of Science and Technology of China, Hefei 230093, China
{lxz123, cbtang, zhuoyuanli}@mail.ustc.edu.cn, {lil1, dongeliu}@ustc.edu.cn

Contents

1. Detailed Experimental Setup

1.1. Traditional Codec Configuration

For the traditional video coding baseline, we use VTM-13.2 under the standard low-delay setting, with Quantization Parameter (QP) values set to 25, 29, 33, and 37. The exact command-line parameters used for VTM encoding are listed below:

```
-c encoder_lowdelay_vtm.cfg
--InputFile={input_file}
--InputBitDepth=8
--OutputBitDepth=8
--OutputBitDepthC=8
--InputChromaFormat=420
--FrameRate={frame_rate}
--DecodingRefreshType=2
--FramesToBeEncoded={frames}
--SourceWidth={width}
--SourceHeight={height}
--IntraPeriod={intra_period}
--QP={qp}
--Level=6.2
--BitstreamFile={bitstream_file}
```

1.2. Data Processing and Metric Calculation

All evaluated codecs inherently operate in the YUV color space. The original uncompressed YUV420 sequences (`in.yuv`) are directly fed into the encoders to produce the reconstructed sequences (`out.yuv`) in the exact same format. Quality metrics are computed based on these two sequences.

For objective quality assessment, we compute PSNR and SSIM for each individual channel (Y, U, V). Following standard committee practice [2], the overall YUV PSNR and YUV SSIM are derived via a weighted average of the three channels with a ratio of 6:1:1. To calculate VMAF, we directly process the YUV sequences using the official command-line tool:

```
vmaf -r in.yuv -d out.yuv --width W
--height H --pixel_format 420 --bitdepth 8
```

Since perceptual metrics (LPIPS, DISTS, KID, FID) are designed for the RGB domain, the YUV420 frames are converted to RGB via the BT.709 standard prior to evaluation.

2. Additional Experimental Results

2.1. 96 Frames with Intra-Period -1

Table 1 and Figure 1 supplement Table 1 and Figure 3 of the main paper, respectively, under the setting of 96 frames with intra-period = -1. Specifically, Table 1 further reports the BD-Rate results on U PSNR, V PSNR, Y SSIM, U SSIM, and V SSIM. Correspondingly, Figure 1 visualizes the rate-distortion (RD) curves for these additional metrics evaluated on the HEVC-B dataset.

2.2. 96 Frames with Intra-Period 32

Table 2 reports the BD-Rate comparison under the setting of 96 frames with intra-period = 32, where VTM-13.2 is used as the anchor.

2.3. All Frames with Intra-Period -1

Table 3 reports the additional BD-Rate comparison under the all-frame setting with intra-period = -1, where DCVC-DC is used as the anchor.

Table 1. Additional BD-Rate (%) comparison supplementing the main paper Table 1. The anchor is VTM-13.2. 96 frames with intra-period = -1. **Red** indicates the best performance.

