HyTIP: Hybrid Temporal Information Propagation for Masked Conditional Residual Video Coding
Supplementary Materials

Yi-Hsin Chen[1] Yi-Chen Yao[1] Kuan-Wei Ho[1] Chun-Hung Wu[1] Huu-Tai Phung[1]
Martin Benjak[2] Jörn Ostermann[2] Wen-Hsiao Peng[1]

[1] National Yang Ming Chiao Tung University, Taiwan
[2] Leibniz Universität Hannover, Germany

This supplementary document provides the following additional materials and results to assist with the understanding of our HyTIP.

- Additional results on buffer size in Section A1;
- Results on long-sequence training under BT.709 color space conversion in Section A2;
- Rate-distortion comparisons to state-of-the-arts in Section A3;
- Network architecture in Section A4;
- Training details in Section A5;
- Configurations of HM 16.25 and VTM 17.0 in Section A6;

A1. Additional Results on Buffer Size

Table A1 and Table A2 provide additional results on the buffer size configurations reported in Table 2 of the main paper, using BT.601 and BT.709 color space conversions, respectively. As shown, the conclusions are consistent with those in the main paper, where the hybrid buffering strategy demonstrates strong robustness and efficiency. It achieves better performance than both the explicit and implicit buffering strategies and exhibits greater resilience to buffer size reduction compared to the implicit buffering strategy. According to the results in Table A1 and Table A2, our final design buffers 1 explicit decoded flow map and 0.125 full-resolution feature map (i.e., 2+0.125 full-resolution feature maps) for motion coding, and 1 explicit decoded frame and 2 full-resolution feature maps (i.e., 3+2 full-resolution feature maps) for inter-frame coding to balance coding performance and complexity.

A2. Results on Long-sequence Training under BT.709 Color Space Conversion

Table A3 presents the results of the same study as Table 3 in the main paper, but employs BT.709 color space conversion, following [5, 6], instead of BT.601 for the YUV420 to RGB444 conversion. As shown, the conclusions remain consistent with those in the main paper, where training with longer sequences generally improves all buffering strategies. The explicit buffering strategy shows a smaller performance gain due to its dual constraint on the output frame, while the implicit and hybrid strategies, which propagate implicit features without such constraints, are better able to leverage long-sequence training.

A3. Rate-distortion Comparisons to State-of-the-arts

Fig. A1 and Fig. A2 present the rate-distortion comparisons between our method and state-of-the-art approaches in terms of PSNR-RGB, using BT.601 and BT.709 for color space conversion, respectively. Similarly, Fig. A3 and Fig. A4 show the comparisons in terms of MS-SSIM-RGB under BT.601 and BT.709 color space conversion, respectively. The corresponding BD-rates using BT.709 as the color space conversion are summarized in Table A4 and Table A5.

A4. Network Architecture Details

Fig. A5, Fig. A6, and Fig. A7 provide the network architecture details in Fig. 2 of the main paper. The base motion and inter-frame codec are adapted from [6] while the channel transform module in the inter-frame codec is from [3] and the checkerboard context model in the inter-frame codec is from [7] .

A5. Training Details

Table A6 summarizes our HyTIP training procedure, adapted from [3]. The first six phases follow [3], using explicit temporal reference $\hat{x}_{t-1}$ as the temporal reference in-

formation only in the inter-frame codec. Subsequently, the implicit related module is incorporated for further training.

A6. Configurations of HM 16.25 and VTM 17.0

Following the recommendation from [5], we encode videos in YUV444 format. We use the *encoder_lowdelay_vtm.cfg* of VTM [2] with the following parameters:

–c {config file name}
–InputFile={input file name}
–InputBitDepth=8
–InputChromaFormat=444
–ChromaFormatIDC=444
–InternalBitDepth=10
–OutputBitDepth=8
–DecodingRefreshType=2
–FrameRate={frame rate}
–FrameSkip=0
–SourceWidth={width}
–SourceHeight={height}
–FramesToBeEncoded=96
–Level=4.1
–IntraPeriod=32
–QP={qp}
–BitstreamFile={bitstream file name}
–ReconFile={reconstruction file name}

Similarly, We use the *encoder_lowdelay_main_rext.cfg* of HM [2] with the following parameters:

–c {config file name}
–InputFile={input file name}
–InputBitDepth=8
–InputChromaFormat=444
–ChromaFormatIDC=444
–InternalBitDepth=10
–InternalBitDepthC=10
–OutputBitDepth=8
–OutputBitDepthC=8
–FrameRate={frame rate}
–FrameSkip=0
–SourceWidth={width}
–SourceHeight={height}
–FramesToBeEncoded=96
–Level=4.1
–IntraPeriod=32
–QP={qp}
–BitstreamFile={bitstream file name}
–ReconFile={reconstruction file name}

References

[1] https://vcgit.hhi.fraunhofer.de/jvet/HM/. 4

[2] https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM. 2, 4

[3] Yi-Hsin Chen, Hong-Sheng 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 (TCSVT)*, 34(11):11980–11992, 2024. 1, 4, 9

[4] 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 (ACM MM)*, pages 1503–1511, 2022. 4

[5] 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 (CVPR)*, pages 22616–22626, 2023. 1, 2, 4

[6] 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 (CVPR)*, pages 26099–26108, 2024. 1, 4, 9

[7] Fangzheng Lin, Heming Sun, Jinming Liu, and Jiro Katto. Multistage spatial context models for learned image compression. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023. 1

[8] Guo Lu, Chunlei Cai, Xiaoyun Zhang, Li Chen, Wanli Ouyang, Dong Xu, and Zhiyong Gao. Content adaptive and error propagation aware deep video compression. In *European Conference on Computer Vision (ECCV)*, pages 456–472, 2020. 9

[9] Xihua Sheng, Jiahao Li, Bin Li, Li Li, Dong Liu, and Yan Lu. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia (TMM)*, 25:7311–7322, 2023. 4

Table A1. BD-rate (%) comparison of different buffering strategies with different buffer sizes, using BT.601 for color space conversion. The anchor employs explicit buffering in both motion and inter-frame coding. The values in parentheses, as in Table 2, indicate the number of full-resolution feature maps buffered for coding one input flow map or frame (explicit + implicit).

