Traditional vs. Neural Video Codecs: Compression Efficiency, Visual Artifacts, and Quality Analysis Beyond PSNR

Leandro Tavares
lwtavares@inf.ufpel.edu.br
Federal University of Pelotas
Pelotas, RS, Brazil

Ruhan Conceição
radconeicao@inf.ufpel.edu.br
Federal University of Pelotas
Pelotas, RS, Brazil

Victor Costa
vcrcosta@inf.ufpel.edu.br
Federal University of Pelotas
Pelotas, RS, Brazil

Luciano Agostini
agostini@inf.ufpel.edu.br
Federal University of Pelotas
Pelotas, RS, Brazil

Marcelo Porto
porto@inf.ufpel.edu.br
Federal University of Pelotas
Pelotas, RS, Brazil

Guilherme Corrêa
gcorrea@inf.ufpel.edu.br
Federal University of Pelotas
Pelotas, RS, Brazil

Figure 1: Visual comparison of compression artifacts on a selected region of the *YachtRide* video. The left tile shows the original uncompressed frame. The remaining tiles display the reconstructions produced by HEVC and VVC (at QP 47), and DCVC-FM and DCVC-RT (at quality level 0), respectively. This teaser highlights the residual distortions introduced by each codec in challenging textured areas, revealing differences in artifact patterns between traditional and neural compression approaches.

ABSTRACT

Video compression technology is rapidly evolving, with end-to-end learned Neural Video Codecs (NVCs) emerging as powerful alternatives to traditional, block-based standards. This paper presents a comparative analysis of two traditional codecs, HEVC and VVC, against two leading NVCs, DCVC-FM and DCVC-RT. We conduct a thorough rate distortion analysis using multiple objective metrics (PSNR, SSIM, VMAF, and LPIPS) and introduce a formal method to identify rate-matched operating points for fair visual comparison. Our results demonstrate that NVCs offer substantial gains not only in signal fidelity but also in perceptual quality, achieving up to 37.82% bitrate savings over HEVC. Furthermore, our analysis reveals fundamental differences in distortion patterns: NVCs excel at preserving structural and color fidelity, producing visually pleasing results free of blocking artifacts, but can sometimes smooth over fine textures. Conversely, traditional codecs are prone to blockiness but can occasionally retain more high-frequency detail. These findings confirm the superior efficiency of NVCs and highlight the need for evaluation methodologies that look beyond PSNR.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

KEYWORDS

video compression, neural networks, neural codecs, quality assessment, HEVC, VVC, DCVC-FM, DCVC-RT

1 INTRODUCTION

With the rapid growth of digital video content due to the popularization of streaming platforms, video conferencing, and social media, the need for efficient video compression tools has become increasingly critical. Traditional video coding standards, such as H.264/AVC [1], H.265/HEVC [2], and the most recent H.266/VVC [3], were developed to drastically reduce the amount of data required for video representation. These codecs are built on manually engineered algorithms and rely on hybrid block-based architectures, combining many different techniques. While each new standard improves compression efficiency, they follow the same core design principles established decades ago.

In contrast, advances in deep learning have enabled the exploration of entirely new paradigms for video compression. NVCs replace hand-crafted modules with end-to-end trainable neural networks that automatically learn compact representations from data. One of the pioneering works in this area is Deep Video Compression (DVC) [11], which mimics the traditional hybrid coding structure using neural components for motion estimation, residual compression, and entropy modeling. Since then, the field has rapidly

evolved, giving rise to more sophisticated architectures. A notable example is Deep Contextual Video Compression (DCVC) [8], which introduces temporal context modeling to enhance compression performance. The original DCVC framework has inspired several extensions aimed at improving different aspects of neural video compression. DCVC Feature Modulation (DCVC-FM) [9] enhances compression efficiency by introducing multi-scale temporal context mining and refined motion modeling, achieving state-of-the-art rate-distortion performance. In contrast, DCVC Real Time (DCVC-RT) [7] focuses on real-time feasibility by eliminating sequential dependencies such as motion estimation and autoregressive entropy models, enabling fully parallel decoding while maintaining competitive compression quality.

Recent studies have shown that NVCs can outperform traditional codecs in terms of rate-distortion performance, especially at low bitrates. However, these two classes of codecs differ not only in architecture but also in the characteristics of the compression artifacts they produce (as can be seen in Fig. 1), and how such artifacts are perceived by the Human Visual System (HVS). This raises important questions about how to fairly compare traditional and neural codecs across objective and perceptual metrics, and under which conditions one approach may be preferable to the other.

In this context, this work aims to provide a comprehensive comparison between state-of-the-art traditional and neural video codecs, with a particular focus on their compression efficiency, visual quality, and artifact characteristics. First, we evaluate how well traditional codecs—such as HEVC and VVC—compare against recent learned approaches like DCVC-FM and DCVC-RT in terms of rate-distortion (RD) performance. By analyzing RD curves across several datasets, we identify representative operating points where both codec classes yield either similar bitrates, expressed in bots per pixel (bpp), or similar objective quality (e.g., Peak Signal-to-Noise Ratio – PSNR). These aligned points serve as the foundation for deeper cross-class comparisons.

At these matched operating points, we further investigate how codecs compare under alternative objective metrics that better reflect human perception, including the Structural Similarity Index (SSIM), the Video Multi-method Assessment Fusion (VMAF), and the Learned Perceptual Image Patch Similarity (LPIPS). We analyze per-dataset averages and present side-by-side visual comparisons to highlight differences in perceptual quality, even when PSNR or bpp are nearly identical. Building upon these observations, we explore the nature of artifacts introduced by each codec class. Through spatial and frequency-domain analysis, we characterize the distortions and identify patterns unique to neural or traditional compression.

