ECVC: Exploiting Non-Local Correlations in Multiple Frames for Contextual Video Compression

Wei Jiang, Junru Li, Kai Zhang, Li Zhang✉
Bytedance
{jiangwei.lvc, lijunru, zhangkai.video, lizhang.idm}@bytedance.com

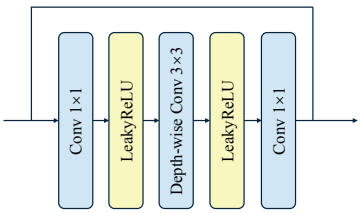

Figure 1. Architecture of the depth-wise Res Block.

A. Network Structure

Our ECVC is based on DCVC-DC [4], but focuses more on exploiting non-local correlations in multiple frames to boost the rate-distortion performance. Here we describe implementation details in the introduced components to DCVC-DC.

A.1. Embedding Layers and Depth-wise Res Block

In the multi-head linear cross attention layer, the depth-wise res block are employed for embedding and point-wise interactions. The structure of the depth-wise res block is depicted in Figure 1.

A.2. Multi-Scale Refine Module

The architecture of multi-scale refine module is depicted in Figure 2.

A.3. Channel numbers

The channel numbers of multi-scale features are $\{d_f^0, d_f^1, d_f^2\} = \{48, 64, 96\}$.

The channel numbers of mid-features during encoding are $\{d_e^0, d_e^1, d_e^2\} = \{3, 64, 96\}$.

The channel numbers of mid-features during decoding are $\{d_d^0, d_d^1, d_d^2\} = \{32, 64, 96\}$.

B. ECVC-FM

We introduce ECVC-FM, a new variant of ECVC that adopts DCVC-FM [5] as its backbone while incorporating key techniques from ECVC. Additionally, we modify the channel dimensions of temporal contexts, resulting in a model with 53.05M parameters.

C. Test Settings

To conduct comparisons, we compare the LVCs and traditional codecs in both RGB color space. BT.601 is employed to convert the frames in YUV color space to RGB color space. To obtain the RD data of VTM-13.2 LDB [2], following commands are employed:

```
EncoderApp
-c encoder_lowdelay_vtm.cfg
--InputFile={input file name}
--BitstreamFile={bitstream file name}
--DecodingRefreshType=2
--InputBitDepth=8
--OutputBitDepth=8
--OutputBitDepthC=8
--InputChromaFormat=444
--FrameRate={frame rate}
--FramesToBeEncoded={frame number}
--SourceWidth={width}
--SourceHeight={height}
--IntraPeriod={IP}
--QP={qp}
--Level=6.2
```

For DCVC-DC and DCVC-FM [5], we use the official weights and code to evaluate the performance.

D. Rate-Distortion Results

We provide the peroformances of our baseline model, reproduced DCVC-DC* in Table 1, Table 2, Table 3 and Table 4. The bpp-PSNR curves are presented in Figure 3, Figure 4 and Figure 5. The bpp-MS-SSIM curves are presented in Figure 6.

References

[1] Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *ITU-T SG16 Q*, 6, 2001. 2, 3

