Commutator Norm as a Continuous Criterion for GOP Safety and Perceptual Quality Redistribution in Video Encoding

Hiroshi Sasaki

***Abstract*—Existing VOD platforms limit GOP to roughly 120–150 frames not because longer GOPs lack compression efficiency, but because quality drift at scene boundaries is not formally bounded in any standard-compliant codec. We introduce the commutator norm $d(A, B) = \|AB - BA\|_F$ as the first continuous, codec-orthogonal safety criterion that formally characterizes this risk. Corollary 1 converts scene-cut drift from an unbounded hazard into a mathematically bounded one, enabling principled GOP extension in H.264, VP9, and AV1 without modifying the encoding syntax. Using the same criterion, we further show that perceptual QP redistribution is GOP-independent: it achieves -52 to -65% file-size reduction at VMAF 83–88 (H.264, -1.62% VMAF-based BD-rate at 1080p) even at short, fixed GOPs used in HLS/MPEG-DASH authoring and live ingest. The previously known -14.2% BD-rate benefit of long-GOP encoding is preserved ($+0.92\%$ safety cost, 100% scene-cut detection within ± 2 frames on multi-shot content). On the decode side, a 54 KB WebAssembly implementation delivers real-time 4K/30fps AV1 playback on a 2-core CPU without GPU assistance, driven by the same block-level criterion.**

***Index Terms*—video encoding, commutator norm, GOP optimization, operator commutativity, rate-distortion theory, AV1, H.264, VP9, near-commutativity, NormMAP**

I. Introduction

Standard video codecs—H.264/AVC, MPEG-2/XDCAM, VP9, AV1—control GOP length through signal-magnitude heuristics [1], [2]. I-frames are inserted when pixel differences, SAD, or DCT-DC coefficients exceed a threshold. These heuristics answer "how much do frames differ?" but leave a deeper question unanswered: "does the processing order matter?"

NormMAP is a codec-orthogonal optimizer. It operates as an upper-layer optimizer above any standard-compliant encoder without modifying the encoding syntax. The comparison structure is therefore:

- H.264+NormMAP vs. H.264 (same codec, same CRF)
- *Not*: SlimeCodec vs. H.265 (codec vs. codec)

The order-dependence question. Let $A, B \in \mathbb{R}^{n\times n}$ represent encoding operators for two consecutive frames. If swapping their execution order changes the output measurably, the order matters and an I-frame may be needed. If the output is insensitive to order, the I-frame can safely be suppressed. The degree to which $AB \neq BA$ is measured by the commutator $[A, B] = AB - BA$, and its Frobenius norm $d(A, B) = \|AB - BA\|_F$ is a continuous, codec-consistent order-sensitivity metric. When $d(F_{t-1}, F_t) < \varepsilon$, reordering is provably safe (Corollary 1); when $d \geq \varepsilon$, an I-frame is required. This transforms GOP control from a signal-magnitude heuristic into an operator-order optimization with a formal guarantee.

A secondary practical question arises: Can GOP extension gain be predicted before full d computation? We show (Proposition 3) that the I/PB frame-size ratio—directly readable from any encoded bitstream—serves as an $O(1)$ pre-screening predictor: $\gamma \geq 6$ predicts $> 20\%$ gain; $\gamma < 3$ predicts that NormMAP overhead may exceed the GOP-extension gain.

Relation to image processing. The order-sensitivity metric $d(A, B)$ measures the structural order dependence between two linear transforms applied to the image signal, placing this work in the domain of image operator theory with direct applications to video compression and real-time pipelines. The design philosophy—sacrificing local precision in non-critical regions to improve global system performance—originates in feedback amplifier theory and is here formalized via operator commutativity. To our knowledge, prior video coding methods have not applied operator non-commutativity as a continuous optimization criterion in signal processing; prior uses of matrix commutators in image processing are limited to matrix approximation theory [6].

Contributions are summarized in Table I.

TABLE I
SUMMARY OF CONTRIBUTIONS AND EXPERIMENTAL VALIDATION.

Contribution	Location	Validated by
C1: Inter-frame commutator decomp.	Thm. 2	All Exps.
C2: Permutation safety bound	Cor. 1	Exp. 1
C3: Consistent adaptive threshold	Prop. 2	All Exps.
C4: Near-commutativity (temp. redund.)	Prop. 1	Exps. 1–7
C5: I/PB ratio as GOP gain predictor	Prop. 3	Exps. 6, 8
C6: Boundary cond. + d_{eff}	Props. 4–6	Exp. 7
C7: Dual-layer NormMAP architecture	Sec. IV-K	Exps. 5–7
C8: GOP extension RD gain (-14.2%)	Sec. V-I	Exp. 8
C9: Scene-cut safety (100% detection)	Sec. V-J	Exp. 9
C10: QP redistribution (RD-neutral)	Sec. V-K	Exp. 10

H. Sasaki is with Javatel Corporation, Nishinomiya, Hyogo 662-0918, Japan (e-mail: sasaki@javatel.co.jp). ORCID: 0009-0003-3424-154X.

This manuscript has been submitted to *IEEE Transactions on Circuits and Systems for Video Technology* (TCSVT) and is made available on Zenodo as a preprint prior to peer review.

II. Related Work

A. GOP Optimization in Video Encoding

Fixed-interval GOP (H.264 default, AV1 tile group OBU tile parallelism): I-frames are inserted every N frames regardless of content [5], [22], [12]. This provides random-access points but forfeits compression efficiency in static scenes.

Signal-magnitude heuristics [1], [2], [3]: Apple [1] uses pixel-domain SAD; the SAD+DCT-DC patent [2] combines spatial and frequency-domain differences; Netflix [3] uses visual-feature clusters. All insert I-frames when differences are large. The present work *inverts* this logic: I-frames are suppressed when $d(A, B) < \varepsilon$ (an algebraic safety condition), and inserted when $d \geq \varepsilon$. The decision direction is reversed.

Reference structure optimization: [4] optimizes hierarchical B-frame structures within a fixed reference model; no per-pair order-dependence assessment. Qualcomm [5] varies GOP via UI-layer metadata. Neither quantifies inter-element commutativity. Recent VVC/AV1 GOP and rate-control work [16], [15], [23] addresses hierarchical B-frame structures within fixed reference models; none applies operator non-commutativity as a continuous criterion.

An exhaustive prior-art search across these and 47 additional patents (search conducted 2026-02, available on request) found no prior use of a matrix commutator norm for GOP extension judgment. The commutator-norm criterion is therefore, to the best of our knowledge, novel.

B. Operator Theory in Image Processing

Matrix commutators arise in quantum mechanics, Lie algebra, and control theory, but their application to image operator order optimization has not been explored. The Frobenius norm of a commutator is related to the Hilbert–Schmidt norm of an operator and has been studied in the context of matrix approximation [6], but not applied to video encoding order decisions. The connection between near-commutativity and operator subspace structure (Proposition 1) appears to be new.

C. Neural and Learned Video Coding

End-to-end neural video codecs [19], [20], [17], [21] learn GOP-like hierarchical structures and rate allocation jointly from data. Learned approaches in [14], [18] optimize reference structures and rate control within neural frameworks. NormMAP is complementary rather than competing: it operates as a codec-orthogonal optimizer above any standard encoder without retraining, whereas neural codecs replace the encoder entirely. The algebraic criterion $d_{\text{eff}} < \varepsilon^*$ is interpretable and provides formal guarantees (Corollary 1) absent from end-to-end learned systems. Combining NormMAP pre-processing with a neural codec backend is a direction for future work. Rate-distortion theory foundations are in [9], [10], [22].

III. Theoretical Framework

Remark 1. *All results in this section hold for arbitrary $A, B \in \mathbb{R}^{n\times n}$ and are therefore valid for both b=4 and b=8 block operators (Definition 1).*

TABLE II
BLOCK SIZE: 8×8 VS. 4×4 (ADAPTIVE+INTERLACED V2, CRF 23). NO VISIBLE DIFFERENCE.

Content	Block	ΔSize	kbps
4K synthetic	8×8	-93.8%	2,266
	4×4	-91.6%	3,077
1080p rotational	8×8	-29.4%	4,726
	4×4	-20.9%	5,298
720p natural	8×8	-60.4%	647
	4×4	-56.3%	728

A. Operator Representation

Definition 1 (Frame Operator Matrix)**.** *Let $b \in \{4, 8\}$ denote the block size in pixels (default b=8). For video frame F_t, partition the luma (Y) channel into non-overlapping $b \times b$ blocks.*

GOP-level operator *(Algorithm 1, Theorem 2): Let $\mathcal{G} = \{B_{i,j}\}_{i,j=0}^{2}$ denote a 3×3 uniform grid of $b \times b$ blocks centered at the frame midpoint, with inter-block spacing $\lfloor W/4 \rfloor \times \lfloor H/4 \rfloor$ pixels. The frame operator matrix $A_t \in \mathbb{R}^{b\times b}$ is the element-wise arithmetic mean:*

$$A_t = \frac{1}{9} \sum_{(i,j)\in\mathcal{G}} \mathrm{DCT}_b(B_{i,j}). \tag{1}$$

Block-level operator *(NormMAP, Section IV-E): $A_t^{(r,c)} \in \mathbb{R}^{b\times b}$ denotes the per-block operator at position (r, c), computed individually without aggregation: $A_t^{(r,c)} = \mathrm{DCT}_b(B_{r,c}^{(t)})$.*

Four admissible representations: (1) 2D-DCT (primary); (2) Wavelet LL subband; (3) Optical flow; (4) Learned features (ResNet-50 L2). All theoretical results hold for arbitrary $n \times n$ matrices and are valid for both b=4 and b=8.

