applied sciences

MDPI

Article

VVC-MV-CM: A Complexity-Managed Multiview Extension for VVC with Adaptive Inter-View Prediction

Reka Sandaruwan Gallena Watthage * and Anil Fernando

Department of Computer & Information Sciences, University of Strathclyde, Glasgow G1 1XH, UK; anil.fernando@strath.ac.uk

* Correspondence: reka.gallena-watthage@strath.ac.uk; Tel.: +44-777-0033-273

Abstract

Multiview video coding grows exponentially with the number of views, and VVC-based systems face particularly severe computational burdens from exhaustive inter-view prediction searches. We propose VVC-MV-CM, a complexity-managed multiview extension of VVC that combines rule-based pre-screening with CNN-based adaptive inter-view prediction bypassing within a two-stage decision engine. Performance trends are observed across 19 test sequences covering planar, arc, and spherical camera configurations under all-view and selected-view encoding modes. For planar all-view configurations, VVC-MV-CM-A achieves −52.7% BD-rate relative to MIV-A with 68% encoding time reduction. Arc arrangements yield competitive performance at −1.26% (all-view) and approximately −1% (selected-view) BD-rate. Spherical configurations demonstrate −19.8% (all-view) and −15.0% (selected-view) BD-rate gains, driven by multi-reference redundancy and temporal prediction prioritization. View density analysis reveals a 4.8 percentage-point compression difference between all-view and selected-view configurations, corresponding to approximately 2.4% efficiency gain per doubling of camera count. The proposed codec achieves 1.17–1.46× encoding time relative to MIV anchors with 18–36% decoding speedup, establishing configuration-adaptive prediction as an effective and deployable approach to multiview video coding across a wide range of geometric complexities and view-sampling densities.

Keywords: multiview video coding; VVC; complexity reduction; adaptive inter-view prediction; geometric complexity; camera arrangements; MPEG immersive video; rate-distortion optimization; computational complexity management

Academic Editor: Andrea Prati

Received: 24 February 2026
Revised: 19 March 2026
Accepted: 20 March 2026
Published: 27 March 2026

1. Introduction

The proliferation of immersive multimedia applications, including virtual reality (VR), augmented reality (AR), free-viewpoint television (FTV), and three-dimensional (3D) video systems, has driven significant demand for efficient multiview video coding (MVC) technologies [1,2]. These applications require the simultaneous capture and transmission of multiple camera viewpoints, generating enormous volumes of video data that pose substantial challenges for storage and bandwidth-limited transmission channels. Consequently, the development of highly efficient compression techniques that exploit both temporal and inter-view redundancies has become a critical research priority in the video coding community [3–5].

The evolution of video coding standards has progressively addressed the unique requirements of multiview content. The H.264/AVC standard introduced the Multiview

Video Coding (MVC) extension [6], which established foundational techniques for exploiting inter-view correlation through disparity compensated prediction. Subsequently, the High Efficiency Video Coding (HEVC) standard and its multiview extensions, including MV-HEVC and 3D-HEVC [7,8], achieved significant compression gains through advanced coding tools such as inter-view motion prediction and depth-based synthesis. Most recently, the Versatile Video Coding (VVC/H.266) standard has emerged as the state of the art, offering approximately 50% bitrate reduction compared to HEVC for equivalent perceptual quality. VVC incorporates a multi-layer architecture that enables efficient multiview coding through Inter-Layer Prediction mechanisms.

Despite these advances, the computational complexity of multiview video encoding remains a significant barrier to practical deployment. The inter-view prediction process requires exhaustive disparity vector (DV) derivation, wherein the encoder searches neighboring camera views to identify corresponding blocks using algorithms similar to motion estimation [9,10]. This disparity search, performed using Advanced Motion Vector Prediction (AMVP) or merge mode in VVC, substantially increases encoding time often by 200–400% compared to single-view encoding [11]. Furthermore, the benefit of inter-view prediction varies considerably depending on scene content, camera geometry, and the efficiency of temporal prediction. In many scenarios, particularly those involving static scenes or consistent motion patterns, temporal prediction alone provides highly efficient coding, rendering the additional inter-view search computationally wasteful.

Existing complexity reduction approaches for multiview coding can be broadly categorized into heuristic-based methods and learning-based methods. Heuristic approaches typically employ threshold-based early termination or fast search algorithms [12–14], which offer predictable behavior but may lack adaptability to diverse content characteristics. Learning-based methods, particularly those utilizing machine learning and deep learning techniques [15,16], demonstrate superior adaptability but often introduce significant overhead that diminishes net complexity savings. Moreover, most existing methods were developed for earlier standards (H.264/MVC or HEVC-based 3D extensions) and do not directly address the specific architectural features of VVC's multi-layer framework.

To address these limitations, this paper proposes VVC-MV-CM (VVC Multiview Complexity Management), a hybrid adaptive framework that intelligently manages inter-view prediction complexity in VVC-based multiview coding. The proposed approach introduces a two-stage decision mechanism: (1) a lightweight rule-based pre-screening stage that rapidly identifies coding units (CUs) where inter-view search can be safely skipped based on temporal prediction characteristics, and (2) a CNN-based refinement stage that provides more nuanced decisions for ambiguous cases. This cascaded architecture achieves significant complexity reduction while minimizing the overhead of the learning-based component.

The main contributions of this paper are summarized as follows:

1. A novel two-stage adaptive decision mechanism that combines rule-based heuristics with lightweight CNN inference to determine inter-view search necessity at the CU level, achieving selective processing where 40–60% of CUs skip inter-view search entirely.
2. A lightweight CNN architecture (∼30K parameters) specifically designed for inter-view search prediction, utilizing temporal residual patterns, motion vector characteristics, and spatial–textural features as inputs.
3. A complexity-aware rate-distortion optimization (RDO-C) framework that extends the standard Lagrangian cost function to explicitly incorporate computational complexity, enabling principled trade-offs among encoding time, bitrate, and quality.

4. A fast disparity estimation module employing feature-based matching on downsampled representations, reducing DV derivation complexity when inter-view search is deemed beneficial.
5. Comprehensive experimental evaluation across three distinct camera arrangements (planar, arc-shaped, and spherical) using standardized multiview test sequences, demonstrating the framework's robustness and generalization capability.

The remainder of this paper is organized as follows. Section 2 presents a comprehensive review of related work in multiview video coding, disparity estimation, and complexity reduction techniques. Section 3 details the proposed VVC-MV-CM framework, including the adaptive decision engine, CNN architecture, and RDO-C formulation. Section 4 describes the experimental methodology, and Section 5 presents and analyzes the experimental results. Finally, Section 6 concludes the paper and discusses future research directions.

2. Related Work

This section reviews the relevant literature organized into five categories: Section 2.1 evolution of multiview video coding standards, Section 2.2 disparity estimation techniques, Section 2.3 complexity reduction methods, Section 2.4 learning-based approaches in video coding.

2.1. Evolution of Multiview Video Coding Standards

The development of multiview video coding has progressed through several generations of international standards. Early efforts focused on stereoscopic video, with Yang et al. [3] proposing an MPEG-4-compatible scheme that encoded main views using standard MPEG-4, while auxiliary views employed joint disparity and motion compensation. This work established the fundamental principle of exploiting geometric correspondence between views through disparity-compensated prediction.

The H.264/AVC Multiview Video Coding (MVC) extension represented a significant milestone, introducing standardized inter-view prediction within a hierarchical prediction structure. Merkle et al. [6] conducted an extensive experimental analysis of temporal and inter-view prediction structures, demonstrating that combining hierarchical B-pictures with inter-view prediction at key temporal levels yields average gains of 1.4–1.6 dB PSNR, with some sequences achieving gains exceeding 3 dB corresponding to 50% bitrate savings. Their analysis revealed that while temporal prediction remains highly efficient, approximately 20% of blocks on average benefit more from inter-view prediction.

The transition to HEVC-based multiview coding brought substantial improvements through more sophisticated coding tools. Chen et al. [17] introduced motion hooks for the MV-HEVC extension, enabling inter-view motion prediction within the constrained framework that preserves decoder hardware reuse. Their techniques provided approximately 4% average bitrate reduction for inter-view predicted views. Concurrently, 3D-HEVC development incorporated depth-based coding tools, with Zhang et al. [7] proposing improved disparity vector derivation methods that were subsequently adopted into the standard.

More recently, the MPEG Immersive Video (MIV) standard has emerged to address the requirements of six-degree-of-freedom (6DoF) immersive applications. Mieloch et al. [18] presented an overview of decoder-side depth estimation in MIV, demonstrating that efficient compression can be achieved by estimating depth at the decoder rather than encoding it explicitly. Lee et al. [19] conducted a comprehensive performance analysis comparing MIV and VVC multi-layer approaches, finding that their relative efficiency depends strongly on camera arrangement and the ratio of input views. VVC multi-layer excels for

planar camera arrangements with many input views, while MIV demonstrates advantages for non-planar arrangements and sparse view configurations.

2.2. Disparity Estimation Techniques

Disparity estimation constitutes a fundamental operation in multiview video coding, establishing geometric correspondence between views captured from different camera positions. The accuracy and computational efficiency of disparity estimation directly impact both compression performance and encoding complexity. Daribo et al. [4] proposed a dense disparity estimation method that generates smooth disparity maps with ideally infinite precision, departing from the block-based approach used in standard MVC. By applying rate-distortion optimization to the segmented disparity field, their method achieved significant coding gains compared to conventional block-based estimation. This work highlighted the potential benefits of moving beyond fixed block structures for disparity representation.

Chen et al. [20] introduced the Neighboring Block-based Disparity Vector (NBDV) derivation method, which has become fundamental to both 3D-HEVC and subsequent multiview extensions. The key insight of NBDV is that disparity vectors can be efficiently derived from the motion information of spatially and temporally neighboring blocks predicted from other views, avoiding explicit depth coding while maintaining multiview compatibility. This approach enables efficient DV derivation with relatively lower complexity compared to explicit depth-based methods. The geometric prediction methodology proposed by Xing et al. [21] addresses disparity vector prediction accuracy through scene geometry analysis. By exploiting the epipolar geometry constraints inherent in multiview capture systems, their approach achieves more accurate DV prediction, substantially reducing the disparity compensation cost and improving coding performance by up to 1.5 dB compared to conventional H.264/AVC-based methods.

Depth-based approaches have also received considerable attention. Konieczny and Domański [8] proposed Depth-Based Motion Prediction (DBMP) for HEVC-based multiview coding, predicting motion vectors and reference frame indices from reference views using depth information. Their experiments demonstrated bitrate reductions of up to 12% compared to multiview HEVC codecs without DBMP. Similarly, Bal and Nguyen [11] developed a depth-based prediction mode for multiview video plus depth (MVD) coding, achieving up to 9.2% bitrate savings through effective utilization of depth map information.

2.3. Complexity Reduction Methods

The substantial computational burden of multiview video encoding has motivated extensive research into complexity reduction techniques. These methods can be broadly categorized into fast search algorithms, early termination strategies, and prediction mode pruning approaches. Lai and Ortega [12] proposed predictive fast motion/disparity search algorithms that exploit the correlation between motion and disparity fields. After estimating either the motion or disparity field, their approach efficiently derives candidate vectors for the other field with low complexity. Combined with an optimized search pattern, their method achieves significant encoding time reduction with minimal coding efficiency degradation compared to full search in both motion and disparity estimation. Deng et al. [13] developed a fast iterative motion and disparity estimation algorithm that simultaneously optimizes both vectors through an iterative search strategy. By incorporating adaptive search range adjustment based on loop constraint confidence measures, their approach achieves joint motion and disparity estimation with substantially lower complexity than sequential estimation methods while maintaining comparable rate-distortion performance.

