Neural Video Compression with Reference Hierarchy

Chuanbo Tang, Zhuoyuan Li, Li Li, Dong Liu*, Feng Wu

MOE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition,
University of Science and Technology of China
{cbtang, zhuoyuanli}@mail.ustc.edu.cn, {lil1, dongeliu, fengwu}@ustc.edu.cn

Abstract

Efficient reference structures are essential in video compression, enabling the exploitation of temporal dependencies across frames to reduce redundancy. In this paper, we delve into the inter-frame reference management mechanism in neural video codecs (NVCs). Previous schemes have inherited the reference propagation mechanism with the guidance of predefined reference structure, but the reference modeling across diverse reference sources remains underexplored. Moreover, the mismatch between the reference structure used for motion estimation and motion compensation limits the effectiveness of inter-frame prediction. To address the above limitations, we propose the unified reference hierarchy that integrates a learned hierarchical reference structure into the existing inherent reference propagation mechanism. Specifically, we first propose the hierarchical reference structure (HRS) to manage the multiple temporal contexts in the propagated reference feature, where a hierarchy-aware reference modulation module is integrated to select the most relevant reference features across different quality levels under the guidance of the reference balance loss. In addition, we propose the HRS-guided feature-wise inter-frame prediction that learns the low-rank approximation of the selected reference feature for ensuring the consistency and improving the inter-frame prediction performance. We conduct experiments on a state-of-the-art NVC, DCVC-DC. Experimental results show that our codec achieves an average 26% bitrate saving over H.266/VVC, and a 28.2% bitrate reduction compared to DCVC-DC without increasing the decoding complexity.

Introduction

In video compression, a more effective reference structure typically leads to greater bitrate savings (Wiegand, Zhang, and Girod 1999; Li et al. 2012; Bross et al. 2021b). Motivated by this, we rethink the inter-frame reference management mechanism in neural video codecs (NVCs).

Recent advances in neural video compression have continuously evolved the design of inter-frame reference mechanisms, shifting from simple reconstructed-frame references to more sophisticated feature-based approaches. The pioneering work DVC (Lu et al. 2019) firstly incorporated deep neural networks into the traditional residual coding-based video compression framework, where the adjacent reconstructed frame is used as the reference for both motion estimation and motion compensation during temporal inter-frame prediction. FVC (Hu, Lu, and Xu 2021) further extended it by extracting pixel-level reference representations in the feature space to enhance the inter-frame reference. DCVC (Li, Li, and Lu 2021) proposed the conditional coding-based framework, and also applied feature-based reference extraction for motion compensation. Notably, the context learned during motion compensation was directly utilized as the condition for the contextual coding.

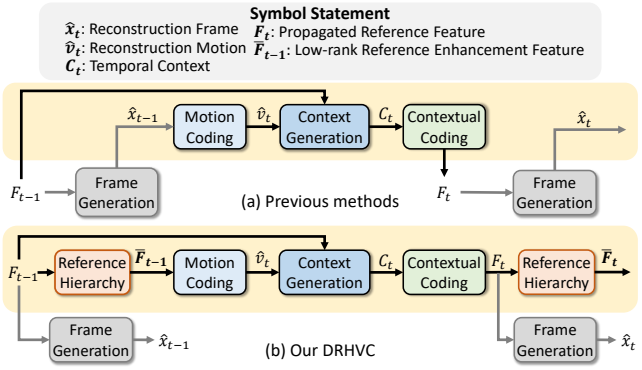

Figure 1: Framework comparison for our DRHVC (Deep Reference Hierarchy for Video Compression) with previous conditional coding-based schemes.

Since the introduction of the reference feature propagation mechanism in DCVC-TCM (Sheng et al. 2022), a series of subsequent works (Li, Li, and Lu 2022, 2023, 2024; Qi et al. 2023; Tang et al. 2024, 2025; Jiang et al. 2025) have adopted and improved upon this approach, where the reference source for motion compensation is the propagated feature from the adjacent frame. As illustrated in Fig 1 (a), in these schemes, the propagated reference feature F_{t-1} is firstly mapped into the pixel domain to generate the reference frame $\hat{x}_{t-1}$, which is then used as the sole reference for motion estimation. Simultaneously, F_{t-1} also directly serves as the reference for context generation in motion compensation. To further leverage temporal correla-

*Corresponding Author.

tions, DCVC-DC (Li, Li, and Lu 2023) inherited the reference feature propagation mechanism and introduced the hierarchical weights. By assigning periodically varying weight to each frame during cascaded training, the propagated feature F_{t-1} can exploit multi-frame temporal contexts. In addition, recent studies (Li, Li, and Lu 2024; Tang et al. 2025) also adopted the reference mechanism and incorporated the reconstructed reference frame $\hat{x}_{t-1}$ into context generation to alleviate the error propagation.

Despite recent advances in inter-frame reference modeling, the reference modeling across diverse reference sources remains underexplored. Specifically, motion estimation in NVC still relies solely on the adjacent decoded frame, overlooking the potential relevance between multiple temporally decoded frames and the current to-be-encoded frame. By revisiting the reference mechanisms across different codecs, the traditional codec (Bross et al. 2021a) typically performs motion estimation using multi-frame references, where the set of reference frames is manually selected using predefined strategies to enhance inter-frame prediction efficiency. Building upon the insight into the effectiveness of the multi-frame-based reference mechanism, we identify two key limitations in the reference mechanism for NVC. Firstly, the reference sources used for motion estimation are typically the sole reconstructed frame $\hat{x}_{t-1}$ derived from the propagated feature F_{t-1}, and the reconstruction quality is constrained by a single-frame distortion term in the rate-distortion (RD) loss function, leading to a single-frame reference loop. As a result, even with predefined hierarchical weights, the multi-frame correlation of the reference structure cannot be fully exploited. Moreover, there is an asymmetric reference structure between the motion estimation and motion compensation in the current prediction chain, where motion compensation benefits from the propagated reference features F_{t-1} with multi-frame temporal modeling. The inconsistency weakens the coherence of inter-frame prediction and limits the overall efficiency of reference modeling.