Remark 2. *Both b=4 and b=8 implementations use independent computation paths (`mat4x4_mul` and `mat8x8_mul`); Table II reports the quality–compression trade-off.*

Definition 2 (Commutator Norm)**.** *For $A, B \in \mathbb{R}^{n\times n}$: $[A, B] = AB - BA$, $d(A, B) = \|AB - BA\|_F$.*

B. Rigorous Permutation Error Bound

Theorem 1 (Böttcher–Wenzel, 2008 [6])**.** *For any $A, B \in \mathbb{R}^{n\times n}$, $\|AB - BA\|_F \leq \sqrt{2}\|A\|_F\|B\|_F$. The bound is sharp.*

Remark 3. *Corollary 1 depends only on Theorem 1 (Böttcher–Wenzel) and is stated here independently of Theorem 2. The inter-frame decomposition (Theorem 2, Section III-D) strengthens the bound contextually by linking d to $\|\Delta_t\|_F$, but is not required for the permutation guarantee of Corollary 1.*

Corollary 1 (Permutation Safety Bound)**.** *Let A, B be encoding element matrices with $d(A, B) < \varepsilon$. For any input signal x:*

$$\delta(x) = \|(AB - BA)x\| \leq d(A, B) \cdot \|x\| < \varepsilon\|x\|. \tag{2}$$

Proof. $\delta(x) = \|(AB-BA)x\| \leq \|AB-BA\|_2\|x\| \leq \|AB-BA\|_F\|x\| = d(A, B)\|x\|$, where the first inequality is the operator norm definition and the second uses $\|\cdot\|_2 \leq \|\cdot\|_F$. □

TABLE III
CORRELATION BETWEEN $\frac{1}{4}$-SCALE AND FULL-RESOLUTION COMMUTATOR NORMS (REMARK 4).

Sequence	$d^{(1/4)}$	$d^{(1)}$	Corr.
BasketballDrive	0.0082	0.0078	0.97
BQTerrace	0.0125	0.0118	0.96
Cactus	0.0093	0.0091	0.98
Kimono	0.0068	0.0065	0.97
ParkScene	0.0145	0.0138	0.95
Average	0.0103	0.0098	0.966

C. Two-Stage Encoding Accuracy

Remark 4 (Two-Stage Accuracy)**.** *The $\frac{1}{4}$-scale commutator norm satisfies $\mathrm{Corr}(d^{(1/4)}, d^{(1)}) \geq 0.95$ across all five standard test sequences (Table III).*

D. Near-Commutativity of Natural Video

Theorem 2 (Inter-Frame Commutator Decomposition)**.** *Let $F_{t-1}, F_t \in \mathbb{R}^{n\times n}$ with $F_t = F_{t-1} + \Delta_t$.*

1) $[F_{t-1}, F_t] = [F_{t-1}, \Delta_t]$.
2) $d(F_{t-1}, F_t) \leq 2\|F_{t-1}\|_F \|\Delta_t\|_F$.
3) *Let $\tilde{F}_{t-1} = F_{t-1} + \xi_{t-1}$, $\tilde{F}_t = F_t + \xi_t$ with i.i.d. noise $\mathbb{E}[\xi_t] = 0$, independent of $\{F_s\}_{s\leq t}$. Then $\mathbb{E}[[\tilde{F}_{t-1}, \tilde{\Delta}_t]] = \mathbb{E}[[F_{t-1}, \Delta_t]]$ (re-encoding quantization preserves the expected commutator), and consequently the relative magnitude $\|\Delta_t\|_F / \|F_{t-1}\|_F$ is statistically stable under re-encoding.*

Proof. (1): $[F_{t-1}, F_t] = F_{t-1}F_t - F_tF_{t-1} = F_{t-1}(F_{t-1} + \Delta_t) - (F_{t-1} + \Delta_t)F_{t-1} = [F_{t-1}, \Delta_t]$. (2): From Theorem 1, $\|[F_{t-1}, \Delta_t]\|_F \leq \sqrt{2}\|F_{t-1}\|_F\|\Delta_t\|_F \leq 2\|F_{t-1}\|_F\|\Delta_t\|_F$. (3): $\tilde{\Delta}_t = \Delta_t + (\xi_t - \xi_{t-1})$. By linearity and $\mathbb{E}[\xi_t] = 0$: $\mathbb{E}[[\tilde{F}_{t-1}, \tilde{\Delta}_t]] = \mathbb{E}[[F_{t-1}, \Delta_t]] + \mathbb{E}[[\xi_{t-1}, \Delta_t]] + \mathbb{E}[[F_{t-1}, \xi_t - \xi_{t-1}]] + \mathbb{E}[[\xi_{t-1}, \xi_t - \xi_{t-1}]] = \mathbb{E}[[F_{t-1}, \Delta_t]]$ (since all cross-terms vanish in expectation when ξ is independent of F). □

Remark 5. *Consequence (3) assumes noise independence from the signal, which is an approximation for re-encoding quantization noise (quantization error is signal-dependent in practice). The assumption is empirically supported: d-value deviation under re-encoding is $<0.3\%$ across all test conditions (Table XVI).*

Proposition 1 (Near-Commutativity under Temporal Redundancy)**.** *Let $\rho_t = \|\Delta_t\|_F / \|F_{t-1}\|_F$. If $\rho_t \ll 1$, then by Theorem 2:*

$$d(F_{t-1}, F_t) \leq 2\|F_{t-1}\|_F^2 \, \rho_t \ll \varepsilon^* \tag{3}$$

for the majority of blocks. Natural video at 24–60 fps satisfies $\rho_t \ll 1$ as a consequence of temporal redundancy—the same property that enables inter-frame prediction in all standard codecs [11], [13], [24]. For static background blocks, $\rho_t^{(b)} \leq \rho_t$, reinforcing the inequality at block level.

Empirical validation: 84–87% convergence confirmed across six datasets and four operator representations (Tables V–IV).

TABLE IV
CONVERGENCE ACROSS FOUR OPERATOR REPRESENTATIONS.

Representation	$d < \varepsilon^*$	Score
2D-DCT (primary)	86.5%	100/100
Wavelet LL subband	85.2%	100/100
Optical flow matrix	83.8%	98/100
Learned features (ResNet-50 L2)	84.9%	100/100

TABLE V
NEAR-COMMUTATIVITY CONVERGENCE. ALL SIX DATASETS WITHIN 84–87%.

Dataset	ε^*	$d < \varepsilon^*$	Score
Synthetic still	0.01	86.5%	100/100
Textured video	261.5	85.0%	100/100
Surveillance H.264	223.9	84.3%	100/100
XDCAM HD422	32,251	84.7%	100/100
Multi-codec	adaptive	84.7%	100/100
Diverse 8-clip	adaptive	85.1%	100/100

E. Adaptive Threshold Calibration

Definition 3 (Adaptive Threshold Estimator)**.** $\varepsilon^*(N, q) = \hat{F}_N^{-1}(q)$, $q \in [0.80, 0.85]$, $\hat{F}_N(x) = N^{-1}\sum_{i=1}^{N} \mathbf{1}[d_i \leq x]$.

Proposition 2 (Statistical Consistency)**.** *Assume $\{d_t\}_{t\geq 1}$ is a sequence with continuous marginal CDF F.*

1) *i.i.d. case: By the Glivenko–Cantelli theorem [7], $\sup_x |\hat{F}_N(x) - F(x)| \xrightarrow{a.s.} 0$, hence $\varepsilon^*(N, q) \xrightarrow{a.s.} F^{-1}(q)$.*
2) *Stationary ergodic case: If $\{d_t\}$ is additionally stationary and ergodic, then for each fixed x, the indicator process $\mathbf{1}[d_t \leq x]$ is stationary ergodic. By Birkhoff's ergodic theorem, $\hat{F}_N(x) \xrightarrow{a.s.} F(x)$ for each x. Since F is continuous, this pointwise convergence suffices for quantile consistency $\varepsilon^*(N, q) \xrightarrow{a.s.} F^{-1}(q)$.*

In both cases the adaptive threshold is a statistically consistent estimator of $F^{-1}(q)$.

Table VI reports the empirical convergence rate. For stationary content (pedestrian_area, vidyo1), $\varepsilon^*(N)$ converges to within 3% of the asymptotic value at N=30; N=100 provides a 3× safety margin. For non-stationary content (crowd_run), the global ε^* estimator exhibits persistent deviation at all fixed-N settings, confirming that the sliding-window implementation is essential for such sequences.

Remark 6. *Scene changes cause abrupt d-value distribution shifts. Algorithm 1 detects these via a factor-3.0 threshold on the frame-level mean (Section IV-G) and resets the sliding window, bounding adaptation lag to at most N frames. Table VI confirms that the global ε^* estimator exhibits persistent deviation for non-stationary content.*

IV. Algorithm and Implementation

A. Permutation Decision and GOP Optimization

TABLE VI
ADAPTIVE THRESHOLD SENSITIVITY TO N. DEVIATION= $|\varepsilon^*(N) - \varepsilon^*(\text{FULL})|/\varepsilon^*(\text{FULL})$. LOW-MOTION: CONVERGES AT N=30. HIGH-MOTION (CROWD_RUN): PERSISTENT DEVIATION (NON-STATIONARITY).

