Review

Advances in Neural Video Compression: A Review and Benchmarking

Ge Gao *, Chen Feng, Yuxuan Jiang, Tianhao Peng, Ho Man Kwan, Siyue Teng, Chengxi Zeng, Yixuan Li, Changqi Wang, Robbie Hamilton, Zihao Qi, Fan Zhang, David Bull

Posted Date: 1 April 2026

doi: 10.20944/preprints202604.0035.v1

Keywords: learning-based video coding; neural video compression; video codec comparison

 doi:10.20944/preprints202604.0035.v1

Review

Advances in Neural Video Compression: A Review and Benchmarking

Ge Gao *, Chen Feng, Yuxuan Jiang, Tianhao Peng, Ho Man Kwan, Siyue Teng, Chengxi Zeng, Yixuan Li, Changqi Wang, Robbie Hamilton, Zihao Qi, Fan Zhang and David Bull

All authors are with the Visual Information Lab, University of Bristol, Bristol, BS1 5DD, U.K.

* Correspondence: ge1.gao@bristol.ac.uk

† The project has been supported by the UKRI MyWorld Strength in Places Programme (SIPF00006/1).

Abstract

While conventional video coding standards remain predominant in real-world applications, neural video compression has emerged over the past decade as an active research area, offering alternative solutions with potentially significant coding gains through end-to-end optimization. Owing to the rapid pace of recent progress, existing reviews of neural video coding quickly become outdated and often lack a systematic taxonomy and meaningful benchmarking. To address this gap, we provide a comprehensive review of two major classes of neural video codecs—scene-agnostic and scene-adaptive—with a focus on their design characteristics and limitations. More importantly, we benchmark representative state-of-the-art methods from each category under common test conditions recommended by video coding standardization bodies. This provides, to the best of our knowledge, the first first large-scale unified comparison between conventional and neural video codecs under controlled settings. Our results show that neural codecs can already achieve competitive, and in some cases superior, performance relative to VTM and AVM, although they still fall short of ECM in overall coding efficiency under both Low Delay and Random Access configurations. To facilitate future algorithm benchmarking, we will release the full implementations and results at https://nvc-review-2025.github.io, thereby providing a useful resource for the video compression research community.

Keywords: learning-based video coding; neural video compression; video codec comparison

1. Introduction

Video content is the dominant contributor to global internet data traffic, rising from 73% in 2017 to 82% in 2025 [1] in terms of bandwidth usage. This growth has been fueled by the proliferation of video applications, including streaming, live broadcasting, and real-time personal and business communications, as well as by advances in camera technologies that support increased content resolution, frame rate, and dynamic range. This trajectory compounds the urgency for more powerful and potentially revolutionary video compression technologies, which improve coding efficiency while minimizing visual quality degradation. If this can be achieved with manageable complexity, it offers the potential for widespread deployment and future standardization, bringing economic and environmental benefits on a global scale.

Over the past few decades, video coding standards have been driven by enhancements to the classical block-based hybrid coding framework [2], ranging from early standards like H.262/MPEG-2 [3] to widely deployed standards such as H.264/AVC [4]. These have been further enhanced with more sophisticated coding tools, resulting in more recent standards such as H.265/HEVC [5] and H.266/VVC [6]. In parallel with MPEG standardization, a consortium, the Alliance for Open Media (AOM), was formed in 2015 to develop royalty-free standards, primarily for streaming applications. AOM published its first standard, AOM/AV1 [7] in 2015.

All current standardized video codecs are based on the classical block-based hybrid coding architecture, combining transform coding, quantization, and entropy coding with predictive coding

 doi:10.20944/preprints202604.0035.v1

exploiting intra-frame spatial and inter-frame motion redundancies. These codecs typically divide each frame into blocks, predicting content from surrounding areas or neighboring frames and only encoding predictive residuals. Successive standards have incrementally refined this approach, adding new tools and exploiting hardware acceleration to improve compression efficiency.

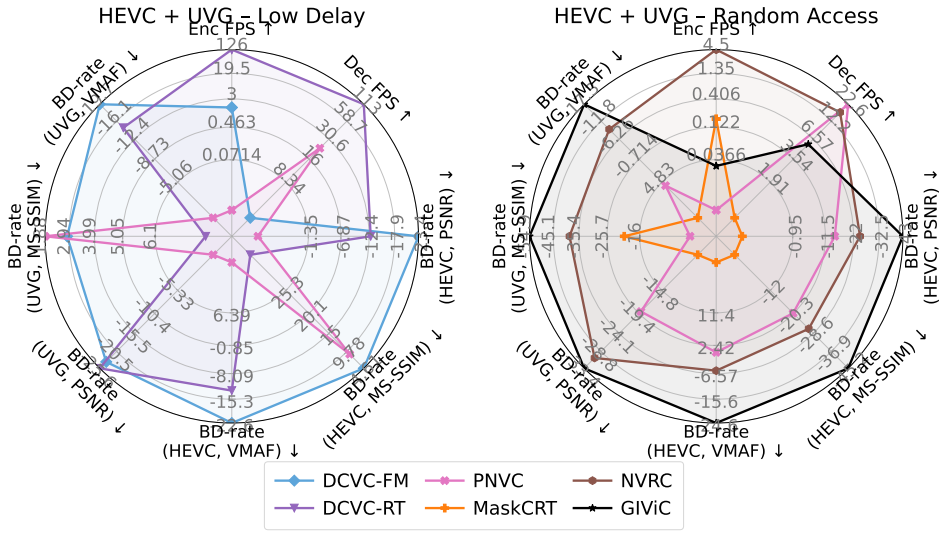


Figure 1. Radar plots summarizing the trade-off between coding efficiency and encoding/decoding complexity for selected neural video codecs (DCVC-RT, DCVC-FM, PNVC, MaskCRT, NVRC, and GIViC) under Low Delay (LD) and Random Access (RA) settings. Coding efficiency is measured by the average BD-rate (lower is better) on HEVC B-E and UVG based on quality metrics, PSNR, MS-SSIM and VMAF. We have also provided encoding and decoding runtime (FPS) figures here. Each axis in the radar plot is normalized across codecs such that a larger radius indicates better performance.

In recent years, the field of video coding has benefited from innovations in machine learning, particularly driven by advances in deep learning. Whereas early interventions focused primarily on the optimization of individual coding tools within conventional coding algorithms, more recently, neural video compression (NVC) techniques have emerged that integrate transform coding, motion estimation/compensation, context-based entropy modeling, and rate allocation into a unified, learnable framework. NVC architectures offer the advantage of being optimized in an end-to-end manner to trade off reconstruction fidelity and bitrate. Since the first neural video codec, DVC [8], was proposed in 2019, compression performance has been significantly enhanced, with methods based on both architectural [9–11] and optimization-related [12–14] innovations.

At the time of writing, state-of-the-art (SOTA) neural video codecs have been reported to match, and in some cases even outperform, conventional codecs under certain coding configurations. This trend reflects a broader shift toward learning-based video compression and opens up new opportunities for future innovations and standardization. However, the rapid evolution of NVC has also made it increasingly difficult to systematically categorize existing approaches and benchmark them fairly. Although several earlier resources [15–19] have provided valuable foundations, their coverage has quickly become outdated, and they do not offer a consistent benchmarking methodology. More recent surveys [20–22] have often focused primarily on innovations in codec architectures, while key aspects such as training strategies, evaluation protocols, and video quality metrics have received comparatively less attention. In this work, we advocate a holistic perspective that treats these elements as integral components of an end-to-end optimized framework. Furthermore, establishing fair and consistent benchmarks remains challenging, as the objective evaluation of new methods is strongly affected by variations in training setups, testing conditions, evaluation criteria, and the reporting of computational complexity.

In light of these challenges, we believe that a **comprehensive survey and benchmarking** of recent advances in neural video compression, together with a unified basis for comparison, are both timely and valuable. To this end, this paper classifies existing works into two major categories, scene-agnostic

 doi:10.20944/preprints202604.0035.v1

and scene-adaptive, tracing their historical development and categorizing key design approaches, optimization techniques, and evaluation criteria. For benchmarking, to ensure fairness, we implement and evaluate representative learned codecs under consistent conditions, based on a common test platform, using the same datasets, evaluation metrics, and codec settings. We include the test models widely used in video standardization as reference points. Throughout this study, we aim to highlight the capabilities and limitations of NVC, while identifying the best practices for rigorous evaluation. The main contributions of our work are summarized below:

- **Comprehensive survey of neural video compression**: We provide a comprehensive review of NVC methods, outlining their progression from early techniques to the latest versions. Our survey establishes a unified taxonomy, summarizing fundamental concepts and highlighting their connections to conventional codecs and broader machine learning methods, ensuring improved accessibility for readers from both domains.
- **Benchmarking under consistent test conditions**: We conduct a standardized benchmark analysis to evaluate the rate-distortion-complexity trade-offs of both conventional and neural video compression methods. By testing all approaches under consistent conditions - including identical datasets, evaluation metrics, and GOP structures - we ensure fair comparisons that provide insights into both coding efficiency and the practical viability of the selected, SOTA baselines.
- **Open framework for continuous benchmarking**: To facilitate further research in NVC and enable fast benchmarking, we design a comprehensive and extensible evaluation suite that accommodates diverse coding settings, hardware environments, and evaluation criteria. It enables unified and reproducible assessment of rate–distortion–complexity performance and serves as a living benchmark that allows the community to track progress and contribute new developments over time.

The remainder of this paper is organized as follows. In Section 2, we introduce the techniques used in conventional video codecs, which serve as guiding principles in the development of many NVC methods. Building on this, Section 3 presents a comprehensive taxonomy of end-to-end optimized NVC models, which we categorize into **scene-agnostic** and **scene-adaptive** methods. We further analyze the progression of their sub-component designs in Section 4. An overview of commonly used test protocols for NVC, including training/testing datasets and evaluation metrics, is then described in Section 5, which is followed by Section 6, presenting the results of a large-scale benchmarking study that systematically evaluates existing video compression techniques. Finally, Section 7 summarizes our empirical findings and discusses promising directions for future research. The primary perspectives related to NVC reviewed in this paper are illustrated by Figure 2.

 doi:10.20944/preprints202604.0035.v1

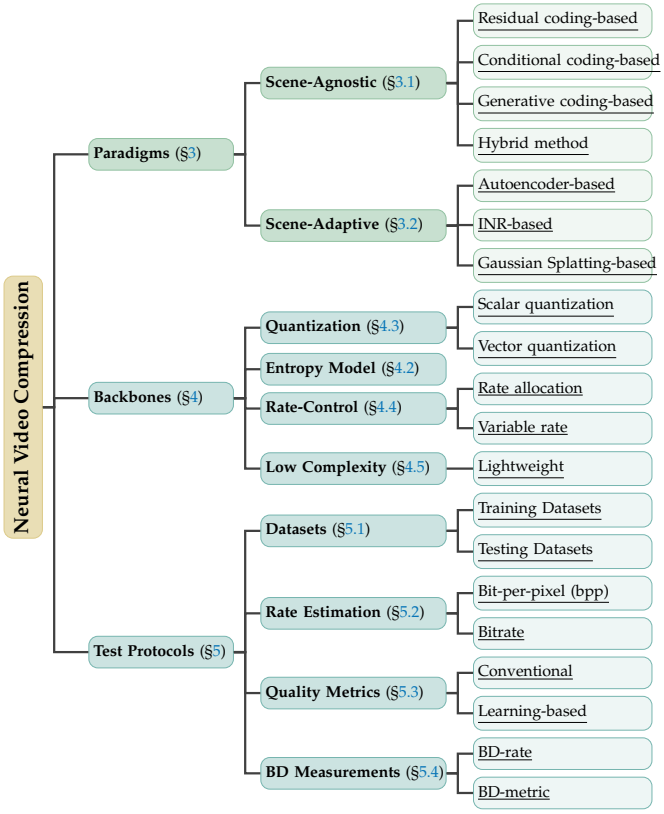


Figure 2. Overall organization of our review/survey. The taxonomy also provides an overview of the paper structure (the review sections) covering dataset, architecture, and visual quality assessment aspects.

2. Conventional Video Codecs

This section provides a brief introduction to video coding standards and their major variants, some of which serve as benchmarks in our comparison experiment. Many key components of these codecs have also inspired the design of neural video coding models. Readers are referred to [2] for a more comprehensive overview.

2.1. Video Coding Standards

Video coding standards built upon classic signal processing theories remain the mainstay of digital video communications. Over the past forty years, a series of standards has been developed that have progressively integrated more advanced coding tools and provided better support for emerging application scenarios. Currently, the most commonly used standards include MPEG H.264/AVC [4], H.265/HEVC [5], H.266/VVC [6], and AOM AV1 [23]. Each of these is associated with a reference model that offers baseline results with full features integrated - JM [24] for H.264/AVC, HM [25] for H.265/HEVC, VTM [26] for H.266/VVC, and libaom [27] for AOM/AV1. Beyond H.266/VVC, MPEG has also started the exploration of next-generation coding standards, with the working codec referred to as ECM (Enhanced Compression Model) [28]. At the time of writing, ECM has already integrated new coding tools, including advanced context-adaptive binary arithmetic coding and refined motion-vector prediction strategies, which provide evident coding gains over VVC VTM.

Similarly, AOM has started the development of AV1's successor by building a new video codec, AVM (AOM Video Model), that includes advanced features such as improved transforms (e.g., Trefftz recursion), enhanced motion-prediction schemes, and better rate-control algorithms. It should be noted that, as ECM and AVM are still evolving models, there are very few studies that comprehensively characterize their performance and computational complexity.

 doi:10.20944/preprints202604.0035.v1

2.2. Commonly Used Coding Structures

To facilitate a wide range of application scenarios, modern video coding standards define multiple commonly used coding structures (coding modes), each of which is associated with different coding latency, frame independence, and coding efficiency characteristics. While different terminologies may have been used for these coding models within the various standard bodies, here we adopt the definitions outlined by MPEG JVET. The test model MPEG VVC VTM [6] defines three primary coding modes: *All Intra*, *Low Delay*, and *Random Access*. In the All Intra coding mode, every frame is encoded independently as an intra frame, eliminating temporal prediction entirely. This mode is effectively equivalent to image compression and thus results in much lower coding efficiency compared to the other two. However, due to the lack of inter frame prediction, it does not introduce any error propagation or coding delay. In contrast, the Low Delay mode adopts an IPPP coding structure, which contains one intra (I) frame in an intra period, with others being (P) frames[1] encoded by further exploiting temporal redundancies. As the frame order for encoding remains the same as the temporal order, this configuration offers minimal coding latency, which renders it suitable for real-time communication applications such as videoconferencing and live streaming. Its coding efficiency is also significantly enhanced compared to the All Intra model due to the exploitation of both spatial and temporal redundancies within/across frames. A further mode, Random-Access, is defined to employ a hierarchical B (bi-directional prediction) frame coding structure, trading off limited coding latency for improved coding performance. This makes it suitable for applications such as on-demand streaming, which do not require real-time encoding. Due to the different coding efficiencies associated with these various coding structures, it is important to ensure that the same latency constraints are defined in the test conditions when benchmarking video codecs.

2.3. Enhancing Conventional Codecs with Deep Learning

Although the primary focus of this paper is on end-to-end optimized coding frameworks (that are independent of conventional codecs), we briefly review how deep learning techniques can also be used to enhance conventional codecs. Rather than building a completely new coding architecture from scratch, this type of approach typically enhances individual coding tools within conventional codecs. These include intra prediction [29,30], motion estimation [31,32], transforms [33], quantization [34], entropy coding [35], loop filtering [36–40], and super resolution[2] [41–43]. The learning-based super resolution tool has been further extended from spatial resolution to include temporal resolution [41] and bit depth [44,45] in the later literature. Moreover, there are recent studies that investigate the joint optimization of preprocessors and postprocessors in the form of neural wrappers for conventional codecs [46–50]. These have been reported to achieve significant coding gains, particularly when assessed using perceptually inspired quality metrics. Although these new methods have shown promise in delivering improved performance over conventional codecs, they are also associated with much higher computational complexity, in particular when the learning-based tool is integrated at the decoder. Hence, the trade-off between performance and complexity remains a challenging issue for this type of approach.

3. NVC Paradigms

A video signal typically contains significant spatiotemporal redundancy, which is exploited by an encoder to achieve substantial compression while maintaining high visual fidelity. As shown in Figure 3, in most cases, a neural video codec comprises an analysis transform, a quantizer, an entropy model (coexisting at both the encoder and the decoder), an inverse quantizer, and a synthesis transform. Specifically, given a video sequence with T frames $\boldsymbol{x}^{1:T} \sim P_{\text{data}}$ and a pre-defined coding structure

[1] Note: The definitions of P and B frames are not exactly the same as in the VVC coding configurations. In our paper, P frames refer to those that only use temporally previous frames in inter prediction, while B frames are those allowing bi-directional motion prediction.

[2] Super resolution is an established feature in AV1 [7], but it is not based on deep learning in libaom.

 doi:10.20944/preprints202604.0035.v1

$[\pi(1), \pi(2), \ldots, \pi(t), \ldots, \pi(T)]$, where $\pi(t)$ denotes the index of a frame to be encoded/decoded at the coding step t, the compression pipeline is described as follows.

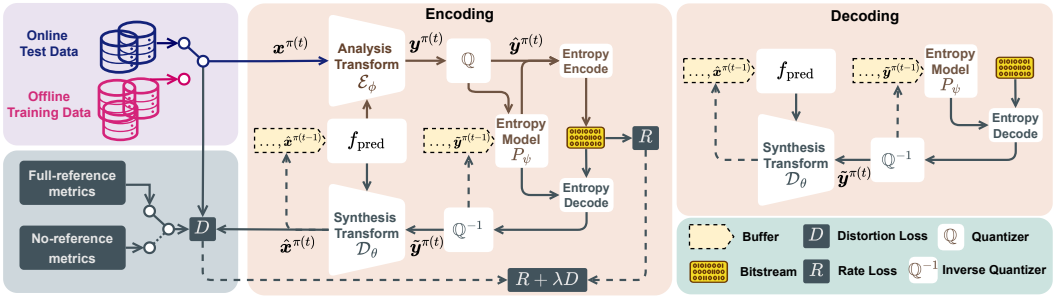


Figure 3. Overview of the neural video compression framework. The neural video codec is (optionally) trained offline to optimize the rate-distortion-(perception) performance (defined by either a reference-based metric, a no-reference-based metric, or a combination of both). For a specific test sequence, the bitstream may include online/iteratively overfitted/fine-tuned parameters encoding instance-specific representations that improve the overall rate-distortion-perception trade-off.

