FLAVC: Learned Video Compression with Feature Level Attention Supplementary Material

Chun Zhang[1], Heming Sun[2]*, Jiro Katto[1]
[1]Department of Computer Science and Communication Engineering, Waseda University, Tokyo, Japan
[2]Facalty of Engineering, Yokohama National University, Kanagawa, Japan
{zhangchun@toki.,katto@}waseda.jp, hemingsun@ieee.org

1. Test Settings

In line with common practice, we use the widely-adopted traditional codec HM-16.26 (HEVC) as the baseline for calculating BD-rate across all deep learning frameworks. Additionally, we include results from the more recent traditional codec VTM-13.2 (VVC), which demonstrates stronger competitive performance. The configurations used for encoding are $encoder\_lowdelay\_main\_rext$ for HM-16.26 and $encoder\_lowdelay\_main\_vtm$ for VTM-13.2. Detailed testing commands for both codecs are as follows:

$-c \quad \{config\}$
$--InputFile = inputfilename$
$--SourceWidth = width$
$--SourceHeight = height$
$--InputBitDepth = 8$
$--OutputBitDepth = 8$
$--OutputBitDepthC = 8$
$--InputChromaFormat = 444$
$--FrameRate = framerate$
$--FramesToBeEncoded = 96$
$--IntraPeriod = 32$
$--DecodingRefreshType = 2$
$--QP = qp$
$--Level = 6.2$
$--BitstreamFile = bitstreamfilename$

2. Data Acquisition and Color Spaces

The most consistently used evaluation metric in LVC research is the Rate-Distortion curve. However, test results can vary depending on the data format. While widely used datasets UVG [10], MCL-JCV [13], and HEVC [1] are provided in the YUV420 color space, a standard practice is to convert these video sequences to the BT.601 RGB color space using FFmpeg. In contrast, recent works, such as DCVC-DC [6], have shifted toward using the BT.709 RGB color space, which generally yields better results compared to BT.601. Additionally, some of the latest studies [7] directly evaluate models in the native YUV420 color space to align more closely with traditional codec performance.

In this paper, we primarily present RD-curves in the BT.601 color space, as it allows for a broader comparison across existing LVC models. To provide a broader perspective, the supplementary material includes comparisons across the BT.601, BT.709, and YUV420 color spaces. As the majority of LVC papers did not have their source code or trained models readily available at the time, we have to reserve to extract data directly from their papers. These extracted results from different studies can be directly compared within respective color space as they all follow the same 32-frame GoP configuration for a total of 96 frames in P-frame compression settings. Table 1 listed the different results available in each paper.

3. BT.601 color space results

To evaluate the performance of video compression models, four widely used datasets—UVG, MCL-JCV, HEVC Class B, and Class C—are provided in the YUV420 format. The standard practice in LVC research involves converting these video sequences into individual frames. Most studies follow the default BT.601 RGB conversion protocol in FFmpeg for consistency.

In the main paper, we adhered to this established practice, presenting the performance of our proposed model tested in the BT.601 RGB color space. This approach enables a direct comparison with a wide range of previous state-of-the-art (SOTA) works. Larger versions of these comparison results are provided in Fig. 1 and Fig. 2 for additional clarity and detail.

*Heming Sun is the corresponding author.

Dataset	UVG		MCL-JCV		HEVC-Class B		HEVC-Class C	
Models	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM
CANF-VC[5]	601 709[6]	601 709[6]	601 709[6]	601 709[6]	601 709[6]	601 709[6]	-	-
VCT[9]	601	601	601	601	601[8]	-	601[8]	-
DCVC-TCM[11]	601[12] 709[6]	601[12] 709[6]	601 709[6]	601 709[6]	601 709[6]	601 709[6]	601[12] 709[6]	601[12] 709[6]
DCVC-DC[6]	601[3] 709 YUV420	709	601[3] 709 YUV420	709	601[3] 709 YUV420	709	709 YUV420	709
ST-XCT[4]	601	-	601	-	601	-	601	-
DHVC[8]	601	601	601	601	601	601	601	601
CANF-VC++[2]	601	-	601	-	601	-	-	-
SDD[12]	601	601	601	601	601	601	601	601
MCRT[3]	601	-	601	-	601	-	-	-
DCVC-FM[7]	YUV420	-	YUV420	-	YUV420	-	YUV420	-
FLAVC(ours)	601 709 YUV420	601 709 YUV420	601 709 YUV420	601 709 YUV420	601 709 YUV420	601 709 YUV420	601 709 YUV420	601 709 YUV420

Table 1. The available data format from each paper. 601, 709 and YUV420 stands for different color space. If a citation follows a specific color space, it indicates that the results for that format were obtained from another cited study rather than directly reported by the authors.