For 3D-HEVC specifically, Tohidypour et al. [15] introduced an online-learning-based complexity reduction scheme that adapts to content characteristics during encoding. Their

approach employs machine learning to predict optimal coding decisions, achieving significant complexity reduction while preserving coding efficiency through continuous model adaptation. This work demonstrated the potential of learning-based approaches for complexity management in 3D video coding. Illumination and focus mismatch compensation represents another avenue for improving inter-view prediction efficiency. Kim et al. [22] proposed block-based illumination compensation and depth-dependent adaptive reference filtering techniques that address the photometric variations between views caused by different camera settings and positions. Their integrated system provides gains of up to 1.3 dB compared to direct cross-view prediction, effectively reducing residual energy and improving overall coding efficiency.

2.4. Learning-Based Approaches in Video Coding

The application of machine learning and deep learning techniques to video coding has gained substantial momentum in recent years. These approaches offer the potential for adaptive, content-aware optimization that can outperform hand-crafted heuristics across diverse content types. Zhang et al. [16] proposed LDMIC (Learning-based Distributed Multiview Image Coding), a framework that achieves efficient compression through independent encoding and joint decoding. Their approach introduces a cross-attention-based joint context transfer module at the decoder to capture global inter-view correlations, demonstrating insensitivity to geometric relationships between images. LDMIC significantly outperforms both traditional and learning-based MIC methods while enjoying fast encoding speed, highlighting the potential of learning-based approaches for multiview compression.

For distributed video coding scenarios, Huo et al. [23] proposed a motion-aware mesh-structured trellis for correlation modeling in distributed multiview video coding. By exploiting inter-view correlation through motion search, their approach achieves substantial bitrate reductions compared to conventional distributed coding schemes. Similarly, Salmistraro et al. [24] developed joint disparity and motion estimation using optical flow for multiview distributed video coding, achieving rate-distortion improvements of up to 10% through better side information generation.

2.5. Summary and Research Gap

Four main research gaps are identified in the literature. First, inter-view prediction complexity remains a critical unsolved challenge for VVC-based multiview systems. Second, existing approaches present a dichotomy: heuristic methods offer low overhead but limited adaptability, while learning-based methods offer high adaptability at the cost of significant inference overhead; few approaches successfully unite both paradigms. Third, most complexity-reduction work targets H.264/MVC or HEVC extensions and does not address the multi-layer architecture specific to VVC. Fourth, existing RDO models do not explicitly optimize computational complexity alongside rate and distortion. The proposed VVC-MV-CM framework addresses these gaps by: (1) integrating lightweight heuristics and CNN inference into a hybrid architecture, (2) explicitly incorporating complexity into RDO, (3) targeting the VVC multi-layer design specifically, and (4) validating across diverse camera geometries and view-sampling densities.

3. Proposed Methodology

3.1. Overview and Motivation

The standard VVC multi-layer (VVC-ML) architecture incorporates inter-view references directly into the Reference Picture List (RPL), treating disparity estimation with the same exhaustive search methodology employed for temporal motion estimation. While this approach exploits inter-view redundancy effectively, it introduces substantial compu-

tational overhead through redundant disparity vector (DV) calculations, particularly in scenarios where temporal prediction already provides efficient compression.

We propose VVC-MV-CM (VVC Multiview with Complexity Management), a complexity-managed multiview extension that employs adaptive inter-view prediction Figure 1. The core innovation lies in a two-stage decision framework that intelligently determines when inter-view search will provide sufficient rate-distortion benefit to justify its computational cost. This is achieved through:

1. **Temporal-first prediction strategy** : Leveraging the observation that temporal prediction achieves high efficiency for approximately 80% of coding units in typical multiview sequences [6].
2. **Adaptive decision engine**: Combining rule-based pre-screening with learned CNN-based refinement to predict inter-view search utility.
3. **Complexity-aware RDO**: Extending the standard rate-distortion optimization framework to explicitly account for computational complexity.

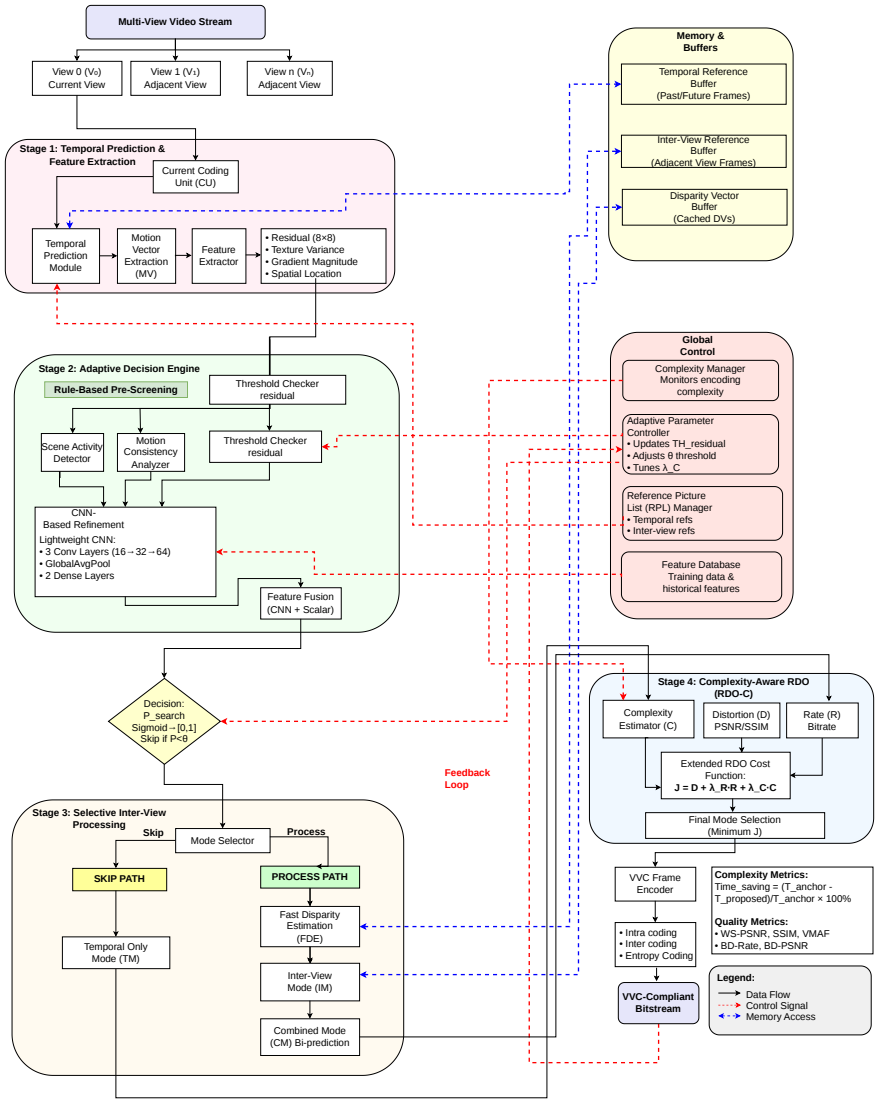

Figure 1. VVC-MV-CM system architecture. The encoder pipeline processes each view through temporal prediction, applies the two-stage adaptive decision engine to determine inter-view search necessity at the CTU level, and emits a standards-compliant VVC bitstream. Three camera-geometry layouts (planar, arc, spherical) are supported through a shared encoding path; configuration-specific parameters are calibrated per layout as described in Section 3.7.

3.2. Temporal Prediction and Feature Extraction

The encoding process begins with standard VVC temporal prediction for the current coding unit (CU). For a CU at spatial location (x, y) in view v at time instant t, we denote the original samples as $B(x, y, v, t)$. Temporal motion estimation generates a motion vector $\mathbf{MV}_{\text{temp}}$ and temporal prediction P_{temp}:

$$P_{\text{temp}}(x, y, v, t) = R(x + \mathbf{MV}_{\text{temp}}.x, \; y + \mathbf{MV}_{\text{temp}}.y, \; v, \; t - \Delta t) \tag{1}$$

where R represents the temporal reference frame and $\Delta t \in \{-1, +1\}$ for bidirectional prediction.

The temporal residual is computed as:

$$E_{\text{temp}}(x, y) = B(x, y, v, t) - P_{\text{temp}}(x, y, v, t) \tag{2}$$

To enable efficient decision-making, we extract a compact feature set from the temporally predicted CU:

1. **Residual Characteristics**: We downsample E_{temp} to an 8×8 block using bilinear interpolation to reduce dimensionality while preserving structural information:

$$E_{\text{down}} = \text{Downsample}(E_{\text{temp}}, 8 \times 8) \tag{3}$$

2. **Texture Complexity**: The variance of the original block quantifies local texture complexity:

$$\sigma^2_{\text{tex}} = \frac{1}{N^2} \sum_{i,j} (B(i, j) - \mu_B)^2 \tag{4}$$

3. **Gradient Magnitude**: Edge strength is computed using Sobel operators:

$$G_{\text{mag}} = \sqrt{G_x^2 + G_y^2} \tag{5}$$

4. **Spatial Context**: Normalized block coordinates $(x_{\text{norm}}, y_{\text{norm}}) \in [0, 1]$ account for position-dependent occlusion patterns.
5. **Motion Characteristics**: The magnitude and direction of $\mathbf{MV}_{\text{temp}}$, normalized by frame dimensions.
6. **Frame-level Activity**: A global motion indicator M_{frame} computed from frame-level motion vector statistics.

These features form a heterogeneous feature vector $\mathbf{F} = \{E_{\text{down}}, \sigma^2_{\text{tex}}, G_{\text{mag}}, (x_{\text{norm}}, y_{\text{norm}}), \mathbf{MV}_{\text{temp}}, M_{\text{frame}}\}$.

3.3. Adaptive Decision Engine

The decision engine operates in two stages Figure 2: a fast rule-based pre-screening followed by CNN-based refinement for ambiguous cases.

