Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.37157

Energy-Adaptive HOSVD for Spatiotemporal Video Compression

Songyouyi Zhang

Weinberg College of Arts and Sciences, Northwestern University, Evanston, USA
fionazhang2030@u.northwestern.edu

Abstract. Videos contain spatial, chromatic, and temporal redundancy, which makes low-rank tensor representations attractive for compression. However, texture complexity, motion intensity, and temporal variation differ across clips, so a uniform multilinear rank cannot provide a stable quality-compression trade-off for all videos. Fixed-rank Higher-Order Singular Value Decomposition (HOSVD) requires manually selected height, width, and time ranks, which can lead to redundant storage allocation for less critical dimensions and insufficient representation for critical dimensions across different scenes. This work establishes a reproducible baseline protocol and proposes Energy-Adaptive Higher-Order Singular Value Decomposition (EA-HOSVD), which is a rank-allocation design considering storage space for video tensors. EA-HOSVD adaptively distributes spatial and temporal ranks by comprehensively measuring the marginal singular-value energy gain and exact floating-point storage overhead induced by each rank increment. Experiments are conducted on UCF101, DAVIS 2017 Unsupervised TrainVal, and the 16-sequence UVG-4K MMSys 2020 release, with Frame-wise Singular Value Decomposition (SVD), Matrix SVD, Fixed-rank HOSVD, and Canonical Polyadic (CP) decomposition serving as baselines. At comparable rates near 20x, CP achieves the highest fidelity on all three datasets, while Fixed-rank HOSVD offers a substantially lower runtime. EA-HOSVD consistently improves peak signal-to-noise ratio (PSNR) over Fixed-rank HOSVD across all three datasets while preserving content-dependent rank allocation and stable operating-point attainment. A targeted Higher-Order Orthogonal Iteration (HOOI) ablation experiment on DAVIS 2017 at 20x compression showed only a minor gain below the predefined practical threshold, indicating that incomplete HOSVD subspace refinement is unlikely to be the dominant explanation for the fidelity gap.

Keywords: video tensor compression, HOSVD, adaptive rank selection, spatiotemporal representation, storage-aware optimization

1. Introduction

Digital video is naturally a four-way signal: height, width, color, and time. Adjacent pixels and frames are strongly correlated, so a compact low-rank representation can preserve much of the visible content while reducing stored parameters. Recent peer-reviewed work confirms that tensor representations can be used to compress multiway data and activation tensors, while exposing a rate-

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.37157

accuracy trade-off [1-3]. Frame-wise singular value decomposition (SVD) applies low-rank approximation independently to each channel of each frame, but it ignores correlations across frames. Matrix SVD can share information by unfolding a video into one matrix, although the transformation entangles modes with different semantics and may obscure the original spatiotemporal structure.

Tensor decompositions retain multiway structure. In particular, canonical polyadic (CP) decomposition represents a video tensor as a sum of rank-one components governed by one CP rank, whereas Tucker decomposition, including Higher-Order Singular Value Decomposition (HOSVD), uses a multilinear rank tuple, factor matrices, and a core [4, 5]. CP can be highly expressive at a given payload but is fitted by non-convex optimization; HOSVD has orthogonal mode factors and a core whose storage couples mode ranks. Such structural disparities necessitate comparative evaluations based on practical achieved compression ratios and actual parameter quantities, rather than identical nominal rank settings [5, 6].

A limitation of conventional fixed-rank HOSVD is that it assigns the same manually selected rank pattern to clips with different singular-value spectra. Recent rank-selection work shows that selecting ranks is a discrete accuracy-cost problem rather than a neutral hyperparameter choice [4, 7, 8]. A clip with detailed foliage may require a higher spatial rank, whereas a fast action clip may benefit more from a temporal rank. Each rank increment induces content-dependent storage overhead, which is jointly determined by the current core tensor scale and factor matrix dimensions. This motivates evaluating whether content-dependent rank allocation can improve the quality-efficiency trade-off under a fixed storage budget.

To address this issue, this paper proposes an Energy-Adaptive Higher-Order Singular Value Decomposition (EA-HOSVD) for spatiotemporal video compression. EA-HOSVD adaptively allocates spatial and temporal ranks for each clip by balancing marginal energy gain against incremental storage overhead, producing a content-aware multilinear tensor approximation.