Motion	Inter	UVG	MCL-JCV	HEVC-B	HEVC-C	HEVC-D	HEVC-E	HEVC-RGB	Average
Explicit (2+0)	Explicit (3+0)	0	0	0	0	0	0	0	0
Implicit (0+4)	Explicit (3+0)	-11.5	-9.5	-12.6	-18.6	-12.6	-11.9	-8.8	-12.2
Implicit (0+2.1875)	Explicit (3+0)	-7.0	-3.6	-8.7	-12.0	-8.4	-9.8	-7.2	-8.1
Implicit (0+2.125)	Explicit (3+0)	-6.5	-3.8	-6.2	-10.4	-6.7	-8.6	-3.0	-6.5
Implicit (0+2.0625)	Explicit (3+0)	-4.0	2.8	-1.9	-2.2	0.8	-2.3	-2.3	-1.3
Hybrid (2+4)	Explicit (3+0)	-15.1	-13.2	-15.4	-23.8	-19.9	-15.8	-9.1	-16.0
Hybrid (2+0.1875)	Explicit (3+0)	-13.6	-9.6	-14.8	-21.4	-18.0	-18.4	-13.1	-15.6
Hybrid (2+0.125)	Explicit (3+0)	-13.0	-11.2	-14.1	-19.1	-16.3	-18.1	-11.0	-14.7
Hybrid (2+0.0625)	Explicit (3+0)	-9.1	-5.4	-9.3	-13.9	-11.7	-13.4	-6.7	-9.9
Hybrid (2+0.125)	Implicit (0+51)	-12.6	-15.5	-20.1	-30.2	-27.8	-20.6	-10.5	-19.6
Hybrid (2+0.125)	Implicit (0+6)	-10.0	-12.6	-15.8	-24.5	-21.9	-13.0	-8.6	-15.2
Hybrid (2+0.125)	Implicit (0+5)	-9.2	-13.0	-14.9	-24.6	-21.4	-14.6	-7.4	-15.0
Hybrid (2+0.125)	Implicit (0+4)	-5.9	-9.0	-13.5	-19.5	-17.6	-1.9	-4.3	-10.2
Hybrid (2+0.125)	Hybrid (3+48)	-17.3	-16.3	-21.0	-30.3	-27.3	-25.7	-15.1	-21.9
Hybrid (2+0.125)	Hybrid (3+3)	-16.5	-15.7	-20.8	-29.1	-26.8	-23.3	-14.7	-21.0
Hybrid (2+0.125)	Hybrid (3+2)	-17.6	-15.9	-20.3	-29.1	-25.9	-26.6	-14.8	-21.5
Hybrid (2+0.125)	Hybrid (3+1)	-12.7	-10.4	-14.8	-23.6	-19.3	-18.9	-9.1	-15.5

Table A2. BD-rate (%) comparison of different buffering strategies with different buffer sizes, using BT.709 for color space conversion. The anchor employs explicit buffering in both motion and inter-frame coding. The values in parentheses, as in Table 2, indicate the number of full-resolution feature maps buffered for coding one input flow map or frame (explicit + implicit).

Motion	Inter	UVG	MCL-JCV	HEVC-B	HEVC-C	HEVC-D	HEVC-E	HEVC-RGB	Average
Explicit (2+0)	Explicit (3+0)	0	0	0	0	0	0	0	0
Implicit (0+4)	Explicit (3+0)	-10.0	-8.1	-10.5	-14.5	-9.0	-10.8	-8.8	-10.2
Implicit (0+2.1875)	Explicit (3+0)	-6.5	-5.9	-8.0	-10.6	-6.0	-9.4	-7.2	-7.7
Implicit (0+2.125)	Explicit (3+0)	-5.5	-3.6	-5.8	-9.3	-5.1	-7.2	-3.0	-5.6
Implicit (0+2.0625)	Explicit (3+0)	-3.7	-0.1	-1.6	-2.2	1.1	-1.5	-2.3	-1.5
Hybrid (2+4)	Explicit (3+0)	-13.3	-11.5	-12.8	-19.1	-15.0	-14.4	-9.1	-13.6
Hybrid (2+0.1875)	Explicit (3+0)	-12.9	-11.4	-13.6	-18.3	-14.6	-17.0	-13.1	-14.4
Hybrid (2+0.125)	Explicit (3+0)	-12.2	-10.3	-12.4	-15.5	-12.2	-18.7	-11.0	-13.2
Hybrid (2+0.0625)	Explicit (3+0)	-8.7	-6.7	-8.3	-10.7	-7.8	-13.6	-6.7	-8.9
Hybrid (2+0.125)	Implicit (0+51)	-12.7	-14.0	-17.1	-26.2	-23.4	-18.9	-10.5	-17.5
Hybrid (2+0.125)	Implicit (0+6)	-9.4	-10.7	-13.0	-20.3	-17.6	-12.3	-8.6	-13.1
Hybrid (2+0.125)	Implicit (0+5)	-9.9	-12.0	-12.9	-21.2	-17.7	-14.1	-7.4	-13.6
Hybrid (2+0.125)	Implicit (0+4)	-6.9	-7.7	-11.3	-16.5	-14.8	-2.6	-4.3	-9.2
Hybrid (2+0.125)	Hybrid (3+48)	-17.7	-15.6	-18.8	-26.6	-24.0	-25.4	-15.1	-20.5
Hybrid (2+0.125)	Hybrid (3+3)	-16.6	-14.3	-18.6	-25.4	-23.3	-23.5	-14.7	-19.5
Hybrid (2+0.125)	Hybrid (3+2)	-17.5	-14.6	-18.3	-25.3	-22.8	-26.8	-14.8	-20.0
Hybrid (2+0.125)	Hybrid (3+1)	-13.0	-9.2	-13.2	-19.8	-16.2	-19.2	-9.1	-14.2

Table A3. BD-rate (%) comparison of longer sequence training impact on three buffering strategies for inter-frame coding, using BT.709 for color space conversion. The anchor is the variant employing explicit buffering in both motion and inter-frame coding. The values in parentheses, as in Table 2, indicate the number of full-resolution feature maps buffered for coding one input flow map or frame (explicit + implicit).