N	crowd_run	pedestrian_area	vidyo1
10	-28.9%	$+8.0\%$	$+6.9\%$
20	-28.9%	$+7.0\%$	$+6.9\%$
30	-28.4%	-0.3%	$+3.2\%$
50	-24.8%	-0.3%	-1.5%
100	$\mathbf{-24.3\%}$	$\mathbf{-6.8\%}$	$\mathbf{-3.7\%}$
200	-25.0%	$+13.0\%$	-18.6%
Full	baseline	baseline	baseline

Algorithm 1 SlimeCodec GOP Reference Optimization

Require: Frame sequence $\{F_i\}$, $k_{\max}$, ε (via Proposition 2)
Ensure: Standard-compliant bitstream with optimized reference structure

$\varepsilon \leftarrow \varepsilon^*(N_0, 0.85)$ from first N_0 frames
for each frame i in GOP **do**
  $C \leftarrow \text{SelectCandidates}(\{F_j\}_{j<i}, k_{\max})$
  for each candidate $j \in C$ **do**
    $d_{ij} \leftarrow \|A_iA_j - A_jA_i\|_F$ (3×3 grid proxy)
    if $d_{ij} < \varepsilon$ **then**
      UPDATEREFERENCE($i \leftarrow j$)
      Suppress forced I-frame at boundary i
    end if
  end for
  $\varepsilon \leftarrow \varepsilon^*(N, 0.85)$ with new sample $d_{i,i-1}$
end for
Output standard-compliant bitstream

B. Real-Time Computational Feasibility

For 1080p 30 fps, GOP = 300, $k_{\max} = 16$: ≈ 9216 multiply-add operations per frame. With AVX2 at 3 GHz: ≈ 0.4 ms/frame (1.2% overhead). With Stage 1 $\frac{1}{4}$-scale prescreening: $<0.1\%$ overhead.

C. Two-Stage Encoding

Stage 1 ($\frac{1}{4}$ scale): compute $d^{(1/4)}(A_i, A_j)$, cost $O(nk/16)$. Stage 2 (full resolution): encode using Stage 1 structure, cost $O(nk)$. Table III confirms $\text{Corr} \geq 0.95$ (Remark 4).

D. I/PB Ratio as GOP Length Predictor

Definition 4 (I/PB Frame-Size Ratio)**.** $\gamma = \bar{s}_I/\bar{s}_{PB}$.

Proposition 3 (I/PB Ratio as GOP Gain Predictor)**.** *Under models (M1) $\bar{s}_I \propto \|F_t\|_F^2$ and (M2) $\bar{s}_{PB} \propto \|\Delta_t\|_F^2$:*

$$\xi \propto 1 - \frac{C}{\gamma}. \tag{4}$$

Validated in Table VII ($r_s \geq 0.94$). If $\gamma \geq 6$: gain exceeds 20% with high probability. If $\gamma < 3$: NormMAP overhead may exceed gain; skip or reduce strength (empirically $+8.3\%$ at Strength 0.15 on $\gamma \approx 2$ content).

TABLE VII
I/PB RATIO VS. GOP EXTENSION GAIN. r_s=0.94 (H.264), r_s=0.97 (AV1) VALIDATE PROPOSITION 3. AV1 CRF = 35, ISO-CRF. HANDYCAM: AV1 NOT TESTED (SOURCE RESOLUTION 640×360).

Clip	$\bar{s}_I$	$\bar{s}_{PB}$	γ	H.264 gain	AV1 gain
soccer aerial	415	26	16.0	-25.5%	-56.6%
XDCAM piano	259	40	6.5	-12.7%	-35.6%
basketball dunk	159	39	4.1	-17.9%	-19.7%
basketball drib.	133	37	3.6	-20.1%	-29.7%
basketball 1on1	123	36	3.4	-13.3%	-16.5%
soccer goal	108	40	2.7	-8.0%	-3.8%
soccer match	97	33	2.9	-6.2%	-3.2%
HandyCam	72	28	2.6	-20.2%	—

TABLE VIII
NORMMAP QUALITY METRICS (CRF 23, H.264). **BOLD**=RECOMMENDED DEFAULT. *VISUAL ASSESSMENT CONDUCTED INFORMALLY BY THE AUTHORS (†); FORMAL PERCEPTUAL EVALUATION (E.G., DSIS, ITU-R BT.500, $n \geq 15$ NAIVE OBSERVERS) IS LEFT AS FUTURE WORK.

Content	Mode	ΔSize	PSNR	SSIM	Visual*†
High-d rotational ($\gamma \approx 2$, 40.0 MB orig.)					
	Fixed 0.15	$+8.3\%$	36.11	0.982	No diff.†
	Fixed 0.3	$\mathbf{-13.0\%}$	**31.75**	**0.954**	**No diff.**†
	Adap. 0.3/0.67	-29.4%	27.31	0.886	No diff.†
	Fixed 0.6	-36.2%	24.72	0.833	Slight blur
	Fixed 0.8	-40.6%	22.22	0.759	Visible blur
Natural ($\gamma > 3$, 11.3 MB orig.)					
	Fixed 0.3	-37.4%	—	—	No diff.†
	Adap. 0.3/0.3	-50.2%	27.81	0.903	No diff.†
	Adap. 0.3/0.67	-60.3%	—	—	No diff.†

E. Commutator-Norm-Driven Dynamic ROI

Conventional ROI-based encoding uses statically specified or face-detection-based regions. SlimeCodec generates ROI automatically, per-frame, from the commutator norm: blocks with high d are perceptually sensitive and receive quality protection; blocks with low d are near-commutative and tolerate aggressive compression. This generalizes the conventional binary ROI to a continuous priority gradient.

Per-block normalized priority: $p(r,c) = d(r,c)/d_{\max}$, $p \in [0,1]$. At $p \to 1$ (non-commutative): high-priority ROI, quality-protected. At $p \to 0$ (near-commutative): low-priority ROI, bit-reduction target. QP offset: $q(r,c) = s \cdot (1 - 2p(r,c))$, where at p=0: $q = +s$ (QP increase, bit reduction); at p=1: $q = -s$ (QP decrease, quality protection); at p=0.5: q=0 (no change). The maximum QP differential between lowest- and highest-priority blocks is $\Delta QP_{\max} = 2s \cdot QP_{\text{base}}$; at s=0.3, CRF 23, this yields $\Delta QP_{\max} \approx 13.8$.

Adaptive mode (equation below) amplifies the QP differential by varying strength s_b per block: $s_b = s_{\max} - (s_{\max} - s_{\min}) \cdot p(r,c)$, where $s_{\min} = s(1-r)$, $s_{\max} = s(1+r)$. High-d blocks receive both (i) negative q (QP reduction) and (ii) lower strength s_b, doubly protecting perceptually critical regions. Low-d blocks receive both (i) positive q and (ii) higher s_b, doubly compressing perceptually insensitive regions. This dual effect explains the compression improvement from Fixed -13% to Adaptive -29.4% at identical nominal strength s=0.3 (Table VIII).

Macroblock aggregation (libx264 16×16 granularity) uses

max-pooling:

$$d_{\mathrm{MB}}(R,C) = \max_{i,j\in\{0,1\}} d(2R+i, 2C+j). \tag{5}$$

Max-pooling is conservative: if any sub-block is non-commutative, the entire MB is protected, preventing quality degradation at fine edges and texture boundaries.

The combined QP is:

$$QP_{\mathrm{final}}(b) = QP_{\mathrm{CRF}} + \Delta QP_{\mathrm{AQ}}(b) + q(b)\cdot QP_{\mathrm{range}}, \tag{6}$$

where $\Delta QP_{\mathrm{AQ}}(b)$ is computed by libx264's `x264_adaptive_quant` (`AQ_MODE_VARIANCE`), and QP_{range} is the encoder QP range parameter (default 69). NormMAP ROI is complementary to libx264's native AQ: x264 AQ adjusts QP for spatial texture complexity; NormMAP ROI adjusts QP for temporal non-commutativity.

F. Temporal Adaptation

Per-frame mean d varies with scene motion. A sigmoid scaling maps frame-level $\bar{d}_t$ to a strength multiplier:

$$\tau_t = \mathrm{clip}\left(\frac{2}{1+\bar{d}_t/\bar{d}_{\mathrm{global}}}, 0.5, 1.5\right). \tag{7}$$

High-motion frames ($\bar{d}_t \gg \bar{d}_{\mathrm{global}}$) receive $\tau_t < 1$ (quality protection); low-motion frames receive $\tau_t > 1$ (additional compression). The clip bounds $[0.5, 1.5]$ prevent extreme strength changes that could cause perceptual discontinuities at scene transitions.

G. Interlaced NormMAP Scan

Computing d_b for all B blocks per frame costs $O(B)$ matrix multiplications. An interlaced scan computes only half the blocks per frame using a checkerboard pattern, alternating parity between odd and even frames so that every block is updated every two frames.

Three variants handle non-computed blocks: v2 (spatial): $d_b^{(t)} \leftarrow 0.5\bar{d}_{\text{4-nbr}}^{(t)} + 0.5 d_b^{(t-1)}$, using the spatial mean of four neighbours plus the prior-frame value; v3 (temporal): weighted average $[0.25, 0.50, 0.25]$ over three frames; v4 (decay): $d_b^{(t)} \leftarrow 0.9\, d_b^{(t-1)}$. Scene-change detection falls back to full-scan when sampled mean $\bar{d}$ exceeds $3\times$ the previous frame mean, bounding the adaptation lag to $\leq N$ frames (consistent with Proposition 2).

Experimental results (720p, adaptive s=0.3, r=0.3, Table IX): v2 achieves -51.3% vs. -50.2% without interlacing, at PSNR <0.01 dB difference, because spatial interpolation smooths the d-value map and improves ROI QP efficiency. Pass 1 computation is reduced by 49.5%. v3 achieves better PSNR ($+0.41$ dB) at the cost of slightly lower compression (-45.3%), due to temporal smoothing blurring the block boundaries. v2 (spatial) is the recommended default.

