Learned Rate Control for Frame-Level Adaptive Neural Video Compression via Dynamic Neural Network

Chenhao Zhang[1] and Wei Gao[1,2⋆]

[1] SECE, Shenzhen Graduate School, Peking University, Shenzhen, China
[2] Peng Cheng Laboratory, Shenzhen, China
chenhaozhang@stu.pku.edu.cn, gaowei262@pku.edu.cn

Abstract. Neural Video Compression (NVC) has achieved remarkable performance in recent years. However, precise rate control remains a challenge due to the inherent limitations of learning-based codecs. To solve this issue, we propose a dynamic video compression framework designed for variable bitrate scenarios. First, to achieve variable bitrate implementation, we propose the Dynamic-Route Autoencoder with variable coding routes, each occupying partial computational complexity of the whole network and navigating to a distinct RD trade-off. Second, to approach the target bitrate, the Rate Control Agent estimates the bitrate of each route and adjusts the coding route of DRA at run time. To encompass a broad spectrum of variable bitrates while preserving overall RD performance, we employ the Joint-Routes Optimization strategy, achieving collaborative training of various routes. Extensive experiments on the HEVC and UVG datasets show that the proposed method achieves an average BD-Rate reduction of 14.8% and BD-PSNR gain of 0.47dB over state-of-the-art methods while maintaining an average bitrate error of 1.66%, achieving Rate-Distortion-Complexity Optimization (RDCO) for various bitrate and bitrate-constrained applications.

Keywords: Neural Video Compression · Rate Control · Rate-Distortion-Complexity Optimization

1 Introduction

Nowadays, video content accounts for over 80% of internet traffic [1], underscoring the critical importance of flexible video compression techniques to handle the enormous data volume. Traditional video compression standards [4, 29, 33] have

⋆ Corresponding Author: Wei Gao. This work was supported by The Major Key Project of PCL (PCL2024A02), Natural Science Foundation of China (62271013, 62031013), Guangdong Province Pearl River Talent Program (2021QN020708), Guangdong Basic and Applied Basic Research Foundation (2024A1515010155), Shenzhen Science and Technology Program (JCYJ20230807120808017), and Sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community (CAAIXSJLJJ-2023-MindSpore07).

been pivotal, incorporating technologies like inter-frame prediction, transform coding, and motion estimation. These methods employ adaptable rate control models to maximize reconstruction quality within the constraints of bandwidth or storage capacities.

Recent advancements in Neural Video Compression (NVC) [2, 5, 8, 9, 13–15, 20, 24, 26, 27, 35, 41] have demonstrated significant rate-distortion (RD) improvements over traditional video compression methods, primarily attributed to their enhanced non-linear transformation capabilities for better mining spatial and temporal redundancy. However, a notable limitation of NVC methods is their training for a specific RD trade-off with a fixed value of λ, which restricts their adaptability in scenarios with bitrate constraints or varying bitrate requirements.

Some works are proposed to solve this issue. These methods mainly adopt two strategies: multi-granularity quantization [14, 15, 42] and feature modulating [18, 28]. The former is a neural network-based strategy that facilitates adaptive quantization by varying the quantization steps across different dimensions in the latent space. This approach allows for a smooth adjustment of bitrate by manipulating the global quantization step. The latter leverages a set of scaling networks to modulate internal feature maps along channel dimension, with variable bitrate achieved by tuning hyper-parameter λ that controls the scaling networks. Both strategies allow for variable bitrate adjustment by either quantization step scaling or feature map modulating. Nevertheless, the final RD performance performed by linear adjustment is sub-optimal when compared to end-to-end non-linear transform coding. Furthermore, they are limited in providing precise rate control due to the absence of rate estimation method, which is crucial for accurately achieving the targeted bitrate. Li *et al*. [16] focused on developing mathematical models that correlate the output bitrate with hyper-parameter λ. However, it relies on previous experiences to iteratively refine its parameters, resulting in low bitrate accuracy, particularly in scenarios where the content of the sequence changes swiftly.

Therefore, in this paper, we introduce an end-to-end coding framework designed to achieve adaptive rate control. The main components in the framework are Dynamic-Route Autoencoder (DRA) and Rate Control Agent (RCA). The DRA incorporates a slimmable autoencoder that navigates video frames through various coding routes, each representing a subset of the overall architecture and leading to a distinct RD trade-off. These routes leverage auto-regressive coding [7, 23] to efficiently explore channel-wise redundancy. To enable precise rate control, we introduce the RCA, a neural network-based agent. For each frame, the agent is tasked with predicting the bitrate for each route. It selects the optimal route for the DRA to achieve the target bitrate, utilizing a sliding window algorithm for this process. Moreover, to achieve a broad spectrum of bitrate and maintain superior RD performance, we propose the Joint-Routes Optimization strategy that collaboratively trains these routes to their corresponding diverging points of the optimal RD curve. Compared to multi-granularity quantization and feature modulating, our method achieves adaptive complexity allocation that hierarchically distributes computational complexity from the lowest-bitrate

route to the highest-bitrate route, enabling an efficient management of computational resources. Experiments on HEVC and UVG datasets demonstrate that the proposed method achieves an average bitrate reduction of 13.1% over leading benchmarks, while maintaining an average bitrate error of 1.66%, outperforming current state-of-the-art NVC rate control methods by approximately two-fold.

The contributions of our work are as follows:

- To the best of our knowledge, this is the first time to achieve Rate-Distortion-Complexity Optimization in the realm of NVC. Our framework dynamically allocates computational complexity while retaining superior RD performance, which is flexible in adjusting bitrates and computational resources in various bitrate applications.
- To realize variable bitrate selection, we propose the Dynamic-Route Autoencoder that navigates the current frame to coding routes with distinct RD trade-offs. To achieve precise rate control, the proposed Rate Control Agent content-adaptively evaluates the bitrates of each route, determining the optimal route via sliding window algorithm.
- To achieve a broad spectrum of bitrate while preserving superior joint RD performance, we propose the Joint-Routes Optimization strategy that iteratively trains each route towards its specific diverging point from the optimal RD curve by decaying the corresponding λ.