For the current frame to be encoded, $\boldsymbol{x}^{\pi(t)}$, the analysis transform $\mathcal{E}(\cdot;\boldsymbol{\phi})$ maps $\boldsymbol{x}^{\pi(t)}$ to the latent/transformed coefficient domain, conditioned by the representation of the previously reconstructed frames, $\hat{\boldsymbol{x}}^{\pi(<t)}$, i.e., $\tilde{\boldsymbol{x}}^{\pi(t)} = f_{\text{pred}}(\hat{\boldsymbol{x}}^{\pi(<t)})$, producing a compact latent representation $\boldsymbol{y}^{\pi(t)}$,

$$\boldsymbol{y}^{\pi(t)} = \mathcal{E}_{\boldsymbol{\phi}}(\boldsymbol{x}^{\pi(t)} | f_{\text{pred}}(\hat{\boldsymbol{x}}^{\pi(<t)})). \tag{1}$$

Here, $\boldsymbol{\phi}$ denotes the encoding parameters. $\tilde{\boldsymbol{x}}^{\pi(t)}$ represents the prediction/estimation of $\boldsymbol{x}^{\pi(t)}$ based on previously reconstructed frames and could be utilized to reduce spatiotemporal redundancies in various ways, as detailed in Section 3.1. The latent $\boldsymbol{y}^{\pi(t)}$ is then quantized by a learnable quantizer $\mathbb{Q}(\cdot)$ into $\hat{\boldsymbol{y}}^{\pi(t)}$,

$$\hat{\boldsymbol{y}}^{\pi(k)} = \mathbb{Q}(\boldsymbol{y}^{\pi(t)}). \tag{2}$$

The discrete-valued representation $\hat{\boldsymbol{y}}^{\pi(t)}$ is then losslessly entropy encoded using methods such as arithmetic coding [51] and transmitted. On the decoding side, $\hat{\boldsymbol{y}}^{\pi(t)}$ is first entropy decoded from the bitstream and inversely quantized,

$$\tilde{\boldsymbol{y}}^{\pi(t)} = \mathbb{Q}^{-1}(\hat{\boldsymbol{y}}^{\pi(t)}), \tag{3}$$

and the synthesis transform $\mathcal{D}(\cdot;\boldsymbol{\theta})$ reconstructs the video frame $\hat{\boldsymbol{x}}^{\pi(t)}$ from $\tilde{\boldsymbol{y}}^{\pi(t)}$, also conditioned on $\tilde{\boldsymbol{x}}^{\pi(<t)}$,

$$\hat{\boldsymbol{x}}^{\pi(t)} = \mathcal{D}_{\boldsymbol{\theta}}(\tilde{\boldsymbol{y}}^{\pi(t)}) | f_{\text{pred}}(\hat{\boldsymbol{x}}^{\pi(<t)})), \tag{4}$$

where $\boldsymbol{\theta}$ represents the learnable synthesis transform parameters. Such a neural video codec is typically optimized based on a rate-distortion objective $R + \lambda D$, where R denotes the expected code length of the compressed representation, λ denotes the Lagrangian parameter controlling the trade-offs, and $D = \mathbb{E}_{\boldsymbol{x}\sim P_{\text{data}}}[d(\boldsymbol{x}, \hat{\boldsymbol{x}})]$ is typically a fidelity-driven or perceptually-driven distortion metric, which quantifies how different the reconstruction is from the ground truth. Historically, this is based on squared error, i.e., $d(\boldsymbol{x}, \hat{\boldsymbol{x}}) \propto \|\boldsymbol{x} - \hat{\boldsymbol{x}}\|^2$, but perceptual metrics such as SSIM or MS-SSIM are also used in many cases. The rate term R, assuming a sufficiently efficient entropy coding technique, could be expressed as:

$$R = \mathbb{E}_{\boldsymbol{x}\sim P_{\text{data}}, \boldsymbol{y}\sim q_{\boldsymbol{y}}}[-\log P_{\boldsymbol{\psi}}(\boldsymbol{y})], \tag{5}$$

where $q_{\boldsymbol{y}}$ and $P_{\boldsymbol{\psi}}$ respectively stand for the real and estimated probability distributions by a parametric entropy model of the quantized latent representation $\hat{\boldsymbol{y}}$. Overall, the entire video coding framework can be optimized by searching for the best parameters to minimize the RD loss:

 doi:10.20944/preprints202604.0035.v1

$$\boldsymbol{\phi}^*, \boldsymbol{\theta}^*, \boldsymbol{\psi}^* = \arg\min_{\boldsymbol{\phi},\boldsymbol{\theta},\boldsymbol{\psi}} R + \lambda D. \tag{6}$$

In this paper[3], we categorize existing NVC frameworks into two primary classes: **scene-agnostic** and **scene-adaptive**. The former aims to build a generic model generalizable to diverse spatiotemporal scenarios, typically by extending the motion-compensated predictive coding design of conventional codecs or by leveraging powerful generative models. In a paradigm shift, the latter takes a different route: they overfit a bespoke, often lightweight, neural network to each individual video. This bypasses the need for a redundant synthesis transform/decoder to generalize across diverse scenarios while still achieving competitive compression performance. Notably, these two types of approaches are not mutually exclusive - the frontier of NVC has started to advance towards a hybridization of both strategies. Typical examples of these NVC methods (their model names, source code, and publication venues) have been summarized in Table 1.

Table 1. Key/milestone methods in chronological order. We categorize the scene-agnostic methods to R (residual coding), C (conditional coding), and H (hybrid coding), scene-adaptive methods to AE (autoencoder), INR (INR), and GS (Gaussian Splatting) based, and the representation type to F (frame), DC (decomposed), P (patch), and T (3D patch/tube). The decoding order could be LD (Low Delay) or RA (Random Access).

	Methods	Paper	Code	Tags	Venue
Scene-agnostic	DVC [8]	Link	Link	R + LD	CVPR'19
	SSF [57]	Link	Link	R + LD	CVPR'20
	HLVC [62]	Link	Link	R + RA	CVPR'20
	FVC [9]	Link	Link	R + LD	CVPR'21
	DCVC [74]	Link	Link	C + LD	NeurIPS'21
	VCT [75]	Link	Link	C + LD	NeurIPS'22
	CANF-VC [76]	Link	Link	R + LD	ECCV'22
	AlphaVC [70]	Link	–	H + LD	ECCV'22
	C2F [68]	Link	–	H + LD	CVPR'22
	MIMT [77]	Link	–	C + LD	ICLR'23
	DCVC-DC [72]	Link	Link	C + LD	CVPR'23
	DCVC-FM [11]	Link	Link	C + LD	CVPR'24
	MaskCRT [78]	Link	Link	H + LD	T-CSVT'24
	DCVC-RT [79]	Link	Link	C + LD	CVPR'25
	ECVC [73]	Link	–	C + LD	CVPR'25
	HyTIP [80]	Link	Link	H + LD	ICCV'25
Scene-adaptive	Lu et al. [81]	Link	–	AE + LD	ECCV'20
	Overfit-FF [82]	Link	–	AE + LD	ICLR'21
	NeRV [12]	Link	Link	INR + RA + F	NeurIPS'21
	E-NeRV [83]	Link	Link	INR + RA + DC	ECCV'22
	FFNeRV [84]	Link	Link	INR + RA + P	ACMMM'23
	Gomes et al. [85]	Link	–	INR + RA	CVPR'23
	HNeRV [86]	Link	Link	AE / INR + RA	CVPR'23
	HiNeRV [87]	Link	Link	INR + RA + P	NeurIPS'23
	NIRVANA [88]	Link	–	INR + RA + T	CVPR'23
	NVRC [13]	Link	Link	INR + RA + P	NeurIPS'24
	NeRV-Boost [89]	Link	Link	INR + RA	CVPR'24
	PNVC [90]	Link	Link	AE / INR + LD / RA + P	AAAI'25
	C3 [91]	Link	Link	INR + RA + P	CVPR'25
	COOL-CHIC-*V* [92]	Link	Link	INR + LD / RA + P	–
	GSVC [93]	Link	Link	GS + LD / RA	ICLR'25
	GIViC [14]	Link	Link	AE / INR + LD / RA + T	ICCV'25

[3] In this work, our focus is on the reconstruction of high-quality video within a specified bandwidth constraint for human consumption. This differs from machine vision–oriented applications, which is a separate research topic that lies beyond the scope of this paper.

 doi:10.20944/preprints202604.0035.v1

3.1. Scene-Agnostic Methods

Scene-agnostic frameworks are the dominant paradigm in NVC, aiming to create a single, universal model that generalizes well to any video content. By predicting the current frame from the previously decoded ones and encoding the resulting representational residual, they learn to exploit temporal redundancies from diverse data. This paradigm is broadly divided into three main sub-classes: (i) motion-compensated residual coding, which mirrors the explicit prediction steps of traditional codecs; (ii) conditional coding, which learns a more generalized, direct model of inter-frame statistical dependencies in the neural network-parameterized latent space; (iii) hybrid coding, which is a combination of the above two.

3.1.1. Motion-Compensated Residual Coding

The first forays into scene-agnostic NVC typically follow a *component-replacement strategy*, substituting the algorithmic modules of the traditional video coding pipeline with trainable neural networks. Seminal works such as DVC [8] and [52] exemplify this approach by employing optical flow networks [53] to perform motion estimation, generating dense motion vector fields between frames, and performing pixel-domain motion compensation by warping the reference frame according to this optical flow. The resulting pixel-level residual frame is subsequently compressed using a learned autoencoder, akin to those used in SOTA learned image compression [54–56].

Subsequent advancements have evolved along two primary trajectories. One line of research has focused on enhancing the motion model itself. This entails, but is not limited to: (i) inducing uncertainty-awareness in motion estimation [57,58] - the Scale-Space Flow model [57] accounts for uncertainty and motion blurriness with an additional scale field, which extends by deploying a cross-scale weighted prediction-based model [59], and [58] adopts an ensemble-based decoding framework; (ii) diversifying motion profiles by exploiting hierarchical [60,61] or multi-frame temporal redundancies [62–65]. An orthogonal yet equally significant line of work shifts the process into a learned feature space [9,66–68], generating feature-level motion estimation, performing feature-level warping via deformation, and compressing the feature-level residuals, which are more compact and hence easier to compress. Furthermore, HDCVC [69] proposes a heterogeneous deformable (HetDeform) network to perform content-adaptive optical flow warping. The recent (and more sophisticated) mechanisms often merge these strategies, employing coarse-to-fine/feature-to-pixel models [70] that refine motion from the feature space to the pixel space and leverage diverse spatiotemporal contexts for greater efficiency [71–73].

3.1.2. Conditional and Generative Coding

The *conditional coding* paradigm is closely coupled with feature-space residual coding, as both operate in the latent domain, whereas the former seeks a more generalized formulation, casting learnable neural networks to automatically capture the conditional dependency between the target frames $\boldsymbol{x}^{\pi(t)}$ and the decoded reference frames $\tilde{\boldsymbol{x}}^{\pi(<t)}$ rather than focusing on the explicit subtraction operation in residual coding schemes, based on the assumption that $H(\boldsymbol{x}^{\pi(t)}|\tilde{\boldsymbol{x}}^{\pi(<t)}) \leq H(\boldsymbol{x}^{\pi(t)} - \tilde{\boldsymbol{x}}^{\pi(<t)})$, where $H(\cdot)$ denotes the entropy measure. DCVC [74] and [94,95] popularize the conditional coding concept, with the DCVC series, in particular, having evolved over multiple generations to further refine the contextual exploitation framework: DCVC-TCM [96], DCVC-HEM [71], DCVC-DC [72], DCVC-FM [11], DCVC-LCG [10], and DCVC-RT [79]. Conditional coding has been further generalized into a spatiotemporally autoregressive transform [97,98], in which the encoding/analysis transform Equation (1) is interpreted as a *progressive whitening process* that converts the input visual signal into uncorrelated noise by gradually removing its spatiotemporal structures, while the decoding/synthesis transform Equation (4) reverses this process. This method is extended from [99,100], which demonstrates that optimizing the RD objective Equation (6) is equivalent to optimizing the variational NELBO (Negative Evidence Lower Bound Objective) of a β-autoencoder [101] (up to a constant):

 doi:10.20944/preprints202604.0035.v1

$$\mathcal{L} := \mathbb{E}_{\boldsymbol{x}\sim P_{\text{data}}}[d_{\text{KL}}[q(\boldsymbol{y}|\boldsymbol{x})\|p(\boldsymbol{y})]] + \lambda \mathbb{E}_{\boldsymbol{x}\sim P_{\text{data}},\boldsymbol{y}\sim q(\boldsymbol{y}|\boldsymbol{x})}[-\log p(\boldsymbol{x}|\boldsymbol{y})] + \text{const}, \tag{7}$$

in which d_{KL} denotes the Kullback–Leibler (KL) divergence, $p(\boldsymbol{x}|\boldsymbol{y})$ is the likelihood/generative model, $q(\boldsymbol{y}|\boldsymbol{x})$ stands for the variational posterior, and $p(\boldsymbol{y})$ represents the variational prior/entropy model.

Building on the above contributions, SOTA dependency modeling techniques have increasingly been adopted for NVC. For instance, hierarchical autoencoding frameworks [102,103] have been applied in [104,105] to represent multi-scale latent variables as a family of flexible priors and posteriors, enabling more accurate predictions of future frames. In contrast, the CANF-VC codecs [106–108] employ hierarchical normalizing flows [109,110] with exactly invertible transforms. VCT [111], MIMT [77], and [112] adopt masked modeling–based transformers [113], an expressive successor to recurrent neural network–based methods [62,114], to capture long-range spatiotemporal redundancies. The subsequent ECVC [73] and GIViC [14] have extended dependency modeling from the GoP level to the entire sequence level, achieving substantial reductions in spatiotemporal redundancy.

To improve the perceptual realism of decoded videos, particularly at ultra-low bitrates, prior work has incorporated generative priors that use learned generative models, such as GANs [115] or diffusion models [116,117], to steer reconstruction and synthesize plausible fine textures beyond distortion-oriented pixel fidelity alone. The early approach [118] incorporates adversarial and perceptual supervision as a generative prior in a learned video codec, demonstrating that perceptual realism can be improved beyond what distortion-only optimization typically yields. Recent methods such as GLC-video [119] and DiffVC [120] exemplify two representative directions: performing transform coding in a perceptually aligned generative latent space (e.g., VQ-VAE latents with spatio-temporal categorical hyperpriors) or using diffusion-based refinement conditioned on decoded representations and temporal context with efficiency-oriented reuse strategies. Building on this trend, GLVC [121] further redesigns the codec in a pretrained latent domain and introduces recurrent memory to stabilize temporal quality, while GNVC-VD [122] leverages a video-native diffusion transformer for sequence-level latent refinement to mitigate flickering under extreme bitrate constraints. In parallel, GIViC [14] introduces a pixel-space coarse-to-fine diffusion prior for INR-based video compression, denoising over a spatiotemporal pyramid to recover global structure first and progressively refine details using restored coarse tokens and decoded references.

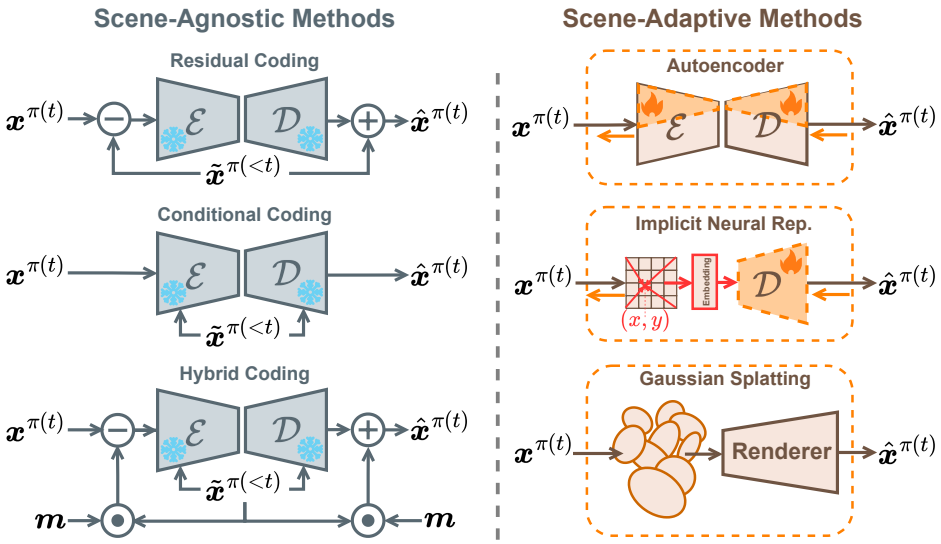


Figure 4. Visualization of Scene-Agnostic and Scene-Adaptive methods.

 doi:10.20944/preprints202604.0035.v1

3.1.3. Hybrid Coding

It has been noted that solely adopting a conditional coding scheme may underperform its residual coding-based counterpart when the cost of recovering from information loss incurred by feature extraction outweighs that of direct pixel-space subtraction - this often occurs for consecutive frames linked by slow motions [78,123]. A solution is to *hybridize* conditional and residual coding adaptively based on the spatiotemporal contents, which could be achieved via learned masking [78,95] or explicit coding mode selection mechanisms [70,124,125].

3.2. Scene-Adaptive Methods

Instead of solely relying on a fixed pretrained model, updating encoder and/or decoder parameters online to suit specific video content has been shown to help bridge the amortization gap [126,127]. Early work involved iteratively updating encoder parameters [81]. GPU [128] inserts low-rank adapters into the encoder-side modules to enable more efficient overfitting and proposes to optimize at the level of patch-based GoPs to alleviate error propagation. In contrast, [82] overfits the entire model by further transmitting quantized decoder updates, using a spike-and-slab prior [129] to simulate the binary decision between updating and not updating for each parameter. Another approach, SRVC [130], uses a separate model stream to guide a super-resolution post-processor, improving reconstruction quality.

However, these methods, which adapt a pre-trained - and, in most cases, large - backbone, are still limited by computationally intensive encoding and slow decoding speeds. More recently, INRs (Implicit Neural Representations) have emerged as a promising alternative. These are typically based on a small neural network (e.g., an MLP) whose weights are optimized to map input coordinates directly to their output values, i.e. $\mathcal{D}_{\boldsymbol{\theta}} : \mathbb{R}^3 \rightarrow \mathbb{R}^3, (x, y, t) \mapsto (R, G, B)$ in the case of videos, avoiding the expensive per-pixel data representation. Initially explored for signal representation in SIREN [131], INRs were soon adapted for compressing static images [132,133] and 3D content [134]. Building on this, [135] proposed a hybrid framework mimicking conventional codecs, in which an INR represents either full frames (I-frames) or prediction residuals (P-frames), encoding its resulting network weights into the final bitstream.

To further enhance the learning and computational efficiency of video signal representation, NeRV [12] was proposed to map frame indices to entire video frames across multiple scales based on convolutional layers, a design that effectively reduces computation compared to MLP-based approaches that generate videos pixel-by-pixel. Noting that neither image- nor frame-wise representations are ideal for video data, as they fail to capture spatio-temporal correlations, subsequent studies explored disentangled spatio-temporal representations (E-NeRV [83]), patch-wise representations (PS-NeRV [136], FFNeRV [84], HiNeRV [87]), chunk-wise representations (NIRVANA [88]), or hybridized representations [86]. Some other methods instead exploit spatiotemporal redundancies by capturing frame-wise correlations in the form of residuals (D-NeRV [137], DS-NeRV [138]), optical flows (FFNeRV [84], DNeRV [139]), or modulation parameters (HNeRV-Boost [89], HiNeRV [87]). Among these, HiNeRV achieves the best representational efficiency (outperforming HEVC/x265 *veryslow*) by leveraging a novel hierarchical encoding technique which facilitates parameter sharing across multi-level spatio-temporal granularities.