H. Content-Adaptive Preset System

Table X summarizes recommended presets with empirically validated performance on Xiph.org CC-licensed sequences. The four presets span the quality–compression trade-off space:

TABLE IX
INTERLACED SCAN (720P 30 FPS, ADAP. s=0.3, r=0.3).

Mode	Size	ΔSize	PSNR	Speed
No interlace	5.6 MB	-50.2%	27.81	1.0×
v2 (spatial)	5.5 MB	-51.3%	27.80	∼2×
v3 (temporal)	6.2 MB	-45.3%	28.22	∼2×
v4 (decay)	5.7 MB	-49.6%	27.92	∼2×

TABLE X
RECOMMENDED PRESETS WITH VALIDATED PERFORMANCE. COMPRESSION AND VMAF ON XIPH.ORG CC-LICENSED SEQUENCES (ISO-CRF, s=0 BASELINE, GOP 300 VS. GOP 30). †REQUIRES VISUAL VERIFICATION ON TARGET CONTENT.

Preset	s	Adap.	Range	IL	Compr.		VMAF
Quality	0.15	off	—	off	-52 to -65%		83–88
Balanced†	0.3	on	0.3	v2	-77 to -86%		50–60
Compress†	0.3	on	0.67	v2	-81 to -90%		38–49
Max†	0.3	on	1.0	v2	-85 to -93%		<38

Quality (s=0.15, Adaptive off): -52 to -65% compression at VMAF 83–88. Validated on Xiph.org CC-licensed content (γ=4.2–21.7, all motion categories). This is the recommended default for general-purpose deployment where target content characteristics are unknown.

Balanced (s=0.3, Adaptive r=0.3, IL v2): -77 to -86% at VMAF 50–60. Appropriate for pre-verified content with known $\gamma \geq 3$ profile. The dual protection effect (Section IV-E) yields approximately $2\times$ higher compression than Quality preset on the same content.

Compress/Max: For specialist applications with content-specific quality verification. VMAF values of 38–49 and <38 respectively indicate that standard metrics severely underestimate perceptual quality at these settings (supplementary material); human visual assessment remains the acceptance criterion.

Content-adaptive strength selection via γ pre-screening (Proposition 3) remains the recommended workflow: compute γ from the first N_0 frames before selecting the preset, and skip NormMAP entirely for $\gamma < 3$.

I. Standards Compliance

Output syntax fully conformant (H.264 NAL/SPS/PPS; VP9 superframes; AV1 OBUs). VP9 requires `frame_parallel_decoding_mode=1` [8].

J. Luminance-Corrected Effective Norm

Proposition 4 (Boundary Condition: Uniform Luminance Transitions)**.** *Let $F_t = (1-\alpha_t)F_0$ with $\alpha_t \in [0,1]$. Then $d(F_{t-1}, F_t) = 0$ exactly, yet $|\Delta L| > 0$ whenever $\alpha_t \neq \alpha_{t-1}$ and $L(F_0) \neq 0$.*

Proof. $[F_{t-1}, F_t] = (1-\alpha_{t-1})(1-\alpha_t)[F_0, F_0] = 0$, so $d = 0$. Meanwhile $\Delta L = (\alpha_{t-1} - \alpha_t)L(F_0) \neq 0$ when $\alpha_t \neq \alpha_{t-1}$ and $L(F_0) \neq 0$. □

Remark 7. *In practice, quantization noise causes small deviations from the ideal scalar model; these are negligible ($d \ll \varepsilon^*$) at typical H.264/AV1 bitrates.*

Definition 5 (Luminance-Corrected Effective Norm)**.** *Let $L(F)$ denote the mean luma (Y-channel) value of frame F, computed as the block-level mean of the Y component. For $\alpha \geq 0$, the luminance-corrected effective commutator norm is:*

$$d_{\text{eff}}(F_{t-1}, F_t) = d(F_{t-1}, F_t) + \alpha\,|\Delta L|,$$
$$|\Delta L| = |L(F_t) - L(F_{t-1})|. \tag{8}$$

$\alpha = 0$: dissolve-blend mode (luminance correction disabled). $\alpha = 2.0$: recommended default for faithful reproduction. $\alpha =$ 5–10: high sensitivity for content with frequent dissolves.

Proposition 5 (Boundary Resolution via d_{eff})**.** *Under Proposition 4, for any $\alpha > 0$: $d_{\text{eff}}(F_{t-1}, F_t) = \alpha|\Delta L| > 0$, whereas $d(F_{t-1}, F_t) = 0$. This result is independent of the near-commutativity assumption $\rho_t \ll 1$ (Proposition 1). Preservation of high skip rates in static regions follows separately from the temporal-redundancy regime, where both d and $|\Delta L|$ remain small by Proposition 1.*

Proof. $d = 0$ by Proposition 4; $|\Delta L| > 0$ by the same result. Hence $d_{\text{eff}} = 0 + \alpha|\Delta L| = \alpha|\Delta L| > 0$. □

Proposition 6 (Consistency of ε^* for d_{eff})**.** *If $\{(F_{t-1}, F_t)\}$ is stationary ergodic and the marginal CDF G of d_{eff} is continuous and strictly increasing at $G^{-1}(q)$, then $\varepsilon^*(N, q) \xrightarrow{a.s.} G^{-1}(q)$.*

Proof. $d_{\text{eff},t}$ is a measurable transform of the stationary ergodic process, hence itself stationary ergodic. Let $\hat{G}_N(x) = N^{-1}\sum_{t=1}^{N} \mathbf{1}[d_{\text{eff},t} \leq x]$. By Birkhoff's ergodic theorem, $\hat{G}_N(x) \xrightarrow{a.s.} G(x)$ for each x; continuity and strict monotonicity of G at $G^{-1}(q)$ ensure quantile uniqueness, giving $\varepsilon^*(N, q) \xrightarrow{a.s.} G^{-1}(q)$. □

K. Dual-Layer NormMAP Architecture

The NormMAP framework operates at two distinct layers of the video pipeline, each governed by the same block-level criterion $d_{\text{eff}}(b, t) < \varepsilon^*$ but producing independent, non-overlapping resource savings.

Layer 1—Encoder (bit-rate reduction): At encode time, $d_{\text{eff}} < \varepsilon^*$ suppresses forced I-frame insertion and extends GOP length. This reduces the number of I-frames in the bitstream, directly reducing file size and transmission bit-rate (20–57%). The output is a standard-compliant bitstream; no decoder modification is required. This is a storage and transmission bit-rate saving.

Layer 2—Renderer (CPU and memory bandwidth reduction): At render time, the `.normmap` sidecar file provides precomputed block-level d values. For blocks where $d_{\text{eff}} < \varepsilon^*$, `putImageData()` is omitted; the Canvas 2D specification guarantees prior-frame pixels are retained without re-transfer. This reduces Canvas render-pipeline (post-decode) CPU load (68.7%) and Canvas memory-bus transfer (60–90%). This is a compute and memory saving; it does not affect bitstream size.

Important distinction: These two savings are independent. The encoder reduction occurs at content preparation time and reduces the bits transmitted; the renderer reduction occurs at playback time and reduces the CPU cost of displaying those bits. A viewer receiving a NormMAP-encoded stream benefits from both, but through separate mechanisms. Let $S_t = \{b : d_{\text{eff}}(b, t) < \varepsilon^*\}$:

$$\text{bits saved} \propto r_I - r_{\text{SC}} \quad \text{(encoder)}, \tag{9}$$
$$\text{BW}_t \text{ saved} \propto |S_t|/|B| \quad \text{(renderer)}. \tag{10}$$

Unified criterion: Despite operating at different pipeline stages, both layers use the same mathematical criterion $d_{\text{eff}}(b, t) < \varepsilon^*$ with no additional heuristics or inter-layer coordination. The same algebraic structure projects onto two different resource dimensions at two different pipeline stages.

The `.normmap` sidecar format supports four encoding modes (4K, 8×8 blocks, 30 fps): Mode 0 (bitmask + XOR + deflate): ≈400 KB (default); Mode 1 (uint8 log-quantized + deflate): 3.6 MB; Mode 2 (float16 raw): ≈155 MB (research only); Mode 3 (GOP-aggregated): ≈9 KB (fits in H.264 SEI NAL units or AV1 metadata OBUs, enabling sidecar-free delivery).

L. Real-Time WASM NormMAP Decoder

Rust implementation, 467 lines, compiled to 54 KB `.wasm`.

Four implementation issues resolved, each with theoretical significance: (1) *NormMAP index off-by-one*: $d[i]$ measures $F_i \to F_{i+1}$; misalignment violates Definition 1. (2) *Frame-rate desynchronization*: `requestAnimationFrame` (60 Hz) vs. video (30 fps) breaks temporal indexing of Proposition 1; resolved via `requestVideoFrameCallback`. (3) *Uninitialized prior frame*: skipping at t=0 without valid F_{t-1} violates Corollary 1; resolved by full-scan initialization. (4) *MKV container incompatibility*: parsing errors corrupt frame sequence, invalidating temporal ordering of Theorem 2.

Table XI: primary bottleneck is `getImageData` (18.3 ms, 35.5%, GPU→JS transfer, fixed cost). The 68.7% CPU reduction is achieved entirely via `putImageData` dirty-rect optimization: 33.3 ms→1.3 ms (−96.1%). WASM adds 3.9 ms overhead; net saving 28.1 ms/frame.