2 Related Works

2.1 Neural Video Compression

Video compression, a critical area of study for decades, leverages core technologies like inter-frame prediction, transform coding, and motion estimation to reduce temporal and spatial redundancy. Foundational codecs such as AVC (H.264) [33], HEVC (H.265) [29], and VVC (H.266) [4] have evolved to significantly improving compression efficiency.

The rise of deep learning in multimedia has led to the advent of NVC, with pioneering frameworks like that of Lu *et al.* [20], which adapted traditional video compression modules into neural network counterparts, achieving comparable RD performance to HEVC. Agustsson *et al.* [2] introduced scale-space flow for enhanced residual coding, while Lin *et al.* [17] utilized multiple reference frames and motion vector refinement networks to reduce residual redundancy. Moving beyond traditional residual coding, Li *et al.* [13–15, 26] explored conditional coding to exploit temporal redundancy more effectively, demonstrating that conditional entropy $H(x_t|x_{ref})$ is no more than residual entropy [11]. Shi *et al.* [27] proposed the pixel-to-feature motion prediction algorithm, gaining superior RD performance over VVC. Additionally, Ho *et al.* [8] and Chen *et al.* [5] introduced Conditional Augmented Normalizing Flow (CANF) for advanced conditional inter-frame coding.

2.2 Rate Control

Rate control is crucial in the realm of video compression to optimize video quality within bandwidth or storage limits [6, 12, 16, 19, 21, 28, 36, 40]. The primary goal of rate control is to minimize distortion loss at the Group of Pictures (GoP) level under the bit constraint R_c as:

$$\min \sum_{i}^{N} R_i + \lambda_i D_i, s.t. \sum_{i}^{N} R_i < R_c, \tag{1}$$

where N is the total number of frames within a GoP, D_i and R_i are the distortion loss and bitrate for the i-th frame, and λ_i is the associated Lagrange multiplier. Traditional rate control methods such as R-Q [21], R-ρ [19], and R-λ [12] models each correlates bitrate with controllable variables of video encoding to balance RD performance.

In the realm of learning-based compression, the model rigidity and complex training parameters pose challenges for variable bitrate applications. Some Neural Image Compression (NIC) methods achieve variable bitrate by applying rate allocation. Cui *et al*. [6] allocates relatively more bits to critical channels, while Song *et al*. [28] achieves rate allocation with a pixel-wise quality map. In NVC, Xu *et al*. [36] utilized Semi-Amortized Variational Inference (SAVI) for iterative latent variable updating. However, the iterative adjustments of bitrate limit their applicability in real-time scenarios due to time constraints.

To achieve real-time rate control, Li *et al*. [16] proposed the R-D-λ model, relating bitrate constraints with compression model parameters. Despite low time-latency, the R-D-λ model made compromise assumptions to achieve an analytical solution. Moreover, it highly depends on coding experiences, hindering its accuracy in scenarios where sequence content rapidly changes.

2.3 Dynamic Neural Network

Dynamic Neural Networks (DNNs) represent a versatile category of neural networks capable of adjusting their architecture [31, 38] or parameters [25, 32] dynamically in response to input data. This adaptability allows DNNs to modify their depth, width, or connectivity based on the input's complexity, context, or content, offering a more flexible approach to data processing. In the realm of NIC and NVC, Yang *et al*. [39] proposed a slimmable autoencoder that transmits images to different RD trade-offs. Tao *et al*. [30] proposed a dynamic autoencoder to facilitate content-adaptive model capacity. Hu and Xu [10] leveraged a slimmable decoder to achieve adaptive decoding complexity. However, the efficient training of joint routes and the precise rate control method remain unresolved challenges.

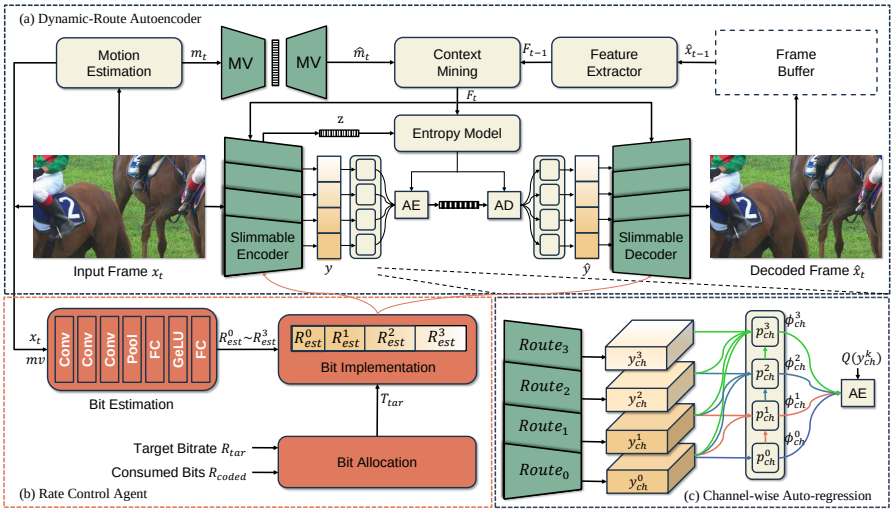

Fig. 1: Overview of the proposed method.

3 Methodology

3.1 Overview

To achieve precise rate control while retaining optimal RD performance in a single model, we propose a dynamic NVC framework with multiple coding routes. The overall architecture is shown in Fig. 1. Specifically, RCA first estimates the bitrate of each coding route and determines the optimal route k for current frame x_t to approach target bitrate R_{tar}. Then, on the encoder side, x_t is transformed into a C_k-channel latent representation $y_{0:C_k} \in \mathbb{R}^{H \times W \times C_k}$ via route k with the help of temporal context x_{ref}. After that, $y_{0:C_k}$ is entropy encoded to bitstream via channel-wise auto-regression. On the decoder side, the bitstream is entropy decoded to latent representation $\hat{y}_{0:C_k}$ following the same auto-regressive procedure as entropy encoding. Finally, decoded frame $\hat{x}_t$ is reconstructed by the decoder via the same route k. To facilitate a broad bitrate spectrum while achieving global optimal RD performance, the coding routes are jointly trained by Joint-Routes Optimization strategy.