Metric	Method	HEVC_B	HEVC_C	HEVC_D	HEVC_E	MCL-JCV	USTC-TD	UVG	Average
U PSNR	DCVC-DC	-5.7	-10.5	-31.1	6.0	-16.0	6.4	-26.4	-11.04
	DCVC-FM	-57.8	-54.6	-70.0	-67.4	-58.7	-46.7	-74.3	-61.36
	DCVC-RT	-58.0	-63.6	-71.5	-62.8	-64.9	-40.2	-75.8	**-62.40**
	PNVC-C-Base	-53.7	-51.8	-63.9	-52.4	-53.7	-49.7	-67.8	-56.14
	PNVC-CR	-35.7	-42.8	-55.2	-26.3	-34.9	-17.4	-54.0	-38.04
V PSNR	DCVC-DC	5.0	-21.7	-45.0	-30.4	-29.8	-8.1	-26.9	-22.41
	DCVC-FM	-43.1	-52.1	-70.2	-71.8	-60.5	-44.8	-71.9	**-59.20**
	DCVC-RT	-34.1	-54.3	-69.1	-70.2	-63.0	-38.3	-72.0	-57.29
	PNVC-C-Base	-42.3	-51.4	-66.5	-64.3	-57.9	-53.8	-66.3	-57.50
	PNVC-CR	-25.3	-42.7	-59.1	-47.2	-45.2	-33.2	-56.1	-44.11
Y SSIM	DCVC-DC	-8.6	-15.7	-28.9	-17.2	-1.2	26.0	-5.8	-7.34
	DCVC-FM	-8.1	-16.2	-28.6	-22.8	0.7	46.5	-7.0	-5.07
	DCVC-RT	-11.3	-13.1	-22.8	-11.6	-0.4	44.5	-8.9	-3.37
	PNVC-C-Base	-18.3	-24.5	-32.5	-22.7	-7.9	28.9	-15.3	**-13.19**
	PNVC-CR	-10.8	-27.1	-35.8	-8.2	-4.3	27.6	-6.4	-9.29
U SSIM	DCVC-DC	-21.4	-21.4	-30.1	-8.8	-11.0	4.8	-23.2	-15.87
	DCVC-FM	-65.0	-66.9	-67.0	-71.2	-59.8	-45.0	-73.0	-63.99
	DCVC-RT	-69.9	-73.0	-74.1	-66.4	-62.5	-33.1	-74.8	**-64.83**
	PNVC-C-Base	-61.0	-62.1	-65.8	-56.0	-52.0	-45.7	-66.8	-58.49
	PNVC-CR	-52.0	-54.0	-59.7	-38.0	-39.1	-21.6	-58.3	-46.10
V SSIM	DCVC-DC	-11.6	-37.6	-42.8	-35.2	-25.2	-5.8	-24.4	-26.09
	DCVC-FM	-58.2	-68.6	-69.4	-74.2	-61.6	-40.9	-73.5	**-63.77**
	DCVC-RT	-59.9	-71.9	-73.7	-72.4	-61.8	-30.2	-72.2	-63.16
	PNVC-C-Base	-54.2	-64.2	-68.5	-64.4	-56.6	-49.1	-67.6	-60.66
	PNVC-CR	-47.1	-57.7	-63.4	-50.0	-48.2	-32.7	-61.1	-51.46

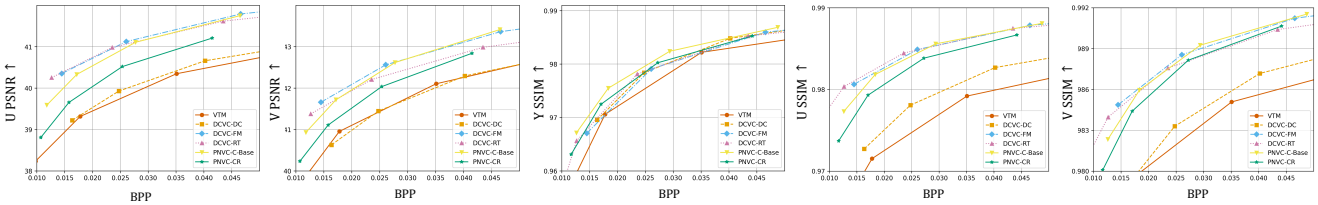

Figure 1. RD curves evaluated on the HEVC-B under the setting of 96 frames with intra-period = -1. This figure supplements Figure 3 of the main paper.

3. Further Analysis and Discussions

3.1. Chroma Bitrate Allocation and Color Fidelity

Q: Does the extremely low chroma bitrate allocation risk color detail loss or color drift, especially in highly saturated scenarios?

A: No, our framework robustly preserves color fidelity. This is supported by both quantitative analysis and the adaptive nature of our architectural design.

First, quantitatively, as shown in Table 1 and Figure 1, our PNVC-CR achieves substantial BD-Rate savings of **38%** and **44%** over VTM in U-PSNR and V-PSNR, respectively. At a bitrate of BPP=0.014, the U-PSNR and V-PSNR reach impressive levels of 38.8 dB and 40.2 dB, respectively, validating our strong color preservation capabilities.

Second, the overall low chroma bitrate is an intentional efficiency gain driven by our Color Fidelity Refinement (CFR) mechanism. Unlike traditional color separation, our end-to-end design enables the Chroma-Net to fully exploit the reconstructed luminance structure to guide chroma generation. This deep cross-component interaction vastly reduces chroma redundancy, yielding much higher color compression efficiency.