To address these limitations, we propose a novel framework, DRHVC (Deep Reference Hierarchy for Video Compression), which enhances inter-frame reference modeling through a unified and feature-domain hierarchical design for contextual coding-based NVC. As illustrated in Fig. 1 (b), DRHVC comprises two key components. First, we propose the hierarchical reference structure (HRS) to efficiently organize the rich temporal context embedded in the propagated reference feature. Unlike prior methods that rely on the single-frame reference loop, HRS learns multi-frame reference structure from the temporally informative propagated reference feature F_{t-1}, where rich temporal contexts are organized and compacted by a hierarchy-aware reference modulation module. The module employs a learnable router to select the most relevant reference feature across different quality levels. To enable both reference diversity and efficient temporal prediction chain in the reference loop, we propose the reference balance loss to supervise the reference router. This selection strategy enables the model to learn the varying contributions of different reference sources along the prediction chain, enhancing its capability of modeling multi-frame temporal dependencies.

Second, given the selected reference feature from the hierarchy-aware reference modulation module, we introduce the HRS-guided feature-wise inter prediction to enhance prediction efficiency and temporal consistency for inter prediction. To avoid extra computational overhead while ensuring consistent references for estimation and compensation, we learn a low-rank approximation $\overline{F}_{t-1}$ of the selected reference feature for motion estimation, which enables compact representations of multi-frame temporal contexts while preserving the context-aligned prediction chain.

The two proposed methods form a unified reference hierarchy that seamlessly integrates learned reference structures into the existing reference propagation mechanism, enabling more coherent inter-frame prediction and significantly boosting overall compression performance. We conduct experiments on a state-of-the-art (SOTA) scheme, DCVC-DC. Experimental results demonstrate the superiority of our proposed DRHVC across a wide range of benchmarks. Under both intra-period 32 and -1 settings, DRHVC consistently achieves SOTA compression performance. Compared to the traditional codec H.266/VVC, our model achieves an average bitrate saving of 26% under intra-period -1. Furthermore, DRHVC achieves a 28.2% bitrate reduction over DCVC-DC without increasing decoder complexity and a 12.5% bitrate saving compared to DCVC-FM, demonstrating the efficiency of our scheme.

Our contributions are summarized as follows:

- We propose a unified reference hierarchy for the contextual coding-based NVC. By integrating the learned hierarchical reference structure into the reference propagation mechanism, our scheme forms a multi-frame feature-domain reference loop to achieve efficient inter-frame prediction.
- We propose the hierarchical reference structure (HRS) to manage the multi-frame temporal contexts embedded in the propagated reference features. Guided by the reference balance loss, the hierarchy-aware reference modulation module selects the most relevant reference feature across different quality levels along the prediction chain.
- We introduce HRS-guided feature-wise inter-frame prediction to improve motion estimation. By learning a low-rank approximation of the selected reference feature, it enables motion estimation to leverage the same temporal context used in motion compensation, ensuring consistency and improving inter-frame prediction performance.
- Our proposed DRHVC achieves an average bitrate saving of 26% over the traditional video codec H.266/VVC. It also outperforms advanced SOTA NVCs, demonstrating the superior compression performance and efficiency.

Related Work

Neural Video Compression

Recently, Neural video compression has advanced rapidly, with most work focusing on P-frame coding for low-delay and real-time streaming. Mainstream P-frame neural video compression approaches can be broadly categorized into two groups. The first is residual coding-based methods, which

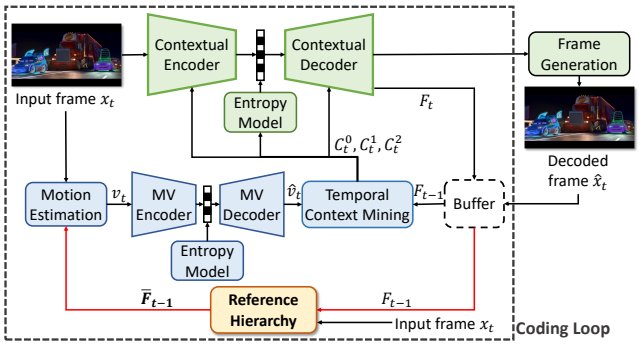

Figure 2: Overview of our proposed framework DRHVC.

reduce inter-frame redundancy by explicitly modeling and compressing residuals between adjacent frames using subtraction (Lu et al. 2019, 2020; Lin et al. 2020; Agustsson et al. 2020; Hu, Lu, and Xu 2021; Hu et al. 2022; Shi et al. 2022; Hu and Xu 2023; Li et al. 2023; Chen et al. 2024). The second is conditional coding-based methods, which learn and exploit temporal dependencies through feature-level context propagation and conditioning on reference information (Li, Li, and Lu 2021; Sheng et al. 2022; Ho et al. 2022; Li, Li, and Lu 2022, 2023; Qi et al. 2023; Li, Li, and Lu 2024; Bian et al. 2024, 2025; Tang et al. 2024, 2025; Jiang et al. 2025). Our proposed scheme builds upon the conditional coding-based framework due to its greater flexibility in modeling complex temporal dependencies.

Reference Modeling in Video Compression

In traditional codecs, multi-frame reference modeling has long been recognized as effective for improving inter-frame prediction (Wiegand, Zhang, and Girod 1999; Li et al. 2024). (Li et al. 2012) formulated reference management as an optimization problem and investigated the reference quality adjustment strategies. (Lin et al. 2018; Huo et al. 2020) leveraged frame extrapolation networks to exploit multi-frame temporal dependencies from previous decoded frames.