4. BT.709 color space results

While traditional BT.601 color space conversion has been the standard in LVC research, the more modern BT.709 color space is specifically designed for high-definition (HD) videos with a spatial resolution of 1920×1080. Consequently, an increasing number of LVC studies, such as [6, 7], have shifted their testing to target the BT.709 RGB color space. Generally, models exhibit better performance when evaluated in BT.709 compared to BT.601. We present our results in the BT.709 RGB color space in Fig. 3 and Fig. 4. Usually, no additional optimization is needed during training as both data formats are in RGB color space.

5. YUV420 color space results

Traditional codecs are inherently designed to compress video data in the YUV420 color space. However, this characteristic was not widely adopted in LVC studies until recently. DCVC-DC [6] was the first to evaluate and present their model performance directly in the YUV420 color space, followed by DCVC-FM [7], which further embraced this approach. In this work, we also evaluated FLAVC in the YUV420 color space, with results presented in Fig. 5. Unlike DCVC-FM, however, we did not incorporate YUV-specific loss functions [7] during training, leading to suboptimal performance in this format. In future studies, we aim to optimize our method specifically for the YUV420 color space to improve its effectiveness.

6. Qualitative Studies

In this section, we provide additional visual comparisons across various compression methods. As illustrated in Fig. 6, FLAVC demonstrates consistently superior visual fidelity compared to the other methods evaluated. These tests were conducted on all four datasets in the BT.601 color space. More descriptions for the visualization in caption.

References

[1] Frank Bossen. Common test conditions and software reference configurations. 2010. 1

[2] Peng-Yu Chen and Wen-Hsiao Peng. CANF-VC++: enhancing conditional augmented normalizing flows for video compression with advanced techniques. *CoRR*, abs/2309.05382, 2023. 2

[3] Yi-Hsin Chen, Hongfu Xie, Cheng-Wei Chen, Zong-Lin Gao, Martin Benjak, Wen-Hsiao Peng, and Jörn Ostermann. Maskcrt: Masked conditional residual transformer for learned video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[4] Zhenghao Chen, Lucas Relic, Roberto Azevedo, Yang Zhang, Markus Gross, Dong Xu, Luping Zhou, and Christopher Schroers. Neural video compression with spatio-temporal cross-covariance transformers. *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. 2

[5] Yung-Han Ho, Chih-Peng Chang, Peng-Yu Chen, Alessandro Gnutti, and Wen-Hsiao Peng. CANF-VC: conditional augmented normalizing flows for video compression. In *Computer Vision - ECCV 2022 - 17th European Conference,*

