# Supplementary Material for
# Real-Time Neural Video Compression with Unified Intra and Inter Coding

## 1. Appendix

This supplementary material document provides additional details of Real-Time Neural Video Compression with Unified Intra and Inter Coding.

## 2. Testing Settings

To ensure a fair comparison with traditional codecs, we employ their optimal configurations to represent their best achievable compression performance. All tested codecs are evaluated in YUV420 color spaces. Our primary focus lies on the YUV420 color space, as it is widely optimized by conventional video codecs.

For the H.266/VVC standard, we utilize VTM (VVC Test Model) version 17.0. The official configuration file, encoder_lowdelay_vtm.cfg, is adopted. The detailed encoding command is specified as follows:

```
-c {config_file_name}
--InputFile={input_video_name}
--InputBitDepth=8
--OutputBitDepth=8
--OutputBitDepthC=8
--FrameRate={frame_rate}
--DecodingRefreshType=2
--FramesToBeEncoded={frame_number}
--SourceWidth={width}
--SourceHeight={height}
--IntraPeriod={intra_period}
--QP={qp}
--Level=6.2
--BitstreamFile={bitstream_file_name}
```

## 3. Implementation Details

### 3.1. Module Structures

The overall architecture of our model is based on DCVC-RT [1]. To accommodate the simultaneous encoding of two frames, we proportionally increase the number of channels in the intermediate convolutional layers. The structure of each module is shown in Fig. 5.

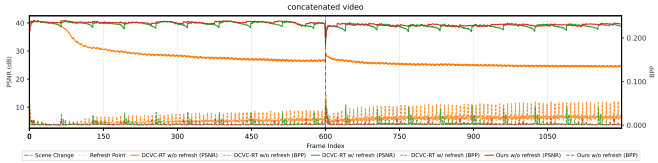

Figure 1. Bitrate and quality variation across frames

### 3.2. Entropy Model

The entropy model employed in our model utilizes a two-stage distribution estimation strategy to balance coding performance and computational efficiency. For simultaneous encoding of two frames, it is equivalent to requiring only one step of distribution estimation per frame. Although more complex entropy models could potentially yield better compression performance, we retain this design due to its faster encoding and decoding speed.

## 4. Experimental Results

### 4.1. Rate-Distortion Curves

Figure 2 illustrates the rate-distortion curves across all test sequences. At lower bitrates, our method performs comparably to or even marginally outperforms DCVC-FM. In the higher bitrate regime, our approach narrows the performance gap observed in DCVC-RT, which typically exhibits inferior performance at high bitrates. However, it is noted that a discernible performance discrepancy with DCVC-FM remains at higher bitrates. As analyzed in [1], this limitation is primarily attributed to the model's relatively low complexity.

### 4.2. Quality and Bitrate across Frames

In Fig. 3, we present the bitrate and quality variation across frames of our model for all sequences of HEVC B, HEVC E, and UVG. In Fig. 4, we present the bitrate and quality variation across frames of our single-frame model as well. It can be observed that after applying unified intra and inter coding, both models achieve more stable quality and bitrate without requiring any refresh, reduce bitrate peaks, and are thus more conducive to practical applications.

### 4.3. Additional Experiments for Scene Change

We concatenated *KristenAndSara* and *FourPeople* (HEVC Class E) into a 1200-frame sequence with a scene cut at

frame 600. As shown in Fig. 1, our method (w/o refresh) achieves 39.71 dB at 0.0027 bpp, outperforming DCVC-RT w/ refresh (39.37 dB at 0.0029 bpp).

### 4.4. Mixed RGB and YUV Loss Results

To enable evaluation in the RGB domain, we further fine-tune the model that was initially trained in the YUV domain for a limited number of iterations by introducing an additional RGB loss term with a weight of $0.2$. During this stage, the color space conversion between YUV and RGB follows the ITU-R Recommendation BT.709 standard.