Finally, the chroma bitrate allocation is not fixed but highly *content-adaptive*. When processing vivid, highly saturated sequences (e.g., **MCL-JCV SRC24**), the chroma bitrate proportion dynamically scales up to **10.7%–16.2%** across the four rate points to preserve fine chromatic textures. Conversely, for plain sequences (e.g., **MCL-JCV SRC01**), it remains minimal at **0.2%–0.6%**. This demonstrates that our framework intelligently allocates bits where color details are most needed.

3.2. Working Mechanism of Rc-GAN

Q: Why doesn't perceptual quality naturally improve with bitrate during adversarial training, and why do quality score inversions occur?

A: Theoretically, higher bitrates provide more capacity and should naturally yield better perceptual quality. However, the root cause of quality inversions lies in the inherent instability of the adversarial discriminator, which acts as an imperfect proxy for human perception. During training, an unconstrained discriminator may erroneously assign higher scores to lower-quality (lower-bitrate) reconstructions. This phenomenon is also observed in ReWaGAN [1], where a critic might even score a distorted image higher than the ground truth. When the discriminator makes such errors, it backpropagates misleading gradients, forcing the high-bitrate model to fit a flawed "high-score" distribution. Our rank chain constraint (Eq. 1) is explicitly designed to calibrate the discriminator's optimization direction. By only passing gradients when the quality hierarchy is correctly recognized, we ensure the encoder is optimized strictly in the correct perceptual direction,

Table 2. BD-Rate (%) comparison. The anchor is VTM-13.2. 96 frames with intra-period = 32. **Red** indicates the best performance.

Metric	Method	HEVC_B	HEVC_C	HEVC_D	HEVC_E	MCL-JCV	USTC-TD	UVG	Average
LPIPS	DCVC-DC	12.1	-9.6	-20.5	10.2	20.7	24.6	14.7	7.46
	DCVC-FM	0.9	-16.3	-29.1	4.1	16.0	44.9	6.8	3.90
	DCVC-RT	8.6	-2.8	-15.2	17.1	19.4	45.7	6.9	11.39
	PNVC-CR	-83.8	-56.2	-57.3	-75.5	-83.9	-71.9	-85.2	**-73.40**
DISTS	DCVC-DC	69.2	23.2	3.8	49.4	56.4	129.0	62.4	56.20
	DCVC-FM	65.4	14.4	-7.4	42.6	54.3	169.3	62.6	57.31
	DCVC-RT	64	20.9	6.1	46.5	56.4	168.1	61	60.43
	PNVC-CR	-48.8	-31.3	-29.4	-26.4	-51.5	-40.6	-52.0	**-40.00**
KID	DCVC-DC	72.9	34.4	8.9	43.1	35.9	72.7	81.5	49.91
	DCVC-FM	71.5	30.3	5.8	28.0	37.5	108.6	80.0	51.67
	PNVC-CR	-52.7	-47.7	-32.3	-67.5	-62.5	-47.3	-40.4	**-50.06**
FID	DCVC-DC	50.4	21.5	8.7	32.5	48.9	95.9	64.7	46.09
	DCVC-FM	48.6	18.4	1.0	21.5	45.4	121.3	68.7	46.41
	PNVC-CR	-52.3	-47.0	-34.2	-51.8	-37.8	-16.5	-30.5	**-38.59**
YUV PSNR	DCVC-DC	-13.6	-10.7	-23.4	-19.8	-14.8	8.6	-20.3	-13.43
	DCVC-FM	-18.5	-18.1	-34.5	-33.5	-19.2	9.1	-28.0	-20.39
	DCVC-RT	-9.7	-1.9	-14.8	-22.4	-13.3	16.7	-23.1	-9.79
	PNVC-C-Base	-21.1	-21.8	-33.5	-29.1	-21.5	-0.9	-30.8	**-22.67**
	PNVC-CR	-12.2	-17.3	-29.1	-17.5	-14.4	12.4	-20.5	-14.09
Y PSNR	DCVC-DC	-14.6	-8.2	-20.0	-17.4	-11.9	13.0	-16.6	-10.81
	DCVC-FM	-11.2	-10.3	-25.9	-21.9	-8.1	29.1	-16.2	-9.21
	DCVC-RT	-0.5	13.7	1.2	-8.1	0.6	37.0	-9.8	4.87
	PNVC-C-Base	-16.1	-15.2	-26.2	-20.8	-13.5	16.7	-22.3	**-13.91**
	PNVC-CR	-9.5	-12.7	-23.9	-12.1	-9.4	22.3	-12.9	-8.31
U PSNR	DCVC-DC	-15.8	-17.6	-30.9	-16.7	-20.8	0.9	-36.8	-19.67
	DCVC-FM	-56.6	-52.4	-66.4	-64.7	-58.4	-47.1	-71.0	**-59.51**
	DCVC-RT	-58.4	-60.3	-66.9	-59.4	-61.0	-39.8	-70.2	-59.43
	PNVC-C-Base	-49.4	-47.6	-58.5	-47.7	-49.6	-46.5	-62.9	-51.74
	PNVC-CR	-31.1	-36.9	-46.8	-24.3	-29.3	-13.6	-49.2	-33.03
V PSNR	DCVC-DC	-1.9	-26.6	-43.9	-37.9	-33.7	-11.7	-37.5	-27.60
	DCVC-FM	-40.8	-49.6	-66.5	-69.5	-60.2	-45.1	-69.1	**-57.26**
	DCVC-RT	-38.2	-51.5	-64.8	-65.5	-59.4	-37.4	-66.0	-54.69
	PNVC-C-Base	-35.3	-47.3	-61.4	-58.2	-53.8	-50.5	-60.9	-52.49
	PNVC-CR	-18.4	-36.7	-51.2	-42.3	-40.7	-29.1	-51.3	-38.53