In the early residual coding-based scheme, DVC (Lu et al. 2019) estimated optical flow using only the nearest reconstructed frame as the reference. Later works (Lin et al. 2020, 2022) further incorporated multi-frame temporal cues into reference modeling. FVC (Hu, Lu, and Xu 2021) further shifted pixel-wise reference frames to the feature domain, and the pioneering conditional coding-based scheme DCVC (Li, Li, and Lu 2021) also adopted feature-domain references for motion compensation. DCVC-TCM (Sheng et al. 2022) proposed the reference feature propagation mechanism, where the feature from the adjacent frame is used directly for motion compensation and projected back to the pixel domain for motion estimation, supervised by single-frame mean squared error (MSE), leading to the single-frame reference loop. DCVC-DC (Li, Li, and Lu 2023) proposed a hierarchical reference weighting-based cascaded training strategy, enabling the propagated reference features to exploit the multi-frame temporal contexts through continuous updates. Subsequent works (Li, Li, and Lu 2024; Tang et al. 2025) further investigated the effectiveness of pixel-wise reference frames in the reference propagation of motion compensation to alleviate the error propagation. Other efforts (Sheng et al. 2024; Jiang et al. 2025) explored explicit multi-frame reference modeling to improve compensation quality. However, most methods emphasize motion compensation, while motion estimation still relies on a single reconstructed frame. In this work, we learn the multi-frame reference structure for motion estimation, aiming to unify motion estimation and compensation under a temporally coherent framework.

Approach

Overview

We implement our reference hierarchy on the contextual coding-based framework DCVC-DC (Li, Li, and Lu 2023) to build our new scheme DRHVC, which is shown in Fig. 2. Instead of relying only on the reconstructed frame $\hat{x}_{t-1}$ as the reference, our coding loop leverages the propagated reference feature F_{t-1} as the primary input for motion estimation. Our model formulates a multi-frame feature-domain reference loop for achieving efficient inter-frame prediction.

The overview architecture of our proposed framework is summarized as follows. First, the input frame x_t and the propagated reference feature F_{t-1}, extracted from the adjacent previous frame x_{t-1}, are input to the proposed reference hierarchy to generate a low-rank reference enhancement feature $\overline{F}_{t-1}$. Given the current frame x_t, $\overline{F}_{t-1}$ is used as the reference for estimating the motion vector (MV) v_t in the form of optical flow. The estimated flow v_t is then compressed via the MV encoder and decoder. The decoded flow $\hat{v}_t$ and the propagated reference feature F_{t-1} are fed into the temporal context mining for learning multi-scale temporal contexts C_t^0, C_t^1, C_t^2. These contexts serve as conditional priors to guide the compression of the input frame x_t, enabling effective temporal redundancy removal through the contextual encoder and decoder. Finally, the reconstructed feature F_t is saved in the buffer for generating the reference features for the subsequent frame with reference hierarchy.

Hierarchical Reference Structure

As recent NVCs adopt the propagated reference feature for motion compensation, which implicitly carries multi-frame temporal contexts across multiple frames (Li, Li, and Lu 2023), we propose the hierarchical reference structure (HRS) to better manage the rich temporal information for improving the efficiency of inter-frame prediction. Inspired by the mature reference modeling in traditional codec (Li et al. 2012) and the constructed concept of large language models (Liu et al. 2024), we formulate reference management as a token-level learning problem and introduce hierarchy-aware reference modulation, which adaptively determines the optimal temporal reference via a mixture-of-experts (MoE) (Zamfir et al. 2024; Wang et al. 2025) mechanism, where each expert models a distinct temporal context representation for different reference quality levels.

As illustrated in Fig. 3, we first perform channel expansion on the propagated reference feature F_{t-1}, then the ex-

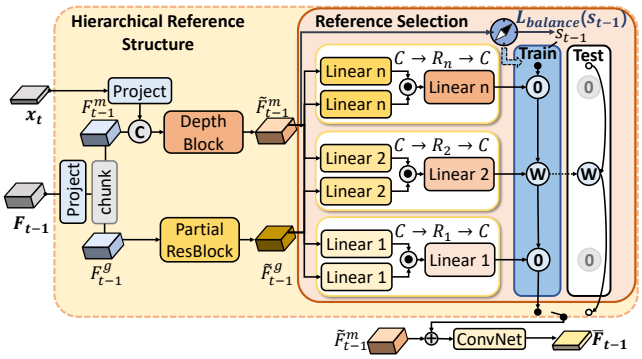

Figure 3: The diagrams of the reference hierarchy. Project denotes feature projection via convolutional layers. Linear represents the linear layer used for channel transformation.

panded feature is split along the channel dimension into two branches. The first branch is used to generate the main reference feature for both routing and expert computation in MoE. Specifically, the initial main feature F^m_{t-1} is concatenated with the feature representation extracted from the current input frame x_t, forming the main reference feature $\tilde{F}^m_{t-1}$ by a depth-wise convolution block. Incorporating information from the current frame enables the router to more accurately determine the optimal reference quality level in the hierarchical structure, while also enriching the main reference feature with current-context cues to enhance temporal alignment and feature interaction in each expert. The second branch generates the gating feature to dynamically modulate the expert responses. The initial gating feature F^g_{t-1} is refined by a lightweight partial residual block (Chen et al. 2023) to produce the calibrated gating feature $\tilde{F}^g_{t-1}$, which improves modulation precision with low computational cost.