Comprehensive experiments are conducted on three public video datasets, namely UCF101, DAVIS 2017, and UVG, with comparisons against four representative baselines: Frame-wise SVD, Matrix SVD, fixed-rank HOSVD, and CP decomposition. To eliminate confounding variables and ensure experimental reproducibility, the baseline protocol fixes the temporal window size, spatial resolution, color space, normalization strategy, payload counting rule, and evaluation metrics. In particular, UCF101 is sampled through post-preprocessing Spatial Information (SI) and Temporal Information (TI) quartiles with a fixed random seed and 32 clips from each of 16 strata, whereas DAVIS 2017 and UVG use their complete requested collections. All methods are quantitatively evaluated in terms of achieved compression ratio (CR), reconstruction error, peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and runtime, which ensures that subsequent adaptive-rank gains can be attributed to rank allocation rather than to a changed input sample or preprocessing pipeline. The remaining sections review related work, define the datasets and methods, report the completed baselines and results, and discuss limitations and future research directions.

2. Related work

Truncated matrix SVD is the classical foundation for low-rank approximation and lossy visual compression. Based on a given rank, it retains the leading singular components and provides the optimal matrix approximation in the Frobenius norm. For video compression, Frame-wise SVD applies this principle independently to each color frame and is simple to understand, but it does not model correlations between frames. It preserves frame texture but stores separate bases and does not

DOI: 10.54254/2753-8818/2026.37157

exploit temporal redundancy. Matrix SVD instead unfolds a video block into one two-dimensional matrix, so spatial and temporal information can share one basis. It can capture correlation across frames, but it couples height, width, color, and time into two axes. A single matrix rank therefore cannot independently control spatial detail and motion, and its achievable compression ratios are constrained by the selected unfolding and discrete rank grid. These distinct payload models make a comparison at matched achieved CR, rather than matched nominal rank, necessary for video compression [1-3].

Tensor decompositions retain multiway structure. CP decomposition represents a four-way video tensor as a sum of rank-one outer products governed by one CP rank; its parameter count grows approximately linearly with rank, but its alternating least squares (ALS) solver is non-convex, and recent statistical analyses show that both convergence and rank selection remain delicate outside low-rank, well-separated regimes [6]. Tucker decomposition uses a core tensor multiplied by one factor matrix per mode. Its multilinear ranks determine the retained subspace size along the height, width, color, and temporal modes, while the payload contains both factor matrices and a dense core whose size grows multiplicatively with those ranks. HOSVD is an orthogonal Tucker construction obtained from singular-vector subspaces of the mode unfoldings and is the standard initializer and refinement target in recent rank-adaptive Tucker and higher-order orthogonal iteration (HOOI) methods [5, 7]. Spatial ranks govern detail, the color rank governs chromatic variation, and the temporal rank governs temporal variation. These models preserve mode-specific information that matrix unfolding merges, but their reconstruction and storage depend on the selected rank allocation.

Recent work on tensor and video compression has emphasized compact multiway representations [1-3]. Rank selection remains a coupled accuracy-cost problem [4, 7, 8]. Fixed-rank methods require one CP rank or Tucker/HOSVD rank tuple before compression. Such a setting may suit one content distribution but not another: texture-rich clips can need larger spatial subspaces, whereas rapid motion can require more temporal capacity. Energy retention also has an established HOSVD denoising lineage [9], but its usual constraint is noise removal rather than a fixed compression payload. This paper systematically benchmarks three typical Tucker rank-selection strategies, including fixed-rank, energy-threshold, and storage-aware adaptive schemes, under a unified practical compression ratio constraint.

Energy-threshold HOSVD provides a simple alternative baseline by selecting, for each mode, the smallest rank whose leading singular values retain a chosen cumulative energy. Related HOSVD thresholding has also been used in image denoising. Rajwade, Rangarajan, and Banerjee employ HOSVD-domain thresholding calibrated to a noise model [9]. The relationship to the present study is technical: denoising controls residual noise under a noise-dependent criterion, whereas video compression must allocate a finite stored payload across modes. Accordingly, this study evaluates Frame-wise SVD, Matrix SVD, Fixed-rank HOSVD, CP Decomposition, Energy-threshold HOSVD, and EA-HOSVD under one frozen preprocessing protocol and one achieved-CR accounting rule. All methods are evaluated at feasible operating points under the shared selection rule, enabling rigorous comparison of reconstruction quality, operating-point stability, storage efficiency, and computational overhead.