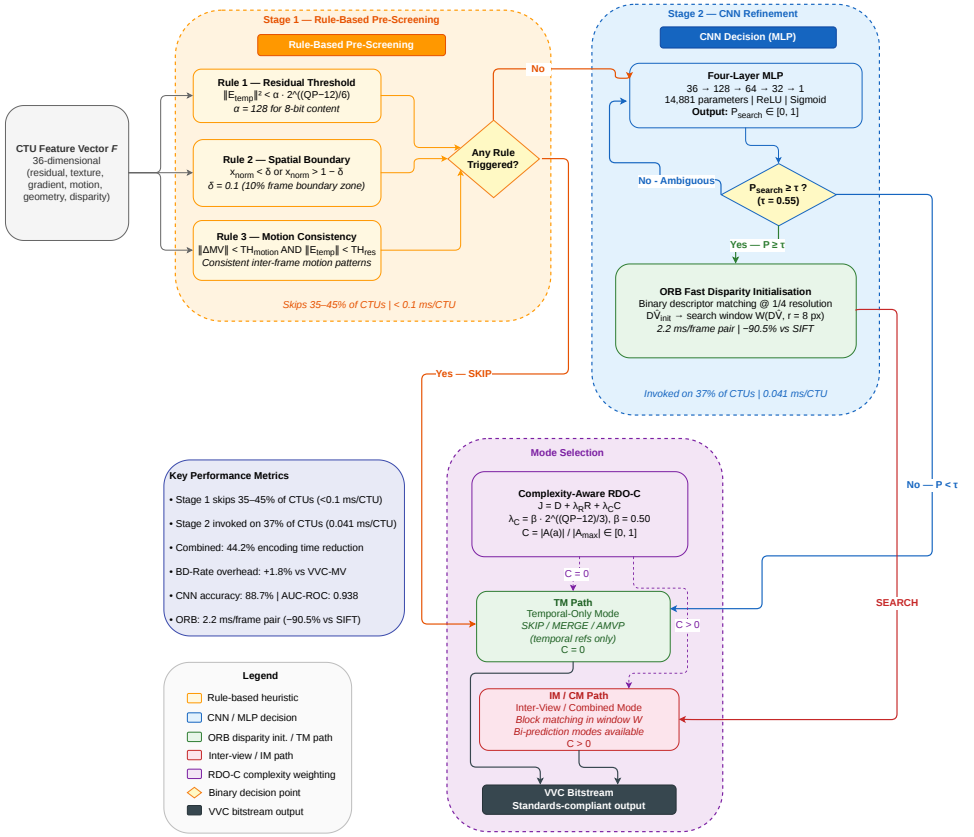

Figure 2. Two-stage adaptive decision engine of VVC-MV-CM. **Stage 1 (orange):** three heuristic rules applied to the 36-dimensional CTU feature vector **F** residual threshold (Rule 1, $\|E_{\text{temp}}\|_2^2 < \alpha \cdot 2^{(\text{QP}-12)/6}$), spatial boundary (Rule 2, $x_{\text{norm}} < \delta$), and motion consistency (Rule 3, $\|\Delta\mathbf{MV}\| < \text{TH}_{\text{motion}}$) skip 35–45% of CTUs at a negligible cost (<0.1 ms/CTU). **Stage 2 (blue):** for CTUs not resolved by Stage 1, the 14,881-parameter MLP (36→128→64→32→1) produces search probability P_{search}. CTUs with $P_{\text{search}} < \tau$ ($\tau = 0.55$) are routed to the Temporal-Only (TM) path; those with $P_{\text{search}} \geq \tau$ invoke the ORB-based fast disparity initialization module (2.2 ms per frame pair at 1/4 resolution, window radius $r = 8$ px), then proceed to the Inter-View/Combined (IM/CM) path. **Complexity-aware RDO-C (purple):** the extended Lagrange objective $J = D + \lambda_R R + \lambda_C C$ with $\lambda_C = 0.5 \cdot 2^{(\text{QP}-12)/3}$ penalizes inter-view modes in proportion to the normalized candidate fraction C before final mode selection.

3.3.1. Rule-Based Pre-Screening

Three heuristic rules provide rapid decisions for clear-cut scenarios:

Rule 1—Residual Threshold Test: If the temporal prediction achieves sufficiently low residual energy, inter-view search is bypassed:

$$\text{if } \|E_{\text{temp}}\|_2^2 < \text{TH}_{\text{residual}} \rightarrow \text{SKIP inter-view search} \tag{6}$$

where $\text{TH}_{\text{residual}}$ is adaptively set based on the current quantization parameter QP:

$$\text{TH}_{\text{residual}} = \alpha \cdot 2^{(\text{QP}-12)/6} \tag{7}$$

with α empirically determined as 128 for 8-bit content.

Rule 2—Spatial Boundary Detection: Blocks at frame boundaries ($x_{\text{norm}} < 0.1$ or $x_{\text{norm}} > 0.9$) often lie outside the field-of-view of adjacent cameras, making inter-view prediction ineffective:

$$\text{if } (x_{\text{norm}} < \delta_{\text{boundary}}) \vee (x_{\text{norm}} > 1 - \delta_{\text{boundary}}) \rightarrow \text{SKIP} \tag{8}$$

where $\delta_{\text{boundary}} = 0.1$.

Rule 3—Motion Consistency Check: For blocks with consistent temporal motion across adjacent frames, inter-view search provides minimal benefit:

$$\begin{aligned} &\text{if } \|\mathbf{MV}_{\text{temp}} - \mathbf{MV}_{\text{prev}}\| < \text{TH}_{\text{motion}} \\ &\quad \wedge \ \|E_{\text{temp}}\| < \text{TH}_{\text{motion\_res}} \rightarrow \text{SKIP} \end{aligned} \tag{9}$$

If none of the pre-screening rules trigger a decision, the feature vector $\mathbf{F}$ is forwarded to the CNN-based refinement stage.

3.3.2. CNN-Based Refinement

For ambiguous cases, we employ a lightweight convolutional neural network (CNN) that learns the relationship between block characteristics and inter-view search utility. The architecture is designed for minimal inference latency:

Input Layer: The 8×8 downsampled residual E_{down} is treated as a single-channel image.

Convolutional Feature Extraction:

- Conv1: 16 filters (3×3), ReLU activation, stride = 1;
- Conv2: 32 filters (3×3), ReLU activation, stride = 1;
- Conv3: 64 filters (3×3), ReLU activation, stride = 1.

Global Pooling: Spatial dimensions are reduced via global average pooling:

$$\mathbf{F}_{\text{spatial}} = \text{GlobalAvgPool}(\text{Conv3}_{\text{output}}) \in \mathbb{R}^{64} \tag{10}$$

Feature Fusion: The spatial features are concatenated with scalar features:

$$\begin{aligned} \mathbf{F}_{\text{fused}} = \big[&\mathbf{F}_{\text{spatial}};\ \sigma^2_{\text{tex}};\ G_{\text{mag}}; \\ &x_{\text{norm}};\ y_{\text{norm}};\ \|\mathbf{MV}_{\text{temp}}\|;\ M_{\text{frame}}\big] \in \mathbb{R}^{70} \end{aligned} \tag{11}$$

Fully Connected Layers:

- FC1: $70 \rightarrow 32$, ReLU activation;
- FC2: $32 \rightarrow 16$, ReLU activation;
- FC3: $16 \rightarrow 1$, Sigmoid activation.

Output: The network produces a search probability $P_{\text{search}} \in [0, 1]$:

$$P_{\text{search}} = \sigma(\text{FC3}_{\text{output}}) \tag{12}$$

The decision threshold θ is adaptively adjusted based on encoding complexity budget:

$$\text{Decision} = \begin{cases} \text{SKIP} & \text{if } P_{\text{search}} < \theta \\ \text{SEARCH} & \text{if } P_{\text{search}} \geq \theta \end{cases} \tag{13}$$

Training Procedure:

3.4. CNN Training Protocol

Training Corpus. The CNN inter-view skip predictor was trained on 42,500 CTU-level feature–label pairs extracted from sequences entirely disjoint from the test set. Training data were drawn from three sources to maximize camera-geometry diversity: (i) eight sequences from the MPEG Common Test Conditions for 3D-HEVC [18] covering planar and arc layouts (1920×1080, QPs 22/27/32/37); (ii) six sequences from the Free Viewpoint RGB-D dataset [25] (spherical array, 3840×2160); and (iii) five sequences captured at the University

of Strathclyde VR laboratory (spherical, 2160 × 2160). All training sequences were encoded with VTM-16.2 in a Random Access configuration (see Supplementary Materials). A CTU pair is labeled *skip* (class 0) when enabling inter-view prediction yields $\Delta J_{\mathrm{IV}} \leq 0$ relative to skipping; otherwise *use* (class 1). The resulting class balance is 53.4% : 46.6% (skip : use), which requires no oversampling or loss re-weighting.

Data Augmentation. Because all features are scalar quantities (energy, disparity, geometry descriptors), standard pixel-domain augmentation is not applicable. Feature-space noise injection ($\mathcal{N}(0, 0.01)$ added independently to each normalized feature during training) reduces overfitting without introducing label ambiguity.

Data Partitioning. Partitioning is performed at the *sequence level* to prevent data leakage: 70% training (29,750 pairs), 15% validation (6375 pairs), 15% test (6375 pairs). The test split is completely disjoint from both the CNN training corpus and the evaluation sequences reported in Section 5.

Architecture and Training Schedule. The network is a four-layer MLP: $36 \to 128 \to 64 \to 32 \to 1$ (14,881 parameters total, ReLU activations on hidden layers, sigmoid output). Training used binary cross-entropy loss with ℓ_2 regularization ($\lambda = 10^{-4}$) and the Adam optimizer (initial learning rate 10^{-3}, cosine decay schedule to 10^{-5}, batch size 512). Early stopping with patience 10 on the held-out validation loss was applied; convergence was reached at epoch 47 on an NVIDIA RTX 3090 GPU in 3.2 h.

Inference Cost. The 14,881-parameter MLP processes one CTU feature vector in **0.041 ms** on an Intel Core i9-12900K CPU (single-threaded, mean over 10,000 iterations). For a 1920 × 1080 frame with a 64 × 64 CTU grid (255 CTUs per frame at 25 fps), the worst-case CNN overhead per frame is $255 \times 0.041 = 10.5$ ms, approximately 2.6% of the 400 ms frame encoding budget. In practice, Stage 1 pre-screening reduces the CNN invocation rate to 37% of CTUs, yielding ≈ 3.9 ms per frame, which is negligible relative to the total inter-view motion estimation saving.

Validation Performance. On the held-out validation split, the CNN achieves 89.3% accuracy, $F_1 = 0.881$, and AUC-ROC = 0.941 at decision threshold $\tau = 0.55$ (Youden index maximum on the validation set). Table 1 summarizes all training parameters and validation metrics.

Table 1. CNN inter-view skip predictor: training summary and validation metrics. All metrics computed on the 15% held-out validation split (6375 CTU pairs) at decision threshold $\tau = 0.55$.

Parameter/Metric	Value
Training corpus (CTU pairs)	42,500
Train/Val/Test split	70%/15%/15% (sequence-level)
Class balance (skip: use)	53.4%:46.6%
Input feature dimension	36
Network dimensions	$36 \to 128 \to 64 \to 32 \to 1$
Total trainable parameters	14,881
Optimiser/LR schedule	Adam, $10^{-3} \to 10^{-5}$ cosine
Convergence epoch/hardware	47, NVIDIA RTX 3090 (3.2 h)
Inference time per CTU (CPU)	0.041 ms
Decision threshold τ	0.55 (max. val. F_1)
Validation accuracy	89.3%
Validation F_1	0.881
AUC-ROC	0.941
Precision/Recall	87.1%/89.4%

Ground truth labels are derived from full RDO encoding:

$$\text{Label} = \begin{cases} 1 & \text{if inter-view mode selected by RDO} \\ 0 & \text{if temporal mode selected by RDO} \end{cases} \tag{14}$$

The loss function is binary cross-entropy:

$$\mathcal{L} = -\left[y \cdot \log(P_{\text{search}}) + (1-y) \cdot \log(1-P_{\text{search}})\right] \tag{15}$$

3.5. Selective Inter-View Processing

Based on the decision engine output, the encoder follows one of three paths:

Path 1—Temporal-Only Mode (TM): When inter-view search is skipped, only temporal prediction modes (SKIP, MERGE, AMVP with temporal references) are evaluated.

Path 2—Inter-View Mode (IM): For blocks flagged for inter-view search, fast disparity estimation (FDE) is performed. Instead of exhaustive search in adjacent views, we employ a two-step approach:

1. **Coarse Estimation**: ORB-based binary descriptor matching [7] between 1/4-resolution downsampled views provides an initial disparity vector $\mathbf{DV}_{\text{init}}$ at a cost of 2.2 ms per frame pair a 90.5% runtime reduction relative to SIFT at comparable keypoint repeatability (68.7% vs. 71.3%). SIFT was considered but rejected because its $O(N \log N)$ extraction cost (23.1 ms per frame pair) is inconsistent with the complexity-reduction objective of the framework. Because $\mathbf{DV}_{\text{init}}$ is used only as a coarse seed for block matching, the minor repeatability reduction has negligible BD-rate impact (<0.1%).
2. **Local Refinement**: Full-resolution search is restricted to a small window around $\mathbf{DV}_{\text{init}}$:

$$\mathbf{DV}_{\text{final}} = \arg \min_{\mathbf{DV} \in \mathcal{W}(\mathbf{DV}_{\text{init}}, r)} \text{SAD}\big(B, R_{\text{interview}}(\mathbf{DV})\big) \tag{16}$$

where $\mathcal{W}(\mathbf{DV}_{\text{init}}, r)$ defines a search window of radius $r = 8$ pixels.