Hierarchy-aware Reference Modulation. Given the reference feature $\tilde{F}^m_{t-1}$ and the calibrated gating feature $\tilde{F}^g_{t-1}$, we propose the hierarchy-aware reference modulation module to decide the optimal temporal reference through the mixture of low-rank expertise. Inspired by the exploration in (Ma et al. 2024), we leverage element-wise multiplication in low-dimensional space, which enables each expert to implicitly model high-dimensional temporal dependencies while maintaining computational efficiency. Specifically, both $\tilde{F}^m_{t-1}$ and $\tilde{F}^g_{t-1}$ are projected to a reduced channel space, and each expert operates with a distinct low-rank representation, introducing diversity in temporal modeling. The gating feature $\tilde{F}^g_{t-1}$ modulates $\tilde{F}^m_{t-1}$ via element-wise multiplication, allowing each expert to selectively emphasize temporal contexts embedded in the propagated reference feature. Finally, the output is projected back to the original channel dimension, yielding a refined reference feature tailored to the expert's perspective. To enable certain selection across different levels of reference quality in the hierarchical structure, we construct a router that maps the main reference feature to a soft distribution over experts. Only the top-1 expert is activated during inference for efficiency, while all experts are trained jointly. This sparse dynamic routing strategy allows the model to flexibly choose the most suitable reference feature based on the input content, enhancing the model's temporal modeling capacity without increasing inference complexity. Moreover, enforcing top-1 expert activation ensures consistency between training and inference, which avoids ambiguity in expert assignments and simplifies the computational pipeline.

Reference Balance Loss. To avoid overly rigid reference selection and enable an efficient temporal prediction chain in the reference loop, we introduce a reference balance loss that promotes less biased expert routing. The supervision preserves the reference modeling capacity to explore varied reference structures during the training. Inspired by the auxiliary regularization strategy in (Fedus, Zoph, and Shazeer 2022), the reference balance loss $\mathcal{L}_{\text{balance}}$ is designed as:

$$\mathcal{L}_{\text{balance}} = \frac{1}{B}\sum_{b=1}^{B}\left(\sum_{e=1}^{n}\text{softmax}_e(r_b)\cdot\text{topK}_e(r_b)\right)\cdot n, \quad (1)$$

where B is the batch size, n is the number of experts, and r_b denotes the b-th sample in the batch of routing terms. $\text{softmax}_e(r_b)$ is the routing probability assigned to expert e by the gating function, and $\text{topK}_e(r_b)$ is an indicator function that is 1 if expert e is selected in the top-K experts for sample b, and 0 otherwise. The multiplication by n serves as a scaling factor to ensure the loss maintains a consistent range when varying the number of experts. The reference balance loss $\mathcal{L}_{\text{balance}}$ penalizes unbalanced expert usage by measuring the overlap between the routing probabilities and the top-K selected experts across the batch.

HRS-Guided Feature-wise Inter-frame Prediction

After managing the multi-frame temporal contexts embedded in the propagated reference features through the proposed HRS, we further address the temporal reference mismatch between motion estimation and motion compensation in prior contextual coding-based frameworks.

Leveraging the selected reference feature from the hierarchy-aware reference modulation module, we propose the HRS-guided feature-wise inter-frame prediction to improve the efficiency and consistency of inter-frame prediction. Specifically, the selected reference feature is first added back to the main reference feature $\tilde{F}^m_{t-1}$ to enhance the spatial structures and temporal coherence. The residual-enhanced reference feature is then projected through a linear layer to produce a compact 3-channel low-rank representation $\overline{F}_{t-1}$ as input to the motion estimation module:

$$\overline{F}_{t-1} = \mathcal{P}_{\mathcal{C}\to 3}\left(\sum_{i=1}^{n} G(\tilde{F}^m_{t-1})\cdot E_i(\tilde{F}^m_{t-1}, \tilde{F}^g_{t-1}) + \tilde{F}^m_{t-1}\right), \quad (2)$$

where $G(\cdot)$ and $E_i(\cdot)$ represent the learned routing function and the i-th expert, respectively. $\mathcal{P}_{\mathcal{C}\to 3}$ represents a linear projection layer that reduces the feature dimensionality from C to 3 channels. This low-rank approximation compresses the residual-enhanced reference feature into a compact 3-channel feature $\overline{F}_{t-1}$, which not only enables seamless integration into the motion estimation module, but also retains

	UVG	MCL-JCV	HEVC B	HEVC C	HEVC D	HEVC E	USTC-TD	Average
VTM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DCVC	314.1	185.1	291.7	266.9	224.6	2083.2	221.2	512.4
DCVC-TCM	109.8	69.9	106.5	132.2	81.3	492.7	99.3	156.0
DCVC-HEM	23.8	10.5	27.5	33.4	11.3	139.1	24.2	38.5
SDD	19.1	5.7	30.7	35.3	–6.7	234.6	15.4	47.7
DCVC-DC	–7.8	–10.0	–0.9	6.1	–15.0	33.3	9.4	2.2
DCVC-FM	–19.8	–10.9	–12.2	–12.9	–27.1	–27.7	16.3	–13.5
DCMVC	–25.2	–20.9	–17.8	–18.0	**–31.4**	–23.5	**–6.3**	–20.4
Ours	**–36.2**	**–24.0**	**–22.6**	**–24.2**	–31.3	**–38.8**	–5.0	**–26.0**

Note: The quality indexes of DCVC-FM are set as 36, 45, 54, 63 to match the bit-rate range of DCVC-DC.