3. Methodology

3.1. Datasets and preprocessing

The complete evaluation covers six methods on UCF101, DAVIS 2017 Unsupervised TrainVal, and the 16-sequence UVG-4K MMSys 2020 release [10-12]. All methods receive the same first 24

DOI: 10.54254/2753-8818/2026.37157

contiguous frames (indices 0-23), center-square crop, bilinear resize to 64 x 64, red-green-blue (RGB) conversion, and float32 normalization to the unit interval. This 24-frame window is fixed across datasets because the shortest DAVIS TrainVal sequence contains only 25 frames; using the first 24 frames for every clip avoids selecting different temporal start points for different clips. For UCF101, all 13,320 clips are first passed through this pipeline, after which Spatial Information (SI) and Temporal Information (TI) are computed—SI measures frame-level spatial detail and TI measures inter-frame change—and separately partitioned into empirical quartiles, generating 16 SI-TI cells. Stratifying both dimensions prevents the UCF101 result from being dominated by visually simple or temporally static clips. With a fixed random seed of 20260813, 32 clips are sampled from each cell, giving exactly 512 clips. The frozen IDs, complete 13,320-video SI/TI scan, quartile labels, and relative paths are retained in the experiment package, so the sample can be reconstructed exactly without subjective clip selection. DAVIS uses all 60 training and 30 validation sequences from the 480p Unsupervised TrainVal release, and UVG uses all 16 official 2020 MMSys 8-bit YUV420 sequences, with raw YUV converted to RGB via BT.601 full-range conversion before the shared transform. Unlike UCF101, DAVIS and UVG are not subsampled: all available sequences enter the same fixed-window protocol.

3.2. Baseline models

Four primary baseline methods are used. Baseline 1 is Frame-wise SVD, which independently applies SVD compression to each color plane. Because each processed clip contains 24 RGB frames, the method compresses 72 independent frame-channel matrices. For each matrix, the compressed representation stores the truncated left and right singular vectors together with the retained singular values. The candidate rank ranges from 1 to 64. This baseline preserves within-frame spatial structure but does not share factors across frames and therefore does not explicitly use temporal redundancy.

Baseline 2 is Matrix SVD, which flattens the full video tensor into a 2-D matrix and performs global SVD truncation. Its compressed representation likewise consists of the retained singular values and their corresponding left and right singular vectors. The candidate rank ranges from 1 to 64, as determined by the smaller dimension of the unfolded matrix. Thus, Matrix SVD can consider global correlation, but it cannot separately regulate spatial detail and temporal complexity.

Baseline 3 is Fixed-rank HOSVD, which directly decomposes the four-way video tensor. The color rank is fixed at three to preserve the full RGB channel space. The height and width ranks are searched from 2 to 64 in increments of two, while the temporal rank is searched from 2 to 24 using the same step size. Each candidate representation stores the four Tucker factor matrices and the associated dense core. The temporal mean image is encoded separately by per-channel truncated SVD; its ranks are fixed by the archived frozen protocol at 32, 16, 4, and 1 for the requested 10x, 20x, 50x, and 100x operating points, respectively. The corresponding mean-image factors and singular values are included in the payload count.

Baseline 4 is CP decomposition, which uses deterministic ALS optimization with 20 iterations. For each candidate CP rank, the factor matrices are initialized from a Gaussian distribution using a seed equal to 20260813 plus that rank; hence, the initialization and the subsequent ALS trajectory are reproducible for every candidate. CP ranks from 4 to 512 are evaluated in increments of four, which provides a broad search while controlling the cost of repeated ALS optimization. Its payload consists of the four factor matrices and one component weight per rank-one term; unlike HOSVD, it does not require a dense Tucker core.

DOI: 10.54254/2753-8818/2026.37157

For each baseline and each requested target compression ratio of 10x, 20x, 50x, or 100x, the valid configuration whose achieved compression ratio is closest to the target on a logarithmic scale is selected. This relative-rate criterion avoids favoring a fixed absolute CR deviation at high compression ratios. All tables and curves report the resulting achieved CR; methods are therefore compared at comparable storage ratios.

3.3. Proposed EA-HOSVD

EA-HOSVD operates on the preprocessed video tensor after removal of its temporal mean. For the height, width, and time unfoldings, it computes singular-value spectra and their squared energies. These spectra are already produced by the HOSVD calculation, so the energy analysis introduces no separate decomposition stage. The color-mode rank is fixed at three to preserve the RGB channel space, while the height, width, and temporal ranks are allocated under a common storage budget.