2 BACKGROUND

2.1 Quality Metrics

To comprehensively evaluate the visual quality of compressed video, we employ a combination of traditional signal fidelity metrics and perceptual quality metrics. Specifically, we use PSNR, SSIM, LPIPS, and VMAF, each capturing different aspects of distortion and perceptual quality. In what follows, we briefly describe each of these metrics.

2.1.1 Peak Signal-to-Noise Ratio (PSNR). PSNR is one of the most widely used objective quality metrics in video compression research due to its simplicity and computational efficiency. It measures the logarithmic ratio between the maximum possible signal value and the mean squared error (MSE) between the original and reconstructed frames. For an image of size $W \times H$, MSE is defined as:

$$\text{MSE} = \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} (x(i,j) - \hat{x}(i,j))^2, \tag{1}$$

where $x(i,j)$ and $\hat{x}(i,j)$ represent the pixel values of the original and reconstructed images, respectively. The PSNR is then given by:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX^2}{\text{MSE}} \right), \tag{2}$$

where MAX is the maximum possible pixel value of the image (e.g., 255 for 8-bit images). Although PSNR is mathematically convenient, it does not correlate well with human visual perception, especially when different types of distortions are present [16].

2.1.2 Structural Similarity Index (SSIM). The Structural Similarity Index (SSIM) [16] improves upon PSNR by incorporating perceptual phenomena such as luminance masking and contrast sensitivity. It measures the similarity between two image patches based on luminance, contrast, and structural information. SSIM values range from 0 to 1, where 1 indicates perfect structural similarity. Unlike PSNR, SSIM is better aligned with the Human Visual System (HVS), particularly in preserving local structural details. It is commonly computed over luminance channels and averaged spatially across an image or video frame.

2.1.3 Learned Perceptual Image Patch Similarity (LPIPS). LPIPS [17] is a deep learning-based perceptual metric designed to capture visual similarity as perceived by humans. It computes the distance between deep feature representations extracted from a pretrained convolutional neural network (typically VGG or AlexNet) that processes the two compared images. LPIPS has been shown to correlate better with human judgments than traditional metrics like PSNR or SSIM, particularly in cases where distortions are more perceptual than signal-based. LPIPS values range from 0 to 1, with lower values indicating higher similarity.

Since LPIPS is a dissimilarity metric where lower values indicate higher perceptual quality, it cannot be used directly with the standard Bjøntegaard Delta (BD) Rate calculation, which assumes that higher metric scores correspond to better quality. To align LPIPS with this framework, in this work we transformed the scores by computing $1 - \text{LPIPS}$ for each sequence. This inverted metric, where higher values now signify better quality, was used for all BD-LPIPS calculations and corresponding Rate-Distortion curves presented in this work.

2.1.4 Video Multi-method Assessment Fusion (VMAF). VMAF [13] is a perceptual video quality metric developed by Netflix that combines multiple quality indicators—such as detail loss and temporal artifacts—using a machine learning model trained on subjective human ratings. It operates on a fusion of handcrafted and data-driven features to predict quality scores that align with viewer perception. VMAF values typically range from 0 to 100, where higher values indicate better perceived quality. Due to its strong correlation with

subjective opinion scores, VMAF has become a widely accepted benchmark for perceptual video quality evaluation in industry and academia.

2.2 Related Works

2.2.1 Traditional Codecs. Pakdaman et al. [14] provide a detailed investigation into the complexity of VVC standard. The authors decompose both the encoding and decoding pipelines of VVC to analyze the contributions of its individual coding tools and modules. By evaluating processing time, memory consumption, and algorithmic behavior across configurations, the study highlights the primary computational bottlenecks introduced by VVC. The results show that, although VVC achieves superior compression efficiency compared to earlier traditional video compression standards, this improvement comes at the cost of significantly increased complexity, particularly due to advanced tools such as multiple transform selection, affine motion compensation, and enhanced intra prediction modes. The findings in [14] are essential for codec designers and practitioners seeking to implement or deploy VVC in real-time or resource-constrained environments.

The work by Siqueira et al. [15] presents a thorough assessment of the Versatile Video Coding (VVC) standard, focusing on both its compression efficiency and computational complexity. The authors structure their analysis into three parts: (i) the impact of enabling SIMD (Single Instruction/Multiple Data) optimizations, (ii) the effect of restricting VVC's flexible partitioning structures on encoding performance, and (iii) a detailed profiling of the computational cost of individual tools within the VVC encoder. Experimental results show that SIMD optimizations reduce encoding time by 62% on average, while limiting partitioning depths can lower complexity by up to 72% at the cost of significant compression efficiency loss (e.g., 255.5% BD-rate increase in the most constrained case). The profiling in [15] identifies motion estimation and motion compensation – particularly VVC's advanced affine and optical flow-based modes – as the most computationally expensive modules. This paper provides valuable insights into the design of fast, optimized implementations of the VVC standard and highlights critical areas where complexity can be traded off for acceptable coding efficiency.

2.2.2 Neural Frameworks. The paper by Liao et al. [10] presents a comprehensive review of recent advances in video compression systems based on deep neural networks (DNNs). It surveys a wide range of neural video compression methods, categorizing them based on architectural design, such as hybrid models, end-to-end learned frameworks, and enhancement modules integrated into traditional codecs. The authors analyze core components of neural video compression systems, including motion estimation, residual compression, and entropy modeling, highlighting how DNNs have enabled more flexible and effective alternatives to traditional hand-crafted tools. Additionally, the paper presents multiple case studies that empirically compare compression performance and complexity between learned and traditional codecs. This review provides valuable insights into the trade-offs between compression efficiency and computational cost, positioning deep learning as a promising direction for future video coding standards.