Motion	Inter	# Frame	UVG	MCL-JCV	HEVC-B	HEVC-C	HEVC-D	HEVC-E	HEVC-RGB	Average
Hybrid (2.125)	Explicit (3)	5	-12.2	-10.3	-12.4	-15.5	-12.2	-18.7	-11.0	-13.2
		10	-18.5	-15.2	-14.5	-17.6	-14.0	-19.9	-13.2	-16.1
Hybrid (2.125)	Implicit (5)	5	-9.9	-12.0	-12.9	-21.2	-17.7	-14.1	-7.4	-13.6
		10	-20.6	-20.4	-18.3	-25.3	-19.3	-12.8	-12.6	-18.5
Hybrid (2.125)	Hybrid (2)	5	-17.5	-14.6	-18.3	-25.3	-22.8	-26.8	-14.8	-20.0
		10	-24.5	-21.7	-22.5	-29.5	-25.8	-28.8	-20.3	-24.7

Table A4. BD-rate (%) comparison between our HyTIP and the state-of-the-art methods in terms of PSNR-RGB, using BT.709 for color space conversion. The anchor is VTM 17.0. Negative BD-rates suggest bitrate savings.

	UVG	MCL-JCV	HEVC-B	HEVC-C	HEVC-D	HEVC-E	HEVC-RGB	Average
VTM [2]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HM [1]	26.0	35.7	31.9	29.9	29.8	31.9	29.6	30.7
MaskCRT [3]	6.4	19.1	9.2	29.0	4.7	26.1	-7.2	12.5
DCVC-TCM [9]	35.0	39.3	34.1	62.5	25.5	70.4	19.2	40.9
DCVC-HEM [4]	-6.2	-0.7	-0.4	19.4	-5.3	9.3	-14.2	0.3
DCVC-DC [5]	-24.0	-18.4	-16.4	-11.6	-29.0	-25.4	-32.1	-22.4
DCVC-FM [6]	-24.0	-15.0	-16.5	-14.9	-31.1	-32.0	-23.1	-22.4
HyTIP (Ours)	-22.3	-9.7	-16.3	-6.8	-25.6	-13.9	-27.5	-17.4

Table A5. BD-rate (%) comparison between our HyTIP and the state-of-the-art methods in terms of MS-SSIM-RGB, using BT.709 for color space conversion. The anchor is VTM 17.0. Negative BD-rates suggest bitrate savings.

	UVG	MCL-JCV	HEVC-B	HEVC-C	HEVC-D	HEVC-E	HEVC-RGB	Average
VTM [2]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
HM [1]	21.4	31.5	28.9	29.0	29.7	29.6	26.1	28.0
MaskCRT [3]	-23.0	-30.8	-38.6	-29.0	-43.4	-31.9	-43.8	-34.4
DCVC-TCM [9]	-9.5	-22.5	-26.5	-17.4	-33.8	-18.7	-27.7	-22.3
DCVC-HEM [4]	-28.8	-43.3	-47.7	-40.4	-52.8	-53.1	-46.3	-44.6
DCVC-DC [5]	-36.4	-50.4	-55.7	-51.4	-61.2	-66.0	-56.4	-53.9
HyTIP (Ours)	-38.9	-50.5	-56.4	-50.2	-59.5	-62.6	-58.5	-53.8