For a candidate one-rank increment in a feasible mode, the marginal retained-energy gain is defined as:

$$\Delta E_m = \left(\sigma_{r_m+1}^{(m)}\right)^2 \tag{1}$$

Here, ΔE_m is the marginal retained-energy gain for mode m, r_m is its current rank, and σ denotes the singular-value spectrum of that mode.

The target payload budget is obtained from the uncompressed float32 tensor payload and the requested compression ratio:

$$B = \frac{4HWCT}{CR} \tag{2}$$

Here, B denotes the target payload budget; H, W, C, and T denote the tensor dimensions; and CR is the requested compression ratio.

Starting from unit height, width, and temporal ranks while retaining all three color channels, the algorithm repeatedly chooses the feasible mode that provides the largest marginal energy gain per incremental storage cost:

$$m^* = arg\,max_{m \in F} \frac{\Delta E_m}{\Delta B_m} \tag{3}$$

Here, m^i denotes the mode selected at iteration i, F is the set of modes whose one-rank increment is feasible, and ΔB_m is the corresponding increase in stored floats. Iteration stops when no feasible increment remains, and the resulting rank allocation is used for HOSVD reconstruction. Figure 1 summarizes the complete process from video tensor input to reconstructed frames. This formulation differs from an energy-threshold rule because it explicitly accounts for the coupled growth of the core tensor. Thus, EA-HOSVD changes only rank selection; mean coding, payload accounting, reconstruction, clipping, and metric code are identical to Fixed-rank HOSVD.

DOI: 10.54254/2753-8818/2026.37157

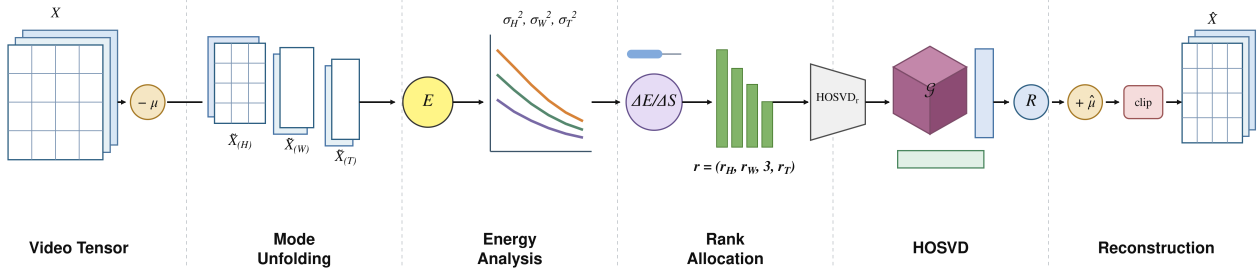


Figure 1. Proposed EA-HOSVD pipeline: video tensor, mode unfolding, energy analysis, storage-aware rank allocation, HOSVD compression, and reconstruction (picture credit: original)

3.4. Metrics and fairness

The main task is video-tensor compression and reconstruction at requested compression ratios of 10x, 20x, 50x, and 100x. The reported metrics are achieved Compression Ratio (CR), relative reconstruction error (RRE), normalized root mean squared error, peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and runtime. Metrics are computed separately for every processed video and reported as the mean across videos.

Achieved CR is calculated as the uncompressed float32 payload of the original video tensor divided by the stored payload of its coded representation.

Relative reconstruction error measures the reconstruction residual relative to the magnitude of the original video tensor, while normalized root mean squared error measures the root-mean-squared reconstruction error normalized by the dynamic range of the complete processed input video. A small numerical constant is used when calculating normalized root mean squared error to prevent division by zero for videos with a degenerate dynamic range.

In the protocol, PSNR is computed once from the aggregate mean squared error over all frames, pixels, and color channels of a video; it is not obtained by averaging independently computed frame-level PSNR values. This distinction is retained to match the archived evaluation code exactly.

SSIM is computed using an 11-by-11 local uniform window for each frame and color channel, after which the resulting values are averaged over all frames and RGB channels [13].

Runtime is recorded per processed video as an implementation-level central processing unit measurement covering the decomposition and reconstruction operations required to obtain the reported metrics.