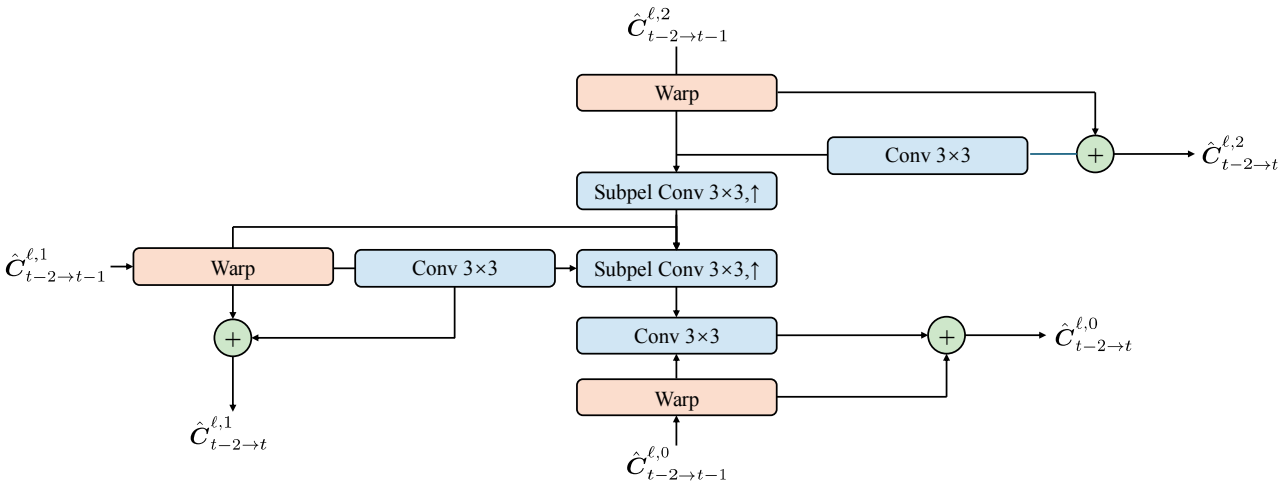

Figure 2. Architecture of multi-scale refine module. "Subpel Conv" means sub pixel convolution.

Method	Venue	BD-Rate (%) w.r.t. VTM-13.2 LDB [2]						
		HEVC B	HEVC C	HEVC D	HEVC E	UVG [6]	MCL-JCV [8]	Average
DCVC-TCM [7]	TMM'22	+28.5	+60.5	+27.8	+67.3	+17.1	+30.6	+38.6
DCVC-HEM [3]	ACMMM'22	−5.1	+15.0	−8.9	+7.1	−18.2	−6.4	−2.8
DCVC-DC [4]	CVPR'23	−17.4	−9.8	−29.0	−26.0	−30.0	−20.0	−22.0
DCVC-FM [5]	CVPR'24	−12.5	−10.3	−26.5	−26.9	−24.0	−12.7	−18.8
DCVC-DC*	Reproduced	−20.5	−6.8	−27.1	−12.1	−29.0	−19.3	−19.1
ECVC	Ours	**−28.3**	**−19.6**	**−36.7**	**−27.1**	**−37.6**	**−26.3**	**−29.3**
ECVC-FM	Ours	**−30.9**	**−22.9**	**−40.7**	**−23.4**	**−42.5**	**−33.6**	**−32.3**

[1] The quality indexes of DCVC-FM are set to match the bit-rate range of DCVC-DC.
[2] Please note that the ECVC is based on our reproduced DCVC-DC* since training scripts of DCVC series are not open-sourced.

Table 1. BD-Rate (%) [1] comparison for PSNR (dB). The anchor is **VTM-13.2 LDB**. **The Intra Period is 32 with 96 frames**.

[2] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J Sullivan, and Jens-Rainer Ohm. Overview of the versatile video coding (vvc) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021. 1, 2, 3

[3] Jiahao Li, Bin Li, and Yan Lu. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1503–1511, 2022. 2, 3

[4] 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, 3

[5] 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, 2, 3

[6] 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. 2, 3

[7] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*, 25:7311–7322, 2022. 2, 3

[8] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. Mcl-jcv: a jnd-based h. 264/avc video quality assessment dataset. In *IEEE International Conference on Image Processing*, pages 1509–1513. IEEE, 2016. 2, 3

[9] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, pages 1398–1402. IEEE, 2003. 3

Method	Venue	BD-Rate (%) w.r.t. VTM-13.2 LDB [2]						
		HEVC B	HEVC C	HEVC D	HEVC E	UVG [6]	MCL-JCV [8]	Average
DCVC-TCM [7]	TMM'22	−20.5	−21.7	−36.2	−20.5	−6.0	−18.6	−20.6
DCVC-HEM [3]	ACMMM'22	−47.4	−43.3	−55.5	−52.4	−32.7	−44.0	−45.9
DCVC-DC [4]	CVPR'23	−53.0	−54.6	−63.4	−60.7	−36.7	−49.1	−52.9
DCVC-DC*	Reproduced	−53.5	−52.8	−61.6	−48.8	−39.1	−52.2	−51.3
ECVC	Ours	**−57.7**	**−58.2**	**−65.6**	**−60.5**	**−42.7**	**−54.9**	**−56.6**

[1] The MS-SSIM [9] optimized weights of DCVC-FM are not open-sourced.