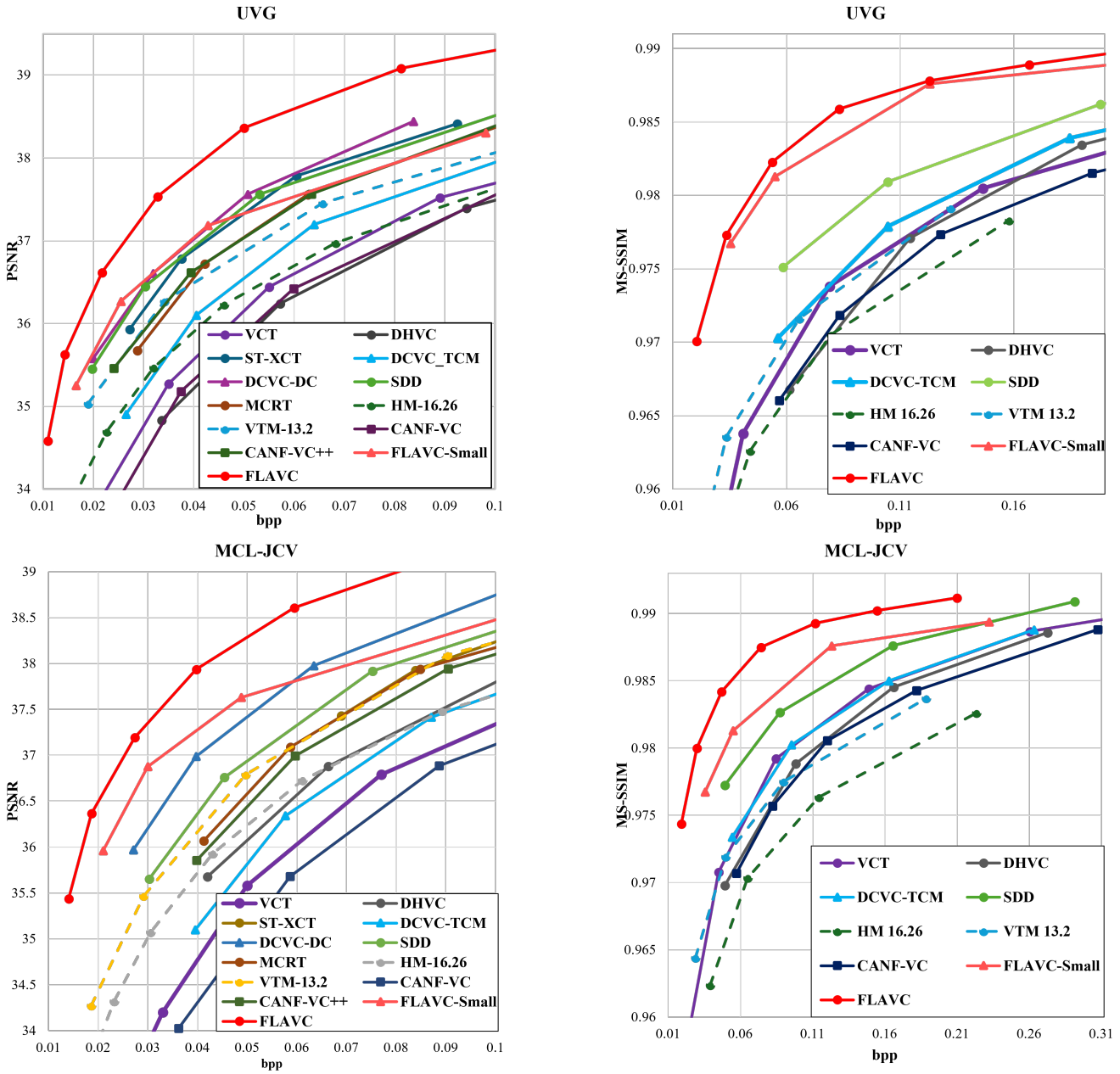

Figure 1. Overall Rate-Distortion performance comparison on UVG and MCL-JCV datasets: **Left** : PSNR, **Right** : MS-SSIM. The two traditional methods are marked with dashed lines. RD-curve in BT.601 color space.

Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XVI, pages 207–223. Springer, 2022. 2

[6] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with diverse contexts. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22616–22626, 2023. 1, 2

[7] Jiahao Li, Bin Li, and Yan Lu. Neural video compression with feature modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26099–26108. IEEE, 2024. 1, 2

[8] Ming-Tse Lu, Zhihao Duan, Fengqing Maggie Zhu, and Zhan Ma. Deep hierarchical video compression. In *AAAI Conference on Artificial Intelligence*, 2023. 2

[9] Fabian Mentzer, George Toderici, David C. Minnen, Sung Jin Hwang, Sergi Caelles, Mario Lucic, and Eirikur Agustsson. Vct: A video compression transformer. *ArXiv*, abs/2206.07307, 2022. 2

[10] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. Uvg dataset: 50/120fps 4k sequences for video codec analysis and