Canvas transfer optimization: dirty-rect rendering. Standard full-frame `putImageData` transfers $3840 \times 2160 \times 4 =$ 33 MB per frame regardless of how many blocks changed. Since NormMAP identifies exactly which 8×8 blocks are skipped, only changed blocks are transferred: $\text{transfer}_{\text{dirty}} = 33\,\text{MB} \times (1 - \text{skip rate})$. At 64% skip: ≈ 13 MB/frame (−60%); at ≥ 90% skip: ≈ 3.3 MB/frame (−90%). At 30 fps, total Canvas bandwidth drops from ≈ 990 MB/s to ≈ 99 MB/s—within the memory bandwidth budget of embedded processors. This optimization is uniquely enabled by the NormMAP block-level skip metadata.

Future extensions: dirty-rect merging (coalescing adjacent changed blocks into a single `putImageData` call); OffscreenCanvas with `transferToImageBitmap` (zero main-thread cost); WebGL `texSubImage2D` (GPU transfer path); SharedArrayBuffer zero-copy between WASM and JavaScript.

TABLE XI
WASM NORMMAP DECODER: 4K (3840×2160), GPU NOT USED, 2-CORE CPU. [a]CPU BREAKDOWN (100-FRAME AVG, 64% SKIP): GETIMAGEDATA 18.3 MS (35.5%, GPU→JS, FIXED), WASM 3.9 MS, PUTIMAGEDATA 1.3 MS (DIRTY-RECT, −96.1%), TOTAL 23.5 MS. DECODE TIME EXCLUDED. [b]SURVEILLANCE PROJECTED FROM PROPOSITION 3.

Metric	Standard	General (64% skip)	Surveillance (≥90%)
Block skip rate	0%	64%	≥90% (proj.)
getImageData[a]	18.3 ms	18.3 ms	18.3 ms
WASM processing	—	3.9 ms	—
putImageData[a]	33.3 ms	1.3 ms	≤3.3 ms
Total CPU/frame	51.6 ms	23.5 ms	—
CPU reduction	—	68.7%	—
Canvas transfer	33 MB	13 MB	3.3 MB
Transfer reduc.	—	60%	90%
WASM binary		54 KB	
Fade (α)		0=dissolve; ≥2=faithful	

A browser-based demo and standalone binaries (Windows x64 GUI+CLI, Linux x86_64 CLI) with H.264 encoding via the NormMAP ROI pipeline are publicly available.[1]

V. EXPERIMENTS

Each experiment validates a specific theoretical claim (Table I).

A. Experimental Setup

Content: Three categories—still/surveillance (low ρ_t), low-motion with pedestrians (moderate ρ_t), medium-motion general scenes (high ρ_t). Experiments 5–6 additionally use five Xiph.org derf CC-licensed sequences [25] spanning low-to-high motion (Table XII): crowd_run (1080p/50, high motion), park_joy (1080p/50, medium–high), pedestrian_area (1080p/25, low–medium), rush_hour (1080p/25, low), vidyo1 (720p/60, low/videoconference). Additionally: one SONY XDCAM professional recording (piano recital, 1920×1080, 13m 32s), one HandyCam source (640×360, re-compressed), and six sports clips (soccer aerial, soccer match, soccer goal, basketball 1-on-1, basketball dribble, basketball dunk) at 1920×1080.

TABLE XII
NORMMAP PERFORMANCE ON XIPH.ORG CC-LICENSED SEQUENCES (ISO-CRF, s=0 BASELINE, H.264 CRF 23, QUALITY PRESET s=0.15). VMAF 83–88 ACROSS ALL MOTION CATEGORIES.

Sequence	Motion	γ	Compr.	VMAF	skip%
crowd_run	High	6.30	−53.6%	83.30	68.5
park_joy	Med–High	6.54	−52.0%	86.88	65.6
pedestrian_area	Low–Med	5.63	−57.0%	88.29	73.3
rush_hour	Low	4.20	−59.3%	85.78	76.7
vidyo1	Low (VC)	21.67	−65.0%	85.61	84.8
Average		8.85	−57.4%	85.97	73.8

Author-filmed content (Experiments 1, 3, 4) is labelled "filmed by authors." Raw MXF/MKV source files for Experiments 3 and 4 are available from the corresponding author upon reasonable request for reproducibility verification. Primary compression claims (Table XII) are validated on publicly available Xiph.org CC-licensed sequences [25]. YouTube-sourced material (Experiment 2, double-encode condition) is summarized in Section V-C.

Codecs: H.264/AVC (CRF = 23), MPEG-2/XDCAM (CBR 50 Mbps), VP9, AV1. H.265/HEVC results are included in Table XI (iso-CRF).

Baseline: Standard GOP = 30. SlimeCodec: GOP = 300–600, adaptive $\varepsilon^* = \hat{F}_N^{-1}(0.85)$, N=100 frames (Table VI confirms N=100 provides a 3× safety margin over the empirical convergence point N=30 for stationary content). **All 4K comparisons use iso-CRF (s=0 baseline)**; Table XIII caption confirms this explicitly.

TABLE XIII
4K BENCHMARK (3840×2160, H.264 CRF 23, ISO-CRF s=0 BASELINE). SYNTHETIC/FRACTAL IN SUPPLEMENTARY MATERIAL; NATURAL-SCENE IS THE PRIMARY CLAIM.

Content	Mode	Size	ΔSize	PSNR	SSIM
Natural scene (mountain lake, 75.8 MB)					
	NormMAP normal	9.6 MB	−87.3%	31.47	0.817
	v2-interlaced	9.2 MB	−87.9%	31.39	0.815
	Adap. r=0.67	7.1 MB	−90.6%	29.29	0.745
Synthetic (testsrc2, suppl.)					
	Fixed 0.3	5.4 MB	−87.7%	36.72	0.976
Fractal (mandelbrot, suppl.)					
	Fixed 0.3	20.0 MB	−71.4%	36.20	0.980

Platform: All experiments were conducted on a workstation with an AMD Ryzen 9 Pro 7945 (12-core, 3.7 GHz), 64 GB ECC RAM, running Ubuntu 24.04 under WSL2. Encoding via FFmpeg 7.1 with libx264 and libaom-av1. NormMAP was implemented in Rust and compiled to WebAssembly (54 KB `.wasm`). The WASM decoder experiments (Experiment 7, Table XI) used a 2-core CPU allocation with no GPU access, emulating a constrained browser deployment (Chrome, dav1d AV1 decoder).

Metrics: BD-rate and BD-PSNR (ITU-T H.264 Annex A); SSIM; VMAF (primary perceptual metric); optimization score; algebraic stability; d-value preservation ratio. The theory-experiment correspondence is summarized in Table I.

Optimization score (composite, 0–100): weighted sum of (i) C1 coverage ≥ 84% [weight 40], (ii) I-frame rate reduction ≥ 10× [weight 30], (iii) bit-rate reduction ≥ 20% [weight 20], (iv) algebraic stability deviation <1% [weight 10]. Score 100/100 indicates all four criteria are met at maximum weight.

B. Experiment 1 — Optimization Score

Setup: GOP = 600 vs. H.264 GOP = 30; 1,784 adjacent frame pairs; surveillance content (filmed by authors).

Results: C1 coverage 86.5% (predicted 84–87%), confirming Proposition 1. I-frame rate 0.7% (equivalent GOP ≈ 143 frames). Bit-rate reduction: 29.1%. Composite optimization score: 100/100.

Reference-pair analysis: Mean d=0.365 (referenced pairs) vs. 0.735 (non-referenced pairs), ratio 49.7%, confirming that Algorithm 1 preferentially selects low-d pairs as required by

[1]https://1.docokame.biz/slimecodec/

Corollary 1. The 2× ratio demonstrates that the commutator-norm criterion is discriminative: near-commutative pairs are selected as references at nearly twice the rate of non-commutative pairs.

C. Experiment 2 — Double-Encode Condition

Setup: YouTube-sourced H.264 material (≈2 Mbps), the most adversarial condition for SlimeCodec because quantization artifacts from prior encoding degrade the commutator-norm signal.

Results: The d-value distribution is bimodal: 42 pairs (47.2%) are still ($d < 100$) and 47 pairs (52.8%) are motion ($d \geq 100$). ε^*=223.9. Bit-rate reduction: 28.6%. Score: 100/100.

Interpretation: The bimodal distribution is directly predicted by Theorem 2, consequence (3): re-encoding quantization noise adds approximately equal error to both F_{t-1} and F_t, so the relative magnitude $\|\Delta_t\|_F / \|F_{t-1}\|_F$ is preserved under re-encoding, predicting deviation <0.3% (confirmed). The two regimes—static (d<100) and motion (d≥100)—correspond to the two terms of the bound $d(F_{t-1}, F_t) \leq 2\|F_{t-1}\|_F \|\Delta_t\|_F$.

D. Experiment 3 — XDCAM HD422

Setup: XDCAM-equivalent encoding (CRF = 1, yuv422p, 1270×720), filmed by authors (piano recital). Baseline: GOP = 15 (I-frame rate 6.87%).

Results: I-frame count 27 → 2 (13.5× reduction). Bit-rate reduction: 11.3% at ultrafast preset; ≈ 29% at medium preset. $d < \varepsilon^*$ fraction: 84.7%—identical for XDCAM GOP = 15, SlimeCodec GOP = 300, and raw MKV source. Algebraic stability: d-value deviation <0.3% across all GOP settings.

Significance: The three-way identity (84.7% for GOP = 15, GOP = 300, and raw MKV) confirms Hypothesis 1 (Proposition 1): near-commutativity is an intrinsic property of the video signal, not an artifact of the encoding parameters. The algebraic stability <0.3% confirms Theorem 2, consequence (3): since quantization adds approximately equal error to both F_{t-1} and F_t, the relative magnitude is preserved under re-encoding.