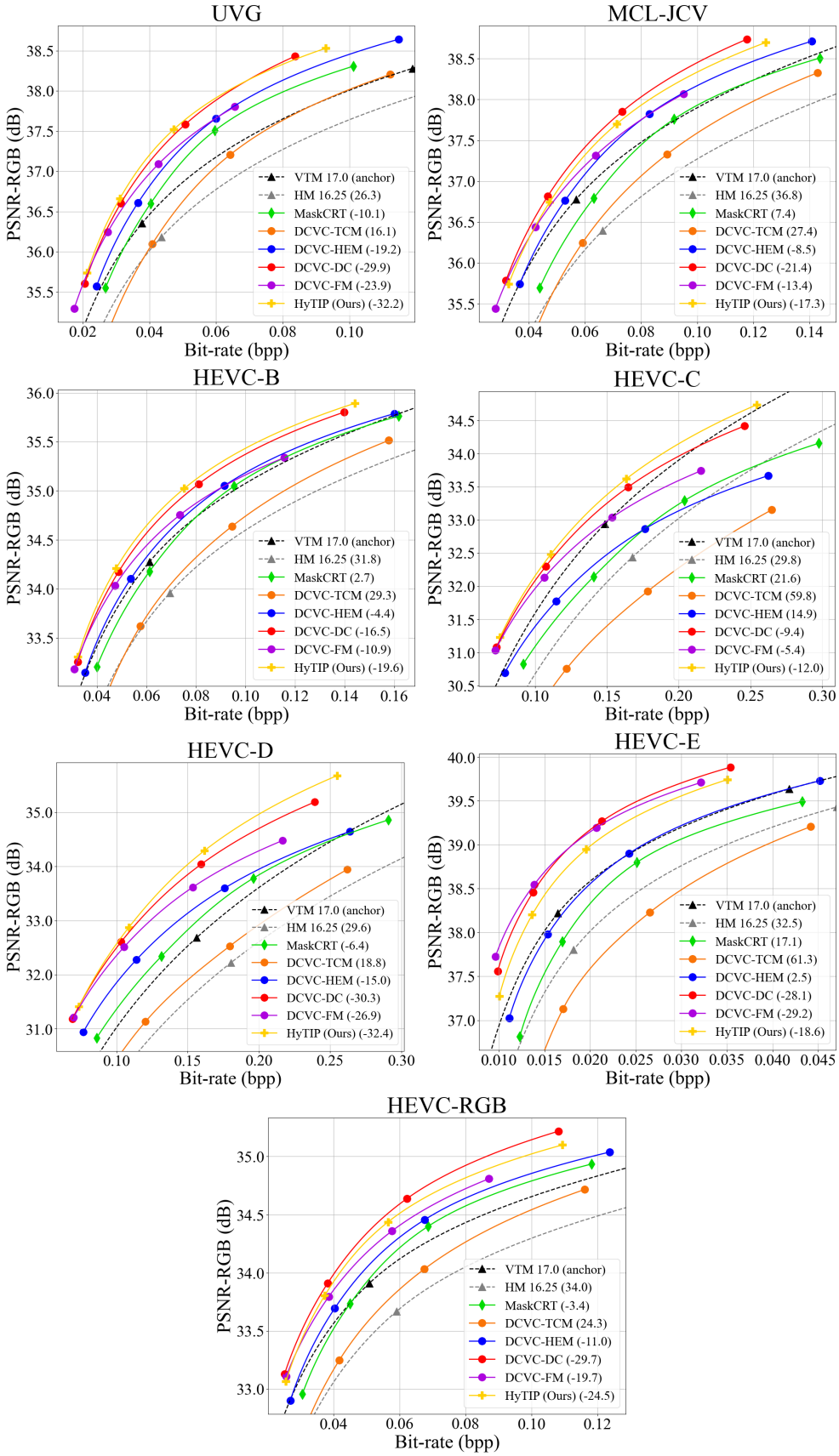

Figure A1. Rate-distortion comparison with state-of-the-art methods in terms of PSNR-RGB, using BT.601 for color space conversion. The values in parentheses represent BD-rates, with VTM 17.0 serving as the anchor.

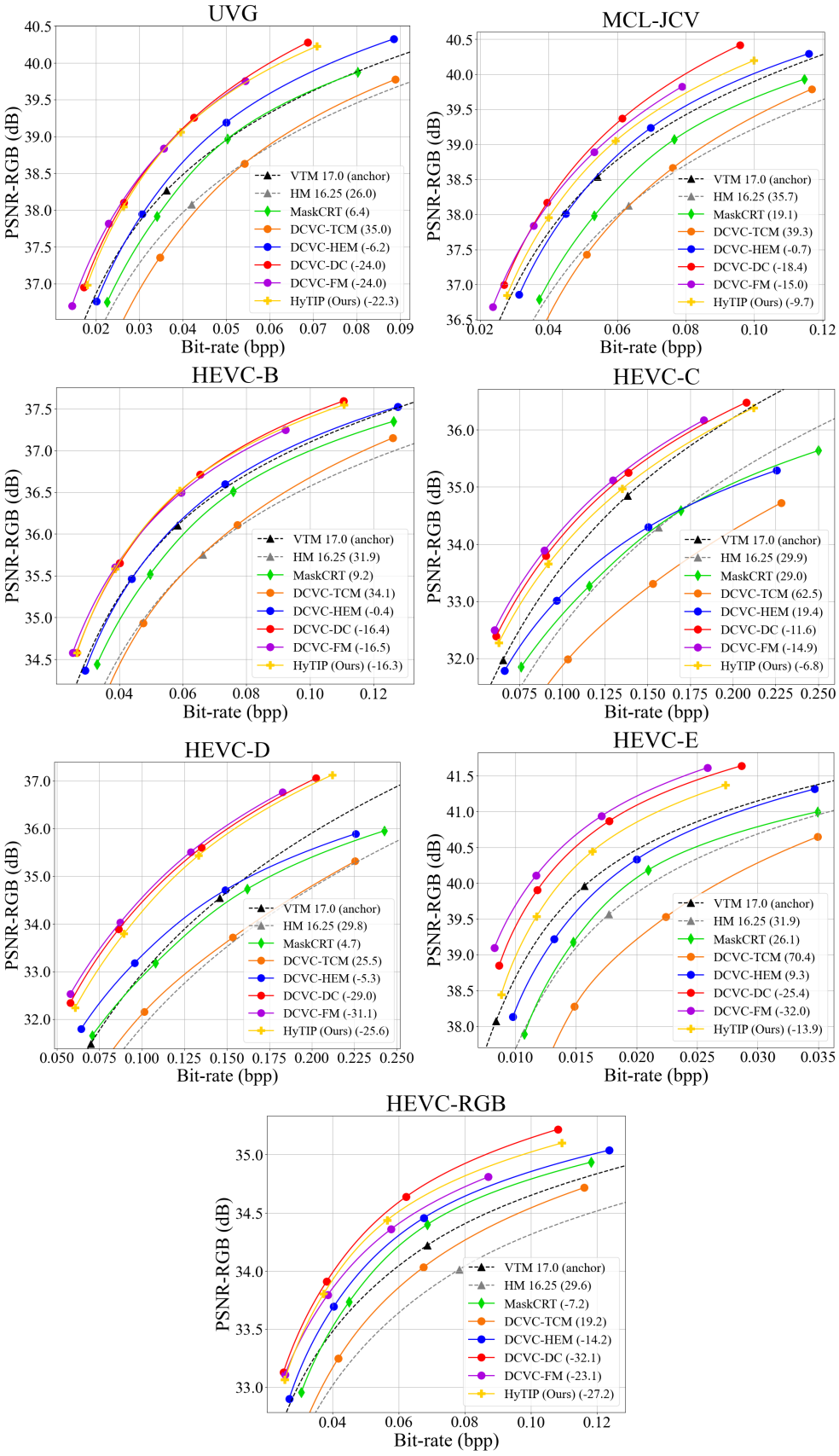

Figure A2. Rate-distortion comparison with state-of-the-art methods in terms of PSNR-RGB, using BT.709 for color space conversion. The values in parentheses represent BD-rates, with VTM 17.0 serving as the anchor.