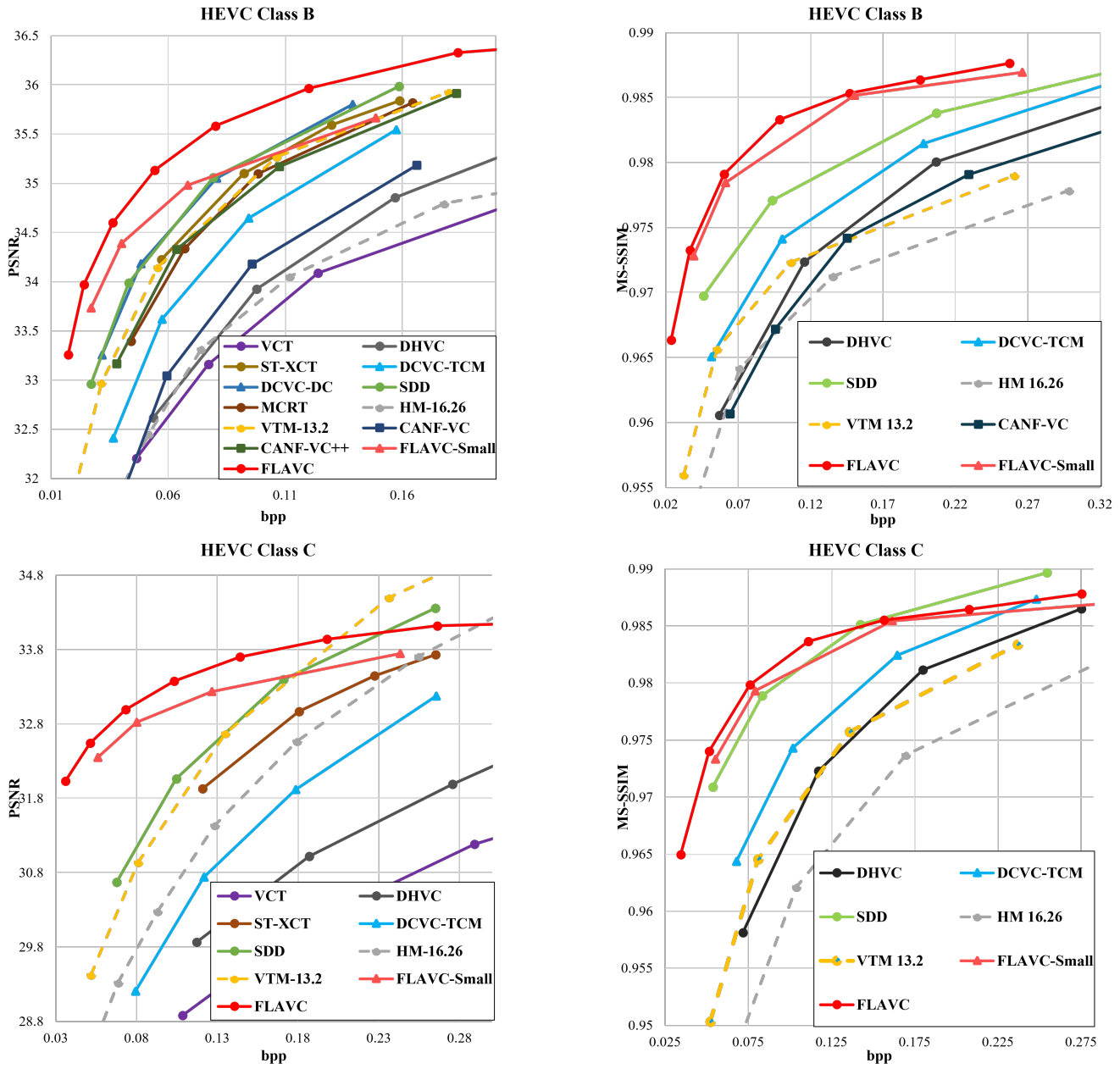

Figure 2. Overall Rate-Distortion performance comparison on HEVC Class B and HEVC Class C datasets: **Left** : PSNR, **Right** : MS-SSIM. The two traditional methods are marked with dashed lines. RD-curve in BT.601 color space.

development. *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020. 1

[11] 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, 2021. 2

[12] Xihua Sheng, Li Li, Dong Liu, and Houqiang Li. Spatial decomposition and temporal fusion based inter prediction for learned video compression. *IEEE Transactions on Circuits and Systems for Video Technology*, 34:6460–6473, 2024. 2

[13] 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. *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1509–1513, 2016. 1

