High Resolution Neural Video Coding with Bi-directional Confidence-Guided Reference Information Modeling

Supplementary Material

A. Test Configuration

We compare the traditional codecs HM-16.20-LDB, HM-16.20-RA and VTM-23.0-RA. For a fair evaluation of compression performance, we adopt the experimental setup from [19]. All YUV420 inputs are converted to RGB using the BT.709 standard and treated as the reference sequences for every method. Because the traditional codecs do not support RGB input directly, they encode the sequences in YUV444 and PSNR is computed after reconverting the decoded output back to RGB with BT.709. To obtain optimal traditional codec performance, we use a 10-bit YUV444 colorspace. The codecs share most command-line options; their configuration files are set as follows:

- HM-LDB: *encoder_lowdelay_main_rext.cfg*
- HM-RA: *encoder_randomaccess_main_rext.cfg*
- VTM-RA: *encoder_randomaccess_vtm.cfg*

The command line is as follows:

$$
\begin{aligned}
&\text{-}c\ config\_file\\
&\text{--}InputFile = input\_file\\
&\text{--}InputBitDepth = 10\\
&\text{--}OutputBitDepth = 10\\
&\text{--}OutputBitDepthC = 10\\
&\text{--}InputChromaFormat = 444\\
&\text{--}FrameRate = frame\_rate\\
&\text{--}FramesToBeEncoded = 97\\
&\text{--}SourceWidth = width\\
&\text{--}SourceHeight = height\\
&\text{--}IntraPeriod = 32\\
&\text{--}QP = qp\\
&\text{--}BitstreamFile = bitstream\_file
\end{aligned}
$$

B. Module-wise Computational Complexity

To better understand how each component contributes to the overall computational complexity, we break down the multiply–accumulate operations (MACs) of the major modules in both DCVC-B and our HR-NVC system, as summarized in Table 7. For *flow estimation*, HR-NVC reports two numbers because the virtual reference frame is activated only for 1080p and higher-resolution sequences. Even with this dual-path design, the overall cost of *flow estimation* is reduced compared with DCVC-B. This improvement is largely attributed to the multi-resolution motion estimation design: although the model no longer relies on full-resolution reference flow, the progressively refined low-scale motion cues remain sufficiently accurate, enabling strong prediction quality while reducing computational cost.

Table 7. Module-wise computational complexity (kMACs/pixel) comparison between DCVC-B and the proposed HR-NVC. HR-NVC reports two values for *flow estimation* since the virtual reference frame is used only for $\geq 1080p$ sequences. The results show that confidence-aware motion compression lowers the cost of context refinement. Other modules remain comparable or slightly lighter.

	DCVC-B	HR-NVC
Flow Estimation	1284	1005/1056
Mition Compression	106	272
Motion Compensation	55	62
Context Refinement	533	378
Contextual Compression	508	572
Feature Extraction & Reconstruction	461	399

For *motion compression*, HR-NVC introduces an additional decoding-side branch to compute confidence maps. Although this brings moderate overhead to the motion codec itself, it substantially lowers the cost of *context refinement*. With reliable confidence modeling, the system can directly combine bi-directional predictions through a linear fusion mechanism, alleviating the burden on the refinement network.

In *motion compensation* and *feature extraction & reconstruction*, HR-NVC maintains a similar or slightly lower computational load. This indicates that the proposed hierarchical structure mainly affects modules directly tied to motion modeling, without causing unnecessary overhead on the rest of the pipeline.

Overall, the results show that HR-NVC redistributes complexity rather than simply increasing it: motion-related operations become stronger but more efficient, while downstream modules benefit from reduced refinement cost and simplified processing.

C. Rate-Distortion Results

Table 8 presents the performance comparison under GOP 64 and intra-period 64 using 197 frames. For DCVC-B[34], we follow its hierarchical quality control strategy, where the 32-nd frame within each GOP is treated as the quality-layer 0 frame, with weighting factors={1.4,1.4,1.4,0.7,0.5,0.5}

Table 8. BD-rate(%) comparision for PSNR(dB) with 193 frames with intra-period=64 and GOP=64. The anchor is DCVC-B.

	JCT-VC Class B	JCT-VC Class C	JCT-VC Class D	JCT-VC Class E	UVG	Overall Average
B-CANF[4]	48.47	58.81	67.15	106.79	36.03	63.45
HR-NVC	-7.70	-10.13	-7.96	-8.05	-9.08	-8.58

applied across hierarchical levels. With DCVC-B as the baseline, our HR-NVC achieves substantial gains across all test classes. Compared to GOP=32, this improvement becomes especially pronounced, highlighting the strengths of our proposed components—Hierarchical Motion Representation (HMR), Bi-Contextual Asymmetric Harmonization (CAH), Spatial Anchor (SA), and Temporal Virtual Anchor (TVA).