Table 3. BD-Rate (%) comparison. The anchor is DCVC-DC. All frames with intra-period = -1. **Red** indicates the best performance.

Metric	Method	HEVC_B	HEVC_C	HEVC_D	HEVC_E	MCL-JCV	USTC-TD	UVG	Average
LPIPS	DCVC-FM	-25.1	-27.8	-25.5	-52.9	-11.6	5.3	-35.5	-24.73
	DCVC-RT	-23.7	-16.5	-12.6	-48.2	-8.6	6.3	-36.7	-20.00
	PNVC-CR	-88.2	-69.0	-62.7	-82.4	-87.3	-79.6	-89.0	**-79.74**
DISTS	DCVC-FM	-35.1	-29.6	-25.3	-51.8	-13.1	4.0	-25.0	-25.13
	DCVC-RT	-39.0	-26.5	-16.8	-52.3	-15.6	0.2	-31.5	-25.93
	PNVC-CR	-82.8	-66.2	-49.3	-77.5	-75.6	-77.2	-80.8	**-72.77**
YUV PSNR	DCVC-FM	-26.1	-25.3	-23.7	-66.4	-10.7	-6.7	-27.0	-26.56
	DCVC-RT	-18.1	0.1	11.1	-58.6	-4.3	1.3	-24.9	-13.34
	PNVC-C-Base	-31.8	-27.0	-19.6	-63.1	-16.8	-19.4	-32.4	**-30.01**
	PNVC-CR	-18.6	-23.1	-15.6	-49.8	-8.4	-9.8	-16.2	-20.21
Y PSNR	DCVC-FM	-12.5	-14.0	-11.9	-52.9	-1.2	10.1	-15.3	-13.96
	DCVC-RT	-0.3	19.5	33.4	-42.8	9.3	19.6	-9.7	4.14
	PNVC-C-Base	-21.7	-17.9	-8.6	-50.6	-9.4	-4.4	-23.9	**-19.50**
	PNVC-CR	-9.5	-16.4	-7.8	-36.7	-3.5	0.8	-6.3	-11.34
U PSNR	DCVC-FM	-74.0	-72.3	-76.6	-96.5	-50.0	-53.6	-68.7	-70.24
	DCVC-RT	-76.1	-71.2	-69.6	-94.4	-60.1	-55.0	-74.7	**-71.59**
	PNVC-C-Base	-68.6	-65.2	-63.7	-94.6	-47.1	-57.3	-64.4	-65.84
	PNVC-CR	-53.6	-53.2	-49.6	-89.3	-28.1	-36.8	-50.1	-51.53
V PSNR	DCVC-FM	-66.4	-65.9	-66.5	-81.3	-40.1	-43.6	-60.7	-60.64
	DCVC-RT	-65.4	-59.7	-58.6	-91.6	-48.5	-41.8	-67.3	**-61.84**
	PNVC-C-Base	-63.5	-59.0	-57.1	-77.4	-41.1	-52.6	-59.0	-58.53
	PNVC-CR	-48.5	-46.1	-43.6	-85.6	-26.5	-36.2	-46.8	-47.61

effectively avoiding GAN-induced quality inversions.