As shown in Table 1, when evaluated in the RGB domain, our method (Ours RGB BT709) achieves consistent coding efficiency improvements over the DCVC-RT anchor across most datasets. For comparison, we also report the results obtained directly in the YUV domain (Ours YUV), which similarly demonstrate clear bitrate savings with respect to DCVC-RT. These results indicate that the proposed mixed RGB-YUV training strategy maintains strong compression performance in both color spaces while providing additional flexibility for RGB-domain evaluation.

### 4.5. Further Ablation of the Hybrid Ref.

We incorporate the Hybrid Ref. method into our reproduced DCVC-RT baseline for training. The evaluation results across all frames for the whole sequence are presented in Table 2.

Compared to our reproduced DCVC-RT baseline, the average performance improves by 2.6% in terms of PSNR BD-Rate under the IP=-1 setting. Furthermore, when utilizing the Hybrid reference approach, we no longer employ the refresh strategy.

### 4.6. Test Results on VVC Class A1/A2

Our model achieves excellent results across most HEVC resolutions, and it particularly outperforms DCVC by a large margin on smaller resolutions. However, our evaluation on the 10-bit 4K resolution datasets (VVC Class A1/A2) indicates that the performance is not yet optimal, as shown in Table 3. This limitation arises primarily because our current training methodology does not explicitly account for large-resolution scenarios. We plan to address and improve upon this aspect in future work.

### 4.7. Comparison with ECM

For the H.266/VVC standard, we further compare our method against ECM (Enhanced Compression Model) version 11.0. The official configuration file, `encoder_lowdelay_ecm.cfg`, is adopted. The detailed encoding command is specified as follows:

```
-c {config_file_name}
--InputFile={input_video_name}
--InputChromaFormat=420
--InputBitDepth=8
--FrameRate={frame_rate}
--SourceWidth={width}
--SourceHeight={height}
--IntraPeriod={intra_period}
--FramesToBeEncoded={frame_number}
--QP={qp}
--BitstreamFile={bitstream_file_name}
--ReconFile={recon_file_name}
```

Due to the prohibitive computational cost and extremely slow encoding speed of ECM at high resolutions, we limited our evaluation to the HEVC Class C, Class D, Class E, and MCL-JCV datasets. Table 4 presents the BD-rate results using ECM as the anchor. All tests were conducted on the YUV420 color space, calculating PSNR for all frames with an intra period of $-1$.

## References


[1] Zhaoyang Jia, Bin Li, Jiahao Li, Wenxuan Xie, Linfeng Qi, Houqiang Li, and Yan Lu. Towards practical real-time neural video compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-25, 2024*, 2025.

Table 1. BD-rate (%) performance on YUV and RGB datasets with DCVC-RT as the anchor.

| **Method** | **HEVC B** | **HEVC C** | **HEVC D** | **HEVC E** | **MCL-JCV** | **UVG** |
|---|---|---|---|---|---|---|
| Ours (RGB BT709) | -13.3 | -20.1 | -27.4 | -22.7 | 1.8 | -7.5 |
| Ours (YUV) | -9.3 | -15.1 | -22.9 | -18.4 | 2.6 | -4.4 |