Table 2. BD-Rate (%) [1] comparison for MS-SSIM [9]. The anchor is **VTM-13.2 LDB**. **The intra period is 32 with 96 frames**.

Method	Venue	BD-Rate (%) w.r.t. VTM-13.2 LDB [2]						
		HEVC B	HEVC C	HEVC D	HEVC E	UVG [6]	MCL-JCV [8]	Average
DCVC-TCM [7]	TMM'22	+55.4	+97.4	+50.0	+214.2	+60.4	+50.7	+88.0
DCVC-HEM [3]	ACMMM'22	+3.9	+28.4	−1.2	+66.3	+0.5	+1.7	+16.6
DCVC-DC [4]	CVPR'23	−11.0	+0.2	−23.9	−7.8	−21.0	−13.0	−12.8
DCVC-FM [5]	CVPR'24	−11.7	−7.9	−28.2	−25.8	−23.9	−12.3	−18.3
DCVC-DC*	Reproduced	−9.2	+6.9	−20.0	+45.1	−12.5	−11.7	−0.2
ECVC	Ours	**−27.9**	**−18.9**	**−39.0**	**−26.4**	**−38.3**	**−27.7**	**−29.7**
ECVC-FM	Ours	**−30.9**	**−22.9**	**−40.7**	**−23.4**	**−42.5**	**−33.6**	**−32.3**

Table 3. BD-Rate (%) [1] comparison for PSNR (dB). The anchor is **VTM-13.2 LDB**. **The Intra Period is −1 with 96 frames.**

Method	Venue	BD-Rate (%) w.r.t. VTM-13.2 LDB [2]						
		HEVC B	HEVC C	HEVC D	HEVC E	UVG	MCL-JCV	Average
DCVC-TCM [7]	TMM'22	+107.3	+143.5	+99.2	+835.9	+120.6	+63.7	+228.4
DCVC-HEM [3]	ACMMM'22	+22.8	+32.3	+13.4	+236.9	+33.5	+6.7	+57.6
DCVC-DC [4]	CVPR'23	−7.5	+3.4	−12.0	+83.9	−4.5	−12.9	+8.4
DCVC-FM [5]	CVPR'24	−19.9	−17.4	−25.7	−24.5	−22.5	−13.4	−20.6
DCVC-DC*	Reproduced	+11.9	+11.6	−3.6	+446.4	+16.2	−9.8	+78.8
ECVC	Ours	**−33.4**	**−29.5**	**−38.8**	**−23.5**	**−37.5**	**−29.7**	**−32.1**
ECVC-FM	Ours	**−33.5**	**−32.7**	**−41.3**	**−17.3**	**−41.3**	**−34.3**	**−33.4**

Table 4. BD-Rate (%) [1] comparison for PSNR (dB). The anchor is **VTM-13.2 LDB**. **The Intra Period is −1 with All frames.**