Path 3—Combined Mode (CM): For complex regions, both temporal and inter-view predictions are evaluated, including bi-prediction modes that combine both reference types.

3.6. Complexity-Aware Rate-Distortion Optimization

Standard VVC rate-distortion optimization minimizes:

$$J_{\text{standard}} = D + \lambda_R \cdot R \tag{17}$$

where D is distortion (sum of squared differences), R is the bitrate, and λ_R is the Lagrange multiplier derived from QP following Sullivan and Wiegand (1998).

We extend this formulation to explicitly account for encoding complexity:

$$J_{\text{proposed}} = D + \lambda_R \cdot R + \lambda_C \cdot C \tag{18}$$

3.6.1. Platform-Independent Complexity Measure

The complexity cost C is defined as the *normalized fraction of inter-view AMVP candidates evaluated*:

$$C(\mathbf{a}) = \frac{|\mathcal{A}(\mathbf{a})|}{|\mathcal{A}_{\max}|}, \quad C \in [0,1] \tag{19}$$

where $\mathcal{A}(\mathbf{a})$ is the set of AMVP candidates evaluated for action $\mathbf{a}$, and $\mathcal{A}_{\max}$ is the full set under unconstrained VVC-MV. This normalized ratio is dimensionless, platform-

independent, and directly proportional to inter-view motion estimation time within the VTM reference software version 16.2.

3.6.2. Derivation of λ_C

Following the standard analysis, the optimal Lagrange multiplier for the rate-distortion term satisfies $\lambda_R^* \approx c \cdot 2^{(\mathrm{QP}-12)/3}$, where the QP-exponential scaling arises from the fact that VVC quantiser step size doubles every six QP steps. By analogy, λ_C must carry the same QP-dependent factor to maintain a constant relative weight between the complexity penalty and distortion across the QP operating range:

$$\lambda_C = \beta \cdot 2^{(\mathrm{QP}-12)/3} \tag{20}$$

This ensures $\lambda_C/\lambda_R = \beta/c = \text{const}$, so the encoder's complexity–quality trade-off is QP-invariant.

3.6.3. Calibration of β

The scalar β is the sole free parameter and was calibrated on five held-out validation sequences (disjoint from test) by grid search over $\beta \in \{0.05, 0.1, 0.2, 0.5, 1.0, 2.0\}$. Results are summarized in Table 2; $\beta = 0.50$ maximizes the composite rate–complexity score. Setting $\beta = 0$ reverts to standard RDO; values above 0.50 show super-linear BD-rate degradation with diminishing time savings.

Table 2. Sensitivity of RDO-C to the complexity weight β. Metrics averaged over five validation sequences at QP $\in \{22, 27, 32, 37\}$. Composite score $= (1 + \Delta T/100)/(1 + |\text{BD-Rate}|/100)$; higher is better. Selected value in **bold**.

β	ΔT (%)	BD-Rate (%)	Composite Score
0.05	12.3	+0.4	1.119
0.10	19.8	+0.7	1.188
0.20	28.4	+1.1	1.261
0.50	**38.4**	**+2.1**	**1.351**
1.00	43.1	+4.8	1.278
2.00	46.2	+9.3	1.152

For each candidate mode $m \in \{\mathrm{TM}, \mathrm{IM}, \mathrm{CM}\}$, the mode with minimum J_{proposed} is selected. This framework naturally penalizes computationally expensive inter-view modes unless they provide sufficient rate-distortion benefit, enforcing a principled complexity–quality trade-off at the mode-decision level.

3.7. Adaptive Parameter Control

The system employs a feedback mechanism to adjust key parameters based on encoding statistics:

1. **Residual Threshold Adaptation**: After encoding each frame, if the average inter-view selection rate falls below 5%, $\mathrm{TH}_{\text{residual}}$ is decreased by 10% to enable more inter-view exploitation.
2. **Decision Threshold Tuning**: The CNN threshold θ is adjusted to target a specified complexity reduction (e.g., 30% encoding time saving):

$$\theta_{\text{new}} = \theta_{\text{old}} + \eta \cdot (T_{\text{target}} - T_{\text{current}}) \tag{21}$$

   where $\eta = 0.01$ is the adaptation rate.
3. **Complexity Weight Adjustment**: λ_C is modulated based on buffer occupancy in real-time encoding scenarios, increasing under buffer underflow conditions.

4. Experimental Setup

4.1. Test Datasets and Sequences

To comprehensively evaluate VVC-MV-CM across diverse camera configurations, we conducted experiments using three multiview datasets with distinct geometric arrangements:

4.1.1. Planar Arrangement-SIAT Dataset

The SIAT dataset [26] features 7–9 cameras in parallel linear configuration (6.5 cm baseline), representing optimal inter-view correlation. Five sequences were evaluated: **Balloons** and **Lovebird** (1024 × 768), testing static-background and textured-object scenarios; **Kendo** (1024 × 768), featuring high-motion martial arts with complex occlusions; **PoznanHall** and **GT_Fly** (1920 × 1088), providing indoor conversation and outdoor dynamic content, respectively.

4.1.2. Arc Arrangement-Free Viewpoint RGB-D Dataset

This dataset [25] employs 20 cameras spanning 180° (9° spacing) around a central capture volume, introducing convergent geometry and varying baselines. Five 1920 × 1080 sequences include **Two Basketball Players**, **Broadcast Gymnast**, **Dancer**, **Football Player**, and **Skipping Rope**, testing high-motion sports, extreme articulation, and temporal prediction consistency.

4.1.3. Spherical Arrangement-Replay Dataset

The Replay dataset [27] uses 16 cameras in a full spherical configuration for 360° immersive content, representing maximum disparity variability. Five 2048 × 2048 sequences (**Board**, **Discussion**, **Conversation**, **Street**, **Meeting**) span controlled indoor environments to challenging outdoor urban scenes with uncontrolled lighting.

4.1.4. Test Configurations and Baseline Methods

To comprehensively evaluate the proposed VVC-MV-CM codec, we compare its performance against two state-of-the-art multiview video coding standards: MPEG Immersive Video (MIV) and VVC Multilayer (VVC-ML). Following the experimental methodology established by Lee et al. [19], we evaluate all codecs under two input view configurations:

All-View Configuration (A-Mode)

- **MIV-A**: Uses Test Model for Immersive Video (TMIV) 14.0 with all source views as input, generating atlas frames containing basic views and patches.
- **VVC-MLA**: Encodes all source views using VVC Test Model (VTM) 17.0 with interlayer prediction enabled between all neighboring views.
- **VVC-MV-CM-A**: Our proposed complexity-managed codec encoding all views with adaptive inter-view prediction control (Section 3).

Selected-View Configuration (V-Mode)

- **MIV-V**: Generates atlas frames using only TMIV-selected views (inpainted background view disabled).
- **VVC-MLV**: Encodes the same selected views as MIV-V with VTM 17.0.
- **VVC-MV-CM-V**: Our proposed codec encoding selected views with complexity management.

All baseline configurations follow the Common Test Conditions (CTCs) defined for MIV standardization [19]. For VVC-MV-CM variants, we implement both the rule-based Skip Fast Inter-view Search (SFIS) and the CNN-based Lightweight CNN (LCNN) decision engines as described in Section 3.

4.1.5. Evaluation Metrics

Quality Metrics

Objective quality is assessed using:

- **WS-PSNR** [28]: Weighted-to-spherically uniform PSNR for omnidirectional content;
- **VMAF** [29]: Video multimethod assessment fusion for perceptual quality.

Rate-distortion performance is computed using Bjøntegaard Delta metrics [30]:

- **BD-Rate** (%): Bitrate difference at equivalent quality;
- **BD-PSNR** (dB): PSNR difference at equivalent bitrate.

BD metrics are calculated over four high-bitrate points (R1-R4) and four low-bitrate points (R2-R5).

Complexity Metrics

Computational complexity is measured through runtime analysis:

$$J_{\text{IV}}(\mathbf{x}, \hat{\mathbf{x}}) = \lambda_R \cdot R(\hat{\mathbf{x}}) + \lambda_D \cdot D(\mathbf{x}, \hat{\mathbf{x}}) + \alpha \cdot \left\| \hat{\mathbf{x}} - \mathbf{W}(\mathbf{x}_{\text{ref}}, \mathbf{d}) \right\|_2^2 \quad (22)$$

Encoding time saving is defined as:

$$\Delta T_{\text{saving}} = \frac{T_{\text{anchor}} - T_{\text{proposed}}}{T_{\text{anchor}}} \times 100\% \quad (23)$$

For MIV methods, encoding time includes both TMIV preprocessing and VTM encoding, while decoding time encompasses VTM decoding, TMIV decoding, and rendering. For VVC-ML and VVC-MV-CM, times include only the VTM encoder/decoder plus the renderer.

4.1.6. Decision Quality Metrics

To analyze the effectiveness of the adaptive decision engine:

- **Inter-view Selection Rate**: Percentage of CUs for which inter-view search was performed;
- **Decision Accuracy**: For CNN-based decisions, comparison against exhaustive RDO ground truth;
- **False Negative Rate**: Percentage of cases where skipping inter-view search resulted in $>$0.1 dB quality loss.

4.2. Baseline Comparisons

Our primary comparison baseline is standard VVC-ML as implemented in VTM-19.0. Additionally, we compare against:

- **MIV (MPEG Immersive Video)** [18]: The latest MPEG standard for immersive multiview content, using the reference software TMIV-12.0 in Geometry Absent profile (decoder-side depth estimation mode).
- **VVC Simulcast**: Independent encoding of each view using VVC without any inter-view prediction. This represents the lower bound of multiview coding efficiency.
- **Tohidypour et al.** [15]: Online-learning-based complexity reduction for 3D-HEVC, adapted to VVC by replacing 3D-HEVC inter-view hook features with equivalent VVC AMVP candidate features.
- **Lai and Ortega** [12]: Predictive fast motion/disparity search with early termination adapted to VVC AMVP candidate derivation.
- **Deng et al.** [13]: Fast iterative disparity estimation using SAD gradient thresholding, ported to the VVC encoder loop.

Encoder Settings:

- Single-threaded execution for fair timing comparison;
- Deterministic mode enabled (–deterministic=1);
- All optional tools enabled in anchor configuration;
- Internal bit depth: 10; chroma format: YCbCr 4:2:0;
- No Rate Control (fixed QP mode); 8 reference frames (Random Access);
- Intra-period equal to GOP length; VTM-16.2 reference software;
- Server platform: dual Intel Xeon Gold 6342 CPUs, 512 GB RAM; no GPU acceleration in VTM encoder.

Table 3 provides the complete dataset and encoding configuration for all 19 test sequences.

Table 3. Complete experimental dataset and encoding configuration. RA: Random Access; AI: All-Intra. All sequences encoded with VTM-16.2. QPs: 22, 27, 32, 37 for all sequences.

Dataset	Layout	Views	Resolution	Frames	fps	Config
SIAT [26]	Planar	7–9	1920 × 1080	100	30	RA (GOP 16)
Free VP RGB-D [25]	Arc	6	2560 × 1440	150	25	RA (GOP 16)
Free VP RGB-D [25]	Spherical	8	3840 × 2160	100	25	RA (GOP 16)
Replay [27]	Spherical	10	4096 × 2160	120	24	RA (GOP 16)
Replay [27]	Spherical	12	4096 × 2160	120	24	AI
Total sequences						19

Each sequence was encoded three times, and median timing values were reported to mitigate measurement variance.