Chen et al. [6] investigate the rate-distortion-complexity trade-offs of neural video coding frameworks based on conditional coding. While modern conditional autoencoders explore spatial and temporal information to improve compression efficiency, they often suffer from an information bottleneck, which can reduce performance compared to traditional residual coding. To address this, the authors explore two recently proposed alternatives –conditional residual coding and masked conditional residual coding –both built upon the DCVC framework. These approaches aim to retain or enhance compression efficiency while reducing computational cost and memory usage. Through extensive experiments across multiple datasets, the study demonstrates that both alternatives outperform standard conditional coding in terms of BD-rate savings and complexity, particularly under resource-constrained scenarios. The masked variant further improves adaptability by blending conditional and residual inputs through a learned soft mask, offering better performance in challenging regions like occlusions or complex motion. This research highlights how efficient learned video compression can be achieved when considering practical computational constraints.

3 EXPERIMENTAL SETUP

Here we detail the experimental design used to compare traditional and neural video compression approaches. Our goal is to ensure a fair, thorough, and reproducible evaluation of codec performance across a diverse set of conditions.

3.1 Codecs

To evaluate the performance of neural video compression frameworks against traditional approaches, we selected two representative standards: High Efficiency Video Coding (HEVC) [2] and Versatile Video Coding (VVC) [3]. HEVC (also known as H.265) was finalized in 2013 and is widely deployed across consumer electronics, broadcasting, and streaming platforms due to its substantial compression efficiency. Its widespread adoption has also led to the development of dedicated hardware accelerators, making it a practical and efficient choice in real-time applications. VVC (H.266), on the other hand, represents the current state-of-the-art in traditional video coding. Standardized in 2020, VVC introduces a broad set of advanced coding tools aimed at further reducing bitrate by approximately 30%–50% compared to HEVC, especially for high-resolution (4K/8K) and immersive video content. Given its comprehensive toolset and focus on future media formats, VVC is a natural choice for benchmarking the upper limit of traditional codec performance. By including both HEVC and VVC, we cover a broad spectrum of practical deployment (HEVC) and cutting-edge performance (VVC) in traditional video coding.

For the neural video compression methods, we selected two recent and advanced implementations from the Deep Contextual Video Compression (DCVC) family: DCVC-FM [9] and DCVC-RT [7][1]. Since learned video codecs are not yet widely deployed in commercial applications, we focused on state-of-the-art research models that represent the frontier of what NVCs can achieve. DCVC-FM was chosen for its outstanding compression efficiency, consistently outperforming traditional codecs such as HEVC and even VVC under objective quality metrics like PSNR and MS-SSIM. Its architecture integrates multi-scale temporal context mining and

[1]For brevity, some equations and tables refer to DCVC-FM and DCVC-RT simply as FM and RT, respectively.

re-filling mechanisms, which enhance temporal modeling and significantly improve rate-distortion performance. DCVC-RT, in turn, was selected for its real-time capabilities. While it is optimized for fast encoding and decoding—achieving significant reductions in computational complexity—it maintains high compression efficiency. Notably, it departs from traditional designs by removing components such as motion estimation and the auto-regressive entropy model, enabling a fully parallelized and low-latency decoding pipeline without severely compromising coding performance. Together, DCVC-FM and DCVC-RT provide a comprehensive view of the current capabilities of NVCs, spanning the axes of maximum compression efficiency and real-time feasibility.

3.2 Datasets and Encoder Configuration

To ensure a comprehensive and standardized evaluation across a range of video resolutions and content types, we selected a diverse test dataset composed of sequences from the Ultra Video Group (UVG) dataset [12] and HEVC Common Test Conditions [5] standard test classes (Classes B, C, D, and E). The UVG set includes 7 high-resolution sequences at 1920×1080 resolution and high frame rates (up to 120 fps), while the HEVC sets, composed of 16 videos total, cover a wider range of resolutions, 416x240 (Class D – HEVC-D), 832x480 (Class C – HEVC-D), 1280x720 (Class B – HEVC-B) and 1920x1080 (Class B – HEVC-B), representing both high-motion and complex-texture content. For consistency and computational feasibility, we used the first 96 frames of each sequence in all experiments. All sequences are encoded using an intra period of −1, which corresponds to a single intra frame at the beginning followed by inter-coded frames only. For the learned codecs, DCVC-FM and DCVC-RT, we tested 10 different quality settings (0, 7, 14, 21, 28, 35, 42, 49, 56 and 63), while for the traditional codecs (HEVC and VVC), we evaluated six standard quantization parameters (QPs): 22, 27, 32, 37, 42, and 47. Traditional codecs were configured using the *Low Delay* profile to match the coding structure of the learned models. Additionally, all videos were coded in YUV format, preserving color consistency across all methods. This setup ensures a fair and thorough comparison across codecs with respect to both rate-distortion performance and bitrate scalability.

3.3 Metrics

To assess the performance of the evaluated codecs, we employed a combination of traditional and perceptual quality metrics. These include PSNR, SSIM, LPIPS, and VMAF, each capturing different aspects of visual fidelity and perceptual quality. A detailed explanation of these metrics can be found in Section 2.1.