However, many of these NeRVs are optimized for general video representation tasks rather than specifically for compression. Consequently, they often employ post-training pruning and quantization techniques, which result in sub-optimal compression efficiency compared to the end-to-end optimized autoencoder-based NVCs. To address this, [85] proposed an end-to-end solution that optimizes the model parameters and their associated quantization parameters with entropy-aware penalization, which has been adopted and extended by subsequent works [89,140]. NVRC [13] further proposes to subsume both quantization and entropy models into the end-to-end training regime, using an innovative dual-axis conditional Gaussian model for entropy modeling the quantized network parameters; this was reported to offer superior compression performance compared to VVC VTM-20.0 (*randomaccess*).

 doi:10.20944/preprints202604.0035.v1

Another major focus of INR-based video compression has been on reducing decoding complexity. [141] proposed to cast group convolutions for parallel decoding of multiple sequences; COOL-CHIC-video [92] extended the COOL-CHIC model [142] with an additional inter coding module, achieving comparable performance to AVC x264 *medium* [4] using only 900 multiplications per pixel; C3 [91] was reported to offer comparable performance to NVC models with 1000× its (decoding) complexity (e.g., VCT [111]) in terms of MACs/pixel. This was achieved through improvements in the quantization-aware optimization process and context modeling. Furthermore, NVRC-Lite [143] improves the decoder architecture and adopts an octree-based entropy model, outperforming C3 while achieving 2.5× faster decoding.

Recently, the trend of hybridizing offline pre-training with the online overfitting of INR models has emerged. PNVC [90] first proposed leveraging this strategy to enable frame-wise overfitting of INR-based video codecs. This approach considerably reduces encoding latency by allowing a single set of learnable parameters to represent a group of consecutive frames while still performing on par with VVC VTM-20.0 in both Low Delay and Random Access modes. GIViC [14] further leverages linearized transformer- and diffusion-based backbones that realize dependency modeling across the entire video sequence architecturally and optimization-wise, resulting in coding performance superior to NVRC and VVC VTM-20.0 (under the same coding configurations).

In parallel, Gaussian Splatting (GS) [144] has also been applied with some success to video compression. Initially a 3D reconstruction technique valued for its rapid speed and real-time rendering, GS differs from INRs by using a set of explicit 3D Gaussian points to model a scene. This approach enables new applications, such as representing 4D dynamic videos by moving and deforming these points in a deformation field [145]. GSVC [93] proposes a novel Toast-like sliding window design that exploits considerable temporal redundancy, as well as an end-to-end framework comprising spacetime-adaptive modulations and deformations. It achieves comparable performance to NeRV [12] but with 30% faster rendering speed. In contrast, other studies [146–148] adopt 2D Gaussians based on GaussianImage [149], which require fewer parameters and are thus easier to compress. More recently, GFix [150] targets the perceptual artifacts produced by 3DGS-based codecs with a plug-and-play single-step diffusion enhancer and a compressible modulated-LoRA adaptation, reporting sizable perceptual BD-rate gains over GSVC. However, research into GS-based compression of 2D natural videos is still nascent, and these methods do not yet match the performance of SOTA autoencoder- and INR-based codecs.

4. NVC Key Modules

Underpinning both scene-agnostic and scene-adaptive paradigms is a shared toolkit of fundamental building blocks that constitute the modern NVC backbone. In this section, we review each of these components, tracing their origins and highlighting their main contributions to the NVC framework.

4.1. Entropy Coding

Entropy coding achieves lossless compression by encoding a sequence that represents a discrete random variable, i.e., the message, by assigning codeword lengths to symbols according to their information content or level of surprise. According to Shannon's source coding theorem [151], the length of each codeword in an optimal prefix-free code is roughly proportional to the negative logarithm of its probability.

4.1.1. Huffman Coding

Huffman coding [152] is one of the most widely used prefix-free (i.e., no codeword is a prefix of any other codeword) symbol coding approaches. The method builds a binary tree: starting with all symbols as leaves weighted by their probabilities. It was used in early image and video coding standards, such as H.262/MPEG-2 [3] and H.264/AVC [4], but has been replaced by more advanced entropy coding methods, e.g., arithmetic coding, in more recent standard codecs and neural video coding methods.

 doi:10.20944/preprints202604.0035.v1

4.1.2. Arithmetic Coding

Since probabilities seldom match exact powers of two, Huffman coding can be inefficient, potentially resulting in up to one extra bit per symbol. Arithmetic coding [153] addresses this by encoding the entire message at once, based on the idea of selecting a single real number in $[0, 1)$ that resides within an interval whose size corresponds to the message's probability, so that more probable messages occupy larger intervals and require fewer bits to pinpoint. As the resulting interval's length equals the product of conditional probabilities, the code length is essentially $-\log_2 Q(\boldsymbol{x}^n)$ bits, up to a constant of at most about $+1$ bit. Average under the true source P, this yields at most $H[P, Q] + 2$ bits, which is roughly $2/n$ bits per symbol. Based on the FIFO (first-in-first-out) queue data structure, it also makes arithmetic coding especially effective for long sequences. Due to its advantages, arithmetic coding (and its advanced variants) is the dominant entropy coding approach in current video coding standards (e.g., H.266/VVC [154] and AOM/AV1 [155]); it is also an obvious choice for autoregressive compressors, as discussed in Section 4.2.

4.1.3. Asymmetric Numeral Systems

Asymmetric Numeral Systems (ANS) [156] extend numeral representations to efficiently encode non-uniform symbol distributions. ANS maintains a single integer "state" that evolves during encoding and decoding, and, unlike the FIFO processing in arithmetic coding, it operates in a last-in-first-out (LIFO) manner, recovering the most recently encoded symbol first.

The method generalizes positional numeral systems: in a base-B representation, encoding multiplies the state by B and adds the symbol index, while decoding applies a modulus and division. For uniform probabilities, binary or decimal systems are optimal. ANS adapts this to arbitrary probabilities by mapping symbols to subintervals of the unit interval and discretizing finely; each grid point is assigned to a symbol's subinterval, creating an "effective" alphabet that preserves the target distribution without redundant representations. ANS achieves compression efficiency comparable to arithmetic coding with at most two bits of overhead per message, and some variants admit a "bits-back" interpretation [157], enhancing flexibility in probabilistic modeling.

4.1.4. Relative Entropy Coding

In contrast to conventional entropy coding methods, relative entropy coding (REC) encodes a continuous sample from a target distribution Q using a reference distribution P shared by the encoder and the decoder. This achieves an expected code length equal to the KL divergence $D_{\mathrm{KL}}[Q\|P]$. The method is built on the *shared randomness principle*: both parties generate an identical pseudo-random sequence of samples from P, and the encoder transmits only the index of the first sample that would be drawn from Q under a prescribed acceptance rule. The decoder, using the same randomness, recovers the exact sample without quantization. The detailed description of REC can be found in [158–160].

4.2. Entropy Models

To improve the rate-distortion performance for neural lossy compression, numerous entropy models have been proposed. The earliest and simplest approach is the factorized prior, which assumes that all elements in the latent tensor $\hat{\boldsymbol{y}}$ are statistically independent. The joint probability is therefore simply the product of the marginal probabilities of each element: $P(\boldsymbol{y}) = \prod_i p(y_i)$. While computationally fast and fully parallelizable, this model fails to capture the significant spatial correlations present in the latent space, leading to suboptimal compression performance.

Subsequent contributions focused on applying hierarchical latent-variable modeling and autoregressive modeling, as well as their hybridization, with the goal of increasing the flexibility of the prior density and achieving a better compression bitrate. **Hyperpriors** [161] or **hierarchical priors** are the most common and foundational priors for entropy modeling in both neural image and video compression. Their core idea is to introduce an additional hierarchy of latent variables (called hyper-latents) $\boldsymbol{z}$ to efficiently model the distribution of the main latents $\boldsymbol{y}$. The hyperprior density $p(\boldsymbol{z})$ is usually modeled by a factorized prior, based on which $\hat{\boldsymbol{z}}$ is transmitted to the decoder first and used by

 doi:10.20944/preprints202604.0035.v1

the hyper-decoder as *side information* [161] to estimate the parameters for the conditional probability distribution (e.g., factorized Gaussians) of the main latents, $p(\boldsymbol{y}|\hat{\boldsymbol{z}})$. This has been extended to multiple hierarchies [104].

Drawing inspiration from successful generative models such as PixelCNN [162], another class of entropy models has emerged that directly captures local dependencies (spatial or temporal) using **autoregressive priors**. Here, the probability of each latent element $\hat{y}_i^t$ is made explicitly conditional on its previously decoded spatial neighbors $\hat{y}_{<i}^t$ and, optionally, the temporal neighbors $\hat{\boldsymbol{y}}^{<t}$. This is typically implemented with a masked convolutional network. Although fully autoregressive models excel at capturing fine-grained local details, their sequential processing - encoding one position at a time - is prohibitively slow for real-time applications, often requiring hundreds of steps for a single HD video frame.

To overcome this bottleneck, **semi-autoregressive (or group-wise) priors** have been proposed, which drastically reduce the coding process to just a few steps while maintaining strong performance. A key advantage of this strategy is its ability to access bi-directional (spatiotemporal) contexts, which contrasts with the unidirectional, raster-scan approach of fully spatial autoregression. The techniques employed include spatial partitioning [163–165], channel partitioning [56,166], mixtures of both [56,72], and content-adaptive masked modeling [77,167].

Recent designs have shifted to a more generalized formulation of the **joint autoregressive-hierarchical priors** [168], which exploit spatiotemporally hierarchical and autoregressive contextual information for entropy estimation. Notably, models like NVRC [13], PNVC [90], and GIViC [14], which utilize multi-resolution spatiotemporal latent grids, exploit context from both the previous hierarchy and from regions already decoded at the current hierarchy to achieve optimal performance-complexity trade-offs.

Besides the optimization of decoding patterns, the architecture used for entropy modeling is another important research topic. While masked CNN-based backbones [11,74] still dominate, studies [14,77,111,167] have also explored RNNs and Transformers in order to better capture long-range dependencies. [62] pioneered the adoption of RNNs for entropy estimation; VCT [111] leverages an encoder-decoder vanilla Transformer [169], which performs group-wise semi-autoregressive context modeling, while MIMT [77] and CGT [167] extend this by utilizing the sliding window attention [170]. GIViC [14] instead leverages a gated linear transformer for entropy modeling, whose complexity scales linearly instead of quadratically with the context length.

4.3. Quantization

Quantization is essential in most lossy compression methods because it creates an information bottleneck that limits the precision of the data and allows it to be stored more compactly. The quantization operation creates a many-to-one mapping from input to a discrete set and has zero derivatives almost everywhere; it is thus undefined at points of discontinuity, meaning gradients cannot be backpropagated through the quantizer to the encoder transform. Here, we introduce several types of quantization alongside their corresponding differentiable surrogates. We also cover methods for quantizing neural network parameters.

4.3.1. Vector Quantization

Vector quantization (VQ) is a classical signal processing technique. A VQ of size N is defined as a mapping from $\mathbb{R}^c$ (c is the dimensionality of the latent variable) to the closest codebook vector from a finite set $\boldsymbol{c} = \{\boldsymbol{c}_i \in \mathbb{R}^c | i = 0, 1, \ldots, N-1\}$, i.e., $[\boldsymbol{y}] := \mathrm{VQ}(\boldsymbol{z}, \boldsymbol{c}) = \boldsymbol{c}_j$ where $j := \arg\min_i ||\boldsymbol{y} - \boldsymbol{c}_i||$. VQ is optimal because it partitions the signal space into arbitrarily shaped regions that adapt to the underlying probability distribution, allowing it to allocate higher precision (smaller regions) to more probable signal values and achieve a more geometrically efficient packing of the vector space. However, it is associated with issues of non-differentiability, codebook collapse, and high computational cost. We introduce the progression of different VQ variants designed to mitigate these issues and discuss their respective design trade-offs in the context of image/video compression.

 doi:10.20944/preprints202604.0035.v1

The non-differentiability issue associated with hard codeword assignment and a soft quantization design has been circumvented in [171], which linearly combines codewords based on their distance to $\boldsymbol{y}$: SoftVQ$(\boldsymbol{y}, \boldsymbol{c}) := \sum_i \phi_i \boldsymbol{c}_i$ with $\phi = \phi(\boldsymbol{z}, \boldsymbol{c}) = \text{Softmax}(-\sigma[||\boldsymbol{y} - \boldsymbol{c}_0||^2, \ldots, ||\boldsymbol{y} - \boldsymbol{c}_{N-1}||^2])$. Here $\sigma > 0$ denotes the temperature hyperparameter. By progressively increasing σ toward infinity throughout the course of optimization, the soft approximation asymptotically approaches the hard quantization employed in the final model.

A larger codebook size N is generally expected to enhance representational precision and support higher bitrates. However, in practice, naively increasing the number of centroids (codewords) is computationally prohibitive, and the resulting large codebooks are notoriously difficult to train, as each centroid has a very low probability of being updated. This has been addressed in [164,172] with product quantization (PQ), which partitions the $\boldsymbol{y}$ into M sub-vectors, each of which is quantized using the same small codebook of size N_s, enabling combinatorial generalization to $N = N_s^M$ possible representations while keeping the per-subvector codebook size manageable. The scalability issue has been tackled in [173] by further combining PQ with residual quantization [174], where only the per-hierarchy latent residuals are sliced and product quantized. A similar strategy has also been adopted by VQ-NeRV [175].

Several approaches address the instability and scalability issues associated with vanilla VQs. Rather than performing a full nearest-neighbor search, LVQAC [176] uses structured lattices (such as diamond lattices) that are split into two simple scalar quantizations: one on the standard square lattice and the other on the lattice shifted by half a quantization step. The closest candidate is then selected via arg min. OLVQ [177] casts a learnable VQ that optimizes the lattice basis itself, combining Babai's rounding [178] for efficient, differentiable encoding with orthogonality regularization for accurate entropy modeling.

Moreover, FSQ (Finite Scalar Quantization) [179] and its binary variant LFQ (Lookup-Free Quantization) [180] propose per-dimension scalar quantization (equivalently, product quantization with subvectors of dimensionality equal to 1), where the mapping from $\boldsymbol{x}$ to the discrete codeword is done by applying a bounding function (e.g., $\lfloor L/2 \rfloor \tanh(\boldsymbol{x})$) and then rounding to the nearest integers, where $L = [L_1, \ldots, L_d]$ denotes the number of levels per channel. It is also noted that some other techniques [181–183] have been proposed for optimizing visual tokenizers for generative models, but these are beyond the scope of this paper, which focuses on neural video compression.

4.3.2. Scalar Quantization

Scalar quantization (SQ), sometimes referred to as uniform quantization [184,185], rounds each element of $\boldsymbol{y}$ to the nearest integer. While it is a simplified scalar version of the VQ approach, the analysis transform/encoder $\mathcal{E}_\phi$ learns to untangle and warp the complex distribution of the source data into a latent space where a uniform quantization grid is sufficient [184,186]. Thus, the benefits of SQ, namely its lower computational cost and optimization stability, can be achieved without sacrificing performance. Similarly, SQ is also associated with the non-differentiability issue, which has been addressed in [185] via the Straight-Through Estimator (STE). Moreover, [184] proposes to simulate rounding with additive uniform noise, i.e., $\lfloor \boldsymbol{y} \rceil \approx \boldsymbol{y} + \boldsymbol{u}, \boldsymbol{u} \sim \mathcal{U}([-\frac{1}{2}, \frac{1}{2})^n)$. Here the entropy term 5, originally defined only for integers, is convolved with the uniform noise $\boldsymbol{u}$, i.e., $\tilde{P} := P * \mathcal{U}([-\frac{1}{2}, \frac{1}{2})^n)$, which "smears out" the spike over its corresponding unit width bin, so that the resulting integral of CDF (Cumulative Density Function) agrees with P on all integer points and provides a differentiable surrogate for the discrete entropy. However, it has also been reported that VQ [173] is still beneficial even with the cascaded non-linear transforms.

Instead of mimicking the effect of quantization with additive uniform noise, [187] proposes the replacement of the hard-rounding operator $\lfloor \cdot \rceil$ with a differentiable approximation $s_\alpha(\boldsymbol{y})$ that is universally differentiable:

$$s_\alpha(\boldsymbol{y}) = \lfloor \boldsymbol{y} \rceil + \frac{1}{2} \frac{\tanh(\alpha \cdot (\boldsymbol{y} - \lfloor \boldsymbol{y} \rceil - 1/2))}{\tanh(\alpha/2)}, \tag{8}$$

 doi:10.20944/preprints202604.0035.v1

where α controls the fidelity of the approximation, i.e., $\lim_{\alpha\to 0} s_\alpha(\boldsymbol{y}) = \boldsymbol{y}$ and $\lim_{\alpha\to\infty} s_\alpha(\boldsymbol{y}) = \lfloor \boldsymbol{y} \rceil$. When α is large, the derivatives of s_α tend to have high variance, which they address by taking the derivative of the expectation (instead of the other way around). Over the course of training, it is generally advisable to progressively "anneal" α, or other temperature parameters of similar nature, like ϵ-STE [188] and the sigmoid-based gradient decay [189], to encourage exploration in the earlier stages and reduce train-test mismatch at later stages. A stochastic encoder has been further advocated in [126], which is gradually annealed to a deterministic one by controlling hyperparameter τ, where the posterior encoding transform is defined as follows.

$$q(\boldsymbol{y}|\boldsymbol{x}) := \prod_i q(\boldsymbol{y}_i|\boldsymbol{x}), \tag{9}$$

$$q(\boldsymbol{y}_i|\boldsymbol{x}) \propto \begin{cases} \exp\{-\psi(\boldsymbol{y}_i - \lfloor \boldsymbol{y}_i \rfloor)/\tau\}/C, & \text{if } \hat{\boldsymbol{y}}_i = \lfloor \boldsymbol{y}_i \rfloor \\ \exp\{-\psi(\lceil \boldsymbol{y}_i \rceil - \boldsymbol{y}_i)/\tau\}/C, & \text{if } \hat{\boldsymbol{y}}_i = \lceil \boldsymbol{y}_i \rceil \end{cases}, \tag{10}$$

where C is the normalizing constant and $\psi = \tanh^{-1}$. C3 [91] adopts a similar annealing strategy but replaces the standard uniform noise with samples from the Kumaraswamy distribution [190]. Although akin to the Beta distribution in its flexibility - able to represent uniform, overdispersed, and underdispersed distributions - the Kumaraswamy method offers far greater sampling efficiency.