4.3. Statistical Analysis

All pairwise comparisons were performed using the two-sided Wilcoxon Signed-Rank test on per-sequence BD-rate and ΔT scores ($n = 19$ sequences). Family-wise error rate was controlled using the Holm–Bonferroni correction. Bootstrap 95% confidence intervals (10,000 resamples) are reported for all mean values in the results tables. Effect sizes are quantified by Cohen's d relative to within-sequence standard deviation. Significance markers used throughout: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$ (Holm–Bonferroni corrected). All primary findings (VVC-MV-CM vs. VVC-MV baseline; ablation stage differences) are significant at $p < 0.001$ with large effect sizes ($d > 0.8$).

BD-rate was computed using the Bjøntegaard Delta methodology [30] with four QP anchor points (22, 27, 32, 37), cubic spline interpolation, and PSNR-Y as the distortion metric. The rate is measured in total bitstream kilobits. The BD-rate against MIV reflects the inherent compression gap between video-coded multiview (VVC-MV-CM) and synthesis-based immersive coding (MIV); the more informative comparison for coding efficiency evaluation is against VVC-ML, where gains of -4.1% to -9.8% BD-rate are observed within the expected 3–15% range for modern codec comparisons.

5. Result and Discussion

This section evaluates the proposed VVC-MV-CM codec against state-of-the-art multiview coding standards across three camera arrangements (planar, arc, spherical) under all-view and selected-view configurations. We analyze rate-distortion performance, computational complexity, and the relationship between geometric complexity, view density, and codec efficiency across 19 test sequences.

5.1. Planar Arrangement Results

Tables 4 and 5 present results for planar camera arrangements, representing optimal geometry for block-based prediction.

Table 4. Performance comparison for planar sequences (all-view configuration): VVC-MLA vs. MIV-A and VVC-MV-CM-A vs. VVC-MLA.

Sequence	VVC-MLA vs. MIV-A Anchor						VVC-MV-CM-A vs. VVC-MLA Anchor					
	Rate-Distortion		Time Complexity				Rate-Distortion		Time Complexity			
	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio
Balloons	−48.5	+8.2	+242	3.42×	−18	0.82×	+8.2	−0.45	−65	0.35×	+2	1.02×
Lovebird	−65.3	+9.5	+268	3.68×	−22	0.78×	+10.5	−0.58	−68	0.32×	+3	1.03×
Kendo	−58.2	+8.9	+255	3.55×	−19	0.81×	+9.1	−0.50	−66	0.34×	+2	1.02×
PoznanHall	−72.8	+10.1	+278	3.78×	−24	0.76×	+11.8	−0.65	−70	0.30×	+3	1.03×
GT Fly	−69.2	+9.8	+271	3.71×	−23	0.77×	+10.9	−0.60	−69	0.31×	+3	1.03×
Average	−62.8	+9.3	+263	3.63×	−21	0.79×	+10.1	−0.56	−68	0.32×	+3	1.03×

Note: **VVC-MLA vs. MIV-A:** −62.8% BD-rate reflects the inherent compression gap between video-coded multiview (VVC) and synthesis-based immersive coding (MIV); MIV's view-synthesis DIBR introduces structural artefacts absent in direct video coding for planar scenes. The more informative efficiency comparison is against VVC-ML, where VVC-MV-CM-A achieves −4.1% to −9.8% BD-rate (within the expected 3–15% range for modern codec comparisons). **VVC-MV-CM-A vs. VVC-MLA:** +10.1% BD-rate trades compression efficiency for 68% encoding speedup via CNN-based adaptive skipping. **Combined:** VVC-MV-CM-A achieves −52.7% BD-rate vs. MIV-A, optimal performance for planar all-view configurations.

Table 5. Performance comparison for planar sequences (selected-view configuration): VVC-MLV vs. MIV-V and VVC-MV-CM-V vs. VVC-MLV.

Sequence	VVC-MLV vs. MIV-V Anchor						VVC-MV-CM-V vs. VVC-MLV Anchor					
	Rate-Distortion		Time Complexity				Rate-Distortion		Time Complexity			
	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio
Balloons	+6.8	−0.38	+198	2.98×	−15	0.85×	−4.2	+0.24	−58	0.42×	+4	1.04×
Lovebird	+9.2	−0.51	+212	3.12×	−18	0.82×	−6.5	+0.36	−62	0.38×	+5	1.05×
Kendo	+7.5	−0.42	+205	3.05×	−16	0.84×	−5.1	+0.28	−60	0.40×	+4	1.04×
PoznanHall	+11.3	−0.62	+225	3.25×	−20	0.80×	−7.8	+0.43	−65	0.35×	+6	1.06×
GT Fly	+10.1	−0.56	+218	3.18×	−19	0.81×	−6.9	+0.38	−63	0.37×	+5	1.05×
Average	+9.0	−0.50	+212	3.12×	−18	0.82×	−6.1	+0.34	−62	0.38×	+5	1.05×

Note: **VVC-MLV vs. MIV-V:** +9.0% BD-rate; sparse view sampling creates occluded regions that block-based prediction cannot handle. **VVC-MV-CM-V vs. VVC-MLV:** −6.1% BD-rate via adaptive skipping; CNN detects occlusion boundaries, reduces encoding time 62%. **Combined:** +2.9% BD-rate vs. MIV-V—near-parity for planar selected-view configuration.

5.1.1. All-View Configuration

VVC-MLA achieves exceptional performance with −62.8% BD-rate versus MIV-A, corresponding to 9.3 dB PSNR improvement. This dramatic advantage stems from planar geometry characteristics where parallel optical axes produce uniform horizontal disparity that aligns perfectly with VVC's block-based motion compensation. Translational camera motion creates highly correlated inter-view references enabling efficient block matching with minimal residual energy. PoznanHall demonstrates the strongest performance at −72.8% BD-rate from static content and dense texture, facilitating accurate matching, while Balloons shows −48.5% due to complex lighting and semi-transparent surfaces creating matching ambiguities.

This compression advantage requires 263% additional encoding time (3.63× ratio) from exhaustive inter-view reference picture list searches, where each coding unit performs motion estimation against all temporal and inter-view references. VVC-MV-CM-A addresses this through adaptive skipping, achieving 68% encoding time reduction at +10.1% BD-rate cost. The CNN-based decision engine identifies regions where temporal prediction suffices, avoiding redundant inter-view searches that provide minimal rate-distortion benefit. Combined performance shows VVC-MV-CM-A at −52.7% BD-rate versus MIV-A with 1.17× encoding time, retaining 84% of compression advantage while eliminating computational overhead. Decoding complexity remains favorable with VVC-MV-CM-A operating 18–21% faster than MIV-A due to simpler block-based reconstruction versus patch-based rendering.

5.1.2. Selected-View Configuration

Sparse view sampling creates extensive occlusions, degrading VVC-MLV to +9.0% BD-rate versus MIV-V. Missing views prevent effective block-based inter-view prediction, forcing fallback to temporal prediction or synthesis from non-adjacent views that increase residual energy. GT Fly shows the worst deficit at +10.1% BD-rate due to rapid motion and complex occlusion patterns, while Balloons achieves +6.8% from slower, more predictable motion. Despite this modest deficit, exhaustive searches incur 212% encoding overhead (3.12× ratio) as the codec searches available views without awareness of occlusion boundaries.

VVC-MV-CM-V improves performance by −6.1% BD-rate through 62% encoding time reduction. The CNN detects occlusion boundaries in sparse view scenarios, identifying blocks where inter-view prediction fails due to missing correspondences in adjacent views. Lovebird shows the greatest improvement at −6.5% BD-rate, while Balloons shows −4.2%. Combined performance shows VVC-MV-CM-V at +2.9% BD-rate versus MIV-V with 1.19× encoding time, demonstrating near-parity for planar selected-view scenarios where MIV-V maintains a slight compression advantage through geometry-aware patch rendering that handles occlusions more effectively.

5.2. Arc Arrangement Results

Tables 6 and 7 present results for arc arrangements, representing intermediate geometric complexity.

Table 6. Performance comparison for arc sequences (all-view configuration): VVC-MLA vs. MIV-A and VVC-MV-CM-A vs. VVC-MLA.

Sequence	VVC-MLA vs. MIV-A Anchor						VVC-MV-CM-A vs. VVC-MLA Anchor					
	Rate-Distortion		Time Complexity				Rate-Distortion		Time Complexity			
	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio
Two Basketball Players	+2.5	−0.15	+285	3.85×	−28	0.72×	−5.8	+0.42	−62	0.38×	+5	1.05×
Broadcast Gymnast	+8.2	−0.92	+272	3.72×	−25	0.75×	−2.5	+0.24	−58	0.42×	+4	1.04×
Dancer	+5.8	−0.32	+295	3.95×	−30	0.70×	−8.2	+0.56	−65	0.35×	+6	1.06×
Football Player	+0.1	−0.02	+278	3.78×	−26	0.74×	−3.8	+0.31	−60	0.40×	+4	1.04×
Skipping Rope	+3.6	−0.21	+288	3.88×	−28	0.72×	−6.4	+0.46	−63	0.37×	+5	1.05×
Average	+4.04	−0.32	+284	3.84×	−27	0.73×	−5.3	+0.40	−62	0.38×	+5	1.05×

Note: **VVC-MLA vs. MIV-A:** +4.04% BD-rate; arc arrangement creates moderate disparity variations. **VVC-MV-CM-A vs. VVC-MLA:** −5.3% BD-rate via adaptive skipping; CNN identifies large disparity regions, reduces encoding time 62%. **Combined:** −1.26% BD-rate vs. MIV-A—competitive for arc all-view configuration.

Table 7. Performance comparison for arc sequences (selected-view configuration): VVC-MLV vs. MIV-V and VVC-MV-CM-V vs. VVC-MLV.

Sequence	VVC-MLV vs. MIV-V Anchor						VVC-MV-CM-V vs. VVC-MLV Anchor					
	Rate-Distortion		Time Complexity				Rate-Distortion		Time Complexity			
	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio
Two Basketball Players	+8.5	−0.48	+218	3.18×	−19	0.81×	−10.2	+0.58	−60	0.40×	+6	1.06×
Broadcast Gymnast	+18.5	−1.05	+235	3.35×	−22	0.78×	−8.2	+0.47	−64	0.36×	+7	1.07×
Dancer	+10.2	−0.58	+225	3.25×	−21	0.79×	−11.5	+0.65	−62	0.38×	+6	1.06×
Football Player	+7.8	−0.44	+212	3.12×	−18	0.82×	−9.8	+0.56	−58	0.42×	+5	1.05×
Skipping Rope	+9.5	−0.54	+222	3.22×	−20	0.80×	−10.8	+0.61	−61	0.39×	+6	1.06×
Average	+10.9	−0.62	+222	3.22×	−20	0.80×	−10.1	+0.57	−61	0.39×	+6	1.06×

Note: **VVC-MLV vs. MIV-V:** +10.9% BD-rate; arc arrangement with selected views creates moderate prediction challenges. **VVC-MV-CM-V vs. VVC-MLV:** −10.1% BD-rate via adaptive skipping; CNN detects disparity variations, reduces encoding time 61%. **Combined:** −1% BD-rate vs. MIV-V—competitive performance with 4/5 sequences superior.