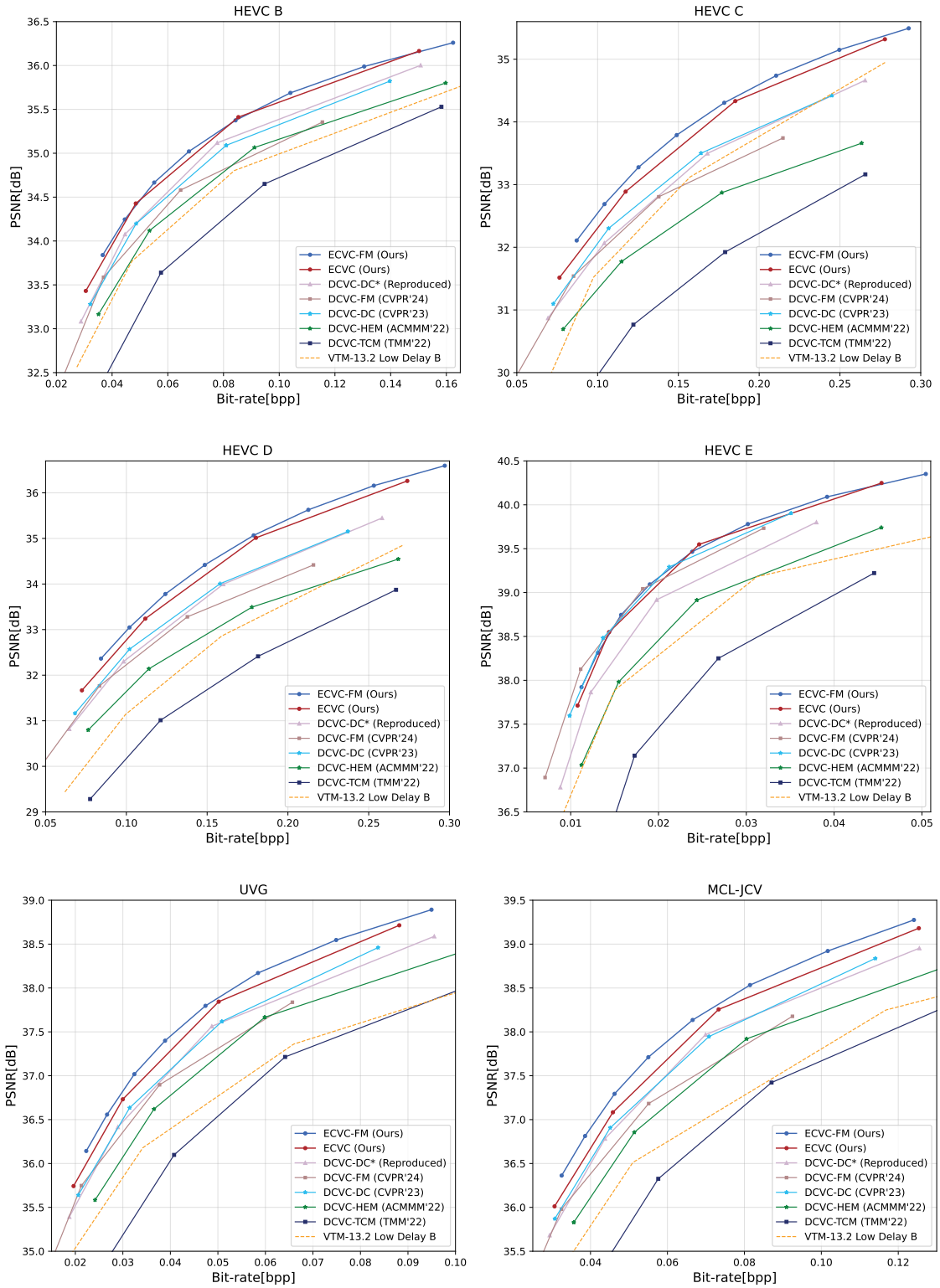

Figure 3. Bpp-PSNR curves on HEVC B, C, D, E, UVG and MCL-JCV dataset. **The intra period is 32 with 96 frames**.