PSNR is a widely used objective metric in video compression research due to its simplicity and mathematical interpretability, SSIM [16] offers an improvement over PSNR by considering structural information in the image, LPIPS [17] leverages deep neural networks to estimate perceptual similarity and VMAF [13], a model developed by Netflix that combines multiple quality indicators using machine learning. Together, these metrics provide a comprehensive evaluation of both fidelity and perceptual quality.

Table 1: Comparison of compression efficiency in terms of BD-rate, computed by averaging the BD-rate values across individual sequences (traditional method).

	HEVC (Anchor)	VVC	FM	RT
UVG	-	-26.06%	**-29.39%**	-24.27%
HEVC-B	-	-27.67%	**-35.44%**	-30.71%
HEVC-C	-	-26.76%	**-39.97%**	-28.49%
HEVC-D	-	-24.12%	**-49.89%**	-37.33%
HEVC-E	-	-29.06%	**-42.43%**	-34.14%
Dataset wide	-	-26.59%	**-37.82%**	-29.97%

Table 2: Comparison of compression efficiency in terms of BD-rate, computed by averaging the rate-distortion points across sequences before applying the BD-rate calculation (NVC method).

	HEVC (Anchor)	VVC	FM	RT
UVG	-	-26.20%	**-40.16%**	-36.74%
HEVC-B	-	-28.05%	**-37.16%**	-33.17%
HEVC-C	-	-27.16%	**-40.87%**	-39.71%
HEVC-D	-	-24.82%	**-50.62%**	-38.61%
HEVC-E	-	-28.93%	**-42.41%**	-33.87%
Dataset wide	-	-26.57%	**-45.55%**	-41.09%

4 RATE-DISTORTION ANALYSIS

This section evaluates the compression efficiency of the selected traditional and neural video codecs. Our primary goal is to assess their rate-distortion (RD) performance using the Bjøntegaard Delta Rate (BD-Rate) metric [4], which quantifies the increase (or decrease) in terms of bitrate at equivalent objective quality. Note that negative BD-Rate indicates better compression efficiency, while a positive value reflects worse performance.

4.1 Compression Efficiency

It is worth mentioning that traditional and neural video compression researchers differ slightly in the way they compute average BD-Rate results. The traditional approach, as specified in the HEVC and VVC Common Test Conditions [5], involves calculating the BD-Rate for each sequence individually and then averaging these values. In contrast, many neural video compression papers report a single BD-Rate figure obtained by first averaging the rate-distortion (RD) curves across all sequences and then computing the BD-Rate from these averaged curves.

To ensure a comprehensive analysis, we report the results using both the traditional and NVC averaging methodologies. The corresponding outcomes are presented in Table 1 and Table 2, respectively. As HEVC is the earliest (and most established) codec among those evaluated, it was selected as the anchor for all BD-Rate comparisons.

These results highlight a consistent trend: NVCs, particularly DCVC-FM and DCVC-RT, significantly outperform traditional codecs (HEVC and VVC) in terms of compression efficiency. Using HEVC

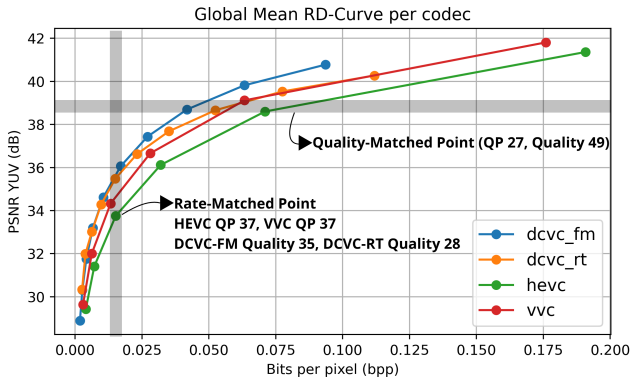

Figure 2: Mean rate-distortion curves per codec reported using NVC method (rate and quality-matched points in shaded area).

as the anchor, the traditional BD-rate calculation method (Table 1) shows that DCVC-FM achieves the highest compression gains, with an average bitrate reduction of 37.82% across all datasets. DCVC-RT also performs well, offering 29.97% average savings, while VVC delivers a 26.59% improvement over HEVC. These results confirm that neural codecs not only surpass their traditional counterparts in RD performance but also maintain a consistent lead across various video classes, with particularly large gains observed for lower-resolution sequences like HEVC-D.

When applying the alternative averaging method (Table 2), commonly used in NVC literature, the performance gap between codecs becomes even more pronounced. DCVC-FM reaches an average bitrate savings of 45.55%, further reinforcing its status as a state-of-the-art solution for learned video compression. The enhanced scores under this method are due to the influence of high-bitrate sequences in the averaging process, which can skew the result in favor of codecs that perform well on such content. Nonetheless, even under this potentially biased aggregation, traditional codecs lag behind both neural alternatives. These findings underline the importance of explicitly stating the BD-rate averaging methodology when comparing codecs and demonstrate that, regardless of the chosen approach, neural codecs provide superior compression efficiency over traditional standards.

Figure 2 depicts these results by plotting the global average RD-curve using the NVC method, corroborating to the advantage of the NVCs frameworks over traditional ones. The additional highlights on this figure will be explained in Section 4.2.

4.2 Rate-Distortion Point Alignment

In order to perform a meaningful and in-depth comparison between traditional and neural video compression frameworks, it is essential to identify a set of rate-distortion (RD) operating points where these codecs produce similar output quality or bitrate. However, as discussed in Section 3.2, traditional and neural codecs define their operating points differently: traditional codecs typically rely on a quantization parameter (QP), while neural video compression frameworks use a quality control variable, often referred to as a "quality level" or "quality point." This distinction introduces challenges when attempting to align the compression settings of both approaches for fair evaluation.