4.3.3. Network Quantization

Network quantization techniques [191] have also been adopted for (i) addressing the cross-platform reproducibility issue, where cross-platform differences in floating-point arithmetic often lead to mismatched probability distributions in entropy models, which can cause the decoder to fail when reading the compressed bitstream from the encoder; (ii) improving compression (in the case of overfitted codecs) and inference efficiency. [192] first identifies the cross-platform reproducibility issue and proposes to perform integer arithmetic with the neural network-based entropy models. [193] proposes an end-to-end image compression method that uses fixed-point weights and activations. After exploring various quantization schemes, they identified a channel-wise non-linear approach as optimal based on a coding gain analysis. An entirely different strategy avoids these arithmetic challenges by using codebook-based vector quantization. This method transmits only the codebook indices, thereby circumventing the use of autoregressive entropy models that are particularly susceptible to catastrophic error accumulations. [194] integerizes the network arithmetic by applying post-training quantization (PTQ) to weights, with symmetric per-channel quantization, and activations, with asymmetric per-tensor quantization. MobileCodec [195] and MobileNVC [196] adopt the same scheme to quantize their codecs to 8 bits without performance loss. The more recent DCVC-FM [11] and DCVC-RT [79] quantize their decoder to 16 bits instead.

4.4. Rate Allocation and Adaptation

For practical video communication usage, video codecs must be able to adapt to a given target bitrate (bandwidth) while achieving optimal reconstruction performance. While most existing NVC models only support a single rate per trained copy of weights, applying one average bitrate to an entire video may cause inconsistent quality, under-allocating bits to complex scenes and wasting them on simple ones. Moreover, this fixed-rate approach adapts poorly to bandwidth changes and requires loading multiple models to compensate, thereby increasing memory usage and latency. This issue can be tackled by rate control algorithms, as discussed below.

4.4.1. Rate Allocation

Rate allocation is the process of distributing a target bitrate budget, formalized as tractably searching for an optimal λ at granularity ranges from frame chunks to pixels. This is typically addressed using empirical models of rate-quality dependency [197] in conventional video codecs. The pioneer of rate allocation for NVC [52,198] builds upon a λ-domain adaptation framework [197] and

 doi:10.20944/preprints202604.0035.v1

proposes a quality dependency model based on inter-frame dependency. Following this direction, more sophisticated rate-quality models have been developed that explicitly factor in content complexity, using native complexity descriptors like region-of-interest [199] and motion variance to achieve more accurate and meaningful bit allocation [200]. This has been extended to learning a neural network, which processes deep spatiotemporal features for rate allocation and implementation [201]. Rate allocation could be formulated as a constrained sequential decision-making problem, such as in [202], which is solved using a MuZero-based [203] rate controller with self-competition. A more recent attempt in [204] leverages a Reinforcement Learning-based agent to determine QPs at macro-block levels that optimize the bit-budget-downstream-tasks-performance trade-offs. Alternatively, it has been proven in [205] that semi-amortized variational inference (SAVI) with a GoP-level likelihood exactly corresponds to optimal pixel-wise bit allocation in NVCs, and this can then be extended into a multi-level latent framework with a practical approximation of optimization-based bit allocation. The formulation is closely related to the sequence-level or GoP-level bitrate allocation via the stochastic optimization line of work, as discussed in Section 3.1.

4.4.2. Rate control

A straightforward solution to rate control (i.e., realizing variable rate in a single pre-trained model) is latent feature modulation [71], which is based on learning a fixed set of spatio-temporal quantization (and inverse quantization) steps and smoothly interpolating between discrete steps to enable wider bitrate points. DCVC-DC [11] improves this by formulating the quantization steps to increase exponentially with a linear increase of quantization parameters (QP), whereas [201] proposes to learn this non-linear mapping via a learnable rate-implementation network. The rate allocation/adaptation design is further extended to the joint rate-distortion-*complexity* trade-offs. For instance, AlphaVC [70] proposes to mask out high-confidence latent elements from being entropy encoded. Another line of works [206–208] utilizes slimmable neural networks [209,210], which dynamically (and structurally) prunes latent channels for jointly reducing bitrate and decoder inference complexity.

4.5. Complexity Reduction

In addition to the challenges of rate adaptation and cross-platform reproducibility, another key practical limitation of existing NVC models is their slow decoding speed. While previous sections have discussed strategies such as overfitting-based acceleration (Section 3.2), semi-/non-autoregressive entropy models (Section 4.2), and slimmable decoders (Section 4.4), here we further introduce some complementary lightweight techniques that could be adopted to further reduce decoder-side complexity and improve runtime efficiency.

Structured pruning [211] prunes regular regions of weights such as neurons, channels, or attention heads, and is a commonly used, model-agnostic technique for neural network inference acceleration. It reduces the width of pre-trained NVCs whilst maintaining the compression performance via a multi-stage knowledge distillation strategy [189]. This distillation scheme is also adopted by PNVC [90] to progressively distill the quadratic-complexity self-attention layers into normalization layers.

Several efficiency-driven neural video codecs [14,79,90,105] have also been proposed to use patchification and omit explicit warping-based motion compensation, thus reducing the operational cost of NVC models. Among these methods, DCVC-RT [79] has conducted a detailed analysis of the computational-operational complexity of typical NVC models, based on which it significantly improves algorithmic efficiency and offers real-time decoding processing on a single commercial GPU.

5. Evaluation Protocols for NVC

The optimization and evaluation of NVCs is a multi-faceted process that must be grounded in a standardized methodology. It begins with the use of common **datasets** for training and testing, respectively, ensuring that comparisons between different models are fair and reproducible. The core performance is then judged on the basis of **rate-distortion trade-off**, which involves measuring the **bitrate** required to represent the video and **assessing the visual quality** of the reconstructed output

 doi:10.20944/preprints202604.0035.v1

using a range of objective metrics from simple PSNR to perceptually-aligned models like VMAF and LPIPS. Finally, the overall efficiency of a codec is typically condensed into a single comparative metric, the **Bjøntegaard-Delta (BD) rate**, which measures the average bitrate savings of one codec over another at equivalent visual quality, establishing a single, standardized score for evaluation.

Table 2. Summary of widely used 2D video compression training and evaluation datasets. *: denotes the image dataset.

Dataset	# Sequences	Tot. # Frames	Resolution	Link
Training dataset				
DIV-2K* [212]	N/A	1,000	2K/1440p	
REDS [213]	300	30,000	HD/720p	
Vimeo-90k [214]	89,800	628,600	256 × 256	
HIF [215]	182	51,335	240p–1080p	
BVI-DVC [216]	800	5,200	270p–2160p	
BVI-AOM [217]	956	61,184	270p–2160p	
TVD [218]	86	5,590	4K/2160p	
Testing Dataset				
JVET-CTC [219]	22	11,994	270p–2160p	
AOM-CTC [220]	48	6,240	270p–2160p	
HEVC [219]	20	300–600	1080p	
UVG [221]	7	3,900	1080p	
MCL-JCV [222]	30	4,115	1080p	

5.1. Datasets

5.1.1. Training Datasets

The **DIV2K** dataset [212] is widely used to optimize video compression performance for I-frames. Originally from a super-resolution challenge, it provides 800 high-fidelity 2K images. Training an I-frame autoencoder on this detail-rich data helps models learn to preserve fine textures, yielding higher-quality anchor frames for subsequent predictions.

Complementing the focus on spatial quality, the **REDS** (Realistic and Dynamic Scenes) dataset [213] targets the challenge of compressing complex motion. It contains 240 training clips (100 frames, 720p) that feature significant motion and camera shake, making the content more demanding than typical training sets like Vimeo-90k. Consequently, its inclusion helps improve motion estimation and compensation modules in a neural video codec, enhancing compression performance on dynamic sequences.

Vimeo-90k [214] is one of the most widely used databases for training NVC models, which consists of 89,800 short video clips. Each sequence contains only 7 frames at a resolution of 448×256. The dataset is sourced from Vimeo.com, featuring videos without inter-frame compression to prevent codec artifacts. These videos were then automatically segmented into distinct shots. To ensure content variety, shots with similar backgrounds were removed using GIST features. Finally, only frame sequences with an average motion between 1 and 8 pixels, as calculated by SpyNet, were included in the dataset.

The **HIF** [215] database is designed for training neural in-loop filters, particularly for the HEVC standard. It was generated from 182 raw video sequences, which were encoded using HEVC at four different QP values to produce a range of compression artifacts. The resulting dataset provides pairs of original frames and their distorted counterparts, ideal for training a network to perform artifact reduction and enhance a hybrid codec's decoding loop.

BVI-DVC [216] is specifically created for data-driven video compression research. It contains 800 video sequences with a broader range of resolutions, from 270p up to 2160p. Each sequence is 64 frames long, which is much longer than the clips in Vimeo-90k [214]. This increased temporal length is beneficial for training models that aim to capture longer-term dependencies and improve temporal prediction, potentially leading to better performance on longer video sequences during inference. More recently, its extension, **BVI-AOM** [217], has been introduced to further improve content diversity

 doi:10.20944/preprints202604.0035.v1

and, more importantly, address the limited copyright issues within BVI-DVC. These two databases have been reported to offer evident coding gains compared to DIV2K, REDS, and HIF when used for training multiple coding tools. Recently, they have been adopted by MPEG JVET [223] and the Alliance for Open Media (AOM) [224] for developing neural network based coding modules.

The Tencent Video Dataset (TVD) [218] is a key training resource developed to advance video compression and machine vision technologies, playing a formal role in standardization activities such as MPEG's Video Coding for Machines (VCM) and JVET's Neural Network Video Coding (NNVC). Composed of 86 high-resolution (4K) video sequences, each 65 frames long, its primary purpose is to support both the training of neural network-based coding tools and the testing of machine vision tasks, such as object detection and tracking.

5.1.2. Test Datasets

For conventional video codecs, the most widely used benchmark is the sequences defined in **MPEG JVET/JCT-VC Common Test Conditions (CTC)** [219]. While this official benchmark includes high-resolution sequences up to 3920×2160 (Class A), most evaluations in neural video compression, as noted in recent works [11,72,74], are based on lower resolutions, e.g., Class B (1920×1080), Class C (832×480), Class D (416×240), and Class E (1280×720). These sequences, typically with 300 to 600 frames, feature a wide range of content, including fast and slow motion, complex textures, and camera pans, providing a rigorous test for any compression algorithm.

Similarly, the Alliance for Open Media (**AOM**) defines its Common Test Conditions (**CTC**) for evaluating codecs such as AV1 [220,225]. This features a diverse set of sequences, including not only camera-captured video but also extensive screen content and gaming clips, spanning a wide range of resolutions from 2160p down to 360p.

In the wider literature, commonly used test video datasets also include **UVG** [221], which contains seven high-quality, 1080p video sequences. Each of these sequences has 300 or 600 frames at a frame rate of 50 or 60 fps. The **MCL-JCV** [222] dataset is another widely used benchmark, which includes 30 diverse video sequences at 1080p resolution, with lengths varying from 120 to 150 frames. It provides a larger set of test cases than UVG or individual HEVC classes, offering a more comprehensive evaluation across various modern, realistic scenes.

5.2. Rate Estimation

In video compression, the rate (Eq. 5) refers to the code length used to encode the video signal. This is primarily measured using two metrics: bits per pixel (bpp) and bitrate. Bpp is a normalized measure representing the average number of bits required to store a single pixel, making it ideal for comparing compression efficiency across different spatial resolutions: a lower bpp signifies more effective compression. Bitrate, typically measured in bits per second (bps), quantifies the data rate of the video stream over time. It directly corresponds to the final file size and the bandwidth needed for transmission, making it a crucial metric for practical applications like online streaming. While these two can be easily interconverted, bpp is more commonly used in image compression, and bitrate is typically used in video coding.

5.3. Quality Measures

The performance of a video codec can be measured by the perceived visual quality of the reconstructed content. In principle, perceptual video quality can be assessed through controlled subjective experiments, where human observers provide quality judgments on distorted videos by comparing them to their uncompressed reference versions under standardized protocols (e.g., ITU-R BT.500). The collected ratings are subsequently aggregated into Mean Opinion Scores (MOS) or, in degradation-category designs, Differential Mean Opinion Scores (DMOS), which are widely regarded as the ground-truth proxy for human visual quality. Despite their fidelity to perception, subjective studies are inherently resource-intensive and difficult to scale, which limits their practicality for iterative and up-to-date codec development.

 doi:10.20944/preprints202604.0035.v1

To enable efficient video quality assessment, objective video quality assessment (VQA) models are employed to complement subjective testing, which approximates perceptual quality through algorithmic predictors. Objective quality metrics are canonically classified into three categories based on the availability of the source content: full-reference (FR), reduced-reference (RR), and no-reference (NR) models. In video codec evaluation, FR models are most frequently adopted to allow direct quantification of distortion relative to the source. Among FR measures, pixel-wise fidelity metrics such as Peak Signal-to-Noise Ratio (PSNR) remain widely reported due to their simplicity and analytical convenience, yet their correlation with human judgments is often limited, particularly for structured artifacts. To better account for perceptual characteristics, structural similarity based measures, exemplified by SSIM [226] and its variants [227–230], compare the reference and the reconstruction in luminance, contrast, and structure components, yielding improved, albeit still suboptimal, agreement with human perception. Beyond structural similarity, a line of perceptually motivated FR models explicitly incorporates human visual system principles: Visual Information Fidelity (VIF) [231] formulates quality as information loss under an information-theoretic view, Most Apparent Distortion (MAD) [232] models distortion visibility via contrast sensitivity and related perceptual mechanisms [233], and MOVIE extends quality modeling to the temporal domain by assessing distortions along motion trajectories [234].

Inspired by advances in machine learning, a prominent line of work formulates objective VQA as a supervised regression problem, where a set of carefully designed quality-aware features are fused by a learned regressor calibrated against subjective ground truth. A representative instance is Netflix's Video Multi-Method Assessment Fusion [235], which performs full-reference quality prediction by regressing spatial and temporal features. Subsequent studies have shown that explicitly incorporating efficient temporal evidence and model ensembling can further improve VMAF's sensitivity to diversified temporal distortions, enabling better exploiting motion-related cues and artefact adaptivity [236,237]. A well-documented caveat of learned fusion is vulnerability to hacking by pre- or post-processing that boosts the predicted score without commensurate perceptual improvement [238], leading to the "No Enhancement Gain" (NEG) VMAF [235].

Learning-based quality predictors have increasingly shifted from hand-crafted feature engineering to end-to-end models adapted with subjective annotations. While early deep perceptual metrics such as LPIPS [239] and DISTS [240] demonstrate that distances in deep feature spaces can better align with human judgments than pixel fidelity, they are primarily formulated for images and do not explicitly model temporal perception. Modern VQA models, therefore, incorporate spatiotemporal representations via 3D CNNs or transformer-style video backbones, together with efficient temporal sampling or fragment aggregation to balance accuracy and computational cost. Representative ranking-inspired models such as RankIQA [241] and RankDVQA [242] have brought in generalization ability on diversified distortions and contents with significant efficiency. Very recent efforts further explore unifying heterogeneous distortion domains within a single model via mixture-of-experts designs [243].

Recently, leveraging large multimodal models (LMMs) for perceptual quality assessment has attracted growing interest because of their flexibility in extracting visually grounded descriptions enabled by cross-modal pretraining. In this paradigm, LMMs are typically used as first-hand quality raters [244–250] or language-conditioned supervisors to produce weak supervisions that can be exploited by downstream quality predictors [249,251–254]. Nevertheless, translating these advances to routine video compression evaluation remains non-trivial: beyond computational overhead, the key difficulties lie in the intersection of fine-grained visual awareness, temporally consistent, and scale-calibrated judgments under diverse content and coding conditions. Accordingly, the practical adaptation of LMM-based quality assessment measures for codec evaluation still requires dedicated designs and remains far from mature.

5.4. BD Measurements

The Bjøntegaard-Delta (BD) measurements [255] are standard metrics that quantify the coding gain achieved by one compression method over another (the anchor), based on the rate-distortion

 doi:10.20944/preprints202604.0035.v1

or rate-quality curves fitted to data points from each method. The BD-rate represents the average bitrate difference between these curves for the same quality levels; a negative value indicates superior performance. Similarly, the BD-metrics (e.g., BD-PSNR, BD-SSIM) use the same principle to measure the average quality gain within an overlapped bitrate range. In practice, a BD-rate is typically computed for each sequence individually, and the final BD-rate for a database is then obtained by averaging the per-sequence results. When reporting performance, it is therefore important to state whether BD-rate values are calculated per sequence and averaged, or computed once using all data aggregated across the entire database, as the two approaches can yield different interpretations.

6. Benchmarking NVC Methods

To complement our comprehensive overview of NVC methods, in this work, we also conducted a benchmarking experiment evaluating representative neural video codecs, both scene-adaptive and scene-agnostic, against standard video codecs based on fair coding configurations. All evaluated models here are assessed within a unified testbed, analyzing their rate-distortion-complexity trade-offs under the consistent conditions previously described. We begin with the experimental setup, which is then followed by a detailed presentation and analysis of the comparative results.

6.1. Test Conditions

To benchmark the performance of NVC codecs, we have selected *four* standard test coding models, including VTM [26], AV1 [27], ECM [28], and AVM [256]. VTM and AV1 represent the latest video coding standards developed by MPEG JVET and AOM, respectively, while ECM and AVM are their working models, potentially for their successors. Our evaluations are configured in strict accordance with the MPEG JVET Common Test Conditions (CTC) [219] for both Low Delay and Random Access settings. For the neural video codec counterparts, we curated representative models for each configuration (LD and RA) that are **recent** (published within the previous three years), **high-performing** (achieving SOTA results), and **reproducible** (publicly open-sourced with pre-trained models or training code provided). For scene-agnostic models, we include DCVC-DC [72], DCVC-FM [11], and DCVC-RT [79], and MaskCRT [78]. For scene-adaptive models, PNVC [90], and GIViC [14] are evaluated under LD and RA configurations. In addition, our comparison incorporates C3 [91], HiNeRV [87] and NVRC [13]. These approaches differ from traditional Random Access codecs in that they forgo the GoP-based framework and instead allow for parameter sharing across the entire sequence to maximize compression efficiency. It is noted that DCVC-DC and DCVC-FM may exhibit numerical instability when handling the 10-bit sequences in the AOM A2–A5 dataset, making their evaluation under this setting less reliable. In addition, since MaskCRT is trained in the RGB domain rather than in YCbCr 4:2:0, their performance in our benchmark may also be affected.

To ensure our comparison is both robust and widely applicable, we evaluated the codecs on several standard video test sets, as previously introduced in Section 5.1.2: UVG, MCL-JCV, HEVC Class B-E, and AOM A2-A5. This selection provides a diverse mix of content types, resolutions, and spatiotemporal complexities. It is important to note that 4K/UHD sequences were omitted from this benchmark, a necessary concession due to the significant resource and computational constraints for the majority of neural video codecs.

To ensure a comprehensive assessment of the reconstructed video quality, performance is measured using three quality metrics: PSNR, MS-SSIM [257], and VMAF [235] for each codec. We note that VMAF is evaluated using MSE-optimized NVC models rather than their MS-SSIM-optimized counterparts. For all sequences, the distortions are measured in the YUV 4:2:0 colorspace to align with **MPEG JVET/JCT-VC Common Test Conditions (CTC)** [219]. The Bjøntegaard Delta Rate (BD-rate) [255] is used to measure the relative compression efficiency between codecs. All BD-rate results reported in Table 3 and Table 4 are calculated against VTM-20.0 (LD) as the anchor for a direct and fair comparison.