3.2 Dynamic-Route Autoencoder

The fundamental components of DRA are slimmable operators, which transfer input feature into output feature with dynamic channels, as shown in Fig. 2. Slimmable operators allow for dynamic adaptation of the *supernet* into a series of overlapping sub-networks, or *subnets*. Each subnet embodies a distinct, end-to-end coding route, where slimmable operators facilitate the transformation of a singular input frame into various latent representations. The adaptability in

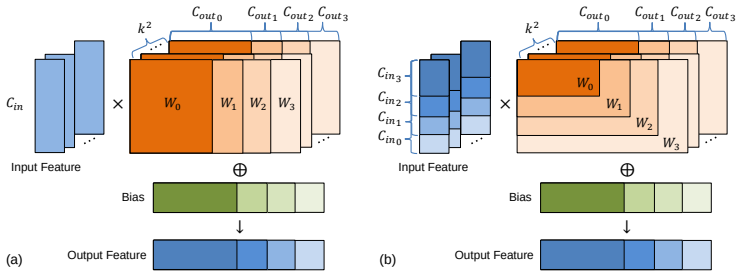

Fig. 2: Illustration of Slimmable Convolution Operators: (a) Conversion from fixed input channels C_{in} to variable output channels $\{C_{out0}, C_{out1}, C_{out2}, C_{out3}\}$; (b) Conversion from variable input channels $\{C_{in0}, C_{in1}, C_{in2}, C_{in3}\}$ to variable output channels $\{C_{out0}, C_{out1}, C_{out2}, C_{out3}\}$

.

channel number empowers the framework with the capacity for variable computational complexity and rate selection. For route k, the coding pipeline can be formulated as follows:

$$y_{ch}^{\leq k} = g_a(x_t | x_{ref}; \theta^{\leq k}), \tag{2}$$

$$\Phi_{ch}^{k} = p_{ch}(Q(FM(y_{ch}^{\leq k-1})), \hat{z}), \tag{3}$$

$$\hat{x}_t = g_s(\hat{y}_{ch}^{\leq k}; \phi^{\leq k}), \tag{4}$$

where $g_a(\cdot;\theta)$ and $g_s(\cdot;\phi)$ represent the contextual encoder and decoder equipped with slimmable parameters θ and ϕ, respectively, $p_{ch}(\cdot)$ denotes the channel-wise auto-regression model that predicts entropy parameters Φ_{ch} with the hyperprior $\hat{z}$, $Q(\cdot)$ refers to the quantization step, and $FM(\cdot)$ is the Feature Modulation network. The current input frame x_t first passes through slimmable encoder $g_a(\cdot;\theta^{\leq k})$ with parameters $\theta^{\leq k} = \{\theta^0, \theta^1, ..., \theta^k\}$, yielding latent representation $y_{ch}^{\leq k} = \{y_{ch}^0, y_{ch}^1, ..., y_{ch}^k\}$, encapsulating frame information in a more condensed format. To fully explore channel redundancy, entropy parameters Φ_{ch}^k of $y_{ch}^{\leq k}$ are auto-regressively derived from the preceding route's output via auto-regression model p_{ch}, as shown in Fig. 1(c). To avoid the high coding latency caused by switching routes in auto-regression, we directly utilize the partial latent representation of the current route, $y_{ch}^{\leq k-1}$, as a replacement for the output of the previous route $k-1$. This approach avoids the need for re-encoding and streamlines the coding process. However, this substitution leads to a sub-optimal RD performance due to the absence of serial-routes optimization. To address this problem, we propose the Feature Modulation network $FM(\cdot)$ that refines $y_{ch}^{\leq k-1}$ to approximate the output of route $k-1$ before quantization and auto-regression.

Through entropy encoding, the current frame is transferred into a bitstream for storage or transmission. The entropy decoding procedure is consistent with that of entropy encoding, which auto-regressively restores the latent representation $\hat{y}_{ch}^{\leq k}$ via p_{ch}. Finally, the decoded frame is reconstructed by slimmable

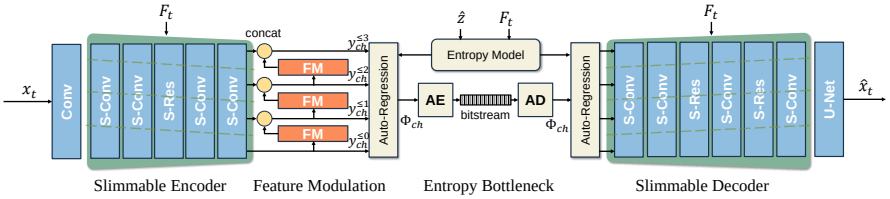

Fig. 3: Illustrations of slimmable autoencoder, including Slimmable Encoder, Feature Modulation, Entropy Bottelneck, and Slimmable Decoder. "S-X" denotes "Slimmable X".

decoder $g_s(\cdot;\phi^{\leq k})$ through the same route k. Implementation details of the slimmable autoencoder are shown in Fig. 3.

Different from other variable rate solutions that maintain stable coding complexity at each RD point, the proposed method is complexity-adaptive, allocating less computation resources to lower-bitrate routes. This reflects the insight that simpler transformations are sufficient for lower-bitrate representations to explore source correlation, enabling effective Rate-Distortion-Complexity Optimization in our framework.

3.3 Rate Control Agent

As shown in Fig. 1(b), the RCA is comprised of three foundational components: the Rate Estimation module, the Bit Allocation module, and the Bit Implementation module. Given the fact that the complexity of frame content significantly influences coded bitrates, with less complex frames requiring fewer coding bits and vice versa, we first develop a lightweight Rate Estimation module, which is designed to correlate the content of current frame x_t with its coded bitrates across various routes. To be specific, we leverage x_t and its associated reference motion vector (mv) to estimate the coded bitrates R^i_{est} of each route i. The final layer of the Rate Estimation module integrates a fully connected (FC) layer with the number of prediction heads equal to the route numbers.