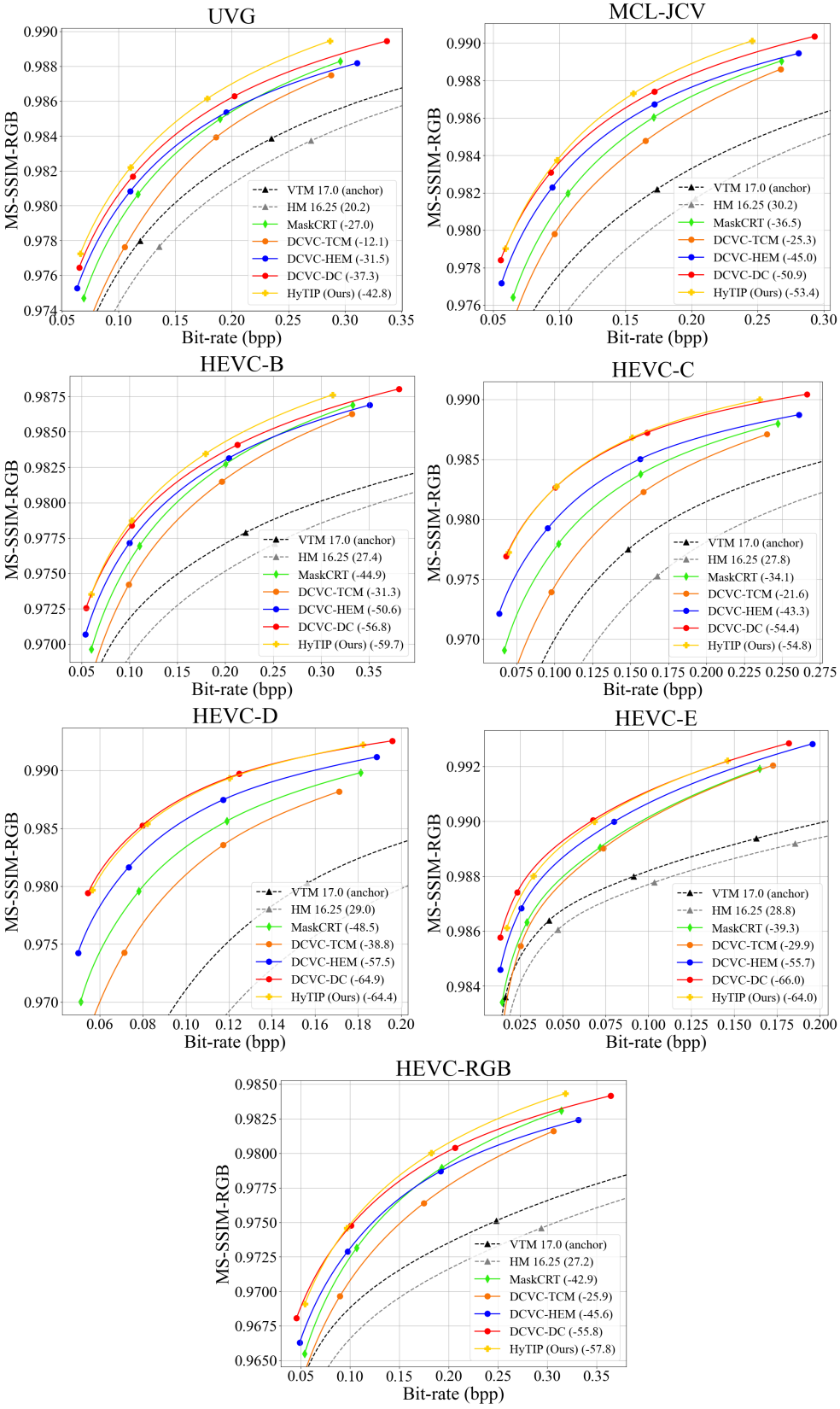

Figure A3. Rate-distortion comparison with state-of-the-art methods in terms of MS-SSIM-RGB, using BT.601 for color space conversion. The values in parentheses represent BD-rates, with VTM 17.0 serving as the anchor.