We further profile the computational complexity of the top-performing codecs per setting (LD/RA) and category, i.e., the entries reported in Table 5 and visualized in Figure 1. For each selected codec, we report three complementary indicators: (i) model size (millions of parameters, M) to capture memory

 doi:10.20944/preprints202604.0035.v1

footprint, applicable to NVCs only; (ii) kMAC/pixel, the estimated multiply–accumulate (MAC) operations per pixel, applicable to NVCs only; and (iii) encoding and decoding throughput (FPS). To reflect realistic deployment, conventional codecs (VTM, ECM, AVM) are timed on CPU (Dual-socket ARM Neoverse-V2), while neural codecs are timed on a single NVIDIA A100 GPU.

Table 3. Low Delay. BD-rate (%, negative = bitrate saving) in YCbCr 4:2:0 colorspace with VTM-20.0 *lowdelay* as the anchor. The best and second-best results for each metric and dataset, excluding the reference VTM row, are highlighted in **bold** and underlined, respectively.

	UVG			MCL-JCV			HEVC B–E			AOM A2–A5		
Codec	PSNR	MS-SSIM	VMAF	PSNR	MS-SSIM	VMAF	PSNR	MS-SSIM	VMAF	PSNR	MS-SSIM	VMAF
Conventional video codecs												
VTM 20.0 (LD) [26]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
AV1 3.8.1 [27]	17.87	37.07	30.83	12.25	19.49	22.18	12.23	54.76	26.50	18.87	20.98	29.24
AVM 2.0.0 [256]	-1.40	-2.94	-6.18	-4.56	-17.52	-16.81	3.24	-10.10	-12.05	16.02	10.62	14.24
ECM 12.0 [28]	-23.44	-12.73	-16.15	-24.77	-17.43	-16.60	**-23.49**	**-25.24**	**-28.35**	**-13.11**	**-14.49**	**-19.94**
Scene-agnostic neural video codecs												
MaskCRT [78]	-10.12	**-20.34**	10.37	8.31	**-35.86**	-13.94	9.57	-3.72	20.41	11.42	32.28	14.63
DCVC-DC [72]	-24.18	-19.83	**-20.73**	**-34.77**	-28.00	**-31.22**	-18.66	-8.67	-25.81	–	–	–
DCVC-FM [11]	-23.62	2.86	-19.72	-29.79	-11.06	-20.63	-23.42	4.61	-22.57	–	–	–
DCVC-RT [79]	**-25.59**	7.16	-14.68	-30.13	-5.06	-10.27	-12.72	30.48	-12.84	–	–	–
Scene-adaptive neural video codecs												
PNVC [90]	-0.27	1.88	-1.40	5.98	-5.66	5.14	3.51	5.27	13.22	15.86	-0.06	16.36

Table 4. Random Access. BD-rate (%, negative = bitrate saving) in YCbCr 4:2:0 colorspace for random-access codecs, with VTM-20.0 *lowdelay* as the common anchor. The best and second-best results for each metric and dataset, excluding the two reference VTM rows, are highlighted in **bold** and underlined, respectively.

	UVG			MCL-JCV			HEVC B–E			AOM A2–A5		
Codec	PSNR	MS-SSIM	VMAF	PSNR	MS-SSIM	VMAF	PSNR	MS-SSIM	VMAF	PSNR	MS-SSIM	VMAF
Conventional video codecs												
VTM 20.0 (LD) [26]	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
VTM 20.0 (RA) [26]	-16.82	-17.54	-15.23	-23.80	-21.39	-19.10	-18.83	-15.09	-12.80	-11.67	-17.83	-12.45
AV1 3.8.1 [27]	-13.38	-4.87	8.14	-15.12	-7.24	-9.57	-12.88	-17.83	2.36	4.83	-12.57	3.18
AVM 2.0.0 [256]	-18.07	-24.46	-16.23	-18.94	-22.37	-14.28	-15.58	-20.43	-11.17	-10.63	-18.16	-10.08
ECM 12.0 [28]	**-44.47**	-48.88	**-19.82**	**-50.63**	**-41.69**	**-23.46**	**-48.03**	-43.06	**-27.28**	**-35.57**	**-47.39**	**-17.09**
Scene-adaptive neural video codecs												
C3 [91]	39.67	18.80	28.31	43.24	22.67	31.54	45.11	25.38	33.27	44.38	24.16	32.48
HiNeRV [87]	4.95	-30.94	-11.72	37.51	13.12	24.48	32.86	7.47	22.31	39.28	14.76	25.93
PNVC [90]	-19.02	-6.25	5.56	-14.24	13.85	-12.08	-12.87	-10.20	0.29	2.21	-13.54	0.31
NVRC [13]	-30.52	-38.89	-9.24	-13.88	-17.07	-8.29	-29.62	-29.61	-15.60	-15.49	-17.36	-4.12
GIViC [14]	-33.42	**-49.86**	-17.34	-37.42	-37.98	-10.46	-43.03	**-45.18**	-24.56	-22.93	-30.53	-2.57

6.2. Rate-Distortion Performance

6.2.1. Rate–Distortion

Across our benchmarks, the current working models of conventional codecs, in particular ECM (and, to a lesser extent, AVM), consistently outperform the MSE-optimized NVC baselines on distortion-based metrics, except in settings where the neural codecs are explicitly optimized for MS-SSIM. This margin becomes even more evident when comparisons are restricted to codecs operating under comparable decoding-speed budgets on modern consumer GPUs, suggesting that the current RD gains of NVCs are often entangled with increased computational and memory costs.

For the Low Delay (LD) configuration (Table 3 and Figure 5), both conventional and neural codecs achieve substantial bitrate savings over the VTM-20.0 LD anchor, but the relative ranking remains highly dependent on the dataset and distortion metric; no single method dominates all cases. Among the conventional codecs, ECM is the strongest baseline overall. Within the conventional set, it achieves the best BD-rate on UVG across all three metrics; on MCL-JCV, it leads in PSNR and VMAF, while AVM is marginally better in MS-SSIM; and on both HEVC B–E and AOM A2–A5, ECM is the best conventional codec across all reported metrics. Among the neural baselines, the picture is more fragmented. On UVG, DCVC-RT attains the best PSNR (-25.59%), MaskCRT the best MS-SSIM (-20.34%), and DCVC-DC the best VMAF (-20.73%). On MCL-JCV, DCVC-DC provides the best neural

 doi:10.20944/preprints202604.0035.v1

PSNR and VMAF (-34.77% and -31.22%), while MaskCRT achieves the best neural MS-SSIM (-35.86%). On HEVC B–E, DCVC-FM attains the best neural PSNR (-23.42%), very close to ECM (-23.49%), whereas DCVC-DC gives the best neural MS-SSIM and VMAF (-8.67% and -25.81%). In contrast, the scene-adaptive PNVC does not show clear advantages under strict LD constraints, suggesting that the benefits of content adaptation are harder to realize when latency and buffering budgets are tightly limited.

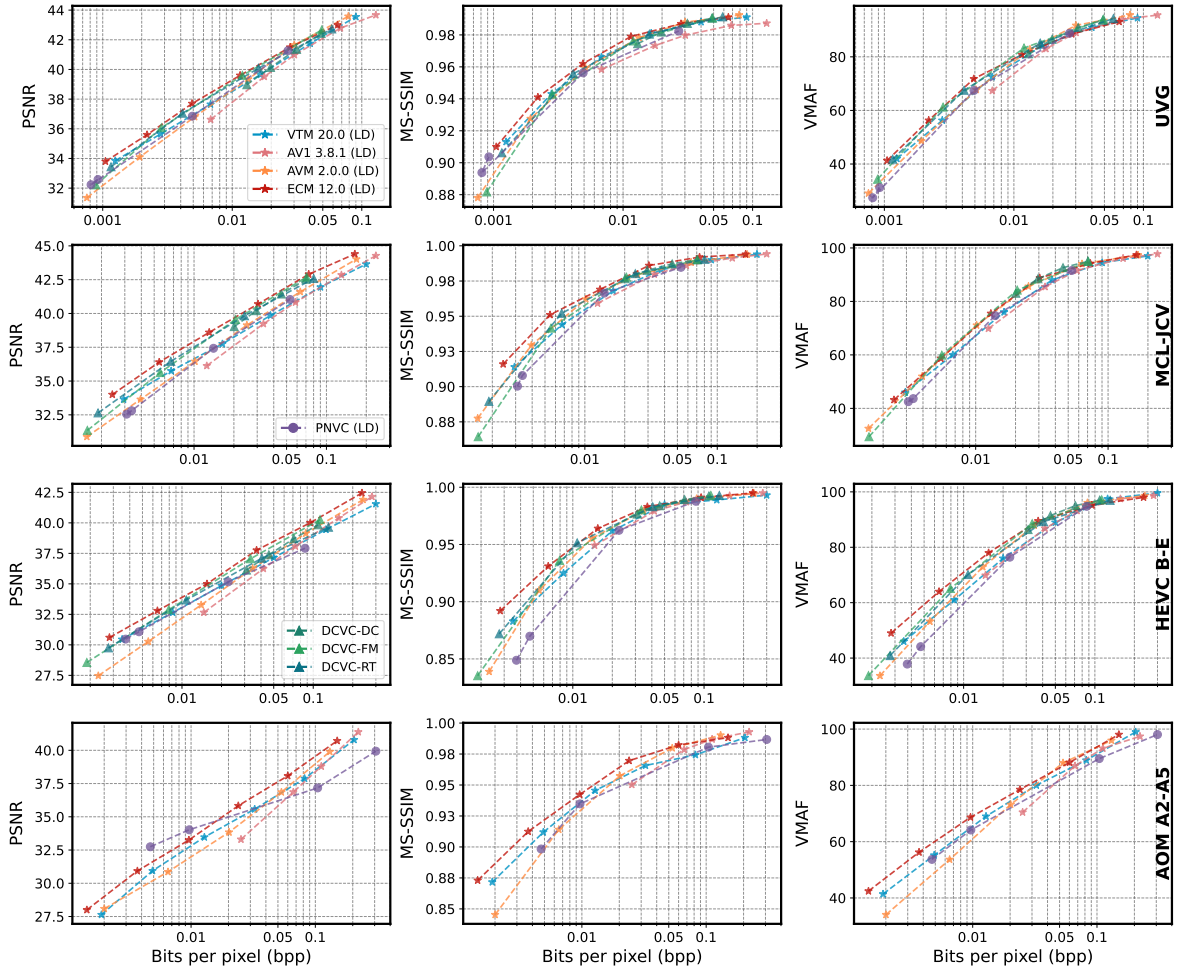


Figure 5. Rate Distortion performance on UVG, MCL-JCV, HEVC B-E, and AOM A2-A5 for selected codecs under the Low Delay setting, where the reconstruction quality is measured by PSNR, MS-SSIM, and VMAF.

For the Random Access (RA) configuration (Table 4 and Figure 6), ECM remains the strongest conventional baseline overall on distortion-oriented metrics. It achieves the best BD-rate on most dataset–metric pairs, with GIViC surpassing it only on UVG MS-SSIM and HEVC B–E MS-SSIM. Scene-adaptive methods are clearly better matched to the RA setting, where bidirectional temporal dependencies and more flexible bit allocation can be exploited. Among the neural baselines, GIViC delivers the strongest overall RA performance, achieving the best neural PSNR and MS-SSIM results on all four test datasets, as well as the best neural VMAF on UVG and HEVC B–E. However, it is not uniformly the best across every RA case: on MCL-JCV, PNVC yields a better neural VMAF (-12.08% versus -10.46%), while on AOM A2–A5, NVRC attains the best neural VMAF (-4.12% versus -2.57%). Overall, NVRC remains competitive on UVG and HEVC B–E, but trails GIViC more clearly on MCL-JCV and AOM A2–A5. As the same anchor is used, the BD-rate values in Table 4 remain numerically comparable across codecs; however, such comparisons should still be interpreted with caution when the underlying coding structures and latency constraints are not fully matched. These results suggest that relaxed latency constraints and stronger temporal dependency modeling are particularly beneficial to scene-adaptive codecs.

 doi:10.20944/preprints202604.0035.v1

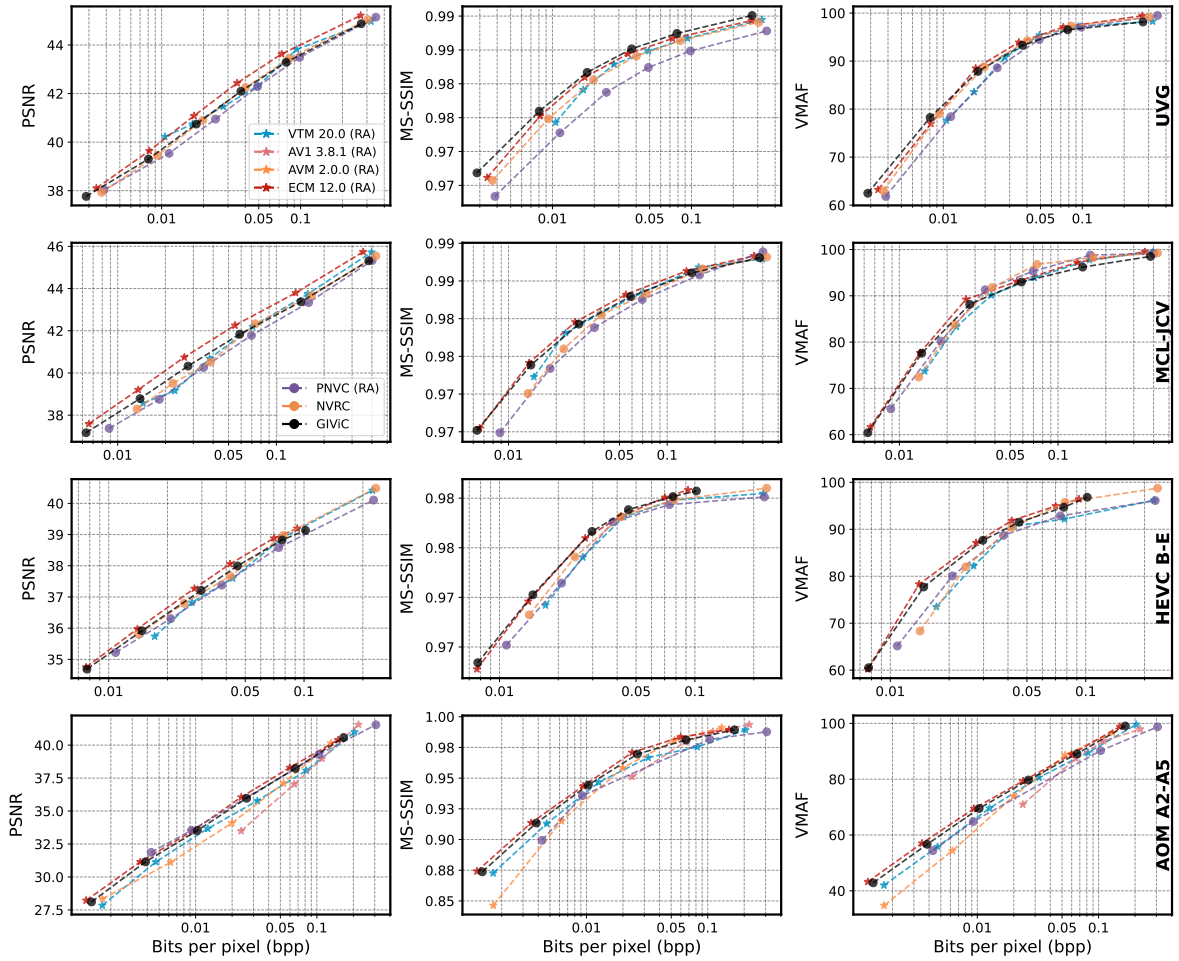


Figure 6. Rate Distortion performance on UVG, MCL-JCV, HEVC B-E, and AOM A2-A5 for selected codecs under the Random Access setting, where the reconstruction quality is measured by PSNR, MS-SSIM, and VMAF.

In summary, the results show that existing NVCs, including both scene-agnostic and scene-adaptive approaches, still do not provide consistent coding gains over the strongest conventional baseline, ECM, under either LD or RA evaluation. Although neural codecs can outperform conventional baselines on specific dataset–metric pairs, especially in MS-SSIM-oriented settings, these advantages are not yet robust across datasets, metrics, and operating conditions, leaving a clear research gap for future video compression research.

6.3. Complexity Analysis

A closer examination of computational complexity reveals systematic differences among conventional, scene-agnostic, and scene-adaptive codecs, as illustrated by the radar plot in Figure 1 and the quantitative results in Table 5. Conventional codecs exhibit diverse rate–distortion–complexity profiles rather than a single uniform operating point. Conventional baselines such as VTM and AVM achieve relatively balanced trade-offs, combining strong rate–distortion performance with practical encoding and decoding throughput. In contrast, ECM prioritizes compression efficiency more aggressively, resulting in substantially higher computational costs, particularly in decoding.

Among neural codecs, scene-agnostic models generally emphasize streamlined and uniform inference pipelines, which lead to more predictable runtime behavior and comparatively stable encoding characteristics, as they typically operate under Low Delay configurations. However, this efficiency does not consistently translate into superior rate–distortion performance, with most scene-agnostic approaches still trailing behind the strongest conventional codecs. DCVC-RT represents a notable efficiency point within this category, achieving competitive compression performance while

 doi:10.20944/preprints202604.0035.v1

maintaining a relatively compact model footprint and stable runtime characteristics compared to other neural baselines.

Scene-adaptive methods, in contrast, demonstrate that strong rate–distortion performance can be achieved with comparatively lightweight models, even without explicit optimization for fast inference. As reflected in Table 5 and the radar plot, several scene-adaptive codecs reach compression efficiency comparable to the most efficient scene-agnostic methods. However, these gains are accompanied by substantially higher encoding costs, stemming from per-content adaptation and additional optimization overhead, which limits their applicability under strict latency constraints.

Overall, while neural video codecs, including both scene-agnostic and scene-adaptive models, have demonstrated competitive performance relative to VTM and, in selected cases, AVM, with promising rate–distortion–complexity trade-offs on consumer GPUs, they still face significant challenges from the strongest conventional working model, ECM. In order to be deployed in practical applications, these learning-based solutions should be further improved significantly in terms of coding performance while maintaining relatively low encoding/decoding complexities.

Table 5. Computational complexity, including model size, estimated kMACs/pixel, and running latency, of the selected baselines on 1920 × 1080 videos. For conventional codecs, results are reported for both LD and RA configurations.

Codec	#params [M]↓	kMACs/px↓	Enc. FPS↑	Dec. FPS↑
VTM 20.0 (LD)	—	—	0.06	30.5
VTM 20.0 (RA)	—	—	0.03	23.2
AVM 2.0.0 (LD)	—	—	0.06	36.3
AVM 2.0.0 (RA)	—	—	0.004	28.1
ECM 12.0 (LD)	—	—	0.003	4.21
ECM 12.0 (RA)	—	—	0.002	2.96
DCVC-FM	18.3	1274.1	1.67	4.35
DCVC-RT	20.7	185.7	126.3	112.6
MaskCRT	27.7	763.0	0.17	1.03
HiNeRV	30.3 ± 27.1	682.8 ± 595.1	0.0157 ± 0.009	19.7 ± 13.5
C3 (Adaptive)	0.01	4.4	0.0015	17.6
PNVC	21.8	101.1	0.011	22.6
NVRC	16.8 ± 14.5	582.1 ± 296.9	4.5 ± 2.3	16.7 ± 6.2
GIViC	225.9	2399	0.03	5.59

7. Discussions and Open Problems

Despite the substantial progress in neural video compression (NVC), our literature review and benchmarking results indicate that several theoretical and practical challenges remain unresolved. In this subsection, we synthesize recurring gaps identified in the literature with failure modes and trade-offs observed empirically and distill them into a set of open problems and future research directions. We hope that this synthesis sheds light on where current methods fall short under realistic constraints and helps sharpen the community's research agenda moving forward.