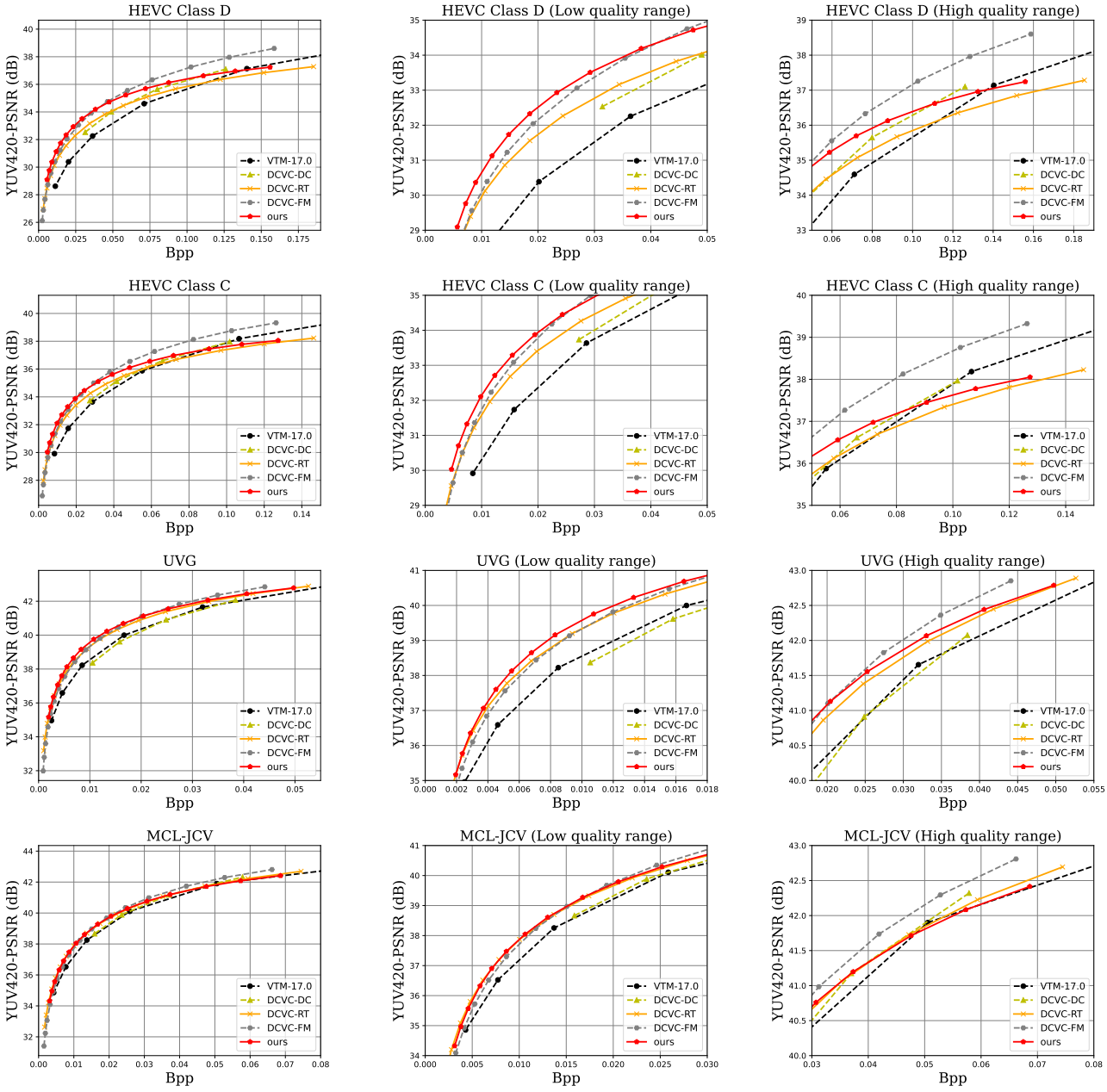


Figure 2. Rate-distortion curves for HEVC-D, HEVC-C, UVG and MCL-JCV. All frames are tested in YUV420 colorspace with intra-period = -1.