To enable a fair comparison between traditional and neural video codecs, we firstly defined a metric to quantify the bitrate discrepancy between any two codecs operating at given quality/QP levels. This metric, denoted as ΔR and shown in Equation 3, represents the relative bitrate difference between two codecs X and Y at operating points m_X and m_Y, respectively. It is computed as the average absolute difference in bitrate (in bits per pixel) across N test sequences, normalized by the bitrate of codec X, which acts as the reference.

$$\Delta R(m_X, m_Y) = \frac{\left| \frac{1}{N} \sum_{i=1}^{N} R^X_{i,m_X} - \frac{1}{N} \sum_{i=1}^{N} R^Y_{i,m_Y} \right|}{\frac{1}{N} \sum_{i=1}^{N} R^X_{i,m_X}} \tag{3}$$

Using this pairwise measure, we then define an aggregate cost function Φ_R, shown in Equation 4. This function computes the total relative bitrate deviation between a selected anchor codec, operating at point m_A, and three other codecs at points m_B, m_C, and m_D. It serves as a proxy for how well-aligned the codecs are in terms of bitrate.

$$\begin{aligned} &\Phi_R(m_A, m_B, m_C, m_D) = \\ &\Delta R(m_A, m_B) + \Delta R(m_A, m_C) + \Delta R(m_A, m_D) \end{aligned} \tag{4}$$

Finally, our goal is to find the combination of operating points for HEVC, VVC, DCVC-FM, and DCVC-RT (denoted as m^*_{HEVC}, m^*_{VVC}, m^*_{FM}, and m^*_{RT}) that minimizes the cost function Φ_R. This optimization, formalized in Equation 5, ensures that the selected RD points yield the most bitrate-aligned configuration across all codecs. Note that the operating points for both HEVC and VVC corresponds to the set of QP for these codecs depicted in Section 3.2, e.g., $m_{\text{HEVC}}, m_{\text{VVC}} \in \{22, 27, 32, 37, 42, 47\}$. Also, the same applies for neural video compression frameworks, meaning that m_{FM} , $m_{\text{RT}} \in \{0, 7, 14, 21, 28, 35, 42, 49, 56, 63\}$.

$$\begin{aligned} &(m^*_{\text{HEVC}}, m^*_{\text{VVC}}, m^*_{\text{FM}}, m^*_{\text{RT}}) \in \\ &\arg\min \Phi_R(m_{\text{HEVC}}, m_{\text{VVC}}, m_{\text{FM}}, m_{\text{RT}}) \end{aligned} \tag{5}$$

Similarly, the same strategy can be applied to identify a set of operating points that yields the most quality-aligned configuration across all codecs. In this case, instead of minimizing the relative difference in bitrate, we minimize the absolute discrepancy in visual quality (Equation 6). To achieve this, we define a quality-based cost function Φ_Q, analogous to Φ_R, and seek the set of operating points m^*_{HEVC}, m^*_{VVC}, m^*_{FM}, and m^*_{RT} that minimizes it, as shown in Equation 7.

$$\Delta Q(m_X, m_Y) = \left| \frac{1}{N} \sum_{i=1}^{N} Q^X_{i,m_X} - \frac{1}{N} \sum_{i=1}^{N} Q^Y_{i,m_Y} \right| \tag{6}$$

$$\begin{aligned} &(m^*_{\text{HEVC}}, m^*_{\text{VVC}}, m^*_{\text{FM}}, m^*_{\text{RT}}) \in \\ &\arg\min \Phi_Q(m_{\text{HEVC}}, m_{\text{VVC}}, m_{\text{FM}}, m_{\text{RT}}) \end{aligned} \tag{7}$$

Once the optimal operating points are identified, they can be used to define specific comparison scenarios that reflect typical usage conditions. Based on that methodology, two scenarios have been selected, as highlighted in Figure 2 and described as follows.

Quality-Matched Point: A configuration where the objective quality—measured in PSNR—is approximately equal across all codecs. This scenario, anchored at HEVC with QP 27, corresponds to the same QP for VVC and to quality level 49 for both DCVC-FM and DCVC-RT. It allows us to assess the relative bitrate efficiency of each codec while maintaining comparable image quality. This case is highlighted in the gray horizontal area of Figure 2.

Rate-Matched Point: A configuration in which all codecs operate at a similar bitrate. It is centered around HEVC/VVC at QP 37 and maps to quality level 35 for DCVC-FM and 28 for DCVC-RT. It is particularly relevant for evaluating the perceptual quality delivered by each codec under a fixed bitrate budget – a common constraint in practical streaming or storage scenarios. This case is highlighted in the gray vertical area of Figure 2.

For the subsequent analysis in this paper, we will focus on the rate-matched operating point. This choice is motivated by its practical relevance to streaming and communication scenarios, where bandwidth is a primary constraint and the main goal is to maximize the perceptual quality delivered to the end-user.

4.3 Rate-Distortion Efficiency with Perceptual Metrics

To provide a more comprehensive view of compression efficiency, this section extends our analysis beyond a single metric. We evaluate the performance of the codecs from two complementary perspectives: (i) bitrate savings for an equivalent quality level (measured by BD-Rate), and (ii) quality gain at an equivalent bitrate (measured by BD-Quality, e.g., BD-PSNR). LPIPS-based metrics are calculated using the 1-LPIPS value, as explained in section 2.1.3.