7.1. Practicality

A key open problem lies in the lack of deployment-grounded evaluation and optimization criteria for neural video compression. Most existing NVCs are assessed under idealized hardware settings and are primarily optimized for rate–distortion performance while overlooking practical constraints such as decoding latency, memory bandwidth, and energy consumption on edge devices. As a result, models that appear competitive in the RD space often become impractical under realistic decoding budgets.

Beyond the lightweight architectures discussed in Section 4.5, this limitation also manifests in several related challenges. **Adaptive rate–distortion–complexity control** remains difficult, as achieving

 doi:10.20944/preprints202604.0035.v1

stable trade-offs across heterogeneous content and hardware often relies on heuristic or platform-specific tuning. **Algorithm–hardware co-design** is another important direction, since neural codecs are rarely designed with concrete hardware constraints in mind, leading to mismatches between model structure and accelerator data flows. Finally, **error resilience** is a practically relevant concern, given the strong temporal dependencies and potential error propagation in predictive learned codecs.

7.2. Realism vs. Fidelity

As neural video codecs become increasingly generative, a central open problem is to define training and evaluation criteria that properly balance *fidelity* (faithful reproduction) and *realism* (perceptual plausibility). Classical distortion metrics (e.g., PSNR/MS-SSIM) provide weak guidance once the decoder can synthesize plausible textures; they may penalize visually pleasing reconstructions and favor over-smoothed outputs. Conversely, optimizing only for perceptual realism can incentivize hallucinated details, identity drift, or temporally inconsistent artifacts. The core challenge, therefore, is to develop criteria that are simultaneously (i) aligned with human preference, (ii) explicitly sensitive to temporal coherence, and (iii) robust against semantic or structural corruption.

Learned perceptual objectives, including VQA-based signals, are natural candidates to better approximate human judgments. However, using VQA directly as a primary optimization target remains difficult in practice: it is often too slow for large-scale codec training and systematic ablations, and its reliability for diverse video failure modes (e.g., subtle flicker, motion-compensated inconsistencies, and long-range temporal drift) is still imperfect. a number of region-adaptive metrics/losses have been proposed for image compression that prioritize distortion in visually salient regions while allowing more perceptual synthesis elsewhere [258]; whether such ROI-aware principles can be made stable and effective for video, where saliency could potentially be highly dynamic and temporal consistency is critical, remains an open problem.

Finally, beyond better metrics and objectives, another line of research for improving perceptual quality is to introduce *stronger generative priors*, such as diffusion/score-based models, to enhance fine details and realism at low bitrates. However, diffusion-style methods face a major practicality hurdle: their multi-step sampling (and any associated objective evaluation) is computationally expensive, making them difficult to deploy as part of the decoding pipeline or to use directly in large-scale codec training. Although recent advances, such as flow matching/consistency-style formulations or common randomness-based methods, can reduce sampling steps and variance, an open problem remains regarding how to amortize or distill diffusion-level perceptual gains into standard rate–distortion training for video under strict compute and latency budgets without introducing temporal instability.

7.3. Robustness and Adaptation under Distribution Shifts

Despite steady progress on benchmark datasets, the generalization behavior of neural video compression (NVC) remains insufficiently understood. Many codecs are trained and evaluated under relatively controlled content distributions and distortion types; yet real-world deployment inevitably encounters out-of-distribution (OOD) shifts, such as domain changes (animation, gaming, surveillance), sensor noise and ISP pipelines, editing artifacts, extreme motion patterns, or unseen resolutions and frame rates. These shifts can lead to unstable rate-distortion performance, perceptual failures (e.g., flicker or hallucination), or unexpected bitrate spikes, suggesting that generalization should be treated as a first-class design objective rather than a by-product of scale.

A common practical strategy is *pretrain-then-overfit* (content-adaptive finetuning), which can recover substantial gains by specializing the codec to a target video or domain. However, this paradigm is not necessarily optimal; it can be computationally expensive, introduce additional engineering complexity, and may require transmitting adaptation side information or synchronizing encoder/decoder states. More fundamentally, it does not address the broader question of how to build codecs that adapt efficiently and reliably *over time* as the content distribution evolves.

This motivates online adaptation and continual learning for NVC, where the codec incrementally updates to new content streams, device characteristics, or user preferences. Yet continual learning

 doi:10.20944/preprints202604.0035.v1

introduces its own open challenges, most notably catastrophic forgetting: updates that improve performance on new domains can degrade compression efficiency or perceptual quality on previously seen content. Designing adaptation mechanisms that are lightweight, decoder-safe, and robust, whilst balancing plasticity (fast adaptation) and stability (retaining past capabilities) under strict bitrate and latency constraints remains an open and practically important research direction.

8. Conclusion

This paper presents a comprehensive and unified survey and benchmarking of recent neural video compression methods under consistent evaluation conditions. Our analysis shows that the best-performing conventional codecs continue to offer excellent rate–distortion–complexity trade-offs, while scene-agnostic neural codecs generally provide lower encoding latency at the expense of rate–distortion performance, and scene-adaptive methods achieve competitive compression efficiency with higher computational costs (at the encoder). Closing the remaining gap to conventional codecs at practical decoding speeds remains an open challenge, and we hope that the benchmarking framework introduced in this work will support more rigorous evaluation and guide future research in neural video compression.

References

1. C. Systems, "VNI complete forecast highlights," Cisco, Tech. Rep., 2022, accessed: 2025-03-14.
2. D. Bull and F. Zhang, *Intelligent image and video compression: communicating pictures*. Academic Press, 2021.
3. ITU-T Rec. H.262, *Information technology - Generic coding of moving pictures and associated audio information: Video*, ITU-T Std., 2012.
4. T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on circuits and systems for video technology*, vol. 13, no. 7, pp. 560–576, 2003.
5. G. J. Sullivan, J. Ohm, W. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.
6. B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the Versatile Video Coding (VVC) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.
7. J. Han, B. Li, D. Mukherjee, C.-H. Chiang, A. Grange, C. Chen, H. Su, S. Parker, S. Deng, U. Joshi *et al.*, "A technical overview of AV1," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1435–1462, 2021.
8. G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "DVC: An end-to-end deep video compression framework," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 006–11 015.
9. Z. Hu, G. Lu, and D. Xu, "Fvc: A new framework towards deep video compression in feature space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1502–1511.
10. L. Qi, Z. Jia, J. Li, B. Li, H. Li, and Y. Lu, "Long-term temporal context gathering for neural video compression," in *European Conference on Computer Vision*. Springer, 2024, pp. 305–322.
11. J. Li, B. Li, and Y. Lu, "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*, 2024.
12. H. Chen, B. He, H. Wang, Y. Ren, S. N. Lim, and A. Shrivastava, "Nerv: Neural representations for videos," *Advances in Neural Information Processing Systems*, vol. 34, pp. 21 557–21 568, 2021.
13. H. M. Kwan, G. Gao, F. Zhang, A. Gower, and D. Bull, "NVRC: Neural video representation compression," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37. Curran Associates, Inc., 2024, pp. 132 440–132 462.
14. G. Gao, S. Teng, T. Peng, F. Zhang, and D. Bull, "Givic: Generative implicit video compression," *arXiv preprint arXiv:2503.19604*, 2025.
15. S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and Video Compression with Neural Networks: A Review," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1683–1698, 2019.
16. D. Liu, Y. Li, J. Lin, H. Li, and F. Wu, "Deep Learning-Based Video Coding: A Review and A Case Study," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–35, 2020.

 doi:10.20944/preprints202604.0035.v1

17. D. Ding, Z. Ma, D. Chen, Q. Chen, Z. Liu, and F. Zhu, "Advances in video compression system using deep neural network: A review and case studies," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1494–1520, 2021.
18. H. M. Yasin and S. Y. Ameen, "Review and evaluation of end-to-end video compression with deep-learning," in *2021 International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI)*. IEEE, 2021, pp. 1–8.
19. Y. Yang, S. Mandt, L. Theis *et al.*, "An introduction to Neural Data Compression," *Foundations and Trends® in Computer Graphics and Vision*, vol. 15, no. 2, pp. 113–200, 2023.
20. Z. Chen, H. Sun, L. Zhang, and F. Zhang, "Survey on Visual Signal Coding and Processing with Generative Models: Technologies, Standards and Optimization," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2024.
21. H. Keunen, M. Wijnants, and J. Liesenborgs, "A survey of implicit neural representations for video compression," *TechRxiv*, 2025.
22. J. S. Gomes, M. Grellert, F. L. Ramos, and S. Bampi, "End-to-end neural video compression: A review," *IEEE Open Journal of Circuits and Systems*, 2025.
23. J. Han, B. Li, D. Mukherjee, C.-H. Chiang, A. Grange, C. Chen, H. Su, S. Parker, S. Deng, U. Joshi *et al.*, "A technical overview of av1," *arXiv preprint arXiv:2008.06091*, 2020.
24. Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, *JM Reference Software*, ITU-T and ISO/IEC JTC1, 2003, the specific version of the software (e.g., JM 8.6, JM 19.0) is often mentioned in the text.
25. C. Rosewarne, K. Sharman, R. Sjoberg, and G. Sullivan, "High efficiency video coding (HEVC) test model 16 (HM 16) improved encoder description update 16," in *the JVET meeting*. ITU-T and ISO/IEC, 2022.
26. Joint Video Experts Team (JVET), "Versatile Video Coding Test Model (VTM) 20.0 Library ," https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM, 2024, accessed: 2024-04-10.
27. Alliance for Open Media, "AV1 3.8.1 Codec Library," https://aomedia.googlesource.com/aom/, 2024, accessed: 2024-04-10.
28. Joint Video Experts Team (JVET), "Enhanced Compression Model (ECM) 12.0 Library ," https://vcgit.hhi.fraunhofer.de/ecm/ECM, 2024, accessed: 2024-04-10.
29. I. Schiopu, H. Huang, and A. Munteanu, "Cnn-based intra-prediction for lossless hevc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 7, pp. 1816–1828, 2019.
30. P. Merkle, M. Winken, J. Pfaff, H. Schwarz, D. Marpe, and T. Wiegand, "Intra-inter prediction for versatile video coding using a residual convolutional neural network," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 1711–1715.
31. J. Liu, S. Xia, W. Yang, M. Li, and D. Liu, "One-for-all: Grouped variation network-based fractional interpolation in video coding," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2140–2151, 2018.
32. L. Zhao, S. Wang, X. Zhang, S. Wang, S. Ma, and W. Gao, "Enhanced motion-compensated video coding with deep virtual reference frame generation," *IEEE Transactions on Image Processing*, vol. 28, no. 10, pp. 4832–4844, 2019.
33. D. Liu, H. Ma, Z. Xiong, and F. Wu, "Cnn-based dct-like transform for image compression," in *International Conference on Multimedia Modeling*. Springer, 2018, pp. 61–72.
34. M. M. Alam, T. D. Nguyen, M. T. Hagan, and D. M. Chandler, "A perceptual quantization strategy for hevc based on a convolutional neural network trained on natural images," in *Applications of Digital Image Processing XXXVIII*, vol. 9599. International Society for Optics and Photonics, 2015, p. 959918.
35. R. Song, D. Liu, H. Li, and F. Wu, "Neural network-based arithmetic coding of intra prediction modes in hevc," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.
36. F. Zhang, C. Feng, and D. R. Bull, "Enhancing vvc through cnn-based post-processing," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.
37. Y. Xue and J. Su, "Attention based image compression post-processing convlutional neural network." in *CVPR Workshops*, 2019, p. 0.
38. Y. Zhao, K. Lin, S. Wang, and S. Ma, "Joint luma and chroma multi-scale cnn in-loop filter for versatile video coding," in *2022 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2022, pp. 3205–3209.
39. K. Lin, C. Jia, X. Zhang, S. Wang, S. Ma, and W. Gao, "Nr-cnn: Nested-residual guided cnn in-loop filtering for video coding," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 4, pp. 1–22, 2022.
40. Y. Li, L. Zhang, and K. Zhang, "idam: Iteratively trained deep in-loop filter with adaptive model selection," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 1s, pp. 1–22, 2023.

 doi:10.20944/preprints202604.0035.v1

41. M. Afonso, F. Zhang, and D. R. Bull, "Video compression based on spatio-temporal resolution adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 1, pp. 275–280, 2018.
42. Y. Jiang, J. Nawała, C. Feng, F. Zhang, X. Zhu, J. Sole, and D. Bull, "Rtsr: A real-time super-resolution model for av1 compressed content," *arXiv preprint arXiv:2411.13362*, 2024.
43. Y. Jiang, S. Teng, Q. Zhu, C. Feng, C. Zeng, F. Zhang, S. Zhu, B. Zeng, and D. Bull, "Compressed video super-resolution based on hierarchical encoding," *arXiv preprint arXiv:2506.14381*, 2025.
44. F. Zhang, M. Afonso, and D. R. Bull, "ViSTRA2: Video coding using spatial resolution and effective bit depth adaptation," *Signal Processing: Image Communication*, vol. 97, p. 116355, 2021.
45. C. Feng, Z. Qi, D. Danier, F. Zhang, X. Xu, S. Liu, and D. Bull, "Enhancing hdr video compression through cnn-based effective bit depth adaptation," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2025.
46. A. Chadha and Y. Andreopoulos, "Deep perceptual preprocessing for video coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 852–14 861.
47. O. G. Guleryuz, P. A. Chou, H. Hoppe, D. Tang, R. Du, P. Davidson, and S. Fanello, "Sandwiched image compression: wrapping neural networks around a standard codec," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 3757–3761.
48. ——, "Sandwiched image compression: Increasing the resolution and dynamic range of standard codecs," in *2022 Picture Coding Symposium (PCS)*. IEEE, 2022, pp. 175–179.
49. Y. Hu, C. Zhang, O. G. Guleryuz, D. Mukherjee, and Y. Wang, "Standard compliant video coding using low complexity, switchable neural wrappers," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 1922–1928.
50. M. U. K. Khan, A. Chadha, M. A. Anam, and Y. Andreopoulos, "Perceptual video compression with neural wrapping," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 17 743–17 754.
51. A. Said, "Introduction to arithmetic coding–theory and practice," *arXiv preprint arXiv:2302.00819*, 2023.
52. O. Rippel, S. Nair, C. Lew, S. Branson, A. G. Anderson, and L. Bourdev, "Learned video compression," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3454–3463.
53. Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 402–419.
54. Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized gaussian mixture likelihoods and attention modules," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 7939–7948.
55. G. Gao, P. You, R. Pan, S. Han, Y. Zhang, Y. Dai, and H. Lee, "Neural image compression via attentional multi-scale back projection and frequency decomposition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 677–14 686.
56. D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, "ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5718–5727.
57. E. Agustsson, D. Minnen, N. Johnston, J. Balle, S. J. Hwang, and G. Toderici, "Scale-space flow for end-to-end optimized video compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8503–8512.
58. W. Ma, J. Li, B. Li, and Y. Lu, "Uncertainty-aware deep video compression with ensembles," *IEEE Transactions on Multimedia*, 2024.
59. Z. Guo, R. Feng, Z. Zhang, X. Jin, and Z. Chen, "Learning cross-scale weighted prediction for efficient neural video compression," *IEEE Transactions on Image Processing*, vol. 32, pp. 3567–3579, 2023.
60. H. Liu, H. Shen, L. Huang, M. Lu, T. Chen, and Z. Ma, "Learned video compression via joint spatial-temporal correlation exploration," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 580–11 587.
61. R. Pourreza, H. Le, A. Said, G. Sautiere, and A. Wiggers, "Boosting neural video codecs by exploiting hierarchical redundancy," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5355–5364.
62. R. Yang, F. Mentzer, L. V. Gool, and R. Timofte, "Learning for video compression with hierarchical quality and recurrent enhancement," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6628–6637.

 doi:10.20944/preprints202604.0035.v1

63. M. A. Yılmaz and A. M. Tekalp, "End-to-end rate-distortion optimized learned hierarchical bi-directional video compression," *IEEE Transactions on Image Processing*, vol. 31, pp. 974–983, 2021.
64. J. Lin, D. Liu, H. Li, and F. Wu, "M-lvc: Multiple frames prediction for learned video compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3546–3554.
65. L. Qi, J. Li, B. Li, H. Li, and Y. Lu, "Motion information propagation for neural video compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6111–6120.
66. R. Feng, Y. Wu, Z. Guo, Z. Zhang, and Z. Chen, "Learned video compression with feature-level residuals," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 120–121.
67. B. Liu, Y. Chen, S. Liu, and H.-S. Kim, "Deep learning in latent space for video prediction and compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 701–710.
68. Z. Hu, G. Lu, J. Guo, S. Liu, W. Jiang, and D. Xu, "Coarse-to-fine deep video coding with hyperprior-guided mode prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5921–5930.
69. H. Wang, Z. Chen, and C. W. Chen, "Learned video compression via heterogeneous deformable compensation network," *IEEE Transactions on Multimedia*, 2023.
70. Y. Shi, Y. Ge, J. Wang, and J. Mao, "Alphavc: High-performance and efficient learned video compression," in *European Conference on Computer Vision*. Springer, 2022, pp. 616–631.
71. J. Li, B. Li, and Y. Lu, "Hybrid spatial-temporal entropy modelling for neural video compression," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 1503–1511.
72. ——, "Neural video compression with diverse contexts," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, Canada, June 18-22, 2023*, 2023.
73. W. Jiang, J. Li, K. Zhang, and L. Zhang, "Ecvc: Exploiting non-local correlations in multiple frames for contextual video compression," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 7331–7341.
74. J. Li, B. Li, and Y. Lu, "Deep contextual video compression," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
75. F. Mentzer, G. D. Toderici, D. Minnen, S. Caelles, S. J. Hwang, M. Lucic, and E. Agustsson, "Vct: A video compression transformer," in *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35. Curran Associates, Inc., 2022, pp. 13 091–13 103.
76. Y.-H. Ho, C.-P. Chang, P.-Y. Chen, A. Gnutti, and W.-H. Peng, "Canf-vc: Conditional augmented normalizing flows for video compression," *European Conference on Computer Vision*, 2022.
77. J. Xiang, K. Tian, and J. Zhang, "MIMT: Masked image modeling transformer for video compression," in *The Eleventh International Conference on Learning Representations*, 2023.
78. Y.-H. Chen, H.-S. Xie, C.-W. Chen, Z.-L. Gao, M. Benjak, W.-H. Peng, and J. Ostermann, "MaskCRT: Masked conditional residual transformer for learned video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
79. Z. Jia, B. Li, J. Li, W. Xie, L. Qi, H. Li, and Y. Lu, "Towards practical real-time neural video compression," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-25, 2024*, 2025.
80. Y.-H. Chen, Y.-C. Yao, K.-W. Ho, C.-H. Wu, H.-T. Phung, M. Benjak, J. Ostermann, and W.-H. Peng, "HyTIP: Hybrid temporal information propagation for masked conditional residual video coding," *arXiv preprint arXiv:2508.02072*, 2025.
81. G. Lu, C. Cai, X. Zhang, L. Chen, W. Ouyang, D. Xu, and Z. Gao, "Content adaptive and error propagation aware deep video compression," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 456–472.
82. T. van Rozendaal, I. A. Huijben, and T. Cohen, "Overfitting for Fun and Profit: Instance-Adaptive Data Compression," in *International Conference on Learning Representations*, 2021.
83. Z. Li, M. Wang, H. Pi, K. Xu, J. Mei, and Y. Liu, "E-NeRV: Expedite neural video representation with disentangled spatial-temporal context," in *European Conference on Computer Vision*. Springer, 2022, pp. 267–284.
84. J. C. Lee, D. Rho, J. H. Ko, and E. Park, "FFNeRV: Flow-guided frame-wise neural representations for videos," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7859–7870.