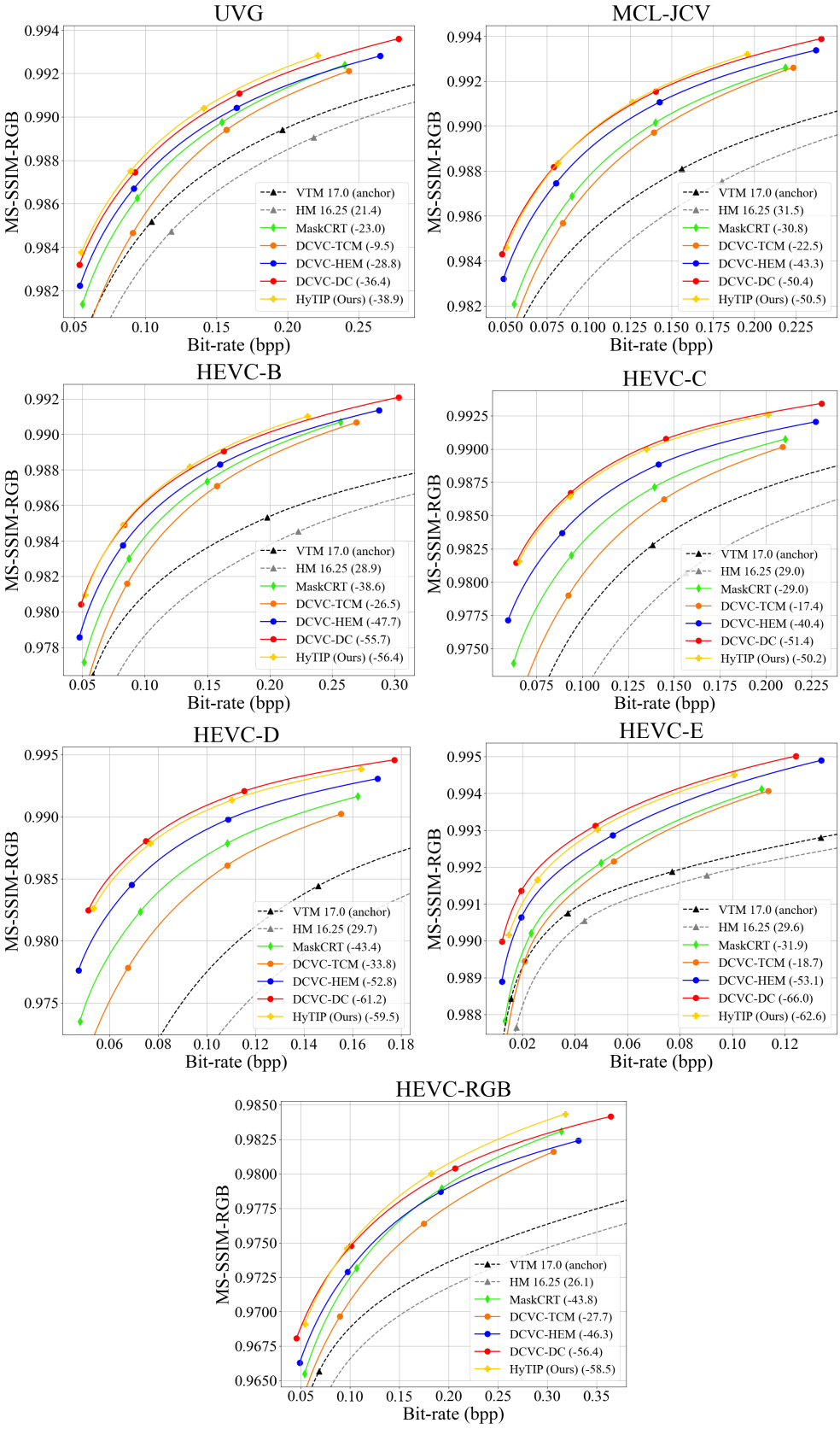

Figure A4. Rate-distortion comparison with state-of-the-art methods in terms of MS-SSIM-RGB, using BT.709 for color space conversion. The values in parentheses represent BD-rates, with VTM 17.0 serving as the anchor.

Table A6. Training procedure. MENet, TCM, FA (frame type adaptation), CTM, and gain units ($s_t^{enc}, s_t^{dec}, s_t^{recon}, s_t^{tcm}$ in Fig. A5, Fig. A7, and Fig. A6) represent the motion estimation network, the temporal context mining module in the inter-frame codec, the hierarchical quality structure [6], the channel transform module [3], and the variable-rate modules, respectively. EPA is the error propagation aware training in [8]. Ref represents the characteristic of reference temporal information in the inter-frame codec.

Phase	# Frames	Training Modules	Loss	lr	Epoch
Motion Coding (Ref: Explicit)	3	Motion codec	$R_t^{motion} + \lambda \times D(x_t, warp(x_{t-1}, \hat{f}_t))$	1e-4	8
Motion Compensation (Ref: Explicit)	3	TCM	$\lambda \times D(x_t, x_c)$	1e-4	10
Inter-frame Coding (Ref: Explicit)	2	Inter-frame codec and Mask Generator	$R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4	2
Motion Compensation (Ref: Explicit)	3	TCM	$R_t + \lambda \times \frac{D(x_t, x_c) + D(x_t, \hat{x}_t)}{2}$	1e-4	3
Inter-frame Coding (Ref: Explicit)	3 5	All modules except MENet, motion codec, and 3x3 Conv in Fig. 2	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4 1e-4	8 5
Fine-tuning (Ref: Explicit)	3 5	All modules except MENet and 3x3 Conv in Fig. 2	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4 1e-4	6 5
Feature Generation (Ref: Hybrid)	3	3x3 Conv in Fig. 2	$R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4	3
Motion Compensation (Ref: Hybrid)	3	TCM	$R_t + \lambda \times \frac{D(x_t, x_c) + D(x_t, \hat{x}_t)}{2}$	1e-4	4
Inter-frame Coding (Ref: Hybrid)	3 5	All modules except MENet and motion codec	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4 1e-4	8 4
Fine-tuning (Ref: Hybrid)	3 5	All modules except MENet	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4 1e-4	3 4
Fine-tuning with EPA (Ref: Hybrid)	5 5	All modules except MENet All modules	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-5 1e-5	4 4
FA (Ref: Hybrid) FA with EPA (Ref: Hybrid)	5 5	Frame Type Conv layers in Fig. A7	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4 1e-5	1 1
Fine-tuning with EPA (Ref: Hybrid)	5 5	All modules except MENet All modules	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-5 1e-5	1 5
CTM (Ref: Hybrid)	5	CTM in Fig. A6	$R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4	1
Fine-tuning with EPA (Ref: Hybrid)	5 5	All modules except MENet All modules	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-5 1e-5	2 7
Context Model Training with EPA (Ref: Hybrid)	5 5	Context Model in Fig. A6 Context Model, hyperprior, and decoder in Fig. A6	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4 1e-4	1 2
Inter-frame Coding with EPA (Ref: Hybrid)	5 5	Inter-frame codec and Mask Generator All modules except MENet and motion codec	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-4 1e-4	3 2
Fine-tuning with EPA (Ref: Hybrid)	5 5	All modules except MENet All modules	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-5 1e-5	3 6
Variable-rate Training with EPA (Ref: Hybrid)	5 5	Gain units in Fig. A5, Fig. A7, and Fig. A6 All modules	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-5 1e-5	8 20
Long-sequence Training with EPA (Ref: Hybrid)	7 10	All modules All modules	$R_t + \lambda \times D(x_t, \hat{x}_t)$ $R_t + \lambda \times D(x_t, \hat{x}_t)$	1e-5 1e-6	1 50