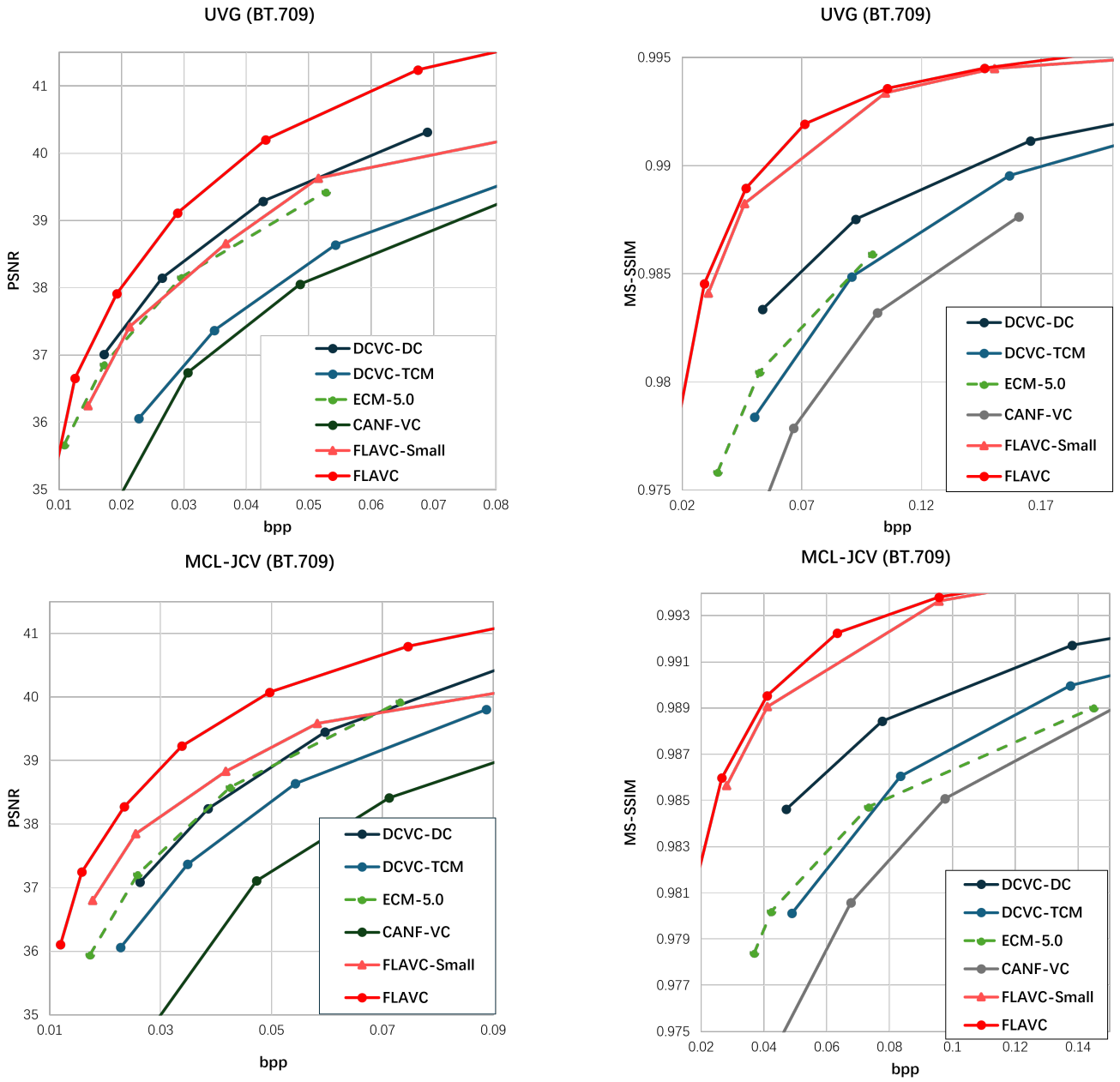

Figure 3. Overall Rate-Distortion performance comparison on UVG and MCL-JCV datasets: **Left** : PSNR, **Right** : MS-SSIM. The traditional method is marked with dashed lines. RD-curve in BT.709 color space.