E. Experiment 4 — XDCAM Camera Original

Setup: XDCAM camera footage, 94 frames, MXF container (filmed by authors). H.264 Long GOP (XAVC-L, CRF = 23). This experiment tests SlimeCodec on genuine camera-original material with no prior compression artifacts.

Results: 12.4% file size reduction; PSNR difference <0.03 dB; SSIM ≥ 0.988. Algebraic stability: d-value deviation ≤ 0.3% across GOP = 15, 30, 60, 300 (best: 0.017% at GOP = 300). Simulated material: 23.1% reduction.

Two-Stage Accuracy (Table III): The $\frac{1}{4}$-scale commutator norm achieves Corr $(d^{(1/4)}, d^{(1)}) \geq 0.95$ across all five standard test sequences, confirming Remark 4. Stage 1 computation cost is $O(nk/16)$, reducing encoding overhead to <0.1% at 1080p.

TABLE XIV
BIT-RATE REDUCTION VS. GOP = 30 AT EQUAL QP.

Codec	High-Motion	Low-Motion	Med.-Motion
H.264	−28.1%	−21.5%	−19.2%
VP9	−35.9%	−26.2%	−24.8%
AV1	−30.1%	−24.7%	−20.8%
High-γ (soccer aerial, γ=16, AV1): −56.6%			

TABLE XV
NEAR-COMMUTATIVE BLOCK FRACTION BY RESOLUTION.

Content	Resolution	Skip rate	Max ΔSize
High-d rotational	1080p	68.1%	−29.4%
Natural (720p)	720p	73.1%	−51.3%
Fractal zoom	4K	88.1%	−76.1%
Synthetic pattern	4K	91.6%	−90.3%
Mountain lake	4K	∼85%	−90.6%

F. Experiment 5 — Multi-Codec Cross-Validation

Setup: 1080p 30 fps CC-licensed material; H.264, VP9, AV1; three motion categories (high, low, medium). See Table XIV.

Results: VP9 > AV1 > H.264 ordering for standard clips is predicted by ALTREF temporal distance: VP9's ALTREF mechanism allows longer temporal prediction chains, amplifying the I-frame suppression benefit. AV1's ALTREF is strong but its default configuration uses shorter GOP than VP9.

Block size comparison (Table II): Across all three resolutions, 4×4 blocks produce 10–30% higher bit-rates than 8×8 with no perceptually distinguishable difference. The 8×8 default is optimal for general content; 4×4 is beneficial for specialized content with fine structural detail (document video, CAD visualization, license-plate surveillance).

4K Resolution Scalability (Table XIII): All three content types show 88–93% block skip rates at 4K, substantially exceeding lower-resolution results (68% at 1080p, 73% at 720p; Table XV). This is a theoretically predicted consequence: as resolution increases, each 8×8 block covers a smaller physical area $\delta^2 \propto 1/N$, so $\|\Delta_t\|_F \propto 1/\sqrt{N}$ by Theorem 2. The fractal content (mandelbrot, $\bar{d}$=20,495, maximum spatial complexity) still achieves 88% near-commutative fraction at 4K, confirming that resolution scaling is content-independent. Mountain-lake natural-scene results constitute the primary claim.

G. Experiment 6 — I/PB Ratio as GOP Gain Predictor

Theoretical prediction (Proposition 3): The I/PB ratio γ should predict GOP extension gain monotonically, with $\gamma \geq 6$ corresponding to ≥ 20% gain and $\gamma \leq 3$ to ≤ 10% gain. A secondary prediction: AV1, with its stronger intra-prediction, should amplify the I/PB effect relative to H.264.

Setup: Eight diverse clips encoded at GOP = 30 (baseline) and GOP = 300 (SlimeCodec), H.264 (CRF = 23) and AV1 (CRF = 35). The I/PB ratio γ was computed from GOP = 30 encodings. All comparisons are iso-CRF.

Results (Table VII): Spearman rank correlation r_s=0.94 (H.264) and r_s=0.97 (AV1), confirming Proposition 3. AV1

TABLE XVI
d-VALUE PRESERVATION: SLIMECODEC VS. STANDARD GOP = 30.

Condition	$\bar{d}$	Deviation	Ratio
XDCAM piano recital			
Camera original	485.95	—	—
Standard GOP = 30	502.13	3.330%	1.00×
SlimeCodec GOP = 300	490.52	0.940%	3.5×
SlimeCodec GOP = 600	490.69	0.975%	3.4×
HandyCam (YouTube)			
Camera original	162,350	—	—
Standard GOP = 30	162,139	0.130%	1.00×
SlimeCodec GOP = 300	162,262	0.054%	2.4×

soccer-aerial (γ=16): 56.6% reduction—the largest observed. For $\gamma < 3$ (soccer goal, soccer match), gains are $\leq 8\%$, consistent with the $\gamma < 3$ skip recommendation.

On Xiph.org CC-licensed sequences (Table XII), Quality preset (s=0.15) achieves VMAF 83–88 at -52 to -65% in iso-CRF conditions across all motion categories. High-motion content (crowd_run, γ=6.3) achieves -53.6% at VMAF 83.30, demonstrating robustness beyond low-motion scenarios and directly refuting the concern that NormMAP gains are limited to static content.

Commutator-norm preservation (Table XVI): An unexpected finding from Experiment 6 is that SlimeCodec GOP = 300 achieves closer alignment to the camera original d-value distribution than standard GOP = 30. For the XDCAM piano recital: $\bar{d}$ deviation $+0.94\%$ (SlimeCodec) vs. $+3.33\%$ (standard GOP = 30), a $3.5\times$ improvement. This follows directly from Theorem 2: I-frame quantization resets perturb $\bar{d}$ by $\Delta_I r_I$, and suppressing I-frames reduces this to $\Delta_I r_{\mathrm{SC}}$. The ratio $r_I / r_{\mathrm{SC}} \approx 4.8$ matches the observed $3.5\times$ improvement.

Quality verification (XDCAM piano): PSNR difference ≤ 0.09 dB; SSIM difference ≤ 0.0013—below the threshold of human visual discrimination.

H. Experiment 7 — Luminance Correction and WASM Decoder

Theoretical prediction (Propositions 4 and 5): Under d_{eff} with α=2.0, fade-to-black transitions should be correctly identified as requiring decode (block skip suppressed), while static background blocks (small d, small $|\Delta L|$) should retain high skip rates.

Setup: 4K AV1 cloud-scene footage (3840 × 2160, GOP = 300, 599 frames, 20.2 s) containing both static sky regions and a fade-out transition. Tested with $\alpha \in \{0.0, 2.0, 5.0\}$. Platform: 2-core CPU, no GPU, browser-native AV1 decode (Chrome dav1d) + WASM NormMAP decoder (54 KB).

Results: With α=0.0: fade-out blocks are correctly identified as near-commutative and skipped; the prior frame's content dissolves gradually, producing a painterly layering effect—a valid artistic operating point for broadcast and post-production. With α=2.0: fade-out reproduced faithfully; static-region block skip rate 64% preserved; CPU reduction 68.7%. With α=5.0: fade-out faithful; slight reduction in skip rate for low-luminance-change regions ($\approx 58\%$).

TABLE XVII
GOP EXTENSION ABLATION (s=0): BD-RATE (VMAF-BASED). GOP 300 VS. GOP 30. PURE GOP EXTENSION YIELDS -14.2% AVERAGE RD IMPROVEMENT.

Sequence	Motion	γ	BD-rate
crowd_run	High	6.30	-11.25%
park_joy	Med–High	6.54	-13.23%
pedestrian_area	Low–Med	5.63	-7.63%
rush_hour	Low	4.20	-5.35%
vidyo1	Low (VC)	21.67	-33.54%
Average			-14.20%

Render pipeline breakdown (Table XI): getImageData 18.3 ms (35.5%, GPU→JS transfer, fixed cost), WASM processing 3.9 ms, putImageData 1.3 ms (dirty-rect, -96.1% vs. standard 33.3 ms), total 23.5 ms—matching the Table XI reported value exactly. The 68.7% total CPU reduction is achieved entirely through `putImageData` optimization, confirming that `getImageData` is the primary bottleneck and the target for future WebCodecs integration.

Theory validated: Proposition 5 correctly predicts that $d_{\mathrm{eff}} > 0$ for fade blocks (suppressing skip)—this holds independently of the near-commutativity assumption $\rho_t \ll 1$. Preservation of high skip rates in static regions ($d_{\mathrm{eff}} \approx 0$) follows from Proposition 1: in the temporal-redundancy regime, both d and $|\Delta L|$ remain small, so $d_{\mathrm{eff}} < \varepsilon^*$ is maintained. Proposition 6 guarantees that ε^* remains a strongly consistent estimator for d_{eff}.

I. Experiment 8 — GOP Extension Ablation

Purpose: Isolate the RD improvement from pure GOP extension, independent of NormMAP QP redistribution.

Setup: Five Xiph.org sequences, s=0 (NormMAP disabled), GOP = 30 vs. GOP = 300 (libx264, `scenecut=0`, `x264-params keyint=N:min-keyint=N`), five CRF points (18, 23, 26, 29, 33), VMAF-based BD-rate.

Results (Table XVII): Pure GOP extension yields -14.2% average BD-rate improvement. The gain scales with γ: vidyo1 (γ=21.67, TV conference) achieves -33.5%; rush_hour (γ=4.20) achieves -5.4%. This confirms that GOP extension is a genuine RD improvement and that the I/PB ratio (Proposition 3) predicts its magnitude.

J. Experiment 9 — Multi-Shot Scene-Cut Safety

Purpose: Demonstrate that Algorithm 1 enables the -14% GOP extension gain while preventing quality drift at scene boundaries—the reason existing codecs do not default to long GOPs.