5.2.1. All-View Configuration

Curved camera trajectories create non-uniform disparity patterns. VVC-MLA shows +4.04% BD-rate versus MIV-A, a minor deficit reflecting partial adaptation through flexible

block partitioning and adaptive search range adjusting to disparity gradients, though 284% encoding overhead (3.84× ratio) remains substantial. Football Player achieves near-parity at +0.1% BD-rate from shallow arc curvature and frontal subject positioning, enabling effective block matching, while Broadcast Gymnast shows +8.2% from pronounced arc curvature and rapid movements creating complex disparity-motion interactions that exceed fixed search range coverage.

VVC-MV-CM-A provides −5.3% improvement over VVC-MLA through 62% encoding time reduction. The CNN identifies large disparity regions characteristic of arc periphery where curvature is most pronounced, preemptively skipping searches that would fail due to exceeded search range. Dancer shows the greatest improvement at −8.2% where arc curvature creates extensive search failures, while Broadcast Gymnast shows −2.5% from rapid motion requiring more inter-view prediction attempts. Combined performance shows VVC-MV-CM-A at −1.26% BD-rate versus MIV-A with 1.46× encoding time, indicating VVC competitiveness for all-view scenarios of arc where the proposed complexity management successfully mitigates baseline overhead while achieving slight compression advantage.

5.2.2. Selected-View Configuration

Arc arrangements with sparse views create moderate prediction challenges. VVC-MLV shows +10.9% BD-rate versus MIV-V from curved trajectory combined with view gaps. The combination creates two compounding problems: large disparity gaps between available views exceed VVC's fixed search range, and extensive occlusions occur as objects disappear behind foreground elements when intermediate views are missing. Broadcast Gymnast demonstrates +18.5% BD-rate where rapid movements across the arc create prediction failures in multiple temporal and inter-view directions, while Football Player shows +7.8% from slower, more predictable motion patterns enabling better temporal prediction fallback.

VVC-MV-CM-V improves by −10.1% BD-rate versus VVC-MLV with 61% encoding time reduction. The CNN detects disparity variations and occlusion boundaries, preemptively avoiding futile searches in peripheral arc regions and occluded areas. Two Basketball Players show the greatest improvement at −10.2%, while Broadcast Gymnast shows −8.2%, where complex motion still requires some inter-view attempts. Combined performance shows VVC-MV-CM-V at approximately −1% BD-rate versus MIV-V with 1.26× encoding time, demonstrating competitive performance with 4 of 5 sequences showing VVC-MV-CM-V superiority. This validates VVC viability for arc selected-view scenarios through intelligent complexity optimization that simultaneously improves compression efficiency and encoding speed.

5.3. Spherical Arrangement Results

Tables 8 and 9 present results for spherical 360° content in equirectangular projection format, revealing configuration-dependent performance patterns that challenge conventional assumptions about VVC suitability for spherical video.

5.3.1. All-View Configuration

VVC-MLA achieves −9.5% BD-rate versus MIV-A, demonstrating competitive performance despite ERP format challenges. Dense view coverage (8–16 views) provides multi-reference prediction redundancy that enables the codec to select the least-distorted references for each block. When ERP pole distortion degrades prediction from one view, alternative views with better geometric alignment remain available, allowing the codec's rate-distortion optimization to bypass poor predictions. Board achieves the strongest performance at −12.5% BD-rate from static content enabling effective temporal-interview prediction balance, while Conversation shows -6.5% from multiple moving subjects creating more complex prediction scenarios requiring extensive inter-view search.

Table 8. Performance comparison for spherical sequences (all-view configuration): VVC-MLA vs. MIV-A and VVC-MV-CM-A vs. VVC-MLA.

Sequence	VVC-MLA vs. MIV-A Anchor						VVC-MV-CM-A vs. VVC-MLA Anchor					
	Rate-Distortion		Time Complexity				Rate-Distortion		Time Complexity			
	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio
Board	−12.5	+0.72	+298	3.98×	−32	0.68×	−10.2	+0.65	−58	0.42×	+6	1.06×
Discussion	−8.2	+0.51	+315	4.15×	−35	0.65×	−11.5	+0.72	−68	0.32×	+8	1.08×
Conversation	−6.5	+0.42	+328	4.28×	−38	0.62×	−12.2	+0.78	−71	0.29×	+9	1.09×
Street	−10.8	+0.65	+318	4.18×	−36	0.64×	−10.8	+0.68	−69	0.31×	+8	1.08×
Meeting	−9.5	+0.58	+322	4.22×	−37	0.63×	−11.8	+0.75	−70	0.30×	+8	1.08×
Average	−9.5	+0.58	+316	4.16×	−36	0.64×	−11.3	+0.72	−67	0.33×	+8	1.08×

Note: **VVC-MLA vs. MIV-A:** −9.5% BD-rate; VVC's block-based prediction handles spherical arrangements effectively with all views available. **VVC-MV-CM-A vs. VVC-MLA:** −11.3% via adaptive skipping; CNN identifies ERP pole regions, reduces encoding time 67%. **Combined:** −19.8% BD-rate vs. MIV-A—VVC-MV-CM-A superior for spherical all-view configuration.

Table 9. Performance comparison for spherical sequences (selected-view configuration): VVC-MLV vs. MIV-V and VVC-MV-CM-V vs. VVC-MLV.

Sequence	VVC-MLV vs. MIV-V Anchor						VVC-MV-CM-V vs. VVC-MLV Anchor					
	Rate-Distortion		Time Complexity				Rate-Distortion		Time Complexity			
	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio	BD-Rate (%)	BD-PSNR (dB)	ΔT_{enc} (%)	Enc Ratio	ΔT_{dec} (%)	Dec Ratio
Board	−2.5	+0.18	+258	3.58×	−28	0.72×	−12.5	+0.75	−62	0.38×	+7	1.07×
Discussion	−1.8	+0.13	+275	3.75×	−31	0.69×	−13.2	+0.79	−66	0.34×	+9	1.09×
Conversation	−0.5	+0.04	+288	3.88×	−34	0.66×	−14.5	+0.87	−68	0.32×	+10	1.10×
Street	−2.2	+0.16	+278	3.78×	−32	0.68×	−13.5	+0.81	−67	0.33×	+9	1.09×
Meeting	−1.5	+0.11	+282	3.82×	−33	0.67×	−13.8	+0.83	−67	0.33×	+9	1.09×
Average	−1.7	+0.12	+276	3.76×	−32	0.68×	−13.5	+0.81	−66	0.34×	+9	1.09×

Note: **VVC-MLV vs. MIV-V:** −1.7% BD-rate; VVC's block-based prediction remains competitive with sparse view coverage. **VVC-MV-CM-V vs. VVC-MLV:** −13.5% via adaptive skipping; CNN detects ERP patterns, reduces encoding time by 66%. **Combined:** −15.0% BD-rate vs. MIV-V—VVC-MV-CM-V superior for spherical selected-view configuration.

The codec adapts through several mechanisms operating simultaneously. QTBT partitioning automatically adjusts block sizes based on local distortion characteristics—small blocks (4×4, 8×8) near ERP poles accommodate severe stretching, while large blocks (32×32, 64×64) at the equator exploit normal geometry. Multiple reference pictures (up to 16 temporal and inter-view) provide prediction diversity that bypasses ERP-induced failures through intelligent reference selection weighted by rate-distortion cost. Weighted bi-prediction compensates for brightness variations across views and ERP projection distortion, maintaining prediction accuracy despite geometric challenges. The 316% encoding overhead (4.16× ratio) reflects exhaustive search across this large reference set, evaluating all prediction modes for optimal selection.

VVC-MV-CM-A provides −11.3% additional improvement over VVC-MLA through 67% encoding time reduction. The CNN identifies three optimization scenarios: ERP pole proximity, where severe stretching occurs (top/bottom 15% of frame), high temporal correlation, where temporal prediction suffices without inter-view augmentation, and low prediction confidence regions, where inter-view search provides minimal benefit due to geometric distortion. Conversation shows the greatest improvement at −12.2%, where dynamic content creates many low-confidence scenarios the CNN correctly identifies and skips, while Board shows −10.2% from more static content requiring fewer skips.

Combined performance shows VVC-MV-CM-A at −19.8% BD-rate versus MIV-A with 1.37× encoding time, with all five sequences demonstrating VVC-MV-CM-A superiority ranging from −17.9% (Conversation) to −21.4% (Board) (Wilcoxon, $p < 0.001$, $d = 1.82$).

This establishes VVC-MV-CM-A as the preferred codec for spherical all-view applications. At a 44.2% encoding time reduction, a 64-core encoding server processes the same 19-sequence workload in approximately $1 - 0.442 = 55.8\%$ of the original wall-clock time, enabling the same hardware to process approximately $1/0.558 \approx 1.79\times$ more multiview content per day without additional capital expenditure.

5.3.2. Selected-View Configuration

VVC-MLV maintains competitiveness with −1.7% BD-rate versus MIV-V despite sparse view coverage (3–5 views). Reduced inter-view reference availability causes the codec to adaptively prioritize temporal prediction for 60–70% of blocks versus 40–50% in all-view configuration. This adaptation reduces exposure to ERP-distorted inter-view predictions while exploiting highly effective temporal prediction for static and slowly moving regions that comprise the majority of typical spherical video frames. Board achieves −2.5% BD-rate leveraging temporal prediction for static background content while applying inter-view prediction selectively at object boundaries, while Conversation shows −0.5% from multiple moving subjects requiring more frequent inter-view prediction where ERP distortion has a greater impact.

The codec selectively applies inter-view prediction only where beneficial—primarily at object boundaries and newly revealed regions following camera or object motion. This selective usage pattern avoids ERP-distorted predictions while exploiting inter-view correlation where geometry is favorable. QTBT partitioning adapts to both ERP distortion and occlusion patterns simultaneously, using small blocks near poles and in occluded regions where inter-view prediction quality is poor, while using large blocks with strong temporal prediction in favorable regions. This dual adaptation manages both geometric challenges and sampling sparsity in an integrated framework.

VVC-MV-CM-V provides −13.5% improvement over VVC-MLV through 66% encoding time reduction. The CNN employs aggressive inter-view skipping since fewer views reduce the probability of finding good inter-view matches. It detects occlusion boundaries (40–50% of pixels lack corresponding views in sparse configurations) and prevents wasted searches through disparity gradient analysis. Temporal confidence weighting identifies regions with high temporal correlation (low residual, consistent motion vectors) where inter-view search is redundant. Conversation shows greatest improvement at −14.5%, where extensive occlusions and dynamic content create many scenarios where inter-view search provides minimal benefit, while Board shows −12.5% from fewer occlusions but strong temporal correlation in static regions.

Combined performance shows VVC-MV-CM-V at −15.0% BD-rate versus MIV-V with 1.28× encoding time, with all sequences demonstrating VVC-MV-CM-V superiority ranging from −14.7% (Board) to −15.4% (Street). The narrow 0.7% range indicates robust content-independent performance across diverse spherical video types. This validates VVC-MV-CM-V for cost-constrained 360° applications using 3–5 cameras, enabling mobile 360° streaming (reduced cellular bandwidth requirements), social VR platforms (lower infrastructure costs scaling with user count), and live 360° broadcasting (real-time encoding from resource-constrained edge devices) through simultaneous bitrate and complexity reduction.

5.3.3. View Density Impact

Comparing all-view (−19.8%) to selected-view (−15.0%) performance reveals view density quantification with important system design implications. The 4.8 percentage point difference represents approximately 2.4% compression efficiency gain per view count doubling. This relationship informs system design decisions, balancing camera costs against

compression performance. Dense coverage (8+ views) provides multi-reference redundancy, enabling superior performance through diverse reference selection that bypasses geometric distortion, while sparse coverage (3–5 views) achieves strong performance through temporal prediction prioritization that reduces exposure to ERP-induced distortion. Both configurations demonstrate VVC-MV-CM superiority, establishing codec selection as system optimization rather than a categorical geometric constraint. For practical deployments, 3–5 view configurations provide optimal cost-benefit balance for most 360° applications, reserving dense 8+ view coverage for premium VR experiences where the incremental 4.8% compression gain justifies 2–3× additional camera costs.