Table 1: BD-Rate (%) comparison in RGB colorspace measured with PSNR. **All frames with intra-period=-1.**

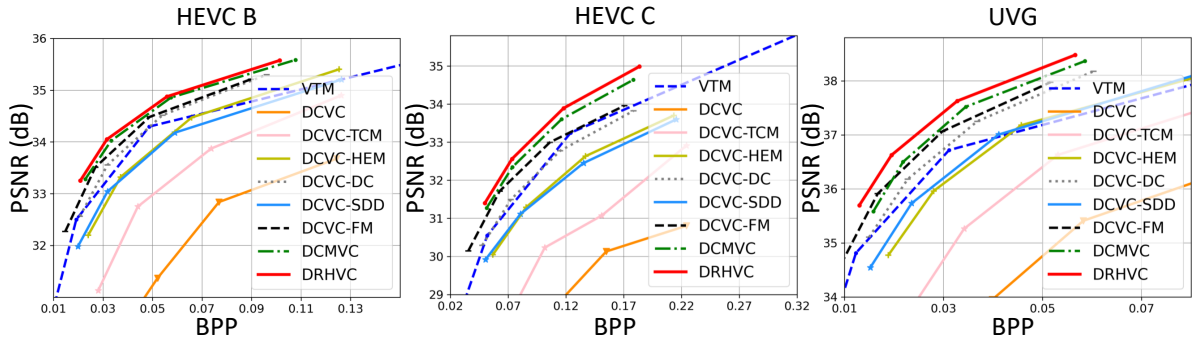

Figure 4: Rate and distortion curve for HEVC HEVC Class B, C, and UVG datasets. The comparison is in RGB colorspace measured with PSNR, and the **intra-period is set as -1 for all frames.**

essential multi-frame temporal dynamics without introducing significant computational complexity.

The proposed HRS-guided feature-wise inter-frame prediction unifies the reference structure for motion estimation and compensation based on multi-frame temporal contexts, which effectively reduces temporal inconsistencies and improves the inter-frame prediction efficiency. Together with the proposed HRS, it forms a unified reference hierarchy for contextual coding-based frameworks. Unlike prior schemes that rely on the single-frame reference loop, our scheme integrates learned reference structure into the reference propagation mechanism to form a multi-frame feature-domain reference loop, which enables more efficient and temporally consistent inter-frame prediction.

Training Loss

We optimize our scheme in an end-to-end manner. The overall training loss of the proposed video compression framework consists of the rate-distortion (RD) loss and the proposed reference balance loss, which is defined as:

$$\mathcal{L} = \lambda \cdot \mathcal{D} + \beta \cdot \mathcal{L}_{\text{balance}} + \mathcal{R}, \tag{3}$$

where $\mathcal{D}$ denotes the distortion between the input and the reconstructed frames, and $\mathcal{R}$ denotes the bitrate required to encode the frame. The coefficients λ and β are hyperparameters that control the weights of the distortion term $\mathcal{D}$ and the reference balance loss $\mathcal{L}_{\text{balance}}$, respectively.

Experiments

Experimental Setup

Datasets. Following the setting of (Tang et al. 2025), we utilize the Vimeo-90k (Xue et al. 2019) for training with 7-frame sequences, and further finetune the models with 32-frame sequences collected from the raw Vimeo videos[1]. To ensure compatibility with the latest NVCs, we evaluate our model on widely adopted benchmarks, including UVG (Mercat, Viitanen, and Vanne 2020), MCL-JCV (Wang et al. 2016), and HEVC Class B, C, D, and E datasets (Bossen 2013), as well as the recently proposed and highly regarded USTC-TD dataset (Li et al. 2025).

Training. Following (Li, Li, and Lu 2023), four base λ values (85, 170, 380, 840) are used to control the rate-distortion (RD) trade-off, and the hyperparameter β is set to 0.01 to weight the proposed reference balance loss. Furthermore, the number of experts n employed in the hierarchy-aware reference modulation is set to 3, based on a trade-off between compression performance and computational complexity. Comprehensive ablation studies on the selection of β and n are provided in the supplementary material. All the training processing is conducted on 8 NVIDIA Ampere Tesla A40 GPUs, with Forward Recomputation Backpropa-

[1]http://toflow.csail.mit.edu

	UVG	MCL-JCV	HEVC B	HEVC C	HEVC D	HEVC E	USTC-TD	Average
VTM	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**	0.0
DCVC	133.9	106.6	119.6	152.5	110.9	274.8	130.4	147.0
DCVC-TCM	23.1	38.2	32.8	62.1	29.0	75.8	73.5	47.8
DCVC-HEM	–17.2	–1.6	–0.7	16.1	–7.1	20.7	21.5	4.5
SDD	–19.7	–7.1	–13.7	–2.3	–24.9	–8.4	6.6	–9.9
DCVC-DC	–25.9	–14.4	–13.9	–8.8	–27.7	–19.1	7.6	–14.6
DCVC-FM	–20.4	–8.1	–10.3	–8.4	–25.8	–21.9	19.6	–10.8
DCMVC	–30.6	–17.3	–14.5	–14.4	–31.6	–28.1	0.9	–19.4
Ours	**–33.5**	**–19.1**	**–18.9**	**–17.8**	**–33.4**	**–30.6**	1.7	**–21.7**

Note: The quality indexes of DCVC-FM are set as 36, 45, 54, 63 to match the bit-rate range of DCVC-DC.