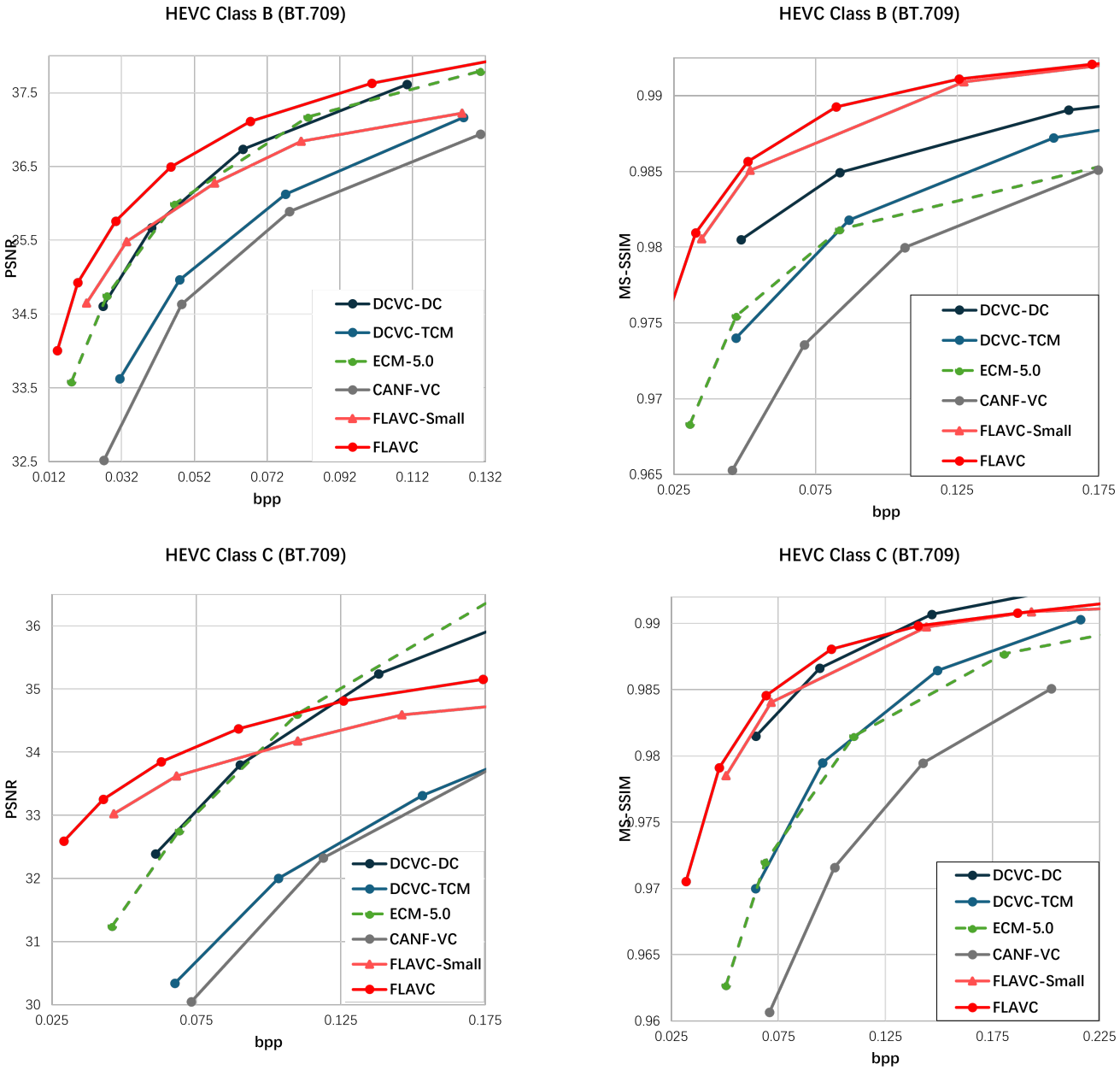

Figure 4. Overall Rate-Distortion performance comparison on HEVC Class B and HEVC Class C datasets: **Left** : PSNR, **Right** : MS-SSIM. The traditional method is marked with dashed lines. RD-curve in BT.709 color space.