For frame-level bitrate allocation, the Bit Allocation module employs the sliding window algorithm. Given a target bitrate R_{tar}, the bitrate of current frame is allocated as follows:

$$T_{tar} = \frac{R_{tar} \times (N_{coded} + SW) - R_{coded}}{SW}, \tag{5}$$

where T_{tar} is the target bitrate allocated to the current frame, N_{coded} denotes the number of frames already coded, R_{coded} is the bits consumed by already coded frames and SW is the sliding window's length, facilitating a smoother rate control process. Guided by Eq. (5), the target bitrate is expected to be achieved within SW frames, ensuring a steady bitrate allocation and consistent video quality. To mitigate potential cumulative errors, the initial frame of a GoP, known as the I frame, is exempt from this bit allocation strategy and

receives a relatively higher bitrate allocation to foster improved RD performance in subsequent frames.

Finally, the Bit Implementation module determines the optimal coding route i^* based on the relationship between bits allocated $R_{tar} \times N_{coded}$ and bits consumed R_{coded}. If more bits are allocated than consumed, it chooses the route with an estimated bitrate just above T_{tar}, defaulting to the highest-bitrate route if none match. Conversely, if fewer bits are allocated than consumed, it chooses the route with an estimated bitrate just below T_{tar}, or the lowest-bitrate route if none route is available, expressed as:

$$i^* = \begin{cases} \arg\min_i R^i_{est} - T_{tar}, s.t. R^i_{est} > T_{tar} & \text{if } R_{tar} \times N_{coded} > R_{coded} \\ \arg\min_i T_{tar} - R^i_{est}, s.t. R^i_{est} < T_{tar} & \text{if } R_{tar} \times N_{coded} < R_{coded} \end{cases} \tag{6}$$

3.4 Joint-Routes Optimization Strategy

NVC's primary aim is to identify the optimal encoding and decoding parameters that minimize the RD loss. The scenario becomes significantly more complex when it comes to dynamic-route NVC optimization due to the presence of multiple RD losses corresponding to various coding routes:

$$\theta^*, \phi^* \leftarrow \arg\min_{\theta,\phi} \sum_{i=0}^{K-1} R_i + \lambda_i D_i, \tag{7}$$

where θ^* and ϕ^* denote the optimal parameters for the encoder and decoder, respectively, R_i and D_i signify the rate and distortion losses of the route i, K is the total number of routes, and λ_i is the Lagrange multiplier for route i. Addressing this optimization is difficult due to the parameter sharing across routes and the complexity of selecting the optimal λ_i for each route. An initial training strategy is detailed in Algorithm 1.

For routes trained under fixed λ values without parameter sharing, Algorithm 1 offers an optimal training strategy. However, when it comes to joint routes with parameter sharing, pre-set fixed λ values result in local-optimal instead of global-optimal RD performance, while hierarchical parameter sharing creates a cascading effect on the performance of higher-bitrate route due to adjustments in lower-bitrate routes. Addressing this, we introduce a diverged RD model that reflects the variation in RD curve behaviors across routes, as shown in Fig. 4. This model leverages the divergence in RD curves at different bitrates, where lower-bitrate routes diverge earlier. By iteratively fine-tuning λ values for lower-bitrate routes to align with higher-bitrate RD curves, we achieve global-optimal RD performance across routes.

To advance the initial training method, we introduce the JRO strategy, outlined in Algorithm 2. This strategy fine-tunes λ_i for each route towards its diverging point, guided by a decay coefficient κ. A diverging point is identified

Algorithm 1 Initial Training Strategy

Input: training iteration N, number of routes K, λ list $[\lambda_0, \cdots, \lambda_{K-1}]$, encoder $Enc(\cdot;\theta)$, decoder $Dec(\cdot;\phi)$, train dataset χ_{train};
Output: optimal coding parameters θ^*, ϕ^*;

for $k = 0, 2, \cdots, K-1$ **do**
  for $j = 1, 2, \cdots, N$ **do**
    $I_{input}, I_{ref} \leftarrow \chi_{train}$
    $R, D \leftarrow Dec(Enc(I_{input}, I_{ref}; \theta_k); \phi_k)$;
    $Loss \leftarrow \sum_{i=0}^{K-1} R + \lambda_i D$;
    update θ_k, ϕ_k;
  end for
end for
return θ_{K-1}, ϕ_{K-1}.

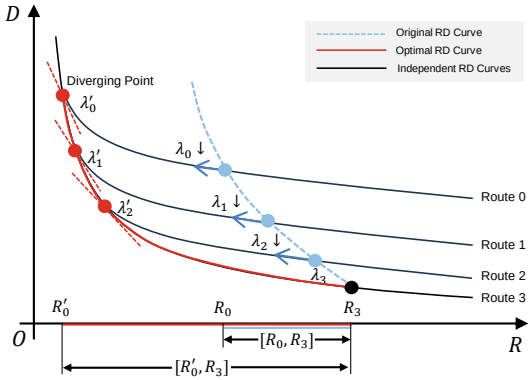

Fig. 4: Illustration of the diverged RD model and the Joint-Routes Optimization (JRO) strategy. The strategy focuses on transferring RD points (blue points) of each route to their corresponding diverging points (red points), shifting the total RD curve from the blue curve to the red curve. This transition broadens the bitrate spectrum and ensures global-optimal RD performance.

when the decline in the slope of adjacent RD points ceases. To mitigate the cascading effect from parameter sharing, adjustments start with the highest-bitrate route and proceed to the lowest. By setting a higher λ_{K-1} for pre-training and a lower λ_0 for post-training, we ensure coverage across a wide bitrate spectrum.