 doi:10.20944/preprints202604.0035.v1

85. C. Gomes, R. Azevedo, and C. Schroers, "Video compression with entropy-constrained neural representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 497–18 506.
86. H. Chen, M. Gwilliam, S.-N. Lim, and A. Shrivastava, "Hnerv: A hybrid neural representation for videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 270–10 279.
87. H. M. Kwan, G. Gao, F. Zhang, A. Gower, and D. Bull, "Hinerv: Video compression with hierarchical encoding-based neural representation," *Advances in Neural Information Processing Systems*, vol. 36, 2023.
88. S. R. Maiya, S. Girish, M. Ehrlich, H. Wang, K. S. Lee, P. Poirson, P. Wu, C. Wang, and A. Shrivastava, "NIRVANA: Neural implicit representations of videos with adaptive networks and autoregressive patch-wise modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 378–14 387.
89. X. Zhang, R. Yang, D. He, X. Ge, T. Xu, Y. Wang, H. Qin, and J. Zhang, "Boosting neural representations for videos with a conditional decoder," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2556–2566.
90. G. Gao, H. M. Kwan, F. Zhang, and D. Bull, "Pnvc: Towards practical inr-based video compression," *arXiv preprint arXiv:2409.00953*, 2024.
91. H. Kim, M. Bauer, L. Theis, J. R. Schwarz, and E. Dupont, "C3: High-performance and low-complexity neural compression from a single image or video," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9347–9358.
92. T. Leguay, T. Ladune, P. Philippe, and O. Déforges, "COOL-CHIC Video: Learned video coding with 800 parameters," in *2024 Data Compression Conference (DCC)*. IEEE, 2024, pp. 23–32.
93. X. Liu, B. Chen, Z. Liu, Y. Wang, and S.-T. Xia, "An Exploration with Entropy Constrained 3D Gaussians for 2D Video Compression," in *The Thirteenth International Conference on Learning Representations*, 2025.
94. T. Ladune, P. Philippe, W. Hamidouche, L. Zhang, and O. Déforges, "Optical flow and mode selection for learning-based video coding," in *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2020, pp. 1–6.
95. ——, "Conditional coding for flexible learned video compression," in *Neural Compression: From Information Theory to Applications – Workshop @ ICLR 2021*, 2021.
96. X. Sheng, J. Li, B. Li, L. Li, D. Liu, and Y. Lu, "Temporal context mining for learned video compression," *IEEE Transactions on Multimedia*, vol. 25, pp. 7311–7322, 2022.
97. R. Yang, Y. Yang, J. Marino, Y. Yang, and S. Mandt, "Deep generative video compression with temporal autoregressive transforms," in *Proc. ICML Workshop Invertible Neural Netw. Normalizing Flows, Explicit Likelihood Models*, 2020.
98. R. Yang, Y. Yang, J. Marino, and S. Mandt, "Insights from generative modeling for neural video compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9908–9921, 2023.
99. A. Habibian, T. v. Rozendaal, J. M. Tomczak, and T. S. Cohen, "Video compression with rate-distortion autoencoders," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7033–7042.
100. J. Han, S. Lombardo, C. Schroers, and S. Mandt, "Deep generative video compression," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019, pp. 9287–9298.
101. I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework," in *International conference on learning representations*, 2017.
102. R. Child, "Very deep vaes generalize autoregressive models and can outperform them on images," *arXiv preprint arXiv:2011.10650*, 2020.
103. A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *Advances in neural information processing systems*, vol. 33, pp. 19 667–19 679, 2020.
104. M. Lu, Z. Duan, F. Zhu, and Z. Ma, "Deep hierarchical video compression," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 8859–8867.
105. M. Lu, Z. Duan, W. Cong, D. Ding, F. Zhu, and Z. Ma, "High-efficiency neural video compression via hierarchical predictive learning," *arXiv preprint arXiv:2410.02598*, 2024.
106. Y.-H. Ho, C.-P. Chang, P.-Y. Chen, A. Gnutti, and W.-H. Peng, "CANF-VC: Conditional augmented normalizing flows for video compression," in *European Conference on Computer Vision*. Springer, 2022, pp. 207–223.
107. P.-Y. Chen and W.-H. Peng, "CANF-VC++: Enhancing conditional augmented normalizing flows for video compression with advanced techniques," *arXiv preprint arXiv:2309.05382*, 2023.

 doi:10.20944/preprints202604.0035.v1

108. M.-J. Chen, Y.-H. Chen, and W.-H. Peng, "B-CANF: Adaptive b-frame coding with conditional augmented normalizing flows," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 4, pp. 2908–2921, 2023.
109. G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed, and B. Lakshminarayanan, "Normalizing flows for probabilistic modeling and inference," *Journal of Machine Learning Research*, vol. 22, no. 57, pp. 1–64, 2021.
110. I. Kobyzev, S. J. Prince, and M. A. Brubaker, "Normalizing flows: An introduction and review of current methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 11, pp. 3964–3979, 2020.
111. F. Mentzer, G. Toderici, D. Minnen, S.-J. Hwang, S. Caelles, M. Lucic, and E. Agustsson, "Vct: A video compression transformer," *arXiv preprint arXiv:2206.07307*, 2022.
112. Z. Chen, L. Relic, R. Azevedo, Y. Zhang, M. Gross, D. Xu, L. Zhou, and C. Schroers, "Neural video compression with spatio-temporal cross-covariance transformers," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 8543–8551.
113. H. Chang, H. Zhang, L. Jiang, C. Liu, and W. T. Freeman, "Maskgit: Masked generative image transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 315–11 325.
114. R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, "Learning for video compression with recurrent auto-encoder and recurrent probability model," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 388–401, 2021.
115. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.
116. J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," *arXiv:2010.02502*, October 2020. [Online]. Available: https://arxiv.org/abs/2010.02502
117. J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
118. F. Mentzer, E. Agustsson, J. Ballé, D. Minnen, N. Johnston, and G. Toderici, "Neural video compression using gans for detail synthesis and propagation," in *European Conference on Computer Vision*. Springer, 2022, pp. 562–578.
119. L. Qi, Z. Jia, J. Li, B. Li, H. Li, and Y. Lu, "Generative latent coding for ultra-low bitrate image and video compression," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
120. W. Ma and Z. Chen, "Diffusion-based perceptual neural video compression with temporal diffusion information reuse," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 21, no. 12, pp. 1–22, 2025.
121. Z. Guo, Z. Jia, J. Li, X. Zhang, B. Li, and Y. Lu, "Generative latent video compression," *arXiv preprint arXiv:2510.09987*, 2025.
122. Q. Mao, H. Cheng, T. Yang, L. Jin, and S. Ma, "Generative neural video compression via video diffusion prior," *arXiv preprint arXiv:2512.05016*, 2025.
123. F. Brand, J. Seiler, and A. Kaup, "Conditional residual coding: A remedy for bottleneck problems in conditional inter frame coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 7, pp. 6445–6459, 2024.
124. Z. Hu, Z. Chen, D. Xu, G. Lu, W. Ouyang, and S. Gu, "Improving deep video compression by resolution-adaptive flow coding," in *European Conference on Computer Vision*. Springer, 2020, pp. 193–209.
125. B. Liu, Y. Chen, R. C. Machineni, S. Liu, and H.-S. Kim, "MMVC: Learned multi-mode video compression with block-based prediction mode selection and density-adaptive entropy coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 487–18 496.
126. Y. Yang, R. Bamler, and S. Mandt, "Improving inference for neural image compression," *Advances in Neural Information Processing Systems*, vol. 33, pp. 573–584, 2020.
127. J. Djelouah and C. Schroers, "Content adaptive optimization for neural image compression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*, vol. 2, 2019, pp. 1–5.
128. Z. Chen, L. Zhou, Z. Hu, and D. Xu, "Group-aware parameter-efficient updating for content-adaptive neural video compression," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 11 022–11 031.
129. V. Ročková and E. I. George, "The spike-and-slab lasso," *Journal of the American Statistical Association*, vol. 113, no. 521, pp. 431–444, 2018.
130. M. Khani, V. Sivaraman, and M. Alizadeh, "Efficient video compression via content-adaptive super-resolution," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4521–4530.

 doi:10.20944/preprints202604.0035.v1

131. V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473, 2020.
132. E. Dupont, A. Golinski, M. Alizadeh, Y. W. Teh, and A. Doucet, "COIN: COmpression with implicit neural representations," in *Neural Compression: From Information Theory to Applications – Workshop @ ICLR 2021*, 2021.
133. E. Dupont, H. Loya, M. Alizadeh, A. Golinski, Y. W. Teh, and A. Doucet, "COIN++: Neural compression across modalities," *Transactions on Machine Learning Research*, 2022.
134. B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
135. Y. Zhang, T. Van Rozendaal, J. Brehmer, M. Nagel, and T. Cohen, "Implicit neural video compression," *arXiv preprint arXiv:2112.11312*, 2021.
136. Y. Bai, C. Dong, C. Wang, and C. Yuan, "PS-NeRV: Patch-wise stylized neural representations for videos," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 41–45.
137. Q. Zhao, M. S. Asif, and Z. Ma, "DNeRV: Modeling inherent dynamics via difference neural representation for videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2031–2040.
138. H. Yan, Z. Ke, X. Zhou, T. Qiu, X. Shi, and D. Jiang, "DS-NeRV: Implicit neural video representation with decomposed static and dynamic codes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23 019–23 029.
139. B. He, X. Yang, H. Wang, Z. Wu, H. Chen, S. Huang, Y. Ren, S.-N. Lim, and A. Shrivastava, "Towards scalable neural representation for diverse videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6132–6142.
140. H. M. Kwan, F. Zhang, A. Gower, and D. Bull, "Immersive video compression using implicit neural representations," in *2024 Picture Coding Symposium (PCS)*, 2024, pp. 1–5.
141. H. Chen, S. Xie, S.-N. Lim, and A. Shrivastava, "Fast encoding and decoding for implicit video representation," in *European Conference on Computer Vision*. Springer, 2024, pp. 402–418.
142. T. Ladune, P. Philippe, F. Henry, G. Clare, and T. Leguay, "COOL-CHIC: Coordinate-based low complexity hierarchical image codec," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 13 515–13 522.
143. H. M. Kwan, T. Peng, G. Gao, F. Zhang, M. Nilsson, A. Gower, and D. Bull, "Ultra-lightweight neural video representation compression," *arXiv preprint arXiv:2512.04019*, 2025.
144. B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.
145. G. Wu, T. Yi, J. Fang, L. Xie, X. Zhang, W. Wei, W. Liu, Q. Tian, and X. Wang, "4D Gaussian Splatting for Real-Time Dynamic Scene Rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 310–20 320.
146. L. Wang, Y. Shi, and W. T. Ooi, "GSVC: Efficient Video Representation and Compression Through 2D Gaussian Splatting," in *Proceedings of the 35th Workshop on Network and Operating System Support for Digital Audio and Video*, 2025, pp. 15–21.
147. M. Liu, Q. Yang, M. Zhao, H. Huang, L. Yang, Z. Li, and Y. Xu, "D2GV: Deformable 2d gaussian splatting for video representation in 400fps," *arXiv preprint arXiv:2503.05600*, 2025.
148. L. Gupta and I. N. Junejo, "Neural Video Compression using 2D Gaussian Splatting," *arXiv preprint arXiv:2505.09324*, 2025.
149. X. Zhang, X. Ge, T. Xu, D. He, Y. Wang, H. Qin, G. Lu, J. Geng, and J. Zhang, "GaussianImage: 1000 fps image representation and compression by 2d gaussian splatting," in *European Conference on Computer Vision*. Springer, 2024, pp. 327–345.
150. S. Teng, G. Gao, D. Danier, Y. Jiang, F. Zhang, T. Davis, Z. Liu, and D. Bull, "Gfix: Perceptually enhanced gaussian splatting video compression," *arXiv preprint arXiv:2511.06953*, 2025.
151. C. E. Shannon, "A mathematical theory of communication," *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.
152. D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, no. 9, pp. 1098–1101, 2007.
153. J. Rissanen and G. G. Langdon, "Arithmetic coding," *IBM Journal of research and development*, vol. 23, no. 2, pp. 149–162, 1979.

 doi:10.20944/preprints202604.0035.v1

154. Y.-K. Wang, R. Skupin, M. M. Hannuksela, S. Deshpande, V. Drugeon, R. Sjöberg, B. Choi, V. Seregin, Y. Sanchez, J. M. Boyce *et al.*, "The high-level syntax of the versatile video coding (VVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3779–3800, 2021.
155. Y. Chen, D. Mukherjee, J. Han, A. Grange, Y. Xu, S. Parker, C. Chen, H. Su, U. Joshi, C.-H. Chiang *et al.*, "An overview of coding tools in AV1: the first video codec from the alliance for open media," *APSIPA Transactions on Signal and Information Processing*, vol. 9, p. e6, 2020.
156. J. Duda, "Asymmetric numeral systems," *arXiv preprint arXiv:0902.0271*, 2009.
157. R. Bamler, "Understanding entropy coding with asymmetric numeral systems (ans): a statistician's perspective," *arXiv preprint arXiv:2201.01741*, 2022.
158. G. Flamich, M. Havasi, and J. M. Hernández-Lobato, "Compressing images by encoding their latent representations with relative entropy coding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 16 131–16 141, 2020.
159. G. Flamich, S. Markou, and J. M. Hernández-Lobato, "Fast relative entropy coding with a* coding," in *International Conference on Machine Learning*. PMLR, 2022, pp. 6548–6577.
160. Z. Guo, G. Flamich, J. He, Z. Chen, and J. M. Hernández-Lobato, "Compression with bayesian implicit neural representations," *Advances in Neural Information Processing Systems*, vol. 36, pp. 1938–1956, 2023.
161. J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018.
162. A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with PixelCNN decoders," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
163. D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 771–14 780.
164. A. El-Nouby, M. J. Muckley, K. Ullrich, I. Laptev, J. Verbeek, and H. Jegou, "Image Compression with Product Quantized Masked Image Modeling," *Transactions on Machine Learning Research*, 2023.
165. F. Lin, H. Sun, J. Liu, and J. Katto, "Multistage spatial context models for learned image compression," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
166. D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3339–3343.
167. J. Tong, W. Zhang, Y. Jin, and X. Shen, "Context guided transformer entropy modeling for video compression," *arXiv preprint arXiv:2508.01852*, 2025.
168. D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in Neural Information Processing Systems*, vol. 31, 2018.
169. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
170. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.
171. E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, "Soft-to-hard vector quantization for end-to-end learning compressible representations," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
172. X. Zhu, J. Song, L. Gao, F. Zheng, and H. T. Shen, "Unified multivariate gaussian mixture for efficient neural image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 612–17 621.
173. R. Feng, Z. Guo, W. Li, and Z. Chen, "NVTC: Nonlinear vector transform coding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6101–6110.
174. D. Lee, C. Kim, S. Kim, M. Cho, and W.-S. Han, "Autoregressive image generation using residual quantization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 523–11 532.
175. G. Zhang, L. Tang, and X. Zhang, "VQ-NeRV: Vector quantization neural representation for video compression," in *2024 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2024, pp. 1–5.
176. X. Zhang and X. Wu, "LVQAC: Lattice vector quantization coupled with spatially adaptive companding for efficient learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10 239–10 248.

 doi:10.20944/preprints202604.0035.v1

177. ——, "Learning optimal lattice vector quantizers for end-to-end neural image compression," *Advances in Neural Information Processing Systems*, vol. 37, pp. 106 497–106 518, 2024.
178. L. Babai, "On Lovász'lattice reduction and the nearest lattice point problem," *Combinatorica*, vol. 6, no. 1, pp. 1–13, 1986.
179. F. Mentzer, D. Minnen, E. Agustsson, and M. Tschannen, "Finite Scalar Quantization: VQ-VAE Made Simple," in *The Twelfth International Conference on Learning Representations*, 2024.
180. L. Yu, J. Lezama, N. B. Gundavarapu, L. Versari, K. Sohn, D. Minnen, Y. Cheng, A. Gupta, X. Gu, A. G. Hauptmann, B. Gong, M.-H. Yang, I. Essa, D. A. Ross, and L. Jiang, "Language model beats diffusion - tokenizer is key to visual generation," in *The Twelfth International Conference on Learning Representations*, 2024.
181. Y. Zhu, B. Li, Y. Xin, Z. Xia, and L. Xu, "Addressing representation collapse in vector quantized models with one linear layer," *arXiv preprint arXiv:2411.02038*, 2024.
182. C. Fifty, R. G. Junkins, D. Duan, A. Iyengar, J. W. Liu, E. Amid, S. Thrun, and C. Re, "Restructuring Vector Quantization with the Rotation Trick," in *The Thirteenth International Conference on Learning Representations*, 2025.
183. Y. Zhao, Y. Xiong, and P. Kraehenbuehl, "Image and Video Tokenization with Binary Spherical Quantization," in *The Thirteenth International Conference on Learning Representations*, 2025.
184. J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *2016 Picture Coding Symposium (PCS)*. IEEE, 2016, pp. 1–5.
185. L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *International Conference on Learning Representations*, 2017.
186. J. Ballé, P. A. Chou, D. Minnen, S. Singh, N. Johnston, E. Agustsson, S. J. Hwang, and G. Toderici, "Nonlinear transform coding," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 2, pp. 339–353, 2020.
187. E. Agustsson and L. Theis, "Universally quantized neural compression," *Advances in neural information processing systems*, vol. 33, pp. 12 367–12 376, 2020.
188. T. Leguay, T. Ladune, P. Philippe, G. Clare, F. Henry, and O. Déforges, "Low-complexity overfitted neural image codec," in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2023, pp. 1–6.
189. T. Peng, G. Gao, H. Sun, F. Zhang, and D. Bull, "Accelerating learnt video codecs with gradient decay and layer-wise distillation," in *2024 Picture Coding Symposium (PCS)*. IEEE, 2024, pp. 1–5.
190. P. Kumaraswamy, "A generalized probability density function for double-bounded random processes," *Journal of Hydrology*, vol. 46, no. 1-2, pp. 79–88, 1980.
191. M. Nagel, M. Fournarakis, R. A. Amjad, Y. Bondarenko, M. Van Baalen, and T. Blankevoort, "A white paper on neural network quantization," *arXiv preprint arXiv:2106.08295*, 2021.
192. J. Ballé, N. Johnston, and D. Minnen, "Integer Networks for Data Compression with Latent-Variable Models," in *International Conference on Learning Representations*, 2019.
193. H. Sun, L. Yu, and J. Katto, "End-to-end learned image compression with quantized weights and activations," *arXiv preprint arXiv:2111.09348*, 2021.
194. D. He, Z. Yang, Y. Chen, Q. Zhang, H. Qin, and Y. Wang, "Post-training quantization for cross-platform learned image compression," *arXiv preprint arXiv:2202.07513*, 2022.
195. H. Le, L. Zhang, A. Said, G. Sautiere, Y. Yang, P. Shrestha, F. Yin, R. Pourreza, and A. Wiggers, "Mobilecodec: neural inter-frame video compression on mobile devices," in *Proceedings of the 13th ACM Multimedia Systems Conference*, 2022, pp. 324–330.
196. T. van Rozendaal, T. Singhal, H. Le, G. Sautiere, A. Said, K. Buska, A. Raha, D. Kalatzis, H. Mehta, F. Mayer *et al.*, "Mobilenvc: Real-time 1080p neural video compression on a mobile device," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 4323–4333.
197. B. Li, H. Li, L. Li, and J. Zhang, "λ domain rate control algorithm for high efficiency video coding," *IEEE transactions on Image Processing*, vol. 23, no. 9, pp. 3841–3854, 2014.
198. Y. Li, X. Chen, J. Li, J. Wen, Y. Han, S. Liu, and X. Xu, "Rate control for learned video compression," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 2829–2833.
199. N. Fathima, J. Petersen, G. Sautière, A. Wiggers, and R. Pourreza, "A neural video codec with spatial rate-distortion control," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 5365–5374.