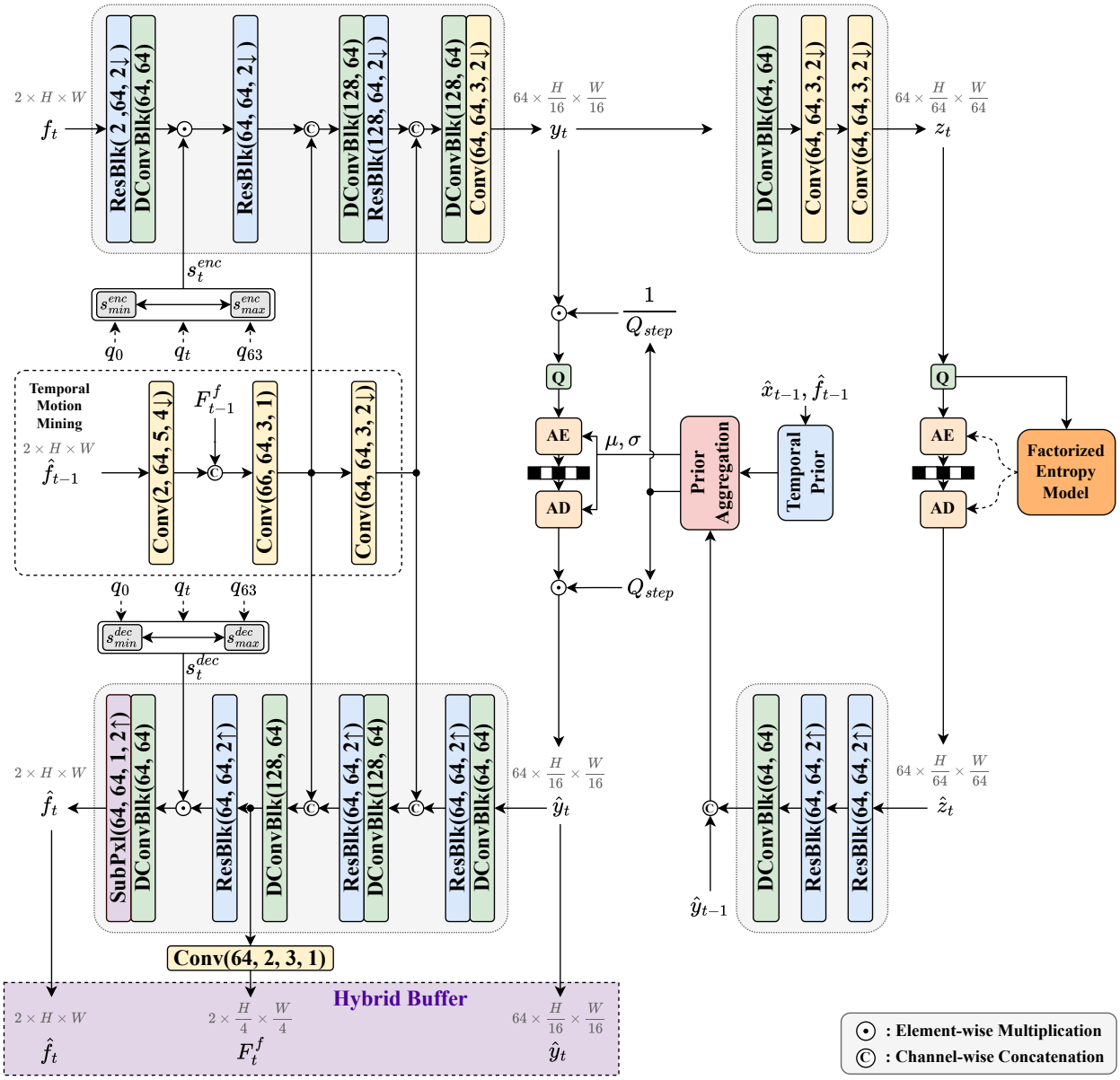

Figure A5. Network architecture detail of our motion codec.