First, we analyze bitrate savings. Table 3 summarizes the BD-Rate results, using HEVC as the anchor. The best result for each metric is shown in bold, and the second-best is underlined. The data confirms that the superiority of NVCs is significant across all metrics. When measured by VMAF, which is highly correlated with user experience, DCVC-FM and DCVC-RT achieve remarkable bitrate savings of 48.95% and 45.61%, respectively, far surpassing VVC's 26.85%. This trend holds for PSNR and LPIPS, where the NVCs consistently offer much larger bitrate reductions than VVC. For SSIM, both NVCs deliver excellent results, with DCVC-RT achieving a slight edge with a 41.77% bitrate reduction.

Next, we analyze the quality gains at a fixed bitrate, with results presented in Table 4. This perspective further solidifies the advantage of the NVCs. For the same bitrate as HEVC, DCVC-FM provides a substantial quality increase of 10.3 VMAF points and 1.86 dB in PSNR. These gains are close to double those provided by VVC. The trend is consistent across all metrics, with DCVC-FM and DCVC-RT consistently delivering a significantly higher quality output than the traditional codecs for the same bitrate. These trends can also be observed on Figure 3, where we present the mean RD-Curve for each codec, highlighting low and high quality ranges for each metric. The charts show the significant difference in compression efficiency of NVCs.

In summary, the results paint a clear and consistent picture. Whether the goal is to minimize bitrate for a target quality (Table 3) or to maximize quality for a fixed bitrate (Table 4), the learned codecs demonstrate a substantial advantage. This efficiency confirms that NVCs offer genuine gains in perceptual compression and motivates a direct visual inspection of codec artifacts in the following sections.

5 QUALITY ASSESSMENT

Having established the rate-distortion efficiency in Section 4.2, we now shift our focus to a deeper analysis of coded video quality. This section moves beyond PSNR to evaluate the codecs using metrics that better correlate with the HVS: SSIM, LPIPS, and VMAF.

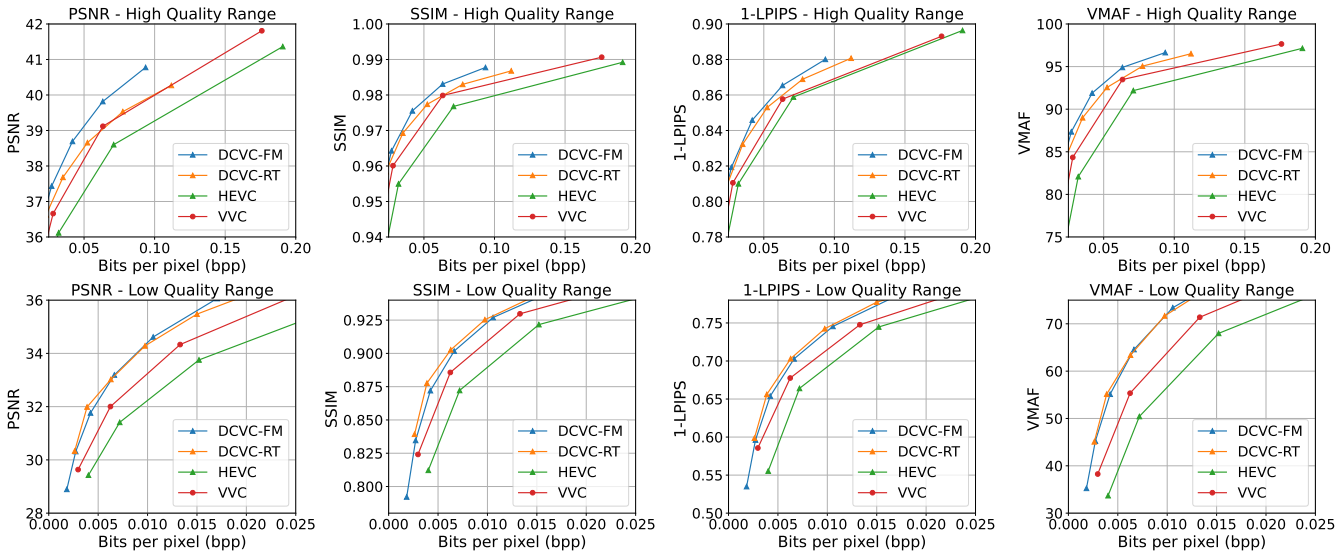

Figure 3: Mean rate-distortion curve per codec for PSNR, SSIM, LPIPS and VMAF, reported using NVC method and highlighting low and high quality ranges.

Table 3: BD-Rate metric using PSNR, SSIM, LPIPS and VMAF as quality values, assuming HEVC as anchor and using NVC method.

	HEVC (Anchor)	VVC	FM	RT
BD-Rate (PSNR)	-	-26.57%	**-45.55%**	-41.09%
BD-Rate (SSIM)	-	-25.71%	-40.84%	**-41.77%**
BD-Rate (1-LPIPS)	-	-17.80%	**-34.22%**	-33.44%
BD-Rate (VMAF)	-	-26.85%	**-48.95%**	-45.61%

Table 4: BD-Quality metric using PSNR, SSIM, LPIPS and VMAF as quality values, assuming HEVC as anchor and using NVC method (in dB for PSNR, score for other metrics).

	HEVC (Anchor)	VVC	FM	RT
BD-PSNR (dB)	-	0.9317	**1.8625**	1.5013
BD-SSIM	-	0.0130	**0.0224**	0.0221
BD-LPIPS*	-	0.0159	**0.0327**	0.0319
BD-VMAF	-	4.8911	**10.3056**	9.2936

The analysis is performed at the **rate-matched operating point** defined previously, where the codecs operate on a similar bitrate.