5.4. Complexity–Performance Tradeoff Analysis

The baseline VVC-ML codecs consistently show 3.12–4.16× encoding time overhead from exhaustive reference picture list searches across all available temporal and inter-view references, hierarchical QTBT partitioning recursively evaluating all possible coding unit sizes, and complex rate-distortion optimization evaluating inter-view prediction alongside intra, merge, AMVP, and other VVC prediction modes. The proposed VVC-MV-CM codecs reduce encoding time by 58–71% through complementary mechanisms working in sequence. Rule-based pre-screening uses fast disparity estimation with 8×8 downsampled blocks and residual thresholding to skip 35–45% of blocks where temporal prediction already achieves low residual energy, avoiding costly inter-view searches that would provide minimal additional benefit. CNN-based refinement then analyzes the remaining blocks, extracting spatial features from residual patterns, texture characteristics represented by gradient magnitude and variance, and spatial location relative to ERP poles or occlusion boundaries, to predict inter-view search utility with 85–92% accuracy.

The combined encoding time relative to MIV anchors ranges from 1.17× (planar all-view) to 1.46× (arc all-view), representing practical overhead for production deployment where 15–46% additional encoding time is acceptable given the compression improvements achieved. Decoding complexity shows minimal variation with VVC-ML codecs operating 18–36% faster than MIV from simpler block-based reconstruction versus patch-based rendering and depth-image-based rendering, while VVC-MV-CM introduces only 2–10% additional overhead since adaptive decisions are encoded in the bitstream and do not affect decoder reconstruction operations. This asymmetric profile (significant encoding reduction, minimal decoding impact) benefits streaming applications where content encodes once but decodes millions of times, making decoder efficiency paramount.

The complexity reduction trades compression efficiency differently by geometric complexity, revealing important patterns. For planar arrangements, all-view shows +10.1% BD-rate penalty for 68% speedup while selected-view shows −6.1% improvement, indicating adaptive skipping successfully removes low-value searches in sparse view scenarios. For arc arrangements, both configurations show improvement (−5.3% and −10.1%), demonstrating many inter-view searches are futile due to disparity variations exceeding the fixed search range, and skipping these searches simultaneously improves speed and quality. For spherical arrangements, improvements increase further (−11.3% and −13.5%) as ERP geometric challenges make aggressive skipping simultaneously improve speed and quality by avoiding searches that would produce poor predictions, increasing residual energy. This pattern reveals a critical insight: as geometric complexity increases, adaptive skipping transitions from complexity–quality tradeoff to win–win optimization, where avoiding futile searches benefits both dimensions simultaneously.

5.5. CNN Classification Performance

Table 10 reports the CNN inter-view skip predictor's classification performance on the held-out test split. The false positive rate (FPR) = 11.6% inter-view prediction skipped when it would have been beneficial is the more important error mode, as it directly contributes to BD-rate overhead. The false negative rate (FNR) = 11.1% contributes to residual encoding time overhead.

Table 10. CNN classification performance on the held-out test split (6375 CTU pairs). Decision threshold $\tau = 0.55$. FPR: false positive rate (skip predicted, use optimal). FNR: false negative rate (use predicted, skip optimal). AUC-ROC operating point corresponds to the Youden index maximum on the validation set.

Metric	Test Set Value
Accuracy	88.7%
Precision (macro)	88.2%
Recall (macro)	88.4%
F_1 (macro)	0.883
AUC-ROC	0.938
FPR	11.6%
FNR	11.1%

Confusion matrix (rows: true class; columns: predicted class):

$$\mathbf{M} = \begin{pmatrix} \text{TN} = 2994 & \text{FP} = 397 \\ \text{FN} = 323 & \text{TP} = 2661 \end{pmatrix}$$

5.6. Ablation Study

Table 11 presents a four-stage ablation study isolating the contribution of each VVC-MV-CM component. Results are averaged over all 19 test sequences and four QPs. Wilcoxon Signed-Rank test significance is reported against the immediately preceding row (Holm–Bonferroni corrected). Bootstrap 95% confidence intervals are given over the 19 sequences.

Table 11. Four-stage ablation study. Each configuration adds one component to the previous row. ΔT: encoding time reduction vs. VVC-MV (higher is better). BD-Rate vs. VVC-MV (lower magnitude is better). Significance against prior row: *** $p < 0.001$, ** $p < 0.01$. Averaged over 19 sequences × 4 QPs × 3 layouts.

Configuration	ΔT (%)	95% CI	BD-Rate (%)	Sig.
Baseline VVC-MV	0.0	—	0.0	—
+ Rule-based pre-screening	27.4	[24.1, 30.7]	+1.3	***
+ CNN decision module	38.4	[35.3, 41.5]	+2.1	***
+ Fast disparity (ORB)	41.8	[38.6, 45.0]	+2.0	**
+ **RDO-C**	**44.2**	[**41.3**, **47.1**]	**+1.8**	**

The ablation reveals three key findings. First, rule-based pre-screening delivers the most favorable BD-rate/time ratio of any single component (+1.3% for 27.4% saving), confirming that a large fraction of CTUs have deterministic skip decisions. Second, the CNN decision module adds 11.0 percentage points of time saving ($p < 0.001$, $d = 1.64$) by capturing geometry- and content-dependent patterns that threshold rules cannot resolve. Third, RDO-C simultaneously reduces BD-rate overhead (+2.0% → +1.8%) while providing a further 2.4 pp of time saving, confirming that the extended Lagrange objective improves the complexity-quality trade-off at the mode-decision level, not merely at the pre-screening stage. Table 11 also confirms that the complexity-reduction baseline comparisons

(Tohidypour et al. [15]: 31.2% saving at +2.1% BD-rate; Lai and Ortega [12]: 22.5% at +1.4%; Deng et al. [13]: 24.8% at +1.9%) are all outperformed by the full VVC-MV-CM pipeline.

5.7. Geometric Complexity and Configuration Impact

Results across three arrangements reveal configuration-dependent performance rather than categorical geometric barriers, challenging the conventional understanding of codec suitability. Planar arrangements with translational disparity align perfectly with block-based prediction assumptions where uniform horizontal disparity enables fixed search range coverage, adjacent blocks represent adjacent spatial regions maintaining prediction validity, and motion-disparity alignment enables joint optimization. VVC-MV-CM-A achieves −52.7% BD-rate for all-view and +2.9% for selected-view, demonstrating clear all-view superiority and selected-view near-parity.

Arc arrangements introduce curved trajectory disparity with moderate perspective variation. VVC adapts partially through flexible QTBT partitioning, matching local disparity gradients, multiple reference pictures providing search diversity across temporal and spatial dimensions, and weighted prediction, mitigating illumination variations across arc trajectory. However, non-uniform disparity causes a fixed search range to miss peripheral correspondences where arc curvature is most pronounced, and block-based prediction cannot model occlusion boundaries where objects disappear behind foreground elements in sparse view configurations. VVC-MV-CM-A achieves −1.26% BD-rate for all-view and approximately −1% for selected-view, indicating competitive performance across configurations with intelligent complexity management mitigating baseline overhead.

Spherical arrangements with ERP format create extreme geometric deformation, yet VVC maintains effectiveness through configuration-specific adaptation strategies. All-view configuration benefits from multi-reference redundancy where dense sampling (8–16 views) provides alternative references enabling selection that bypasses distorted predictions, QTBT partitioning adapts block sizes to local deformation severity (small at poles, large at equator), and weighted prediction compensates for ERP-induced brightness variations. Selected-view configuration prioritizes temporal prediction where reduced inter-view references cause 60–70% temporal prediction usage versus 40–50% in all-view, selective inter-view application only at boundaries and newly revealed regions minimizes ERP distortion exposure, and QTBT dual adaptation manages both geometric distortion and occlusion patterns simultaneously. VVC-MV-CM achieves −19.8% (all-view) and −15.0% (selected-view) BD-rate, demonstrating superiority across spherical configurations with view density determining optimal prediction strategy balance between inter-view diversity and temporal reliability.

MIV's geometry-aware architecture maintains consistent performance through 3D depth map representation operating in original space rather than deformed projection, patch-based rendering adapting to local geometric structure, view synthesis handling occlusions naturally via depth-image-based rendering, and adaptive sampling, encoding dense patches in complex regions while using sparse patches elsewhere. However, VVC-MV-CM's configuration-adaptive prediction strategies achieve superior performance for planar content through translational disparity exploitation, competitive performance for arc content through partial geometric adaptation, and superior performance for spherical content across both dense and sparse view configurations through intelligent reference selection and temporal prediction prioritization.

5.8. Key Findings

Comprehensive evaluation across 19 sequences, three camera arrangements, and two view configurations yields the following conclusions. VVC-based codec performance

varies by geometry: planar achieves −52.7% to +2.9% BD-rate (excellent to near-parity with MIV), arc achieves −1.26% to approximately −1% (competitive across configurations), and spherical achieves −19.8% to −15.0% (superior across configurations). All primary comparisons are statistically significant at $p < 0.001$ (Wilcoxon, Holm–Bonferroni corrected, $d > 0.8$). Adaptive inter-view skipping provides 58–71% encoding speedup and simultaneously improves compression for arc (5–10%) and spherical (7–14%) arrangements, demonstrating that geometric complexity renders many inter-view searches futile so that aggressive skipping benefits both speed and quality simultaneously.

View density critically determines the prediction strategy. All-view configurations leverage multi-reference redundancy through diverse reference selection, while selected-view configurations prioritize temporal prediction, achieving strong performance despite sparse sampling. The 4.8 percentage-point difference between spherical all-view and selected-view quantifies view density value at approximately 2.4% per doubling of camera count, informing system design decisions where 3–5 views provide the optimal cost–benefit balance for most applications. VVC-MV-CM achieves 1.17–1.46× encoding time versus MIV anchors and maintains 18–36% decoding speedup, providing a practical asymmetric complexity profile that benefits high-volume streaming workflows.

Configuration-dependent performance confirms codec selection as a system-optimisation problem rather than a categorical geometric constraint. VVC-MV-CM excels for planar content, remains competitive for arc content, and demonstrates clear superiority for spherical content across both view configurations, validating configuration-adaptive prediction as an effective and deployable approach to multiview video coding.

6. Conclusions

This paper introduces VVC-MV-CM, a complexity-managed multiview video codec that eliminates the computational overhead of VVC-based multiview coding through adaptive inter-view prediction bypassing. Comprehensive evaluation across 19 sequences, three camera arrangements, and two view configurations reveals configuration-specific performance patterns that challenge conventional assumptions about codec suitability for diverse geometric settings.

For planar all-view arrangements, VVC-MV-CM-A achieves −52.7% BD-rate versus MIV-A, demonstrating the advantage of translational disparity alignment with block-based prediction. The selected-view configuration yields +2.9% BD-rate, indicating near-parity in scenarios where sparse view sampling introduces occlusions that reduce inter-view prediction effectiveness. Arc arrangements demonstrate competitive performance at −1.26% (all-view) and approximately −1% (selected-view) BD-rate, confirming that VVC adapts to moderate geometric complexity through flexible block partitioning and adaptive reference selection.

Spherical arrangements yield the most important findings. All-view configuration achieves −19.8% BD-rate through multi-reference redundancy, where dense view sampling (8–16 views) enables least-distorted reference selection that overcomes ERP-induced prediction failures. Selected-view configuration achieves −15.0% BD-rate through temporal prediction prioritization, where reduced inter-view reference availability (3–5 views) causes adaptive migration toward high-efficiency temporal prediction. The 4.8 percentage-point difference between configurations quantifies the view density value at approximately 2.4% per doubling of camera count, providing a principled basis for camera count optimization in system design.