4 Experiments

4.1 Implementation Details

Datasets. We train our model on the Vimeo-90k dataset [37], which comprises 89,800 video clips. To ensure compatibility with the autoencoder framework,

Algorithm 2 Joint-Routes Optimization Strategy

Input: training iteration N, number of routes K, λ of highest route λ_{K-1}, decay coefficient κ, encoder $Enc(\cdot;\theta)$, decoder $Dec(\cdot;\phi)$, train dataset χ_{train}, validation dataset χ_{val};

Output: optimal coding parameters θ^*, ϕ^*;

pre-train θ_{K-1}, ϕ_{K-1} under λ_{K-1}
test $\xi_{cur} \leftarrow \frac{D_{K-1}-D_{K-2}}{R_{K-1}-R_{k-2}}$ on χ_{val}
test R_{cur} on χ_{val}
$\xi_{pre}, R_{pre} \leftarrow 0$
for $k = K-2, K-3 \cdots, 0$ **do**
  repeat
    $[\lambda_0 : \lambda_{k+1}] \leftarrow \kappa[\lambda_0 : \lambda_{k+1}]$
    $\xi_{pre} \leftarrow \xi_{cur}$
    for $j = 1, 2, \cdots, N$ **do**
      $I_{input}, I_{ref} \leftarrow \chi_{train}$
      $R, D \leftarrow Dec(Enc(I_{input}, I_{ref}; \theta_k); \phi_k)$;
      $Loss \leftarrow \sum_{i=0}^{K-1} R + \lambda_i D$;
      update θ_k, ϕ_k;
    end for
    test $\xi = \frac{D_k - D_{k+1}}{R_k - R_{k+1}}$ on χ_{val}
    $\xi_{cur}, R_{cur} \leftarrow \xi, R$
  until $\xi < \xi_{pre}$ and $R < R_{pre}$
end for
post-train by decaying λ_0
return θ_{K-1}, ϕ_{K-1}.

sequences are randomly cropped into patches of size 256×256. For evaluation, we test the performance of our algorithm on the HEVC standard test sequences [34] (Class B, C, D, E) and the UVG dataset [22].

Network Implementation. In our implementation, we set the number of routes, $K = 4$, to achieve accurate rate control across all frames. Increasing the number of routes could further improve rate control precision but at the cost of increased coding latency due to auto-regressive coding. The length of the sliding window, SW, is set to 30. The number of output channels for the Motion Vector compression network and the frame compression network are set to 64 and 96, respectively. The output channel numbers for each coding route are configured as $(24, 48, 72, 96)$.

Evaluation Metrics. The reconstruction quality is assessed using PSNR, while RD performance of different methods is compared using BD-Rate and BD-PSNR [3]. Rate control accuracy is evaluated by bitrate error $\Delta R = \frac{|R_{out}-R_{tar}|}{R_{tar}} \times 100\%$, and coding efficiency is measured by coding time of each frame $T = \frac{T_{total}}{N_{frame}}$.

Benchmark Models. To benchmark our algorithm's RD performance and rate control accuracy, we compare the RD performance against notable NVC algorithms (*i.e.,* DCVC [13], DCVC-HEM [14], DCVC-TCM [26], CANF-VC [8] and

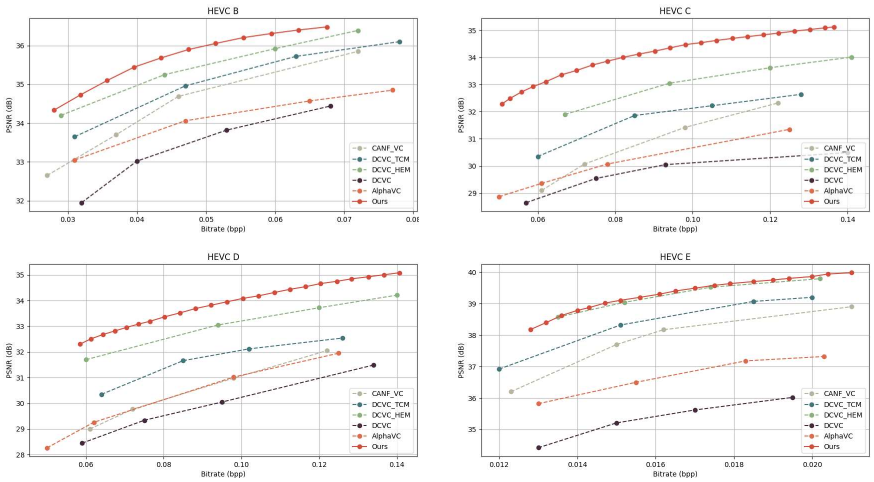

Fig. 5: RD performance comparison of different methods on HEVC datasets.

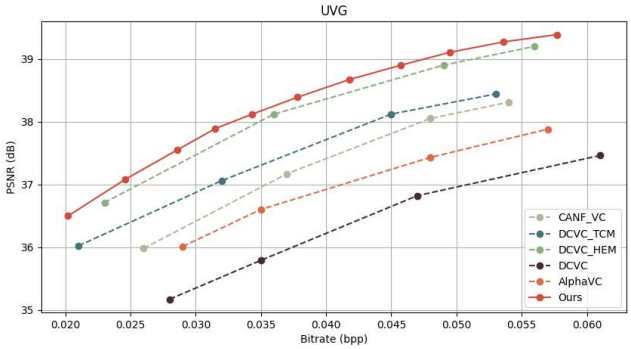

Fig. 6: RD performance comparison of different methods on UVG datasets.

AlphaVC [27]) and the rate control accuracy against the state-of-the-art R-D-λ model [16]. To ensure fairness in comparison, the number of maximum output channels of each model is set to be consistent.

4.2 Experimental Results

RD Performance. Fig. 5 and Fig. 6 illustrate the comprehensive RD performance of our proposed method. Notably, our approach achieves superior RD performance compared to the benchmark methods. Quantitative assessments, including BD-Rate comparisons, are detailed in Table 1. On average, our method facilitates an average BD-Rate reduction of 14.8% and BD-PSNR gain of 0.47dB

Table 1: RD performance comparison of different methods on HEVC and UVG datasets.