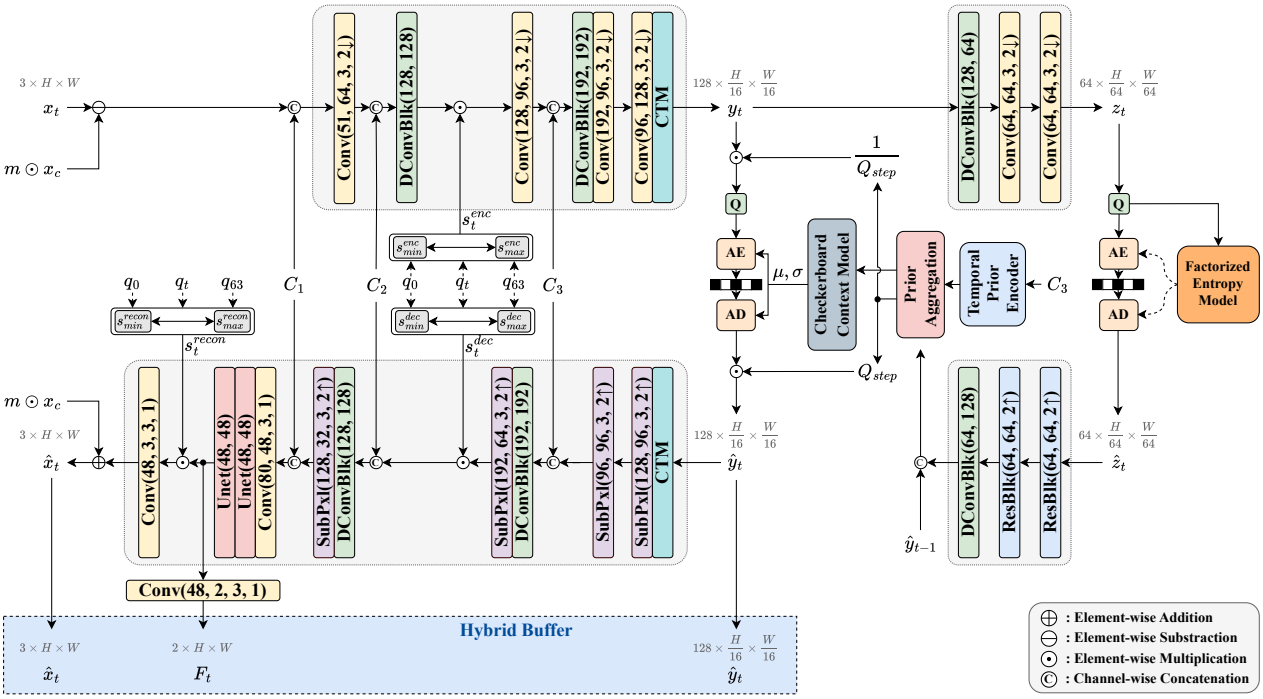

Figure A6. Network architecture detail of our inter-frame codec.

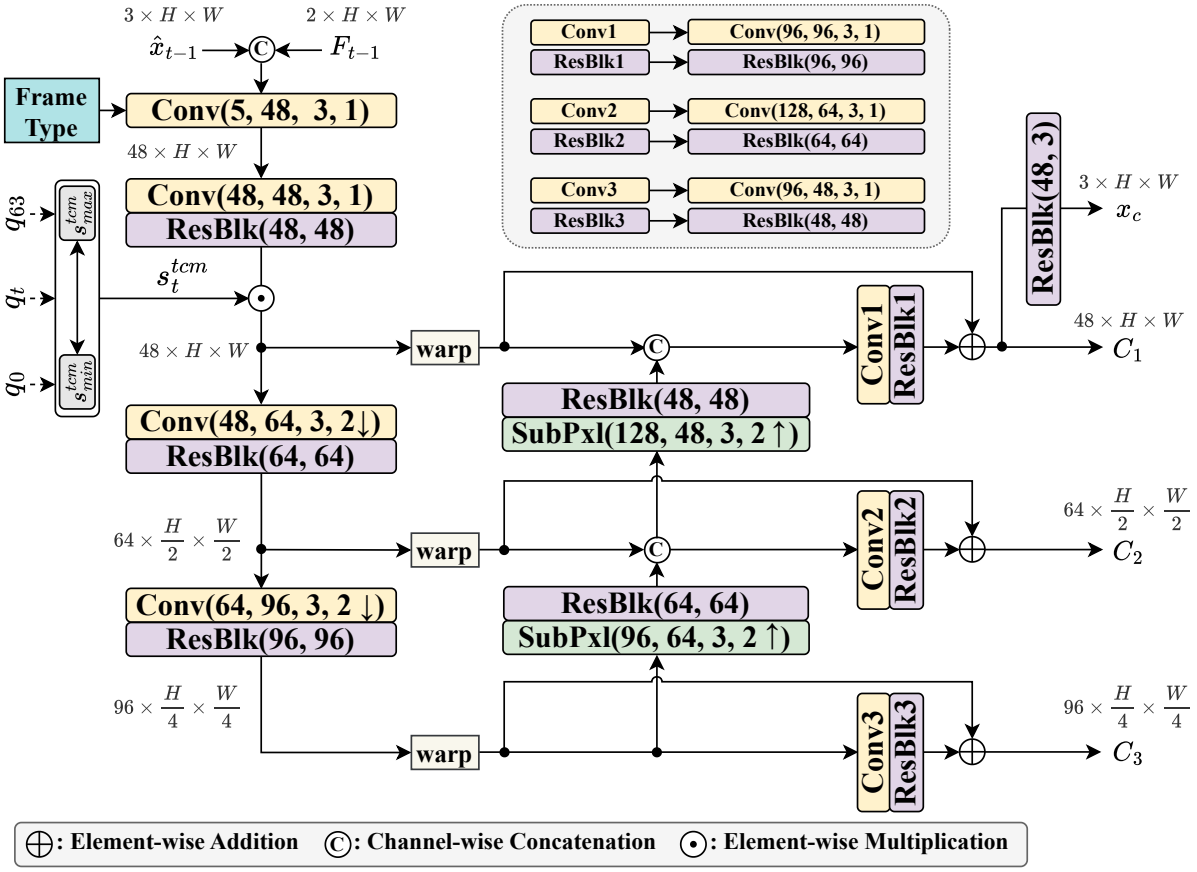

Figure A7. Network architecture detail of our temporal context mining in our inter-frame codec.