Table 2: BD-Rate (%) comparison in RGB colorspace measured with PSNR. **96 frames with intra-period=32.**

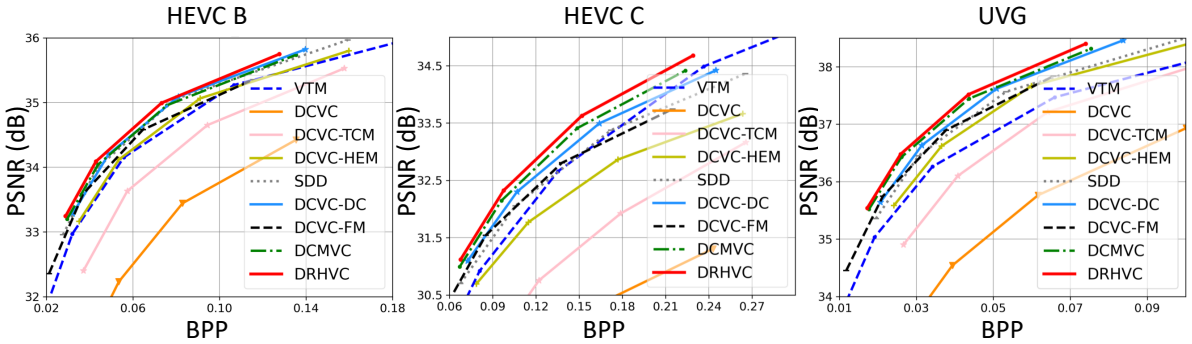

Figure 5: Rate and distortion curve for HEVC Class B, C, and UVG datasets. The comparison is in RGB colorspace measured with PSNR, and the **intra-period is set as 32 for 96 frames.**

gation (FRB)[2] employed for CUDA memory management.

Testing. We performed evaluations under low-delay scenarios using two common configurations with different intra-period (IP) settings: IP32 over 96 frames and IP-1 for all frames in RGB format. The proposed scheme DRHVC is compared with the traditional codec VTM/VVC (Bross et al. 2021a) as well as NVCs DCVC (Li, Li, and Lu 2021), DCVC-TCM (Sheng et al. 2022), DCVC-HEM (Li, Li, and Lu 2022), SDD (Sheng et al. 2024), DCVC-DC (Li, Li, and Lu 2023), DCVC-FM (Li, Li, and Lu 2024), and DCMVC (Tang et al. 2025) to validate its effectiveness.

Comparisons with Previous SOTA Methods

IP-1 Setting. Table 1 reports the BD-Rate (%) results in the RGB format under the challenging IP-1 setting with all frames. Our DRHVC achieves the best compression performance in long-range prediction scenarios, with an average bitrate saving of 26%, significantly surpassing DCMVC (20.4% saving) and DCVC-FM (13.5% saving). The RD curves in Fig. 4 further highlight the significant performance margin of our proposed methods over other schemes. It is worth noting that in long prediction chains, rich temporal contexts accumulate in the propagated reference features. However, previous schemes that rely on the single-frame reference loop suffer from an information bottleneck, limiting the effective exploitation of multi-frame temporal dependencies. In contrast, our DRHVC enables efficient utilization of these temporal contexts by adaptively selecting optimal reference features through the proposed HRS, leading to substantial performance gains.

IP32 Setting. We also provide the BD-Rate (%) comparison results in the RGB format under the IP32 setting with 96 frames in Table 2. Our proposed DRHVC consistently achieves the best compression performance, delivering an average bitrate saving of 21.7% compared to the traditional codec VTM. In addition, the RD curves in Fig. 5, evaluated on three benchmark datasets, further confirm the superiority of our scheme. Additional results for both configurations are provided in the supplementary material.

[2]https://qywu.github.io/2019/05/22/explore-gradient-checkpointing.html

	M_a	M_b	M_c	M_d	M_e	M_f
FIP		✓	✓	✓		✓
HRS			✓	✓		✓
Reference balance loss				✓		✓
Long-sequence training					✓	✓
BD-Rate (%)	-0.0	-3.9	-7.4	-8.8	-4.3	-13.4

Table 3: Ablation study on main techniques (%)

Ablation Study

Table 3 presents the ablation study of each proposed component, with the evaluation measured by average bitrate sav-

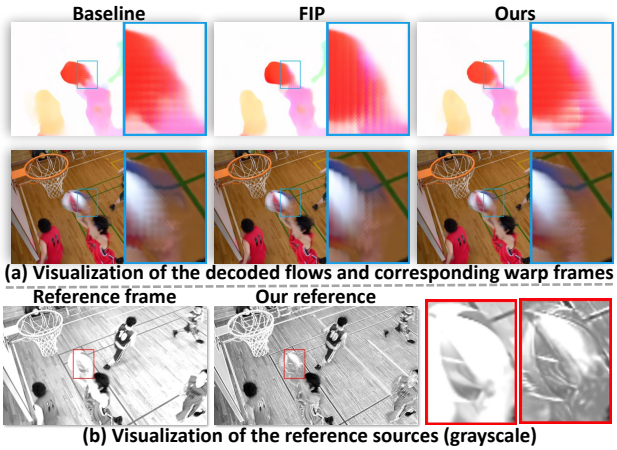

Figure 6: (a) Visualization of the decoded optical flows and the corresponding warped frames for the baseline DCVC-DC, feature-wise inter-frame prediction (FIP), and the proposed reference hierarchy. (b) Grayscale visualizations of the original reference frame and the low-rank reference enhancement feature generated by the reference hierarchy.

	MACs	Params	Encoding Time	Decoding Time
DCVC-DC	2786G	19.78M	663ms	557ms
DCVC-FM	2354G	18.57M	587ms	495ms
SDD	4775G	21.77M	968ms	775ms
DCMVC	4133G	20.98M	932ms	810ms
DRHVC	2953G	19.86M	697ms	558ms

Note: Tested on NVIDIA 3090 with 1080p sequences.

Table 4: Complexity comparison.

ings on the HEVC Class datasets under IP32 configuration. Our study uses the reproduced DCVC-DC (denoted as M_a) as our baseline model. HRS denotes the hierarchical reference structure, while FIP refers to the feature-wise inter-frame prediction module without HRS guidance.