5.1 Objective Quality Analysis

Table 5 presents the performance of each codec at this rate-matched point, averaged per dataset class. In the table, **bolded values** indicate the best result among all codecs for a given metric and dataset, while underlined values denote the second-best performance. The results highlight a clear trend: at a comparable bitrate, the NVC models consistently deliver superior quality.

The SSIM metric shows the smallest gains for NVCs over traditional codecs for the metrics tested, with the best-performing NVC (DCVC-FM) showing a gain of 3.1% over HEVC. Following DCVC-FM, the second best result is shown by DCVC-RT with 2.56% increase on average SSIM. These numbers, although small, are higher than what VVC achieves, showing only a 0.96% increase over the anchor.

The advantages of the learned codecs are most evident with modern perceptual metrics. For LPIPS, where lower scores are better, DCVC-FM and DCVC-RT achieve average score reductions of 16.98% and 13.18% relative to HEVC, respectively, indicating a significant perceptual improvement. In contrast, VVC shows a negligible change, with 1.2% decrease over HEVC. Similarly, the VMAF scores for DCVC-FM and DCVC-RT show an average gain of 19.2% and 15.61%, respectively, against HEVC, overshadowing VVC gains by a large margin, which shows only 5.11% gain over the anchor codec.

The small difference on SSIM might suggests that metrics based primarily on structural similarity, like SSIM, may not be sensitive enough to capture the full benefits of learned codecs, which often excel at preserving perceptual information over pixel-perfect fidelity. This finding underscores the importance of employing learning-based metrics like LPIPS and VMAF for a comprehensive evaluation of modern video codecs.

5.2 Visual Quality Perception

While objective metrics provide a quantitative measure of performance, a qualitative analysis is essential to understand the nature of the artifacts introduced by each class of codec. This subsection provides a visual assessment of the reconstructed frames at the previously established rate-matched operating point, using the examples presented in Figure 4. Our analysis focuses on the distinct ways traditional and neural codecs handle structural integrity, texture detail, and color information.

5.2.1 Structural Coherence and Texture Blurring. The first major difference observed is in the preservation of structures versus detailed textures. NVCs demonstrate a superior ability to maintain the structural coherence of objects. This is evident in the first HEVC-B example, where the athlete's arm and hand retain their shape and form, whereas in the VVC and HEVC reconstructions, these features are distorted and nearly blend into the background. Similarly, for facial reconstructions in HEVC-D, HEVC-E, and UVG examples, the NVCs produce a more recognizable and natural facial structure. Conversely, NVCs tend produce overly smooth regions, sometimes sacrificing fine texture in order to better preserve structures.

5.2.2 Blocking Artifacts and Texture Synthesis. Traditional codecs, due to their block-based coding architecture, are prone to characteristic blocking artifacts, especially at lower bitrates. These are highly visible in the HEVC and VVC reconstructions of faces in HEVC-D and HEVC-E examples, where block boundaries disrupt natural features and make expressions difficult to discern. While these codecs can sometimes preserve high-frequency details more sharply, as seen in the texture of the woman's clothing in HEVC-B second example or the water droplets in UVG examples, this often comes at the cost of these visible block-based distortions in other areas.

NVCs, being free of a rigid block structure, do not exhibit these artifacts. Instead, their primary texture-related artifact is blurring, which can result in the smoothness noted earlier but can also fail on complex, fine textures, leading to a loss of crispness compared to the original material.

5.2.3 Color Fidelity. A consistent advantage of the NVCs at this operating point is their superior color reconstruction. Across multiple examples, such as the first comparisons in HEVC-D and HEVC-E, the colors produced by DCVC-FM and DCVC-RT are visibly closer to the original source frame. The traditional codecs, in contrast, exhibit slight but noticeable color shifts, resulting in a less faithful representation.

6 CONCLUDING REMARKS

In this paper, we conducted a comprehensive comparative analysis between traditional video codecs (HEVC and VVC) and state-of-the-art neural video codecs (DCVC-FM and DCVC-RT). Our evaluation spanned multiple dimensions, including rate-distortion efficiency across various objective metrics and a detailed qualitative assessment of compression artifacts.

Our quantitative results demonstrate the superiority of the learned compression paradigm. From a rate-distortion perspective, NVCs deliver significant bitrate savings for the same quality level, or alternatively, substantial quality gains for the same rate. This advantage

Table 5: Objective quality metrics at the rate-matched operating point (QP 37 for HEVC and VVC, quality 35 for DCVC-FM, and 28 for DCVC-RT). Results are averaged per dataset and reported in terms of structural similarity (SSIM), perceptual similarity (LPIPS), and video quality (VMAF). Arrows indicate the desirable direction for each metric. The last row shows the percentage difference relative to HEVC.

	SSIM ↑				LPIPS ↓				VMAF ↑			
	HEVC	**VVC**	**FM**	**RT**	**HEVC**	**VVC**	**FM**	**RT**	**HEVC**	**VVC**	**FM**	**RT**
UVG	0.9565	0.9593	**0.9685**	0.9681	0.2965	0.2916	**0.2601**	0.2608	65.1388	67.709	**77.0256**	75.4806
HEVC-B	0.9528	0.9586	**0.9709**	0.9705	0.3215	0.3221	**0.2816**	0.2875	66.7996	70.4818	**79.8276**	77.1024
HEVC-C	0.9747	0.9785	0.9867	**0.9878**	0.1823	0.1762	0.1539	**0.1525**	78.1735	81.7795	89.4177	**89.819**
HEVC-D	0.8982	0.9127	**0.9371**	0.9263	0.2103	0.2051	**0.1647**	0.1906	68.3962	72.5441	**82.7787**	78.2811
HEVC-E	0.8053	0.8224	**0.8664**	0.8524	0.1998	0.201	**0.1448**	0.1594	66.3721	69.9996	**82.0548**	78.0288
Avg	0.9175	0.9263	**0.9459**	0.941	0.2421	0.2392	**0.201**	0.2102	68.976	72.5028	**82.2209**	79.7424
Avg (%)	-	0.96	**3.1**	2.56	-	-1.2	**-16.98**	-13.18	-	5.11	**19.2**	15.61