Q: Does the gated rank-chain update mechanism lead to sparse training signals or biased learning in the early stages?

A: No, the gradients are neither sparse nor biased. To mitigate the risk of early-stage sparsity and imbalance, we employ a "warm-up" strategy where the encoder is frozen and only the critic is trained for the first two epochs. As shown in Figure 2, this allows the discriminator to quickly acquire correct ranking capabilities, with the ranking satisfaction rate rapidly climbing and stabilizing at **75%–80%** for Rc-GAN (compared to roughly 60% for an unconstrained WGAN). Consequently, the vast majority of samples provide valid gradients, proving that updates are not sparse. Furthermore, discarding the remaining $\sim$20% of samples acts as a crucial "correction mechanism" rather than a loss of data. Filtering out incorrect discriminator judgments is far better than forcing the model to learn from misleading gradients. Finally, thanks to the robust codec pre-training and this critic warm-up phase, performance fluctuations across random seeds remain negligible ($< 1\%$).

3.3. Alignment with the Human Visual System

Q: Why does explicitly decoupling luminance and chrominance improve perceptual metrics (e.g., LPIPS, DISTS) if these metrics were derived from image classifier embeddings rather than explicitly modeled after the HVS?

A: Although standard perceptual metrics are fundamentally de-

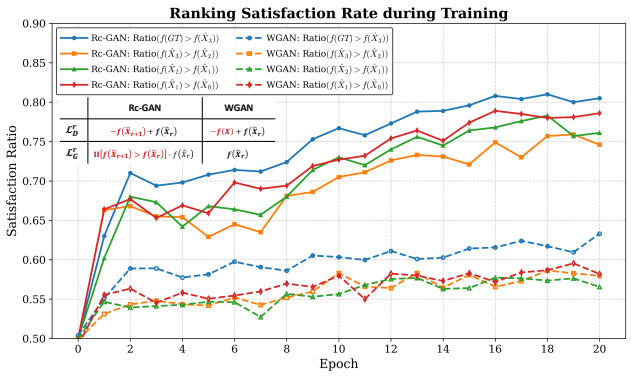

Figure 2. Rc-GAN stabilizes at 75-80% satisfaction rate (vs. WGAN ∼60%), indicating non-sparse gradients. The ∼20% rejection acts as error correction to prevent misleading gradients. A cold-start strategy ensures correct early guidance.

rived from deep feature embeddings, they are strictly calibrated to match human perception. For instance, as validated in Table 4 of the original LPIPS paper [3], distances computed from these pre-trained deep features achieve high consistency with human visual judgments. Therefore, since our explicit Y/UV separation design is directly motivated by the human visual system (HVS), it naturally produces reconstructions that resonate better with human perception, which in turn is inherently reflected as improved scores on these human-aligned metrics.

Furthermore, our decoupled architectural design is supported by a preliminary empirical study. We trained a standard RGB I/O neural video codec (DCVC-DC) using distortion losses calculated in three different color spaces: RGB, YUV, and CIELAB (Lab). We consistently observed a strict hierarchy in perceptual performance: the model optimized for Lab distortion outperformed the YUV-optimized variant on LPIPS and DISTS, which in turn surpassed the RGB-optimized one. This toy experiment empirically confirmed that aligning the optimization domain closer to HVS characteristics intrinsically boosts performance on perceptual metrics, solidly justifying our color-separated coding framework.

References

[1] Haichuan Ma, Dong Liu, and Feng Wu. Rectified Wasserstein generative adversarial networks for perceptual image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3648–3663, 2022. 2

[2] Gary J Sullivan and Jens-Rainer Ohm. Meeting report of the fourth meeting of the joint collaborative team on video coding. In *ITU-T/ISO/IEC JCT-VC Document JCTVC-D500*, 2011. 1

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