Datasets	BD-Rate(%)/BD-PSNR(dB)				
	DCVC [13]	DCVC-HEM [14]	DCVC-TCM [26]	CANF-VC [8]	alphaVC [27]
HEVC B	-57.03/2.36	-14.509/0.37	-28.718/0.87	-32.258/1.18	-53.295/1.75
HEVC C	-98.656/4.27	-36.031/1.2	-56.723/2.31	-58.767/3.22	-74.38/3.68
HEVC D	-59.914/3.81	-24.641/0.84	-52.477/2.04	-53.736/3.08	-57.022/3.04
HEVC E	-77.665/3.88	-5.137/0.13	-16.414/0.69	-29.007/1.13	-88.706/2.59
UVG	-53.555/2.36	-2.826/0.06	-21.842/0.88	-30.533/1.18	-43.795/1.59
Average	-66.700/3.11	-14.808/0.47	-33.184/1.27	-39.014/1.80	-59.443/2.34

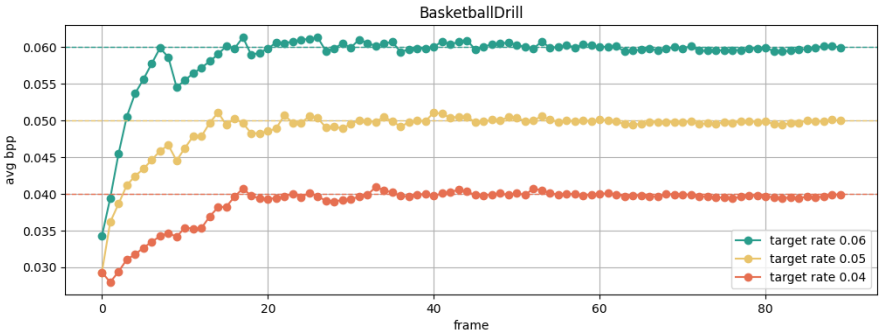

Fig. 7: Bitrate accuracy of proposed method with different target rates on *BasketballDrill* sequence

compared to DCVC-HEM, underscoring the effectiveness of our optimization strategy in enhancing video compression efficiency.

Rate Control Accuracy. Figure 7 demonstrates the rate control performance of our proposed method. In the case of specific test sequences (*BasketballDrill* as an example), the average bpp aligns with the target bpp within the span of SW frames. This performance is contrasted with the state-of-the-art R-D-λ model to highlight improvements in rate control accuracy. The quantitative results, presented in Table 2, affirm our method's capability to adhere to any specified rate within its variable range, achieving an average bitrate error of $\Delta R = 1.66\%$. This level of precision significantly surpasses that of the R-D-λ model, illustrating our approach's enhanced reliability in rate control.

Coding Complexity. In modern video transmission and processing, achieving real-time coding efficiency is crucial. Table 3 displays our method's encoding and decoding times, showing an average encoding time of about 0.28 seconds per frame. For high-definition HEVC Class B (1080p) frames, this time is increased to less than 0.5 seconds due to higher computational needs. Notably, coding time is significantly reduced for low-bitrate routes (like Route 0) thanks

Table 2: Rate control accuracy comparison of the proposed method and R-D-λ model.

Datasets	ΔR%(Ours)	ΔR%(R-D-λ model [16])
HEVC B	**1.26**	5.64
HEVC C	**0.62**	6.92
HEVC D	**2.03**	5.85
HEVC E	**1.02**	5.35
UVG	**2.65**	6.24
Average	**1.66**	6.05

Table 3: Encoding time and decoding time of proposed method on HEVC and UVG datasets.

Datasets	Route 0		Route 3	
	Enc time (s/frame)	Dec time (s/frame)	Enc time (s/frame)	Dec time (s/frame)
HEVC B	0.312	0.276	0.435	0.29
HEVC C	0.104	0.071	0.148	0.103
HEVC D	0.050	0.039	0.063	0.051
HEVC E	0.136	0.16	0.197	0.174
UVG	0.318	0.282	0.376	0.305
Average	**0.216**	**0.192**	**0.280**	**0.212**

Table 4: Complexity comparison between the proposed method and DCVC-HEM (1080p frame).

Methods	MACs	Coding Time (s)	Model Size (MB)
DCVC-HEM [14]	3.3T	0.781	67
Ours(Route 0)	2.1T(DRA)+11.73G(RCA)	0.569	
Ours(Route 1)	2.3T(DRA)+11.73G(RCA)	0.628	61.02(DRA)
Ours(Route 2)	2.5T(DRA)+11.73G(RCA)	0.686	+25.84(RCA)
Ours(Route 3)	2.8T(DRA)+11.73G(RCA)	0.742	

to decreased computational complexity, showing the complexity optimization of our method. Table 4 compares coding complexities with DCVC-HEM, revealing our method's maximum MACs are 15% lower than DCVC-HEM's. The RCA's computational load is marginal compared to the DRA, making our method's larger size a worthwhile trade-off for its superior rate control precision.

4.3 Ablation Study

Impact of RCA on Coding Time Despite the lightweight RCA incurring minimal MACs compared to the DRA, this does not fully reflect on coding time due to parallel processing of GPUs. The evaluation on the HEVC Class B dataset reveals that incorporating the RCA into the DRA framework results

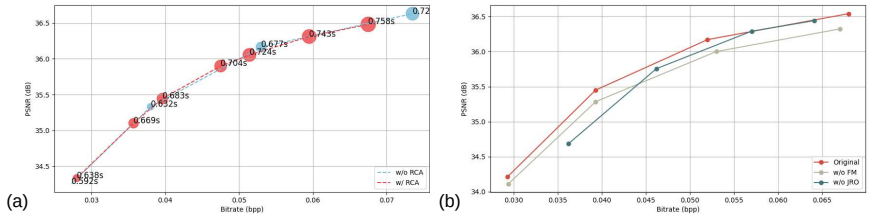

Fig. 8: Ablation study results on HEVC Class B. (a) Coding time comparison between the complete architecture and the architecture excluding RCA. (b) RD performance comparison: the original RD curve, performance without the Capacity Refinement Module, and without implementing the JRO strategy.

in a marginal increase in coding time—merely around 0.05 seconds per frame, as shown in Fig. 8. This insight underscores the proposed methods' capability to improve rate control accuracy without introducing significant time latency, making it a flexible solution for applications requiring both high coding efficiency and RD performance.