Feature-wise Inter-frame Prediction. M_b denotes the baseline model enhanced with feature-wise inter-frame prediction (FIP), where the reference features undergo a low-rank channel mapping through a single convolutional layer. Compared to M_a, the M_b model achieves a 3.9% reduction in BD-rate, demonstrating the potential of adopting a feature-domain reference loop. As shown in Fig. 6 (a), comparing the baseline model, the decoded flow through FIP exhibits higher quality boundaries of moving objects, indicating more accurate inter-frame prediction.

Hierarchical Reference Structure. M_c denotes the model M_b further enhanced by integrating the hierarchical reference structure (HRS) without reference balance loss to guide the FIP, thereby forming the network structure of the proposed reference hierarchy. As shown in Fig. 6 (b), under the HRS guidance, the determined low-rank reference enhancement feature contains more structural details and temporal information through multi-frame temporal contexts management. Compared to M_b, M_c achieves an additional 3.5% bitrate saving. Furthermore, benefiting from a more informative reference source, the decoded flow shown in Fig. 6 (a) further refines the motion boundaries, resulting in higher-quality inter-frame prediction compared to FIP alone.

Training Strategies. Besides improvements to the network structure, M_d denotes M_c further enhanced by incorporating the proposed reference balance loss. As shown in Table 3, M_d achieves an additional 1.4% BD-rate reduction over M_c, confirming the effectiveness of the reference balance supervision in facilitating a more efficient temporal prediction chain within the reference loop. In the M_e model, directly applying long-sequence training (32 frames) to the baseline model (M_a) results in a 4.3% BD-rate reduction. In contrast, M_f, which applies long-sequence training to our proposed reference hierarchy (M_d), achieves a 4.6% reduction over M_d. This indicates that our proposed reference hierarchy not only improves its own compression efficiency but also amplifies the benefits of long-sequence training. The synergy between the two methods contributes to more efficient reference structure modeling.

Complexity Analysis

We conduct a complexity analysis in terms of MACs, network parameters, encoding time, and decoding time, as summarized in Table 4. Compared to SDD (Sheng et al. 2024) and DCMVC (Tang et al. 2025), we have saved 38.2% and 28.6% computational complexity in terms of MACs, respectively. Compared to DCVC-DC (Li, Li, and Lu 2023), our method introduces only 6.0% higher encoding complexity with no additional decoding overhead, while achieving a 28.2% average performance gain. Compared to DCVC-FM (Li, Li, and Lu 2024), which incorporates dedicated complexity optimization, our scheme still exhibits room for further improvement. Nevertheless, considering the significant compression gain achieved (a 12.5% bitrate saving over DCVC-FM), the slight increase in complexity is both acceptable and worthwhile, offering a good balance between computational complexity and compression ratio.

Conclusion

In this paper, we propose the unified reference hierarchy that integrates the learned multi-frame reference structure into the inherent reference propagation mechanism for contextual coding-based NVCs. The hierarchical reference structure (HRS) manages diverse temporal contexts to determine the optimal reference feature, and the HRS-guided feature-wise inter-frame prediction learns its low-rank approximation to further improve the inter-frame prediction accuracy without introducing extra computational overhead. Our DRHVC achieves SOTA compression performance among NVCs with acceptable computational complexity.

Our investigation not only forms a multi-frame feature-domain reference loop but also mitigates the long-standing mismatch between motion estimation and motion compensation. In future work, more advanced motion representations, like implicit motion modeling, can be explored to further exploit the potential of the reference hierarchy in inter-frame prediction.

Acknowledgments

This work was supported in part by the National Key Research and Development Plan under Grant 2024YFF0505702. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC.

References

Agustsson, E.; Minnen, D.; Johnston, N.; Balle, J.; Hwang, S. J.; and Toderici, G. 2020. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8503–8512.

Bian, Y.; Sheng, X.; Li, L.; and Liu, D. 2024. LSSVC: A Learned Spatially Scalable Video Coding Scheme. *IEEE Transactions on Image Processing*.

Bian, Y.; Tang, C.; Li, L.; and Liu, D. 2025. Augmented Deep Contexts for Spatially Embedded Video Coding. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2094–2104.

Bossen, F. 2013. Common HM test conditions and software reference configurations (JCTVC-L1100). *Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG*.

Bross, B.; Chen, J.; Ohm, J.-R.; Sullivan, G. J.; and Wang, Y.-K. 2021a. Developments in international video coding standardization after avc, with an overview of versatile video coding (VVC). *Proceedings of the IEEE*.

Bross, B.; Wang, Y.-K.; Ye, Y.; Liu, S.; Chen, J.; Sullivan, G. J.; and Ohm, J.-R. 2021b. Overview of the versatile video coding (VVC) standard and its applications. *IEEE Transactions on Circuits and Systems for Video Technology*.

Chen, J.; Kao, S.-h.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; and Chan, S.-H. G. 2023. Run, don't walk: chasing higher FLOPS for faster neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12021–12031.

Chen, Z.; Zhou, L.; Hu, Z.; and Xu, D. 2024. Group-aware Parameter-efficient Updating for Content-Adaptive Neural Video Compression. *arXiv preprint arXiv:2405.04274*.

Fedus, W.; Zoph, B.; and Shazeer, N. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120): 1–39.

Ho, Y.-H.; Chang, C.-P.; Chen, P.-Y.; Gnutti, A.; and Peng, W.-H. 2022. Canf-vc: Conditional augmented normalizing flows for video compression. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, 207–223. Springer.

Hu, Z.; Lu, G.; Guo, J.; Liu, S.; Jiang, W.; and Xu, D. 2022. Coarse-to-fine deep video coding with hyperprior-guided mode prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5921–5930.

Hu, Z.; Lu, G.; and Xu, D. 2021. FVC: A New Framework towards Deep Video Compression in Feature Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1502–1511.