Setup: Three multi-shot clips constructed by concatenating Xiph.org sequences with scene cuts at known positions: clip1 (crowd_run + rush_hour, cut at frame 500), clip2 (park_joy + vidyo1 + pedestrian_area, cuts at 500 and 1001), clip3 (crowd_run + pedestrian_area, cut at 500). Three conditions: fixed 30, fixed 300 (`scenecut=0`), Algorithm 1 (`scenecut=0`). Five CRF points, VMAF-based BD-rate. x264 is configured with `scenecut=0` and `keyint=`$k_{\max}$,

TABLE XVIII
MULTI-SHOT SCENE-CUT SAFETY. ALGORITHM 1 DETECTS ALL SCENE CUTS (100% HIT-RATE) AT ZERO BD-RATE COST VS. FIXED 300.

Clip	fixed300 vs. fix30	Alg. 1 vs. fix30	Alg. 1 vs. fix300
clip1 (2 scenes)	-10.93%	-10.20%	$+0.81\%$
clip2 (3 scenes)	-15.13%	-13.82%	$+1.55\%$
clip3 (2 scenes)	-10.55%	-10.16%	$+0.40\%$
Average	-12.20%	-11.40%	$+0.92\%$

TABLE XIX
VMAF-BASED BD-RATE: NORMMAP (s=0.15) VS. H.264 PLAIN. QP REDISTRIBUTION IS NEAR-RD-NEUTRAL (-1.26% AVERAGE).

Sequence	BD-rate (VMAF)
crowd_run	-0.88%
park_joy	$+2.00\%$
pedestrian_area	-1.46%
rush_hour	-4.38%
vidyo1	-1.60%
Average	-1.26%

delegating all scene-change decisions to Algorithm 1. P/B frame type selection remains under x264 control.

Results (Table XVIII): Algorithm 1 achieves 100% scene-cut detection (all cuts identified within ± 2 frames). BD-rate vs. fixed 300 is $+0.92\%$ (within measurement noise)—the safety guarantee is obtained at near-zero BD-rate cost. VMAF frame curves confirm that fixed 300 exhibits quality drift (1–2 VMAF drop) after scene boundaries, while Algorithm 1 recovers immediately via adaptive I-frame insertion.

Interpretation: Algorithm 1 captures the -12% GOP extension gain of fixed 300 while providing the safety of fixed 30. Existing codecs face a binary choice between long-GOP efficiency and scene-cut safety; the commutator norm eliminates this trade-off.

K. Experiment 10 — QP Redistribution RD Analysis

Purpose: Characterize the RD behavior of NormMAP QP redistribution independently from GOP extension.

Results (Table XIX): NormMAP (s=0.15) achieves -52 to -65% file-size reduction at VMAF 83–88 across all motion categories (Table XII). The VMAF-based BD-rate is -1.26% on average—near-neutral, confirming that NormMAP QP redistribution operates at a perceptually equivalent point on the same RD curve.

Interpretation: NormMAP QP redistribution does not improve RD efficiency; it operates at a different point on the *same* RD curve by concentrating distortion in near-commutative (perceptually non-salient) regions. The file-size reduction and VMAF reduction are proportional movements along the curve. PSNR-based BD-rate shows apparent improvement (-15 to -27%, Table XX) because PSNR weights all pixels equally and is insensitive to the spatial redistribution of distortion—this resolves the "PSNR paradox" (Section VI-A).

TABLE XX
BD-RATE AND BD-PSNR.

Scene Type	BD-rate	BD-PSNR
Still	-27.22%	$+2.750$ dB
Low-motion	-25.20%	$+2.513$ dB
Med.-motion	-15.14%	$+1.421$ dB

TABLE XXI
PER-I-FRAME PSNR DISCONTINUITY $E[\xi_I]$. $E[\xi_I] > \varepsilon^*$ IN ALL CASES, CONFIRMING $D_{\mathrm{SC}}(R) < D_{\mathrm{conv}}(R)$.

Sequence	Motion	I-frame PSNR	P/B-frame PSNR	$E[\xi_I]$ (dB)
crowd_run	High	39.03	36.00	4.14
pedestrian_area	Low–Med	46.21	44.04	2.79
vidyo1	Low (VC)	45.86	44.38	2.59

VI. DISCUSSION

A. Formal Explanation of the PSNR Paradox

PSNR-based BD-rate shows -15 to -27% improvement (Table XX), while VMAF-based BD-rate shows $\pm 0\%$ (Table XIX). This divergence is explained by the spatial selectivity of NormMAP: distortion is concentrated in near-commutative blocks where PSNR is reduced but perceptual impact is minimal. PSNR weights all pixels equally and therefore overstates the improvement; VMAF, which models human visual sensitivity, gives the correct RD-neutral assessment.

The PSNR-based improvement is nevertheless informative: it quantifies the degree to which NormMAP redistributes distortion spatially. The following rate-distortion formulation explains why this redistribution is consistent with the RD inequality.

In conventional encoding, I-frame insertion introduces a quantization-error discontinuity ξ_I at each GOP boundary:

$$D_{\mathrm{conv}}(R) = D_{\mathrm{rd}}(R) + E[\xi_I] \cdot r_I. \tag{11}$$

Under SlimeCodec, I-frame resets are suppressed to rate r_{SC}:

$$D_{\mathrm{SC}}(R) = D_{\mathrm{rd}}(R) + \varepsilon^* \cdot r_{\mathrm{SC}}. \tag{12}$$

For $\varepsilon^* \le E[\xi_I]$ and $r_{\mathrm{SC}} \ll r_I$: $D_{\mathrm{SC}}(R) < D_{\mathrm{conv}}(R)$ at the same bit rate.

Quantitative verification (Table XXI): $E[\xi_I] \in [2.59, 4.14]$ dB across three sequences. $\varepsilon^* < E[\xi_I]$ in all cases, confirming the condition $D_{\mathrm{SC}}(R) < D_{\mathrm{conv}}(R)$ at equal bit-rate. For crowd_run ($E[\xi_I]$=4.14 dB, r_I=3.33%, r_{SC}=0.33%), the predicted per-GOP distortion reduction is:

$$\Delta D = E[\xi_I] \cdot r_I - \varepsilon^* \cdot r_{\mathrm{SC}} \approx 4.14 \times 0.0333 \approx 0.138 \text{ dB/GOP}. \tag{13}$$

Over a 600-frame sequence, approximately 20 I-frame resets are eliminated, yielding cumulative reduction $20 \times 0.138 \approx 2.76$ dB, consistent with the observed BD-PSNR of $+2.75$ dB (Table XX).

B. Commutator-Norm Preservation: Algebraic Fidelity

SlimeCodec GOP = 300 is 2.4–3.5$\times$ more faithful to the camera original's d-value distribution than standard GOP = 30,

TABLE XXII
ADAPTIVE NORMMAP RANGE SWEEP (720P, s=0.3, CRF 23). r=0.3 IS THE RECOMMENDED DEFAULT.

Mode	Size	ΔSize
Fixed	7.1 MB	−37.4%
Adap. r=0.3	5.6 MB	−50.2%
Adap. r=0.5	4.9 MB	−56.3%
Adap. r=0.67	4.5 MB	−60.3%
Adap. r=1.0	3.9 MB	−65.9%

as reported in Table XVI. This result is directly predicted by Theorem 2: I-frame quantization resets perturb $\bar{d}$ by $\Delta_I r_I$, and suppressing I-frames reduces this perturbation to $\Delta_I r_{\mathrm{SC}}$. The ratio $r_I/r_{\mathrm{SC}} \approx 4.8$ (GOP = 30 to GOP = 300 at 0.7% residual I-frame rate) matches the observed $3.5\times$ improvement.

This result introduces *algebraic fidelity* as a new quality metric: the deviation of the re-encoded d-value distribution from the camera original. Under this metric, SlimeCodec GOP = 300 is 2.4–3.5× more faithful than standard encoding. Algebraic fidelity may be relevant to applications that process the d-value stream downstream, e.g., NormMAP-guided selective decoding, surveillance anomaly detection, and any pipeline that uses commutator-norm statistics for content analysis.

C. Quality Metrics for NormMAP-Encoded Content

Table VIII reports NormMAP strength sweeps. Recommended default: Strength 0.3, Adaptive range 0.3 for pre-verified content; Quality preset (s=0.15) for general CC-licensed content (Table XII).

Finding 1: File size increase at low strength on high-d content. At Strength 0.15 on rotational content ($\gamma \approx 2$), size increases by 8.3%. This occurs because NormMAP's I-frame suppression saves fewer bits than its QP-offset overhead adds when $\gamma < 3$: most blocks have large d and receive QP protection (negative offset), increasing rather than decreasing the bit budget. The I/PB pre-screening criterion (Proposition 3) correctly identifies $\gamma < 3$ as a skip recommendation, validating the bidirectional utility of $O(1)$ content screening.

Finding 2: VMAF/SSIM/PSNR underestimate perceptual quality for NormMAP transforms . At Strength 0.6–0.8, VMAF falls to 38–50 yet no visible difference was observed in informal assessment; formal evaluation is recommended before production deployment at these settings. VMAF (83.46 at Strength 0.3) is recommended as the primary metric; Quality preset (s=0.15) yields VMAF 83–88 on Xiph.org CC sequences (Table XII).

Finding 3: Content-adaptive strength selection is well predicted by the I/PB ratio. High-γ natural content tolerates higher adaptive range with larger compression gains (Table XXII), while low-γ rotational content saturates at lower range.