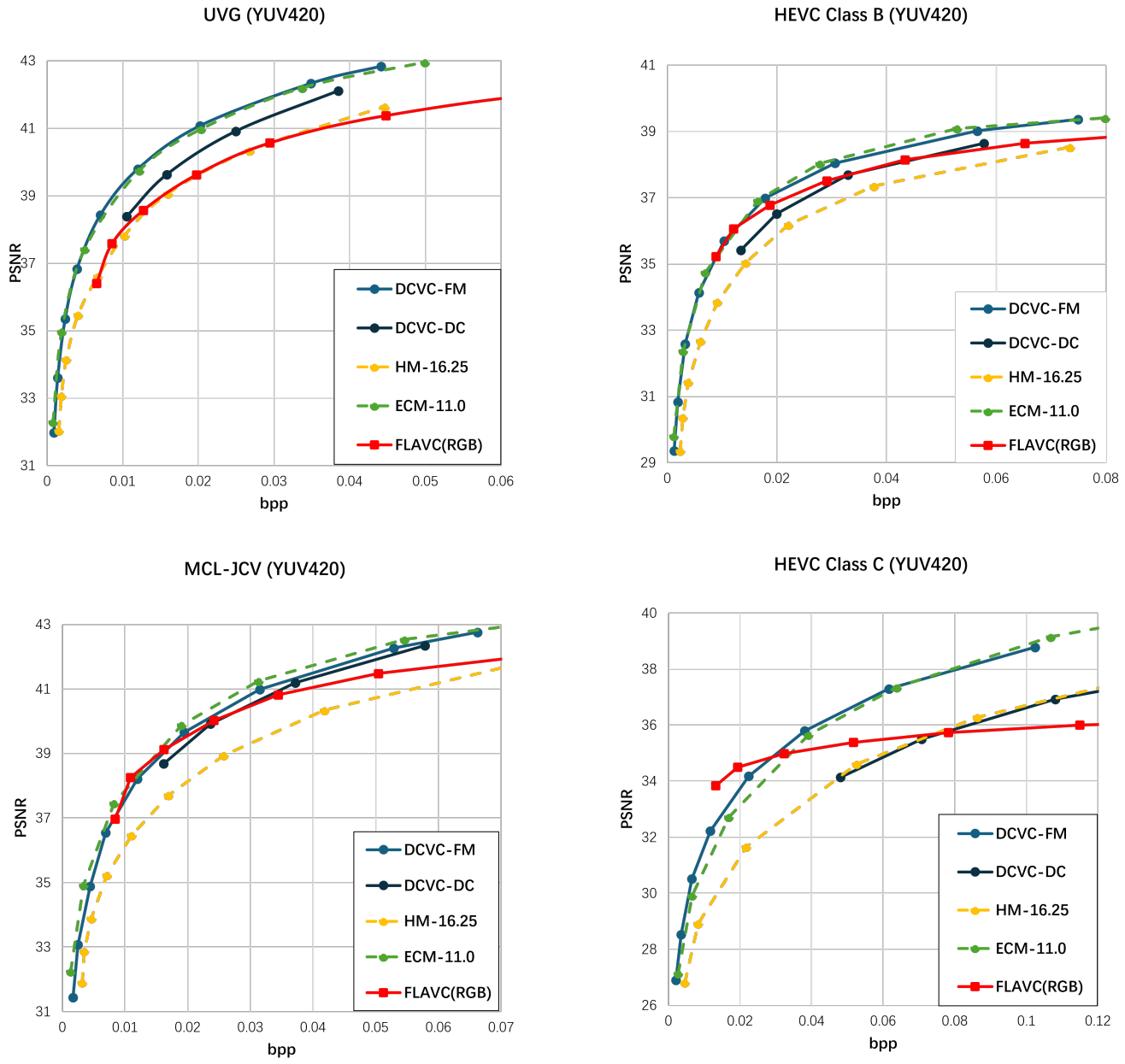

Figure 5. Overall Rate-Distortion (RD) performance comparison on the UVG, MCL-JCV, HEVC Class B, and HEVC Class C datasets in the YUV420 color space. Traditional methods are represented by dashed lines for clarity. Notably, our proposed method is the only approach in this comparison that has **not** been specifically optimized for the YUV420 format, underscoring the potential for further improvements through targeted optimization.

Figure 6. Comparison of different compression models visualized. On the HoneyBee video, FLAVC excels in reconstructing fine details, particularly on the small moving bee. In ShakeNDry, our model effectively preserves the intricate details of water droplets despite their complex motion and noise. In videoSRC16, FLAVC accurately retains the metal structure's pattern, as well as small details like the car's license plate and tail light. For ParkScene and RaceHorseC, our model successfully preserves the texture of the woman's shirt and the horse's mane, even though the motion in these scenes poses significant challenges.