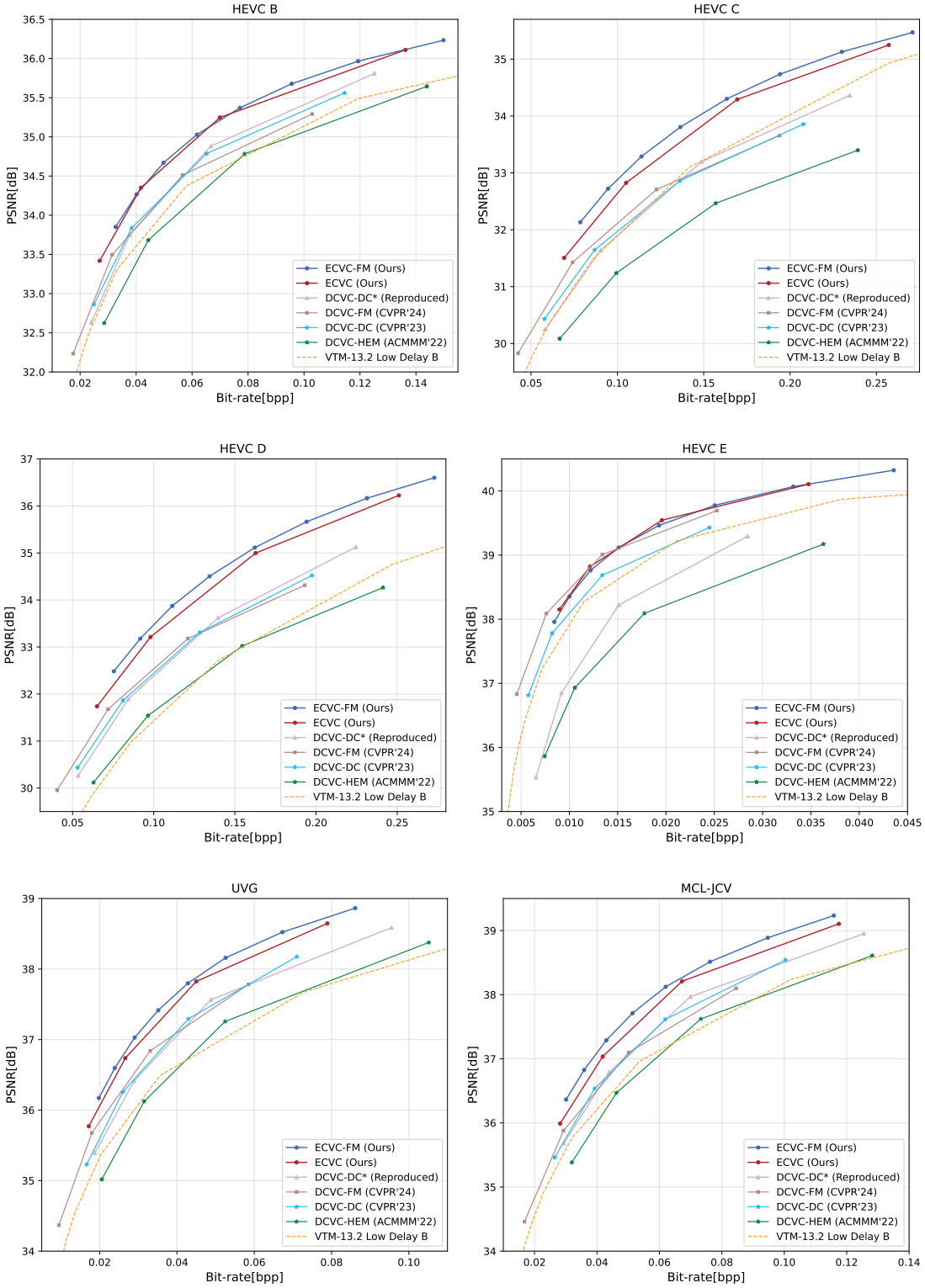

Figure 4. Bpp-PSNR curves on HEVC B, C, D, E, UVG and MCL-JCV dataset. **The intra period is −1 with 96 frames**.