Feature Modulation network and Joint-Routes Optimization The effects of the Feature Modulation network and the JRO strategy were explored by removing the former and substituting the latter with a basic training strategy (Algorithm 1). Results shown in Fig. 8 highlight two key insights: first, RD performance benefits from the FM network more significantly at higher bitrates due to the widening capacity gap between adjacent routes as the bitrate increases; second, excluding the JRO strategy leads to a noticeable decrease in RD performance at lower bitrates and narrows the spectrum of variable bitrates. These findings underscore the essential contributions of both the FM network and JRO strategy in enhancing RD efficiency across a spectrum of bitrates.

5 Conclusion

In this work, we propose a dynamic Neural Video Compression framework combining the DRA and the RCA for precise rate control and superior RD performance. This framework adaptively guides input frames through a series of coding routes, each defined by its complexity, to achieve a diverse range of bitrate. The lightweight RCA estimates bitrates based on frame content, facilitating adaptive rate control. Through extensive experiments, our framework exhibits remarkable RD performance and state-of-the-art precision in rate control, outperforming benchmark methods across various datasets. The proposed framework adaptively balances computational complexity with RD performance, offering a versatile solution for Rate-Distortion-Complexity Optimization in various bitrate and bitrate-constrained scenarios.

References

1. Over 82% of internet traffic will be online videos by 2022. Mediamakersmeet.com (2022), https://mediamakersmeet.com
2. Agustsson, E., Minnen, D., Johnston, N., Balle, J., Hwang, S.J., Toderici, G.: Scale-Space Flow for End-to-End Optimized Video Compression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8500–8509 (2020). https://doi.org/10.1109/CVPR42600.2020.00853
3. Bjøntegaard, G.: Calculation of average psnr differences between rd-curves (2001), https://api.semanticscholar.org/CorpusID:61598325
4. Bross, B., Wang, Y.K., Ye, Y., Liu, S., Chen, J., Sullivan, G.J., Ohm, J.R.: Overview of the versatile video coding (vvc) standard and its applications. IEEE Trans. Circuit Syst. Video Technol. **31**(10), 3736–3764 (2021). https://doi.org/10.1109/TCSVT.2021.3101953
5. Chen, P.Y., Peng, W.H.: CANF-VC++: Enhancing Conditional Augmented Normalizing Flows for Video Compression with Advanced Techniques (2023). https://doi.org/10.48550/arXiv.2309.05382
6. Cui, Z., Wang, J., Gao, S., Guo, T., Feng, Y., Bai, B.: Asymmetric gained deep image compression with continuous rate adaptation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10527–10536 (2021). https://doi.org/10.1109/CVPR46437.2021.01039
7. He, D., Yang, Z., Peng, W., Ma, R., Qin, H., Wang, Y.: ELIC: efficient learned image compression with unevenly grouped space-channel contextual adaptive coding. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5708–5717 (2022). https://doi.org/10.1109/CVPR52688.2022.00563
8. Ho, Y.H., Chang, C.P., Chen, P.Y., Gnutti, A., Peng, W.H.: CANF-VC: Conditional Augmented Normalizing Flows for Video Compression. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Eur. Conf. Comput. Vis. pp. 207–223 (2022). https://doi.org/10.1007/978-3-031-19787-1_12
9. Hu, Z., Lu, G., Guo, J., Liu, S., Jiang, W., Xu, D.: Coarse-to-fine deep video coding with hyperprior-guided mode prediction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5921–5930 (2022)
10. Hu, Z., Xu, D.: Complexity-guided slimmable decoder for efficient deep video compression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14358–14367 (2023). https://doi.org/10.1109/CVPR52729.2023.01380
11. Ladune, T., Philippe, P., Hamidouche, W., Zhang, L., Déforges, O.: Optical flow and mode selection for learning-based video coding. In: IEEE International Workshop on Multimedia Signal Processing. pp. 1–6 (2020)
12. Li, B., Li, H., Li, L., Zhang, J.: λ domain rate control algorithm for high efficiency video coding. IEEE Trans. Image Process. **23**(9), 3841–3854 (2014). https://doi.org/10.1109/TIP.2014.2336550
13. Li, J., Li, B., Lu, Y.: Deep Contextual Video Compression. In: Adv. Neural Inform. Process. Syst. vol. 34, pp. 18114–18125 (2021)
14. Li, J., Li, B., Lu, Y.: Hybrid Spatial-Temporal Entropy Modelling for Neural Video Compression. In: ACM Int. Conf. Multimedia. pp. 1503–1511 (2022). https://doi.org/10.1145/3503161.3547845
15. Li, J., Li, B., Lu, Y.: Neural Video Compression with Diverse Contexts. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22616–22626 (2023). https://doi.org/10.1109/CVPR52729.2023.02166