These results confirm that the proposed techniques are not only effective individually but are especially synergistic under large-GOP conditions, where their advantages in preserving temporal stability and strengthening reference representations become most evident.

D. Visualization

Fig 10 illustrates the effect of our multi-scale flow refinement strategy using the Kimono sequence. The top row visualizes the initial coarse flow, the progressively refined multi-scale flow, and the flow without refinement (rightmost). Without refinement, the background motion field exhibits significant noise and structural inconsistency, especially around tree branches and high-frequency textures. After applying our multi-scale refinement, the flow becomes smoother and more coherent while still capturing fine object boundaries, demonstrating that reliable motion cues can be recovered even without relying on full-resolution reference flow.

The bottom row shows the corresponding warped predictions. When using the unrefined flow, the warped frame contains strong artifacts and distortions in the background, revealing the propagation of noisy motion into the prediction stage. In contrast, the refined flow leads to a much cleaner warp, with better alignment and significantly fewer structural errors. This confirms that the proposed refinement not only reduces computation—by avoiding full-resolution flow—but also improves motion accuracy and prediction quality.

A similar trend is observed on other large-motion, high-detail sequences such as YachtRide (Fig 11), where the refined multi-scale flow consistently suppresses background noise and yields more stable predictions. These results validate the effectiveness of the proposed refinement mechanism across diverse scenes.