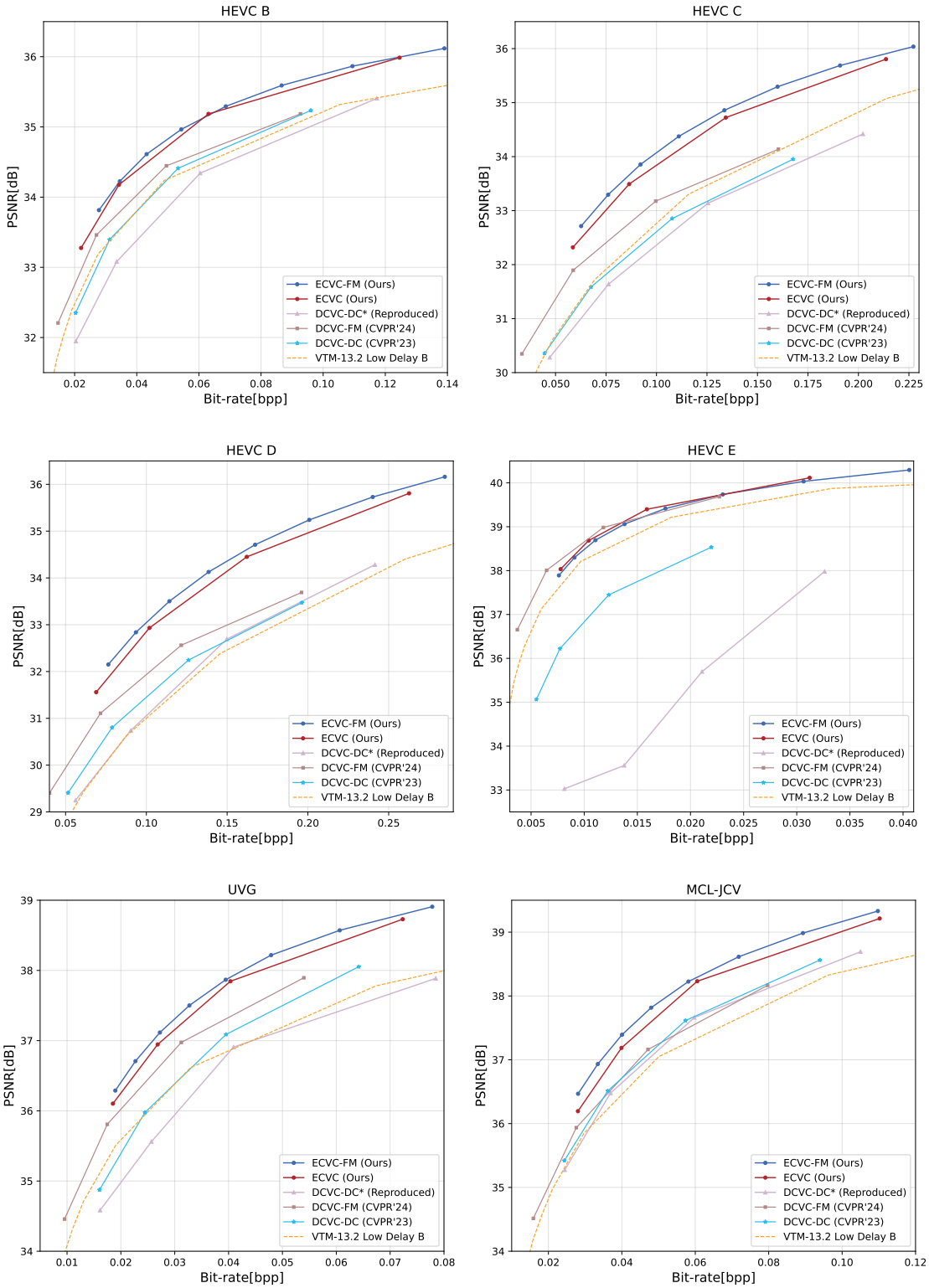

Figure 5. Bpp-PSNR curves on HEVC B, C, D, E, UVG and MCL-JCV dataset. **The intra period is −1 with All frames**.