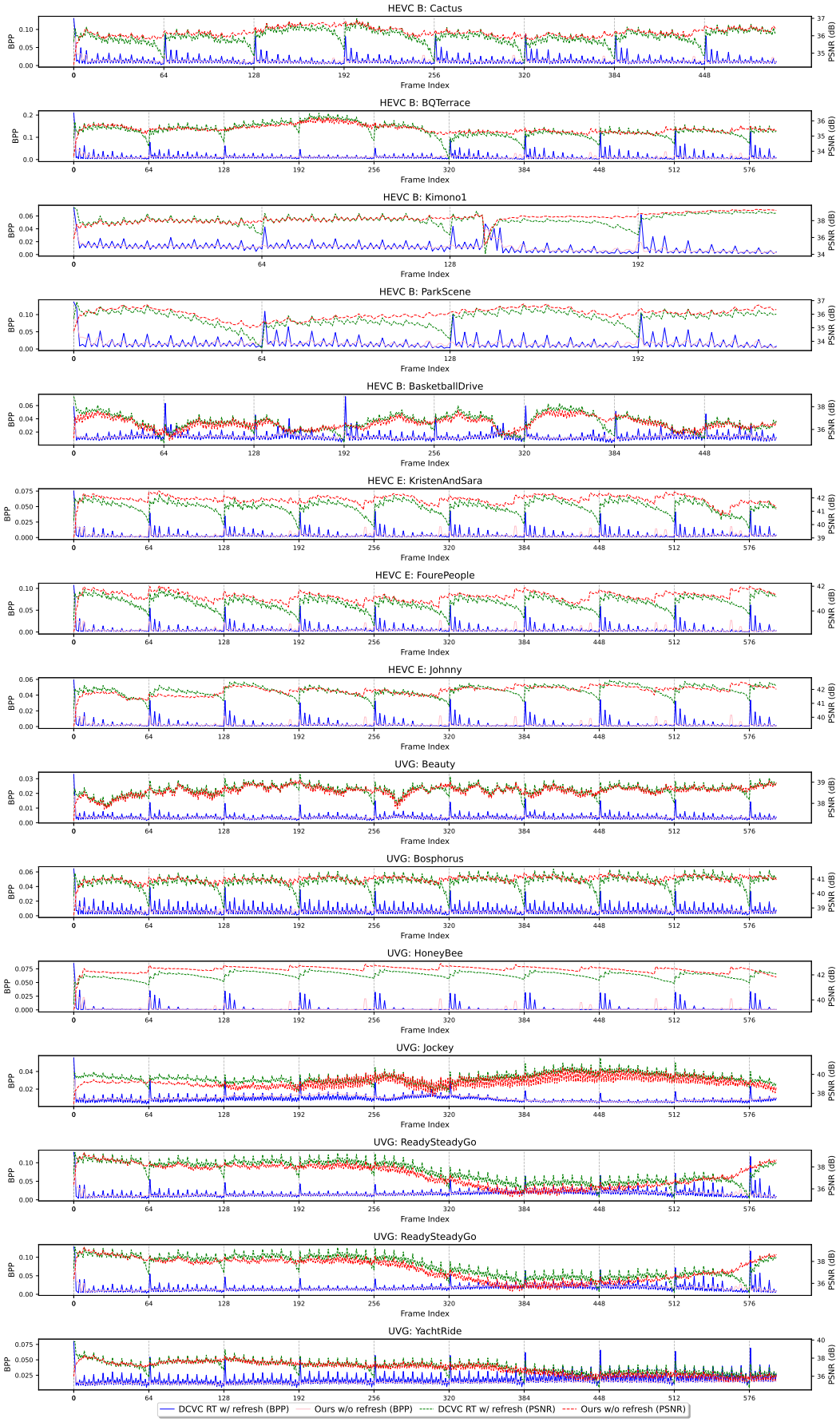


Figure 3. Bitrate and quality variation across frames for all sequences of HEVC B, HEVC E, and UVG.