 doi:10.20944/preprints202604.0035.v1

200. S. Liao, C. Jia, H. Fan, J. Yan, and S. Ma, "Rate-quality based rate control model for neural video compression," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4215–4219.
201. Y. Zhang, G. Lu, Y. Chen, S. Wang, Y. Shi, J. Wang, and L. Song, "Neural rate control for learned video compression," in *The Twelfth International Conference on Learning Representations*, 2023.
202. A. Mandhane, A. Zhernov, M. Rauh, C. Gu, M. Wang, F. Xue, W. Shang, D. Pang, R. Claus, C.-H. Chiang *et al.*, "Muzero with self-competition for rate control in vp9 video compression," *arXiv preprint arXiv:2202.06626*, 2022.
203. J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel *et al.*, "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, 2020.
204. U. Gadot, A. Shocher, S. Mannor, G. Chechik, and A. Hallak, "Rl-rc-dot: A block-level rl agent for task-aware video compression," *arXiv preprint arXiv:2501.12216*, 2025.
205. T. Xu, H. Gao, C. Gao, Y. Wang, D. He, J. Pi, J. Luo, Z. Zhu, M. Ye, H. Qin *et al.*, "Bit allocation using optimization," in *International Conference on Machine Learning*. PMLR, 2023, pp. 38 377–38 399.
206. Z. Liu, L. Herranz, F. Yang, S. Zhang, S. Wan, M. Mrak, and M. G. Blanch, "Slimmable video codec," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1743–1747.
207. Z. Hu and D. Xu, "Complexity-guided slimmable decoder for efficient deep video compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 358–14 367.
208. C. Zhang and W. Gao, "Learned rate control for frame-level adaptive neural video compression via dynamic neural network," in *European conference on computer vision*. Springer, 2024, pp. 239–255.
209. J. Yu, L. Yang, N. Xu, J. Yang, and T. Huang, "Slimmable neural networks," *arXiv preprint arXiv:1812.08928*, 2018.
210. F. Yang, L. Herranz, Y. Cheng, and M. G. Mozerov, "Slimmable compressive autoencoders for practical neural image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4998–5007.
211. Y. He and L. Xiao, "Structured pruning for deep convolutional neural networks: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 46, no. 5, pp. 2900–2919, 2023.
212. E. Agustsson and R. Timofte, "NTIRE 2017 challenge on single image super-resolution: Dataset and study," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.
213. S. Nah, S. Baik, S. Hong, G. Moon, S. Son, R. Timofte, and K. Mu Lee, "NTIRE 2019 challenge on video deblurring and super-resolution: Dataset and study," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.
214. T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, "Video enhancement with task-oriented flow," *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.
215. T. Li, M. Xu, C. Zhu, R. Yang, Z. Wang, and Z. Guan, "A deep learning approach for multi-frame in-loop filter of hevc," *IEEE Transactions on Image Processing*, vol. 28, no. 11, pp. 5663–5678, 2019.
216. D. Ma, F. Zhang, and D. R. Bull, "Bvi-dvc: A training database for deep video compression," *IEEE Transactions on Multimedia*, vol. 24, pp. 3847–3858, 2021.
217. J. Nawała, Y. Jiang, F. Zhang, X. Zhu, J. Sole, and D. Bull, "Bvi-aom: A new training dataset for deep video compression optimization," in *IEEE Visual Communications and Image Processing*, 2024.
218. X. Xu, S. Liu, and Z. Li, "Tencent video dataset (tvd): A video dataset for learning-based visual data compression and analysis," *arXiv preprint arXiv:2105.05961*, 2021.
219. F. Bossen, "Common test conditions and software reference configurations," in *3rd. JCT-VC Meeting, Guangzhou, CN, October 2010*, 2010.
220. X. Zhao, Z. Lei, A. Norkin, T. Daede, and A. Tourapis, "Aom common test conditions v2. 0," *Alliance for Open Media, Codec Working Group Output Document*, 2021.
221. A. Mercat, M. Viitanen, and J. Vanne, "UVG Dataset: 50/120fps 4K Sequences for Video Codec Analysis and Development," in *MMSys*. ACM, 2020, pp. 297–302.
222. H. Wang, W. Gan, S. Hu, J. Y. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C. J. Kuo, "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset," in *ICIP*. IEEE, 2016, pp. 1509–1513.
223. F. Zhang, D. Bull, J. Nawala, Y. Jiang, X. Zhu, J. Sole, and E. Alshina, "[ahg11] response to call for training materials for neural network-based video coding tool development," Joint Video Experts Team (JVET), Tech. Rep. JVET-AK0255, January 2025.

 doi:10.20944/preprints202604.0035.v1

224. J. Nawała, Y. Jiang, F. Zhang, X. Zhu, J. Sole, and D. Bull, "Bvi-aom dataset (cwg-e082)," in *AOM proposal*, 2024.
225. X. Zhao, Z. Lei, A. Norkin, T. Daede, and A. Tourapis, "Aom common test conditions v3. 0," *Document, CWG-C038i*, vol. 5, 2022.
226. Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
227. L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.
228. G.-h. Chen, C.-l. Yang, and S.-l. Xie, "Gradient-based structural similarity for image quality assessment," in *2006 International Conference on Image Processing*, 2006, pp. 2929–2932.
229. Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on image processing*, vol. 20, no. 5, pp. 1185–1198, 2010.
230. A. Rehman, K. Zeng, and Z. Wang, "Display device-adapted video quality-of-experience assessment," in *Human Vision and Electronic Imaging XX*, vol. 9394. International Society for Optics and Photonics, 2015, p. 939406.
231. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. on Image Processing*, vol. 15, pp. 430–444, 2006.
232. E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011 006(1–21), 2010.
233. P. G. Barten, *Contrast sensitivity of the human eye and its effects on image quality*. SPIE Press, 1999.
234. K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. on Image Processing*, vol. 19, no. 2, pp. 335–350, 2010.
235. Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, 2016.
236. C. G. Bampis, Z. Li, and A. C. Bovik, "Spatiotemporal feature integration and model fusion for full reference video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 8, pp. 2256–2270, 2018.
237. F. Zhang, A. Katsenou, C. Bampis, L. Krasula, Z. Li, and D. Bull, "Enhancing vmaf through new feature integration and model combination," in *2021 Picture Coding Symposium (PCS)*. IEEE, 2021, pp. 1–5.
238. M. Siniukov, A. Antsiferova, D. Kulikov, and D. Vatolin, "Hacking vmaf and vmaf neg: vulnerability to different preprocessing methods," in *Proceedings of the 2021 4th Artificial Intelligence and Cloud Computing Conference*, 2021, pp. 89–96.
239. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
240. K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 44, no. 5, pp. 2567–2581, 2020.
241. X. Liu, J. van de Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
242. C. Feng, D. Danier, F. Zhang, and D. Bull, "Rankdvqa: Deep vqa based on ranking-inspired hybrid training," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 1648–1658.
243. C. Feng, T. Peng, F. Zhang, and D. Bull, "Towards unified video quality assessment," *arXiv preprint arXiv:2512.02224*, 2025.
244. H. Wu, Z. Zhang, W. Zhang, C. Chen, C. Li, L. Liao, A. Wang, E. Zhang, W. Sun, Q. Yan, X. Min, G. Zhai, and W. Lin, "Q-Align: Teaching lmms for visual scoring via discrete text-defined levels," *arXiv preprint arXiv:2312.17090*, 2023, equal Contribution by Wu, Haoning and Zhang, Zicheng. Corresponding Authors: Zhai, Guangtao and Lin, Weisi.
245. H. Zhu, H. Wu, Y. Li, Z. Zhang, B. Chen, L. Zhu, Y. Fang, G. Zhai, W. Lin, and S. Wang, "Adaptive image quality assessment via teaching large multimodal model to compare," *Advances in Neural Information Processing Systems*, vol. 37, pp. 32 611–32 629, 2024.
246. Z. You, X. Cai, J. Gu, T. Xue, and C. Dong, "Teaching large language models to regress accurate image quality scores using score distribution," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 14 483–14 494.

 doi:10.20944/preprints202604.0035.v1

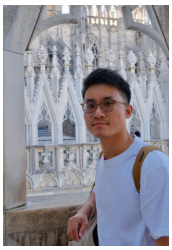
247. Q. Ge, W. Sun, Y. Zhang, Y. Li, Z. Ji, F. Sun, S. Jui, X. Min, and G. Zhai, "LMM-VQA: Advancing video quality assessment with large multimodal models," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
248. W. Wen, Y. Wang, N. Birkbeck, and B. Adsumilli, "An ensemble approach to short-form video quality assessment using multimodal llm," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.
249. W. Li, X. Zhang, S. Zhao, Y. Zhang, J. Li, L. Zhang, and J. Zhang, "Q-insight: Understanding image quality via visual reinforcement learning," *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
250. T. Wu, J. Zou, J. Liang, L. Zhang, and K. Ma, "Visualquality-r1: Reasoning-induced image quality assessment via reinforcement learning to rank," *arXiv preprint arXiv:2505.14460*, 2025.
251. Z. Chen, X. Zhang, W. Li, R. Pei, F. Song, X. Min, X. Liu, X. Yuan, Y. Guo, and Y. Zhang, "Grounding-iqa: Multimodal language grounding model for image quality assessment," *arXiv preprint arXiv:2411.17237*, 2024.
252. C. Chen, S. Yang, H. Wu, L. Liao, Z. Zhang, A. Wang, W. Sun, Q. Yan, and W. Lin, "Q-ground: Image quality grounding with large multi-modality models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 486–495.
253. H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, K. Xu, C. Li, J. Hou, G. Zhai *et al.*, "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 25 490–25 500.
254. Z. You, Z. Li, J. Gu, Z. Yin, T. Xue, and C. Dong, "Depicting beyond scores: Advancing image quality assessment through multi-modal language models," in *European Conference on Computer Vision*. Springer, 2024, pp. 259–276.
255. G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves," in *13th VCEG Meeting*, no. VCEG-M33. Austin, Texas, USA: ITU-T, April 2001.
256. Alliance for Open Media, "AOM Video Model (AVM) Codec 2.0.0 Library," https://gitlab.com/AOMediaCodec/avm, 2024, accessed: 2024-04-10.
257. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. Asilomar Conference on Signals, Systems and Computers*, vol. 2. IEEE, 2003, p. 1398.
258. J. Xu, S. Wang, J. Chen, Z. Li, P. Jia, F. Zhao, G. Xiang, Z. Hao, S. Zhang, and X. Xie, "Decouple distortion from perception: Region adaptive diffusion for extreme-low bitrate perception image compression," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 18 051–18 061.

Short Biography of Authors

Ge Gao (Member, IEEE) received the B.Eng degree in Electrical and Electronic Engineering from the University of Manchester in 2018 and M.Sc. degree in Artificial Intelligence from the University of Southampton in 2019. He is currently a Research Associate with the School of Computer Science, University of Bristol. His research interests focus on low-level computer vision including neural video compression, implicit neural representations, and generative models.

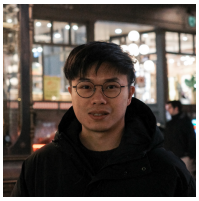
Chen Feng (Member, IEEE) received the B.Sc. degree in automation and electrical engineering from the University of Science and Technology Beijing, China, in 2018, and the M.Sc. and Ph.D. degrees in electrical and electronic engineering from the University of Bristol, U.K., in 2019 and 2025, respectively. His research interests focus on low-level computer vision, video quality assessment, perceptual video compression, and multimodal large language models. Dr. Feng's research is funded by the Amazon Research Awards and UKRI MyWorld. He was the recipient of the First-Place award in the Video Perception track at the 6th Challenge on Learned Image Compression at DCC 2024 and the First Prize in the HDR VQM Grand Challenge at IEEE/CVF WACV 2023.

 doi:10.20944/preprints202604.0035.v1



Yuxuan Jiang (Student Member, IEEE) is currently pursuing the Ph.D. degree with the Visual Information Lab, University of Bristol. He received the B.Eng. degree in Information Engineering from Southeast University in 2020, and the M.Sc. degree in Electrical and Electronic Engineering from Imperial College London in 2021. His research focuses on image and video super-resolution, video compression, perceptual quality assessment, and multimodal models for low-level vision.

Tianhao Peng (Student Member, IEEE) received the B.Sc. degree in Mathematics and Computer Science from the University of Bristol, U.K., in 2024. She is currently pursuing the Ph.D. degree in Computer Science at the University of Bristol. Her research interests focus on low-level computer vision, including image/neural video compression, video quality assessment, implicit neural representations, and data compression.



Ho Man Kwan (Student Member, IEEE) received the B.Eng. degree in Computer Engineering and the M.Phil. degree in Electronic and Computer Engineering from the Hong Kong University of Science and Technology in 2018 and 2021, respectively. He is currently pursuing the Ph.D. degree at the University of Bristol. His main research interests include neural compression for 2-D and volumetric videos using implicit neural representations.

Siyue Teng (Student Member, IEEE) received the B.Eng. degree in Electrical and Electronic Engineering from Xidian University and the M.Sc. degree in Data Science from the University of Bristol, both in 2023. She is currently pursuing the Ph.D. degree at the University of Bristol. Her research interests include low-level computer vision, 3D Gaussian Splatting, and generative models for multimedia processing.

Chengxi Zeng (Member, IEEE) received the M.Eng. degree from the University of Bristol in 2020 and the Ph.D. degree in computer vision from the University of Bristol in 2025. His work varies from high-level computer vision challenges, such as vision-language models for downstream tasks, to low-level video quality metrics and enhancing technologies, such as super-resolution.

Yixuan Li (Member, IEEE) received the B.S. degree from the Central South University, China, in 2018, and the M.S. degree from China University of Mining and Technology, China, in 2021. In 2025, she obtained the Ph.D. degree in the Department of Computer Science at the City University of Hong Kong, Hong Kong SAR. She was a research visitor at Nanyang Technological University, Singapore, and was Postdoctoral Fellow at City University of Hong Kong. She is currently a Postdoctoral Research Associate at University of Bristol, the UK. Her research interests lie in the intersection of computer vision, visual quality assessment, large multimodal models, and multimedia forensics.

Changqi Wang received the B.S. degree in Telecommunication Engineering from Northeastern University, China in 2021, and M.S. degree in Information and Communication Engineering from Northeastern University, China in 2024. He is currently pursuing the Ph.D. degree at the University of Bristol, Bristol, the U.K. His research interests focus on computer vision, low-level tasks, and video compression.

 doi:10.20944/preprints202604.0035.v1

Robbie Hamilton(Student Member, IEEE) received the MPhys degree in Physics from the University of Strathclyde in 2023. He is currently a PhD student with the School of Computer Science, University of Bristol. His research interests focus on applications of deep learning to video quality assessment and generated content.

Zihao Qi (Student Member, IEEE) received the B.S. degree in Electronic Engineering from Tsinghua University, Beijing, China, in 2016, and the M.Phil. degree from the Hong Kong University of Science and Technology, Hong Kong SAR, in 2019. From 2019 to 2020, he was a Research Consultant at the TCL Industrial Research Center, Hong Kong SAR. He is currently pursuing the Ph.D. degree at the University of Bristol, Bristol, the U.K. His research interests focus on computer vision, visual quality assessment, video compression, and natural language processing.

Fan Zhang (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees from Shanghai Jiao Tong University, Shanghai, China, in 2005 and 2008, respectively, and the Ph.D. degree from the University of Bristol, Bristol, U.K., in 2012. He is currently a Senior Lecturer within the School of Computer Science, University of Bristol. He served as an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology (2022-2024), and was a guest editor of IEEE Journal on Emerging and Selected Topics in Circuits and Systems (in 2024) and Frontiers in Signal Processing (in 2022). Fan is also a member of the Visual Signal Processing and Communications Technical Committee associated with the IEEE Circuits and Systems Society, and a Senior Area Editor for IEEE Transactions on Circuits and Systems for Video Technology. His research interests focus on low-level computer vision including video compression, quality assessment, super resolution and video frame interpolation.

David R. Bull (Fellow, IEEE) received the B.Sc. degree from the University of Exeter, Exeter, U.K., in 1980, the M.Sc. degree from the University of Manchester, Manchester, U.K., in 1983, and the Ph.D. degree from the University of Cardiff, Cardiff, U.K., in 1988. He was previously a Systems Engineer with Rolls Royce, Bristol, U.K., and a Lecturer with the University of Wales, Cardiff, U.K. In 1993, he joined the University of Bristol, Bristol, U.K., and is currently its Chair of Signal Processing and the Director of Bristol Vision Institute. He is also the Director of the recently announced £46 m UKRI 'MyWorld' Strength in Places Programme. In 2001, he co-founded a university spin-off company, ProVision Communication Technologies Ltd., specializing in wireless video technology. He has authored more than 850 papers on the topics of image and video communications and analysis for wireless, Internet and broadcast applications, together with numerous patents, several of which have been exploited commercially. He is the author of three books, and has delivered numerous invited/keynote lectures and tutorials. He was the recipient of the two IET Premium Awards for his work. Dr. Bull is a Fellow of the Institution of Engineering and Technology.