16. Li, Y., Chen, X., Li, J., Wen, J., Han, Y., Liu, S., Xu, X.: Rate Control for Learned Video Compression. In: ICASSP. pp. 2829–2833 (2022). https://doi.org/10.1109/ICASSP43922.2022.9746080
17. Lin, J., Liu, D., Li, H., Wu, F.: M-LVC: Multiple Frames Prediction for Learned Video Compression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3543–3551 (2020). https://doi.org/10.1109/CVPR42600.2020.00360
18. Lin, J., Liu, D., Liang, J., Li, H., Wu, F.: A deeply modulated scheme for variable-rate video compression. In: IEEE Int. Conf. Image Process. pp. 3722–3726 (2021). https://doi.org/10.1109/ICIP42928.2021.9506269
19. Liu, M., Guo, Y., Li, H., Chen, C.W.: Low-complexity rate control based on ρ-domain model for scalable video coding. In: IEEE Int. Conf. Image Process. pp. 1277–1280 (2010). https://doi.org/10.1109/ICIP.2010.5653340
20. Lu, G., Ouyang, W., Xu, D., Zhang, X., Cai, C., Gao, Z.: DVC: An End-To-End Deep Video Compression Framework. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10998–11007 (2019). https://doi.org/10.1109/CVPR.2019.01126
21. Ma, S., Gao, W., Lu, Y.: Rate-distortion analysis for h.264/avc video coding and its application to rate control. IEEE Trans. Circuit Syst. Video Technol. **15**(12), 1533–1544 (2005). https://doi.org/10.1109/TCSVT.2005.857300
22. Mercat, A., Viitanen, M., Vanne, J.: UVG dataset: 50/120fps 4k sequences for video codec analysis and development. In: Toni, L., Begen, A.C., Alay, Ö., Timmerer, C. (eds.) ACM Multimedia Systems Conference. pp. 297–302 (2020). https://doi.org/10.1145/3339825.3394937
23. Minnen, D., Singh, S.: Channel-wise autoregressive entropy models for learned image compression. In: IEEE Int. Conf. Image Process. pp. 3339–3343 (2020). https://doi.org/10.1109/ICIP40778.2020.9190935
24. Qi, L., Li, J., Li, B., Li, H., Lu, Y.: Motion information propagation for neural video compression. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6111–6120 (2023). https://doi.org/10.1109/CVPR52729.2023.00592
25. Qi, Y., He, Y., Qi, X., Zhang, Y., Yang, G.: Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. In: Int. Conf. Comput. Vis. pp. 6047–6056 (2023). https://doi.org/10.1109/ICCV51070.2023.00558
26. Sheng, X., Li, J., Li, B., Li, L., Liu, D., Lu, Y.: Temporal Context Mining for Learned Video Compression. IEEE Trans. Multimedia **25**, 7311–7322 (2022). https://doi.org/10.1109/TMM.2022.3220421
27. Shi, Y., Ge, Y., Wang, J., Mao, J.: AlphaVC: High-performance and efficient learned video compression. In: Eur. Conf. Comput. Vis. pp. 616–631 (2022)
28. Song, M., Choi, J., Han, B.: Variable-rate deep image compression through spatially-adaptive feature transform. In: Int. Conf. Comput. Vis. pp. 2360–2369 (2021). https://doi.org/10.1109/ICCV48922.2021.00238
29. Sullivan, G.J., Ohm, J.R., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (hevc) standard. IEEE Trans. Circuit Syst. Video Technol. **22**(12), 1649–1668 (2012). https://doi.org/10.1109/TCSVT.2012.2221191
30. Tao, L., Gao, W., Li, G., Zhang, C.: Adanic: Towards practical neural image compression via dynamic transform routing. In: Int. Conf. Comput. Vis. pp. 16833–16842 (2023). https://doi.org/10.1109/ICCV51070.2023.01548
31. Veit, A., Belongie, S.J.: Convolutional networks with adaptive inference graphs. In: Eur. Conf. Comput. Vis. Lecture Notes in Computer Science, vol. 11205, pp. 3–18 (2018). https://doi.org/10.1007/978-3-030-01246-5_1

32. Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., Hu, X., Lu, T., Lu, L., Li, H., Wang, X., Qiao, Y.: Internimage: Exploring large-scale vision foundation models with deformable convolutions. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 14408–14419 (2023). https://doi.org/10.1109/CVPR52729.2023.01385
33. Wiegand, T., Sullivan, G., Bjontegaard, G., Luthra, A.: Overview of the h.264/avc video coding standard. IEEE Trans. Circuit Syst. Video Technol. **13**(7), 560–576 (2003). https://doi.org/10.1109/TCSVT.2003.815165
34. Wiegand, T., Sullivan, G., Bjontegaard, G., Luthra, A.: Overview of the h.264/avc video coding standard. IEEE Trans. Circuit Syst. Video Technol. **13**(7), 560–576 (2003). https://doi.org/10.1109/TCSVT.2003.815165
35. Wu, Y., Qi, Z., Zheng, H., Tao, L., Gao, W.: Deep image compression with latent optimization and piece-wise quantization approximation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1926–1930 (2021)
36. Xu, T., Gao, H., Gao, C., Wang, Y., He, D., Pi, J., Luo, J., Zhu, Z., Ye, M., Qin, H., Wang, Y., Liu, J., Zhang, Y.Q.: Bit allocation using optimization. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Int. Conf. Mach. Learn. vol. 202, pp. 38377–38399 (2023)
37. Xue, T., Chen, B., Wu, J., Wei, D., Freeman, W.T.: Video enhancement with task-oriented flow. Int. J. Comput. Vis. **127**(8), 1106–1125 (2019). https://doi.org/10.1007/S11263-018-01144-2
38. Yang, C., Wang, X., Yao, L., Long, G., Xu, G.: Dyformer: A dynamic transformer-based architecture for multivariate time series classification. Inf. Sci. **656**, 119881 (2024). https://doi.org/10.1016/J.INS.2023.119881
39. Yang, F., Herranz, L., Cheng, Y., Mozerov, M.G.: Slimmable compressive autoencoders for practical neural image compression. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
40. Yang, Z., Gao, W., Li, G., Yan, Y.: Sur-driven video coding rate control for jointly optimizing perceptual quality and buffer control. IEEE Trans. Image Process. (2023)
41. Zheng, H., Gao, W.: End-to-end rgb-d image compression via exploiting channel-modality redundancy. In: AAAI. vol. 38, pp. 7562–7570 (2024)
42. Çetin, E., Yılmaz, M.A., Tekalp, A.M.: Flexible-rate learned hierarchical bi-directional video compression with motion refinement and frame-level bit allocation. In: IEEE Int. Conf. Image Process. pp. 1206–1210 (2022). https://doi.org/10.1109/ICIP46576.2022.9897455