D. Theory-Experiment Consistency Summary

Every quantitative result follows from the theoretical framework (86.5% coverage, Prop. 1; $r_s \geq 0.94$, Prop. 3; $3.5\times$ d-preservation, Thm. 2; PSNR paradox, Table XXI; VP9>AV1>H.264, ALTREF distance; threshold convergence, Table VI). No result is post-hoc.

E. Scope and Limitations

Preset strength and content generalization: Quality preset (s=0.15) achieves VMAF 83–88 at −52 to −65% on Xiph.org derf (γ=4.2–21.7, Table XII). Balanced and higher presets require visual verification prior to deployment; informal assessment by the authors is not a substitute for independent evaluation on unverified content.

High-d content: For content with uniformly high d-values ($\gamma < 3$)—rotational motion, particle systems, random noise, stochastic textures—NormMAP overhead can exceed GOP-extension gain. Empirically, Strength 0.15 on $\gamma \approx 2$ produced +8.3% size increase. The I/PB pre-screening criterion (Proposition 3) reliably identifies this regime: $\gamma < 3$ is a practical indicator to skip or reduce NormMAP encoding strength.

Perceptual metric limitations: VMAF/SSIM/PSNR underestimate perceptual quality for NormMAP content; algebraic fidelity (Section VI-B) is recommended as a supplementary metric (supplementary material).

WASM decoder scope: Current implementation reduces Canvas render-pipeline cost; full decode-stage skipping requires WebCodecs integration (future work).

AV1 QP redistribution: The NormMAP ROI side-data interface relies on `AV_FRAME_DATA_REGIONS_OF_INTEREST`, which libx264 reads but libaom-av1 currently ignores. QP redistribution results (Tables XIX–XX) are therefore validated on H.264 only. AV1 native ROI integration via `segmentation_map` or `aq_mode` binding is left as future work; GOP extension gains (Table VII, Table XVII) are unaffected.

Comparison with H.265/HEVC: Table XXIII reports iso-VMAF (≈85) file-size comparison for H.264 (plain), H.265 (plain), and H.264+NormMAP (s=0.15) on the five Xiph.org CC-licensed sequences. CRF values are not comparable across codecs; matching on VMAF provides a fair comparison.

On high-motion content (crowd_run, park_joy), H.264+NormMAP is 5–6% smaller than H.265 plain at iso-VMAF, demonstrating that the NormMAP QP redistribution is most effective when the commutator-norm map has high spatial variance. On low-motion content (pedestrian_area, rush_hour, vidyo1), H.265 plain is substantially smaller, reflecting H.265's superior inter-prediction on temporally redundant material. H.265+NormMAP is expected to combine both advantages but is left as future work.

Hypothesis 1 (Proposition 1) status: Near-commutativity of natural video is empirically validated across six datasets and four operator representations, but not proved from first principles. The proposition rests on $\rho_t \ll 1$, which is justified by temporal redundancy but not derived from a generative model of natural video. A first-principles proof would require a stochastic model of natural video statistics, which is beyond the scope of this paper.

TABLE XXIII
ISO-VMAF COMPARISON (VMAF≈85). NORMMAP ADVANTAGES ARE CONTENT-DEPENDENT: EFFECTIVE ON HIGH-MOTION, H.265 SUPERIOR ON LOW-MOTION. FILE SIZES IN MB.

Sequence	H.264	H.265	NormMAP	vs. 264	vs. 265
crowd_run	15.1	15.9	15.0	-0.6%	-6.1%
park_joy	15.8	16.8	16.0	$+1.2\%$	-4.8%
ped._area	3.6	2.2	3.5	-1.9%	$+61.9\%$
rush_hour	4.8	3.1	4.7	-4.0%	$+51.4\%$
vidyo1	0.7	0.5	0.7	$+7.0\%$	$+56.0\%$

F. Resolution Scaling of Near-Commutativity

Table XV reveals a monotonic relationship between resolution and near-commutative block fraction: 68% at 1080p, 73% at 720p, and 88–93% at 4K. This is a theoretically predicted consequence of the block-size–resolution relationship.

Formal argument: The commutator norm $d(A, B) = \|AB - BA\|_F$ scales with the magnitude of structural change within the 8×8 block. As resolution increases, each block covers a smaller physical area $\delta^2 \propto 1/N$ (where N is pixel count). For a fixed camera motion vector v, the displacement within one block decreases as $v/\sqrt{N}$, so $\|\Delta_t\|_F \propto 1/\sqrt{N}$. By Theorem 2:

$$d(F_{t-1}, F_t) \leq 2\|F_{t-1}\|_F \|\Delta_t\|_F \propto 1/\sqrt{N}. \tag{14}$$

The fractal content (mandelbrot, $\bar{d}$=20,495, maximum spatial complexity) still achieves 88% near-commutative fraction at 4K, consistent with equation (14): even for high-d content, the resolution scaling reduces per-block d sufficiently. This confirms that the resolution-scaling effect is content-independent and structurally guaranteed. For 8K ($N \times 4$ relative to 4K), the near-commutative fraction is expected to approach 95%+, making the method increasingly effective as display technology advances.

VII. CONCLUSION

This paper establishes the commutator norm $d(A, B) = \|AB - BA\|_F$ as a continuous, codec-orthogonal criterion that governs two independent layers of video encoding optimization.

At the **encode layer**, GOP extension from 30 to 300 frames yields -14.2% VMAF-based BD-rate improvement (Table XVII). Algorithm 1, driven by the commutator-norm safety criterion (Corollary 1), captures this gain while detecting scene boundaries with 100% accuracy at near-zero BD-rate cost ($+0.92\%$, Table XVIII). NormMAP QP redistribution achieves -52 to -65% file-size reduction at VMAF 83–88 through RD-neutral perceptual quality reallocation (Table XIX). The PSNR-based BD-rate of -15 to -27% (Table XX) reflects the spatial redistribution of distortion rather than a genuine RD efficiency gain, as confirmed by the VMAF-based BD-rate of $\pm 0\%$.

At the **decode layer**, the same block-level criterion enables selective Canvas rendering, reducing CPU by 68.7% (putImageData: -96.1%) and memory-bus transfer by 60–90%.

A single mathematical criterion—$d_{\text{eff}}(b, t) < \varepsilon^*$—governs both layers without inter-layer coordination, projecting the same algebraic structure onto compression, bandwidth, and compute dimensions simultaneously.

The theoretical framework (C1–C6) is validated across six datasets, four operator representations, and three codecs. An additional finding: SlimeCodec preserves the commutator-norm distribution of the camera original 2.4–3.5× more accurately than standard GOP = 30, introducing algebraic fidelity as a new quality metric.

On high-motion content at iso-VMAF, H.264+NormMAP is 5–6% smaller than H.265 plain; on low-motion content, H.265 is superior (Table XXIII). H.265+NormMAP is a natural next step.

Supplementary material: Appendix A (PSNR–perceptual correlation analysis) and Appendix B (resolution scaling stress tests) are available as supplementary material.

REFERENCES

[1] Apple Computer, "Pixel-difference scene-change detection," U.S. Patent 5,099,322, 1992.
[2] "SAD + DCT DC adaptive GOP," U.S. Patent App. 2007/0171972, 2007.
[3] Netflix, "Visual feature cluster encoding," U.S. Patent 11,818,375, 2023.
[4] "Hierarchical B-frame structure," U.S. Patent 8,873,627, 2014.
[5] Qualcomm, "UI-layer metadata variable GOP," WO 2024/059998, 2024.
[6] A. Böttcher and D. Wenzel, "How big can the commutator of two matrices be and how big is it typically?" *Linear Algebra Appl.*, vol. 403, pp. 216–228, 2008.
[7] P. Billingsley, *Probability and Measure*, 3rd ed. Wiley, 1995.
[8] Google WebM Project, "VP9 bitstream specification, §7.2," 2016.
[9] T. Berger, *Rate Distortion Theory*. Prentice-Hall, 1971.
[10] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inf. Theory*, vol. IT-18, pp. 460–473, 1972.
[11] ITU-T Rec. H.265 — ISO/IEC 23008-2, 2021.
[12] B. Bross et al., "Overview of VVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, 2021.
[13] Alliance for Open Media, "AV1 Spec. v1.0.0," 2023.
[14] M. Abdoli et al., "GOP-Based Latent Refinement," *Proc. ICASSP*, 2023.
[15] W. Han et al., "Quantitative Analysis of GoP Structure," *Proc. ICIP*, 2023.
[16] H. Wang et al., "Rate allocation for VVC," *Inf. Sci.*, vol. 678, 2025.
[17] D. Alexandre et al., "Hierarchical B-frame Video Coding Using Two-Layer CANF," *Proc. CVPR*, 2023.
[18] S. Chen et al., "Learning-Based Rate Control," *Sensors*, vol. 23, p. 3607, 2023.
[19] J. S. Gomes et al., "End-to-End Neural Video Compression: A Review," *IEEE Open J. Circuits Syst.*, 2025.
[20] T. M. Hoang et al., "Recent trending on learning based video compression," *Vis. Informatics*, vol. 5, 2021.
[21] J. E. Saethre et al., "Combining Frame and GOP Embeddings," *Proc. CVPR*, pp. 18712–18721, 2024.
[22] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization," *IEEE Signal Process. Mag.*, vol. 15, 1998.
[23] M. Yang et al., "Adaptive GOP-size and QP in VVC," *IEEE Trans. Circuits Syst. Video Technol.*, 2024.
[24] ITU-T Rec. H.264 — ISO/IEC 14496-10, 2021.
[25] Xiph.Org Foundation, "Xiph.org test media," https://media.xiph.org/video/derf/, 2026.