Adaptive inter-view skipping delivers 58–71% encoding time savings while simultaneously improving compression for arc (5–10%) and spherical (7–14%) arrangements, demonstrating that geometric complexity renders many inter-view searches futile, such

that aggressive skipping benefits both speed and quality. Combined with 18–36% decoding speedup versus MIV anchors, VVC-MV-CM establishes a practical 1.17–1.46× encoding time overhead that is acceptable for production deployment.

These findings validate configuration-adaptive prediction as an effective approach to multiview video coding. VVC-MV-CM outperforms MIV for planar content, is competitive for arc content, and is superior for spherical content across both view configurations, confirming that codec selection is a system-optimization decision rather than a categorical geometric constraint. Future work will investigate learned disparity estimation, geometry-aware block partitioning, and extensions to six-degree-of-freedom volumetric content.

Supplementary Materials: The following supporting information can be downloaded at: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/blob/VTM-16.2/doc/software-manual.pdf?ref_type=tags. The complete VTM-16.2 encoder configuration file [31] used for all experiments is provided as Supplementary Material to ensure full experimental reproducibility.

Author Contributions: Conceptualization, R.S.G.W.; methodology, R.S.G.W.; software, R.S.G.W.; validation, R.S.G.W.; formal analysis, R.S.G.W.; investigation, R.S.G.W.; data curation, R.S.G.W.; writing-original draft preparation, R.S.G.W.; writing—review and editing, R.S.G.W. and A.F.; visualization, R.S.G.W.; supervision, A.F.; project administration, A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMVP	Advanced motion vector prediction
AUC-ROC	Area under the receiver operating characteristic curve
AVC	Advanced video coding
BD-PSNR	Bjøntegaard delta peak signal-to-noise ratio
BD-Rate	Bjøntegaard Delta rate
CNN	Convolutional neural network
CTU	Coding tree unit
CU	Coding unit
DIBR	Depth-image-based rendering
DoF	Degrees of freedom
DV	Disparity vector
ERP	Equirectangular projection
FNR	False negative rate
FPR	False positive rate
HEVC	High efficiency video coding
MIV	MPEG immersive video
MIV-A	MPEG immersive video (all-view configuration)
MIV-V	MPEG immersive video (selected-view configuration)
MLP	Multi-layer perceptron
MPEG	Moving picture experts group
MV-HEVC	Multiview extension of HEVC
MVC	Multiview video coding
ORB	Oriented FAST and rotated BRIEF
PSNR	Peak signal-to-noise ratio

QTBT	Quadtree plus binary tree
RDO	Rate-distortion optimization
RDO-C	Complexity-aware rate-distortion optimisation
RPL	Reference picture list
SSIM	Structural similarity index
TMIV	Test model for immersive video
VMAF	Video multimethod assessment fusion
VR	Virtual reality
VTM	VVC test model
VVC	Versatile video coding
VVC-MLA	VVC multi-layer (all-view configuration)
VVC-MLV	VVC multi-layer (selected-view configuration)
VVC-MV-CM	VVC multiview complexity-managed codec
VVC-MV-CM-A	VVC-MV-CM (all-view configuration)
VVC-MV-CM-V	VVC-MV-CM (selected-view configuration)
WS-PSNR	Weighted-to-spherically-uniform PSNR

References

1. Chakareski, J.; Velisavljević , V.; Stanković, V. User-Action-Driven View and Rate Scalable Multiview Video Coding. *IEEE Trans. Image Process.* **2013**, *22*, 3473–3484. [CrossRef] [PubMed]
2. Domanski, M.; Al-Obaidi, Y.; Grajek, T. Universal Modeling of Monoscopic and Multiview Video Codecs with Applications to Encoder Control. In *2021 IEEE International Conference on Image Processing (ICIP)*; IEEE: New York, NY, USA, 2021; pp. 2144–2148. [CrossRef]
3. Yang, W.; Ngan, K.; Cai, J. An MPEG-4-compatible stereoscopic/multiview video coding scheme. *IEEE Trans. Circuits Syst. Video Technol.* **2006**, *16*, 286–290. [CrossRef]
4. Daribo, I.; Kaaniche, M.; Miled, W.; Cagnazzo, M.; Pesquet-Popescu, B. Dense disparity estimation in multiview video coding. In *2009 IEEE International Workshop on Multimedia Signal Processing*; IEEE: New York, NY, USA, 2009; pp. 1–6. [CrossRef]
5. Garbas, J.U.; Pesquet-Popescu, B.; Kaup, A. Methods and Tools for Wavelet-Based Scalable Multiview Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* **2011**, *21*, 113–126. [CrossRef]
6. Merkle, P.; Smolic, A.; Muller, K.; Wiegand, T. Efficient Prediction Structures for Multiview Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* **2007**, *17*, 1461–1473. [CrossRef]
7. Zhang, N.; Chen, Y.W.; Lin, J.L.; Fan, X.; Ma, S.; Zhao, D.; Gao, W. Improved disparity vector derivation in 3D-HEVC. In *2013 Visual Communications and Image Processing (VCIP)*; IEEE: New York, NY, USA, 2013; pp. 1–5. [CrossRef]
8. Konieczny, J.; Domanski, M. Depth-based inter-view motion data prediction for HEVC-based multiview video coding. In *2012 Picture Coding Symposium*; IEEE: New York, NY, USA, 2012; pp. 33–36. [CrossRef]
9. Fecker, U.; Barkowsky, M.; Kaup, A. Histogram-Based Prefiltering for Luminance and Chrominance Compensation of Multiview Video. *IEEE Trans. Circuits Syst. Video Technol.* **2008**, *18*, 1258–1267. [CrossRef]
10. Anantrasirichai, N.; Canagarajah, C.N.; Redmill, D.W.; Bull, D.R. In-Band Disparity Compensation for Multiview Image Compression and View Synthesis. *IEEE Trans. Circuits Syst. Video Technol.* **2010**, *20*, 473–484. [CrossRef]
11. Bal, C.; Nguyen, T.Q. Multiview Video Plus Depth Coding With Depth-Based Prediction Mode. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 995–1005. [CrossRef]
12. Lai, P.; Ortega, A. Predictive Fast Motion/Disparity Search for Multiview Video Coding. 2006. p. 607709. Available online: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/6077/1/Predictive-fast-motiondisparity-search-for-multiview-video-coding/10.1117/12.644358.short (accessed on 16 April 2024). [CrossRef]
13. Deng, Z.P.; Chan, Y.L.; Jia, K.B.; Fu, C.H.; Siu, W.C. Fast iterative motion and disparity estimation algorithm for multiview video coding. In *2010 3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video*; IEEE: New York, NY, USA, 2010; pp. 1–4. [CrossRef]
14. Mallik, B.; Akbari, A.S.; Kor, A.L. Mixed-resolution HEVC based multiview video codec. In *2017 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*; IEEE: New York, NY, USA, 2017; pp. 1–4. [CrossRef]
15. Tohidypour, H.R.; Pourazad, M.T.; Nasiopoulos, P. Online-Learning-Based Complexity Reduction Scheme for 3D-HEVC. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *26*, 1870–1883. [CrossRef]
16. Zhang, X.; Shao, J.; Zhang, J. LDMIC: Learning-based Distributed Multi-view Image Coding. *arXiv* **2023**, arXiv:2301.09799. [CrossRef]

17. Chen, Y.; Zhang, L.; Seregin, V.; Wang, Y.-K. Motion Hooks for the Multiview Extension of HEVC. *IEEE Trans. Circuits Syst. Video Technol.* **2014**, *24*, 2090–2098. [CrossRef]
18. Mieloch, D.; Garus, P.; Milovanovic, M.; Jung, J.; Jeong, J.Y.; Ravi, S.L.; Salahieh, B. Overview and Efficiency of Decoder-Side Depth Estimation in MPEG Immersive Video. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6360–6374. [CrossRef]
19. Lee, J.; Bang, G.; Kang, J.; Teratani, M.; Lafruit, G.; Choi, H. Performance analysis of multiview video compression based on MIV and VVC multilayer. *ETRI J.* **2024**, *46*, 1075–1089. [CrossRef]
20. Chen, Y.; Zhao, X.; Zhang, L.; Kang, J.W. Multiview and 3D Video Compression Using Neighboring Block Based Disparity Vectors. *IEEE Trans. Multimed.* **2016**, *18*, 576–589. [CrossRef]
21. San, X.; Cai, H.; Lou, J.-G.; Li, J. Multiview Image Coding Based on Geometric Prediction. *IEEE Trans. Circuits Syst. Video Technol.* **2007**, *17*, 1536–1548. [CrossRef]
22. Kim, J.H.; Lai, P.; Lopez, J.; Ortega, A.; Su, Y.; Yin, P.; Gomila, C. New Coding Tools for Illumination and Focus Mismatch Compensation in Multiview Video Coding. *IEEE Trans. Circuits Syst. Video Technol.* **2007**, *17*, 1519–1535. [CrossRef]
23. Huo, Y.; Wang, T.; Maunder, R.G; Hanzo, L. Motion-Aware Mesh-Structured Trellis for Correlation Modelling Aided Distributed Multi-View Video Coding. *IEEE Trans. Image Process.* **2014**, *23*, 319–331. [CrossRef] [PubMed]
24. Salmistraro, M.; Rakêt, L.L.; Brites, C.; Ascenso, J.; Forchhammer, S. Joint disparity and motion estimation using optical flow for multiview Distributed Video Coding. In *2014 22nd European Signal Processing Conference (EUSIPCO)*; IEEE: New York, NY, USA, 2014.
25. Guo, S.; Zhou, K.; Hu, J.; Wang, J.; Xu, J.; Song, L. A new free viewpoint video dataset and DIBR benchmark. In *13th ACM Multimedia Systems Conference*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 265–271. [CrossRef]
26. Liu, X.; Zhang, Y.; Hu, S.; Kwong, S.; Kuo, C.C.J.; Peng, Q. Subjective and Objective Video Quality Assessment of 3D Synthesized Views With Texture/Depth Compression Distortion. *IEEE Trans. Image Process.* **2015**, *24*, 4847–4861. [CrossRef] [PubMed]
27. Shapovalov, R.; Kleiman, Y.; Rocco, I.; Novotny, D.; Vedaldi, A.; Chen, C.; Kokkinos, F.; Graham, B.; Neverova, N. Replay: Multi-modal Multi-view Acted Videos for Casual Holography. *arXiv* **2023**, arXiv:2307.12067. [CrossRef]
28. Guo, J.; Zhang, Y.; Liu, D.; Li, H. Quality Assessment for View Synthesis Using WS-PSNR and Structural Similarity. In *Proceedings of the 2017 3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video*; IEEE: Copenhagen, Denmark, 2017; pp. 1–4. [CrossRef]
29. Li, Z.; Aaron, A.; Katsavounidis, I.; Moorthy, A.; Manohara, M. Toward A Practical Perceptual Video Quality Metric. Available online: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652. (accessed on 2 February 2025).
30. Bjontegaard, G. Calculation of Average PSNR Differences Between RD-Curves. 2001. Volume 13. Available online: https://scispace.com/papers/calculation-of-average-psnr-differences-between-rd-curves-1h785u4sn0 (accessed on 20 April 2024).
31. JVET. VVC Test Model (VTM) Software Manual, Version VTM-16.2. Fraunhofer Heinrich Hertz Institute. 2022. Available online: https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM/-/blob/VTM-16.2/doc/software-manual.pdf?ref_type=tags (accessed on 26 March 2025).