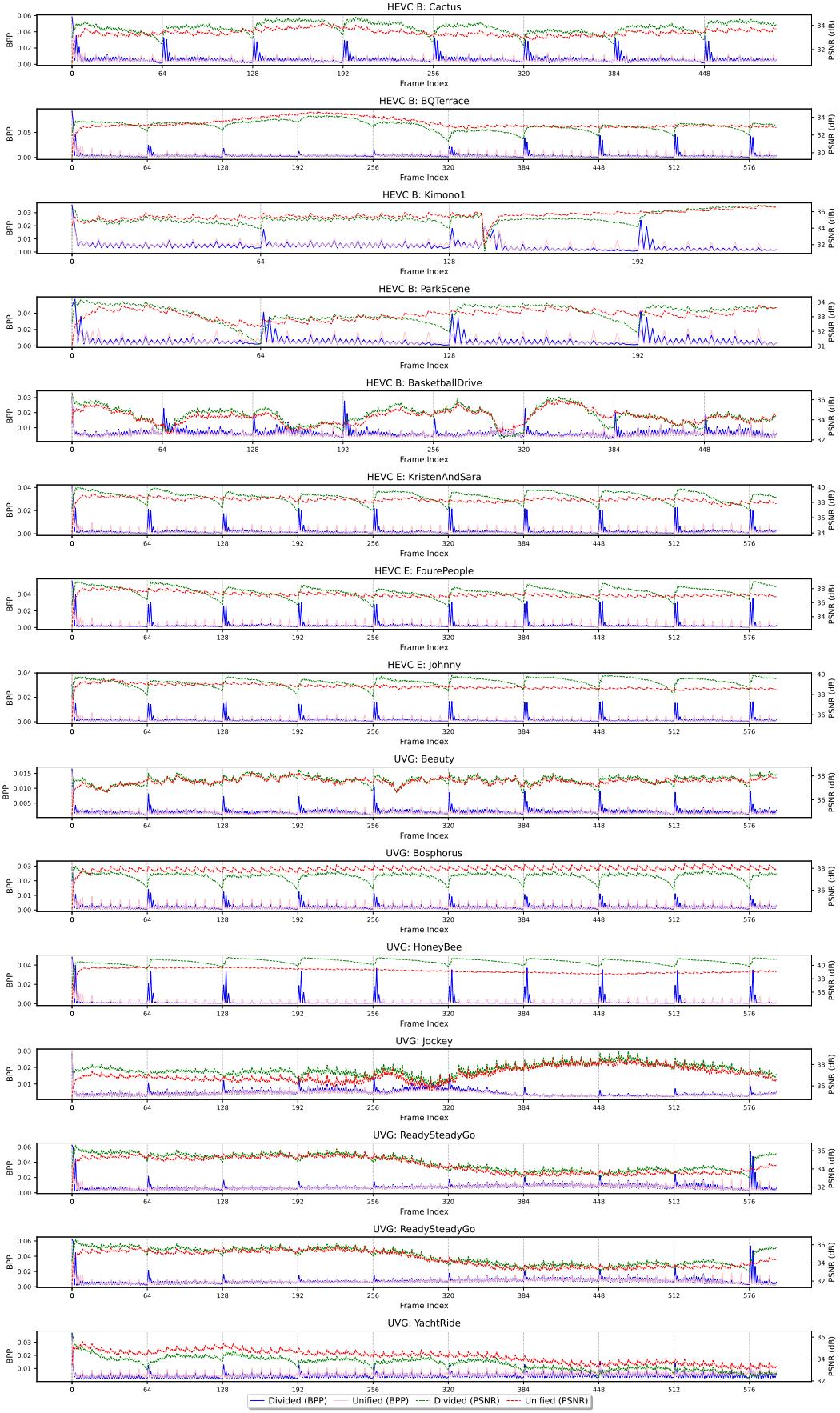


Figure 4. Bitrate and quality variation across frames for all sequences of HEVC B, HEVC E, and UVG. We compare two settings: divided (intra and inter frames use two models, refresh period is 64) and unified (all frames use the same model, no refresh). *Note that we test single-frame compression models in this figure, rather than the proposed two-frame compression scheme.*