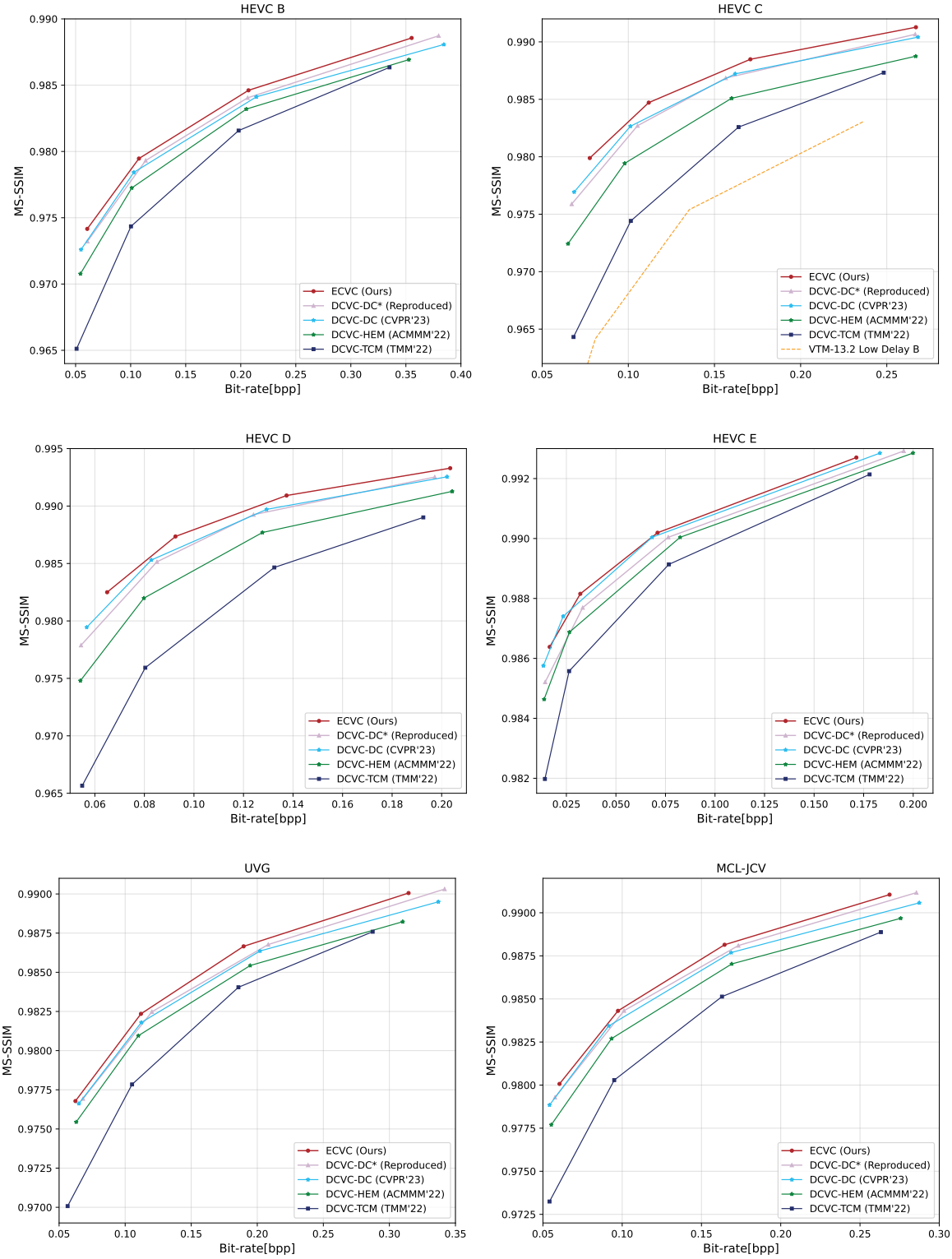

Figure 6. Bpp-MS-SSIM curves on HEVC B, C, D, E, UVG and MCL-JCV dataset. **The intra period is 32 with 96 frames**.