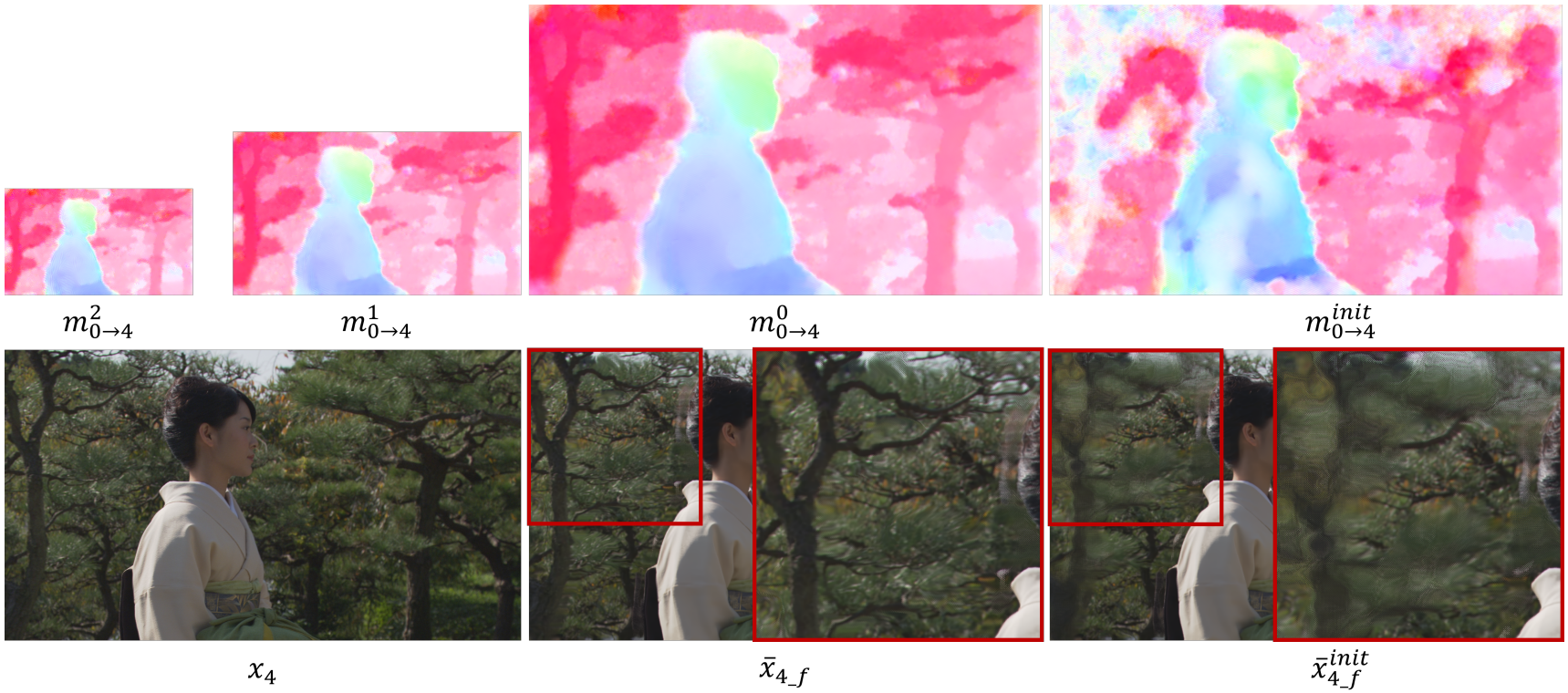

Figure 10. Visualization of the proposed multi-scale flow refinement on JCT-VC Class B Kimono. The refined flow produces cleaner and more coherent motion fields compared with the unrefined version, especially in background regions, leading to more accurate warped predictions. (b) YachtRide: similar improvements are observed, where refinement suppresses noisy background motion and results in more stable frame warping.

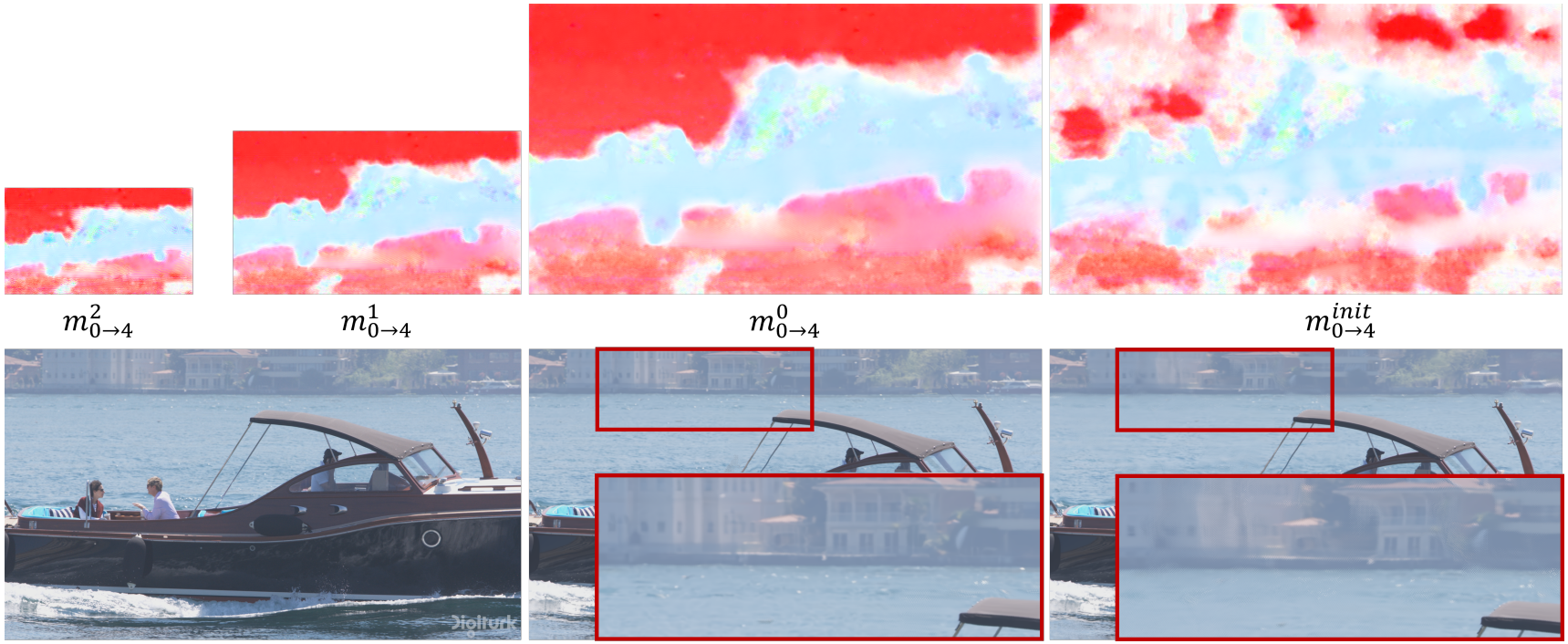

Figure 11. Visualization of the proposed multi-scale flow refinement on JCT-VC Class B YachtRide: similar improvements are observed as Kimono, where refinement suppresses noisy background motion and results in more stable frame warping.