was most pronounced when measured with modern metrics like VMAF and LPIPS, where DCVC-FM achieved bitrate savings of up to 48.95% and quality gains of over 10 VMAF points compared to HEVC. This indicates that the benefits of NVCs extend beyond simple signal fidelity and translate to a better perceptual experience.

Furthermore, our visual analysis explores the trade-offs between the two codec families. We showed that NVCs excel at maintaining the structural integrity of objects and producing smooth, block-free reconstructions. In contrast, traditional codecs often fail on this front, introducing distracting blocking artifacts, but can sometimes better preserve isolated, high-frequency textures.

The results of this work have important implications. They not only validate the studies towards learned codecs but also expose the limitations of relying only on PSNR for codec evaluation. The distinct artifact profiles suggest that future research should focus on developing more sophisticated quality metrics tailored to the unique distortions of NVCs. For future work, we plan to extend this analysis by conducting formal subjective user studies to validate our objective findings. Additionally, exploring the effectiveness of visual quality enhancement (VQE) techniques as post-processing filters to mitigate NVC-specific artifacts.

ACKNOWLEDGMENTS

The authors of this work would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) Finance Code 001, CNPq, and FAPERGS for funding this research.

Figure 4: Comparison between decoded frames from each class in the dataset, processed by each codec.

REFERENCES

[1] 2003. Recommendation H.264: Advanced video coding for generic audiovisual services. International Telecommunication Union (ITU). https://www.itu.int/rec/T-REC-H.264 Available: https://www.itu.int/rec/T-REC-H.264.

[2] 2013. Recommendation H.265: High efficiency video coding. International Telecommunication Union (ITU). https://www.itu.int/rec/T-REC-H.265 Available: https://www.itu.int/rec/T-REC-H.265.

[3] 2020. Recommendation H.266: Versatile video coding. International Telecommunication Union (ITU). https://www.itu.int/rec/T-REC-H.266 Available: https://www.itu.int/rec/T-REC-H.266.

[4] Gisle Bjontegaard. 2001. Calculation of average PSNR differences between RD-curves. *ITU SG16 Doc. VCEG-M33* (2001).

[5] Frank Bossen, Bin Li, Kyeongkeun Lim, and Alexei Norkin. 2013. *Common test conditions and software reference configurations.* Technical Report JCTVC-L1100. Joint Collaborative Team on Video Coding (JCT-VC). Meeting 12, Geneva, Switzerland.

[6] Yi-Hsin Chen, Kuan-Wei Ho, Martin Benjak, Jörn Ostermann, and Wen-Hsiao Peng. 2024. On the Rate-Distortion-Complexity Trade-Offs of Neural Video Coding. In *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*. 1–6. https://doi.org/10.1109/MMSP61759.2024.10743250

[7] Zhaoyang Jia, Bin Li, Jiahao Li, Wenxuan Xie, Linfeng Qi, Houqiang Li, and Yan Lu. 2025. Towards Practical Real-Time Neural Video Compression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-25, 2024.*

[8] Jiahao Li, Bin Li, and Yan Lu. 2021. Deep Contextual Video Compression. *Advances in Neural Information Processing Systems* 34 (2021).

[9] Jiahao Li, Bin Li, and Yan Lu. 2024. Neural Video Compression with Feature Modulation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 17-21, 2024.*

[10] Xiaodong Liao, Yuanyi Yang, Shiqi Li, Kai Huang, Zhenyu Li, Shiqi Lu, Zhan Zhang, and Liang Ma. 2021. Advances in video compression system using deep neural network: A review and case studies. *Neurocomputing* 462 (2021), 364–385. https://doi.org/10.1016/j.neucom.2021.07.045

[11] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, and Zhiyong Gao. 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. https://doi.org/10.1109/CVPR.2019.01126

[12] Aksels Mercat, Mika Vaittinen, and Jarno Vanne. 2020. UVG Dataset: 50/120fps 4K Sequences for Video Codec Analysis and Development. In *Proceedings of the 11th ACM Multimedia Systems Conference (MMSys)*. 297–302. https://doi.org/10.1145/3339825.3394937

[13] Netflix Technology Blog. 2018. VMAF: The Journey Continues. https://netflixtechblog.com/vmaf-the-journey-continues-44b51ee9ed12. Accessed: 2025-07-07.

[14] Farhad Pakdaman, Mohammad Ali Adelimanesh, Moncef Gabbouj, and Mahmoud Reza Hashemi. 2020. Complexity Analysis Of Next-Generation VVC Encoding And Decoding. In *2020 IEEE International Conference on Image Processing (ICIP)*. 3134–3138. https://doi.org/10.1109/ICIP40778.2020.9190983

[15] Icaro Siqueira, Guilherme Correa, and Mateus Grellert. 2021. Complexity and Coding Efficiency Assessment of the Versatile Video Coding Standard. In *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 1–5. https://doi.org/10.1109/ISCAS51556.2021.9401714

[16] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

[17] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2018), 586–595. https://doi.org/10.1109/CVPR.2018.00068