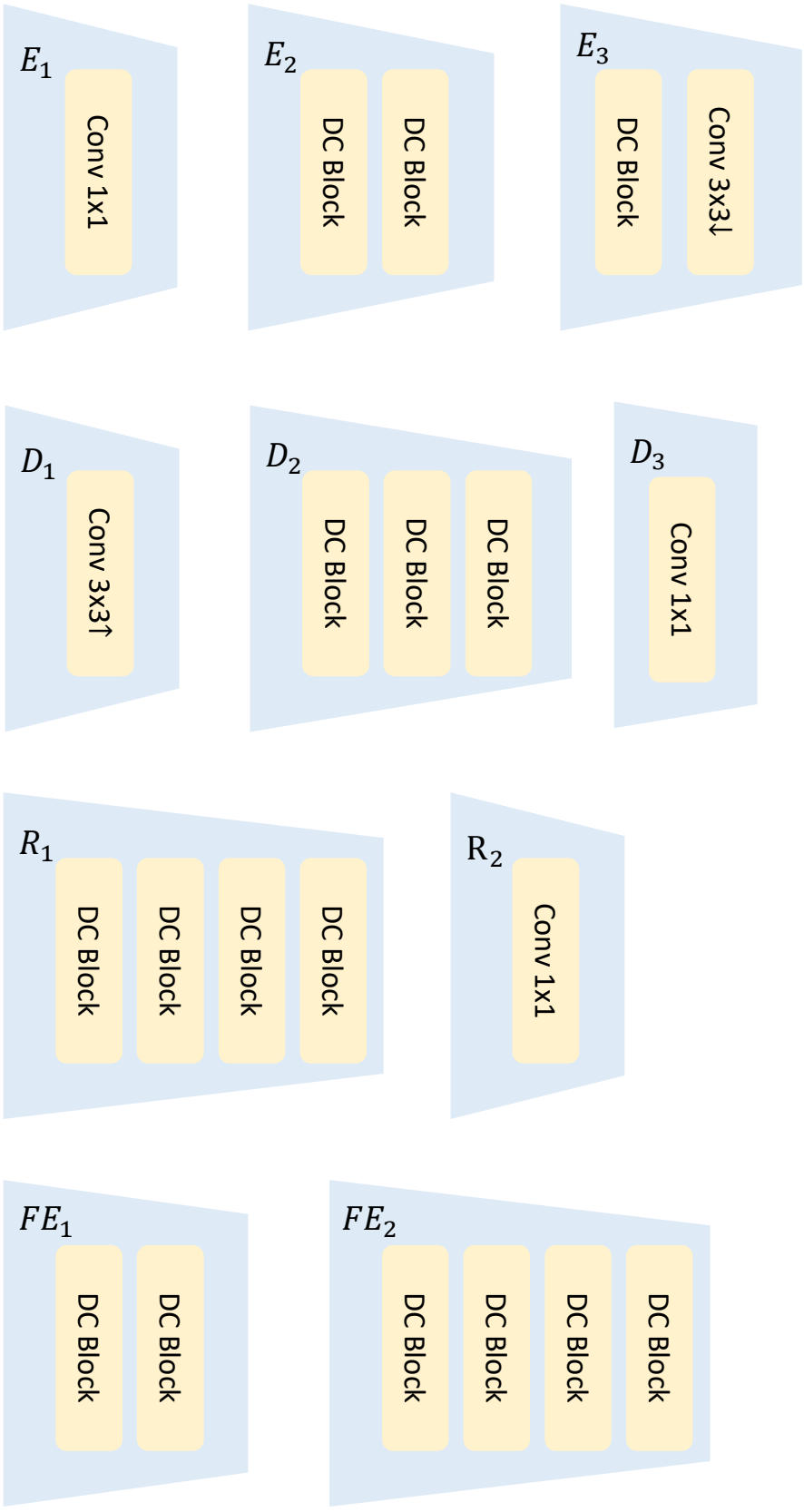


Figure 5. Detailed architecture of our proposed UI$^2$C (unified intra and inter coding) scheme.

Table 2. YUV420-PSNR BD-Rate (%) performance evaluation across all frames under the IP=-1 setting. Negative values indicate bit-rate savings.

| **Method** | **HEVC B** | **HEVC C** | **HEVC D** | **HEVC E** |
|---|---|---|---|---|
| RT-hybrid | $-3.4$ | $-0.7$ | $-0.4$ | $-6.0$ |

Table 3. BD-rate results for YUV420 - PSNR (intra period -1, all frames).

| **Method** | **VVC Class A1** | **VVC Class A2** |
|---|---|---|
| Ours | 10.3 | 3.2 |

Table 4. BD-rate (%) results using ECM as the anchor (YUV420 PSNR, intra period $-1$, all frames). Lower is better.

| **Method** | **HEVC C** | **HEVC D** | **HEVC E** | **MCL-JCV** |
|---|---|---|---|---|
| DCVC-RT | 7.4 | $-3.5$ | $-6.0$ | 11.7 |
| Ours | $-11.1$ | $-25.8$ | $-18.5$ | 12.0 |