Hu, Z.; and Xu, D. 2023. Complexity-guided slimmable decoder for efficient deep video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 14358–14367.

Huo, S.; Liu, D.; Li, B.; Ma, S.; Wu, F.; and Gao, W. 2020. Deep network-based frame extrapolation with reference frame alignment. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3): 1178–1192.

Jiang, W.; Li, J.; Zhang, K.; and Zhang, L. 2025. Ecvc: Exploiting non-local correlations in multiple frames for contextual video compression. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 7331–7341.

Li, H.; et al. 2012. Rate-distortion optimized reference picture management for high efficiency video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(12): 1844–1857.

Li, J.; Li, B.; and Lu, Y. 2021. Deep contextual video compression. *Advances in Neural Information Processing Systems (NeurIPS)*, 34: 18114–18125.

Li, J.; Li, B.; and Lu, Y. 2022. Hybrid spatial-temporal entropy modelling for neural video compression. In *Proceedings of the 30th ACM International Conference on Multimedia*, 1503–1511.

Li, J.; Li, B.; and Lu, Y. 2023. Neural video compression with diverse contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22616–22626.

Li, J.; Li, B.; and Lu, Y. 2024. Neural video compression with feature modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 26099–26108.

Li, M.; Shi, Y.; Wang, J.; and Huang, Y. 2023. High Visual-Fidelity Learned Video Compression. In *Proceedings of the 31st ACM International Conference on Multimedia*, 8057–8066.

Li, Z.; Li, Y.; Tang, C.; Li, L.; Liu, D.; and Wu, F. 2024. Uniformly accelerated motion model for inter prediction. In *2024 IEEE International Conference on Visual Communications and Image Processing (VCIP)*, 1–5. IEEE.

Li, Z.; Liao, J.; Tang, C.; Zhang, H.; Li, Y.; Bian, Y.; Sheng, X.; Feng, X.; Li, Y.; Gao, C.; et al. 2025. USTC-TD: A test dataset and benchmark for image and video coding in 2020s. *IEEE Transactions on Multimedia*.

Lin, J.; Liu, D.; Li, H.; and Wu, F. 2018. Generative adversarial network-based frame extrapolation for video coding. In *2018 IEEE Visual Communications and Image Processing (VCIP)*, 1–4. IEEE.

Lin, J.; Liu, D.; Li, H.; and Wu, F. 2020. M-LVC: multiple frames prediction for learned video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3546–3554.

Lin, K.; Jia, C.; Zhang, X.; Wang, S.; Ma, S.; and Gao, W. 2022. DMVC: Decomposed motion modeling for learned video compression. *IEEE Transactions on Circuits and Systems for Video Technology*.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Lu, G.; Ouyang, W.; Xu, D.; Zhang, X.; Cai, C.; and Gao, Z. 2019. Dvc: An end-to-end deep video compression framework. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, 11006–11015.

Lu, G.; Zhang, X.; Ouyang, W.; Chen, L.; Gao, Z.; and Xu, D. 2020. An end-to-end learning framework for video compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; and Fu, Y. 2024. Rewrite the stars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5694–5703.

Mercat, A.; Viitanen, M.; and Vanne, J. 2020. UVG dataset: 50/120fps 4K sequences for video codec analysis and development. In *Proceedings of the 11th ACM Multimedia Systems Conference*, 297–302.

Qi, L.; Li, J.; Li, B.; Li, H.; and Lu, Y. 2023. Motion information propagation for neural video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6111–6120.

Sheng, X.; Li, J.; Li, B.; Li, L.; Liu, D.; and Lu, Y. 2022. Temporal context mining for learned video compression. *IEEE Transactions on Multimedia*.

Sheng, X.; Li, L.; Liu, D.; and Li, H. 2024. Spatial Decomposition and Temporal Fusion based Inter Prediction for Learned Video Compression. *IEEE Transactions on Circuits and Systems for Video Technology*.

Shi, Y.; Ge, Y.; Wang, J.; and Mao, J. 2022. Alphavc: High-performance and efficient learned video compression. In *European Conference on Computer Vision (ECCV)*, 616–631. Springer.

Tang, C.; Li, Z.; Bian, Y.; Li, L.; and Liu, D. 2025. Neural Video Compression with Context Modulation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 12553–12563.

Tang, C.; Sheng, X.; Li, Z.; Zhang, H.; Li, L.; and Liu, D. 2024. Offline and Online Optical Flow Enhancement for Deep Video Compression. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 38, 5118–5126.

Wang, H.; Gan, W.; Hu, S.; Lin, J. Y.; Jin, L.; Song, L.; Wang, P.; Katsavounidis, I.; Aaron, A.; and Kuo, C.-C. J. 2016. MCL-JCV: a JND-based H.264/AVC video quality assessment dataset. In *2016 IEEE International Conference on Image Processing (ICIP)*, 1509–1513. IEEE.

Wang, Z.; Yan, Z.; Pan, J.; Gao, G.; Zhang, K.; and Yang, J. 2025. DORNet: A Degradation Oriented and Regularized Network for Blind Depth Super-Resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 15813–15822.

Wiegand, T.; Zhang, X.; and Girod, B. 1999. Long-term memory motion-compensated prediction. *IEEE Transactions on Circuits and Systems for Video Technology*, 9(1): 70–84.

Xue, T.; Chen, B.; Wu, J.; Wei, D.; and Freeman, W. T. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision (IJCV)*, 127(8): 1106–1125.

Zamfir, E.; Wu, Z.; Mehta, N.; Zhang, Y.; and Timofte, R. 2024. See more details: Efficient image super-resolution by experts mining. In *Forty-first International Conference on Machine Learning (ICML)*.