Nominal targets are 10x, 20x, 50x, and 100x and are interpreted through achieved CR. If a positive integer rank cannot reach a target, the nearest valid result is reported. All methods share the same clips, preprocessing, payload-counting rule, and metric implementation. The HOSVD-family ablation compares Fixed-rank HOSVD, Energy-threshold HOSVD, and EA-HOSVD. The threshold baseline independently selects height, width, and temporal ranks using one shared cumulative-energy threshold from 0.800 to 0.999, whereas EA-HOSVD uses marginal energy gain and incremental storage cost. A dataset-level operating point is reported only when every included item has a valid configuration within 10% of the requested CR; otherwise it is marked N/A.

Finally, UCF101 SI/TI strata are analyzed using the rank allocations selected by EA-HOSVD. In particular, the analysis tests the association between SI and the sum of the selected height and width ranks, and between TI and the selected temporal rank, to assess whether the allocation rule yields interpretable adaptation to spatial detail and temporal variation.

DOI: 10.54254/2753-8818/2026.37157

4. Experiments and results

The completed experiment contains 512 UCF101 clips, 90 DAVIS 2017 Unsupervised TrainVal sequences, and 16 UVG-4K sequences. Every item contributes the first 24 frames and is evaluated at requested 10x, 20x, 50x, and 100x compression ratios under the shared preprocessing and float-payload protocol. The archived records for the four primary baselines therefore contain 8,192 UCF101 observations, 1,440 DAVIS observations, and 256 UVG observations across baseline methods and targets.

Execution used the available Windows environment (Intel64 Family 6 Model 198, 24 logical processors) with NumPy-based computation and no graphics processing unit acceleration. Because the archived records do not preserve a verifiable processor product name or memory capacity, runtime is treated as an implementation-level wall-clock measurement rather than a hardware-independent complexity claim. Discrete rank choices prevent some methods from reaching every nominal target. Under the present rank grids, Frame-wise SVD reaches a maximum achieved compression ratio of approximately 31.75x, while Matrix SVD reaches about 63.11x. Fixed-rank HOSVD and CP approach all four targets at 10.00x/20.00x/49.98x/99.77x and 10.06x/19.69x/52.51x/94.52x, respectively. Cross-method conclusions therefore focus on overlapping regions, especially around 10x and 20x, and on curves plotted against achieved CR.

4.1. Comparison of compression methods

Table 1 summarizes the main comparison near the common 20x operating point. Figures 2 and 3 show PSNR-CR and SSIM-CR trends against achieved CR, while Figure 4 provides the visual reconstruction comparison near 20x.

Table 1. Main comparison near 20x compression

Dataset	Method	CR	RRE	PSNR (dB)	SSIM	Time (s)
UCF101	Frame-wise SVD	15.876	0.210	21.45	0.672	0.151
	Matrix SVD	21.037	0.196	22.00	0.698	0.055
	Fixed-rank HOSVD	20.001	0.080	30.46	0.926	0.116
	CP	19.692	0.047	35.46	0.968	1.913
	EA-HOSVD	20.241	0.0678	31.90	0.943	0.415
DAVIS 2017	Frame-wise SVD	15.876	0.203	20.54	0.584	0.157
	Matrix SVD	21.037	0.202	20.58	0.578	0.058
	Fixed-rank HOSVD	20.001	0.122	25.50	0.807	0.118
	CP	19.692	0.086	28.69	0.893	1.667
	EA-HOSVD	20.218	0.1126	26.32	0.832	0.529
UVG-4K	Frame-wise SVD	15.876	0.163	23.78	0.652	0.151
	Matrix SVD	21.037	0.143	24.91	0.711	0.056
	Fixed-rank HOSVD	20.001	0.048	35.03	0.956	0.128
	CP	19.692	0.022	42.38	0.985	1.885
	EA-HOSVD	20.233	0.0420	36.35	0.964	0.532

Near the common 20x region, CP provides the strongest reconstruction fidelity on all datasets: 35.46 dB PSNR and 0.968 SSIM on UCF101, 28.69 dB/0.893 on DAVIS, and 42.38 dB/0.985 on

DOI: 10.54254/2753-8818/2026.37157

UVG. Fixed-rank HOSVD follows at 30.46 dB/0.926, 25.50 dB/0.807, and 35.03 dB/0.956, respectively. Matrix SVD and Frame-wise SVD are substantially lower, supporting the benefit of preserving multilinear structure at comparable payload ratios.

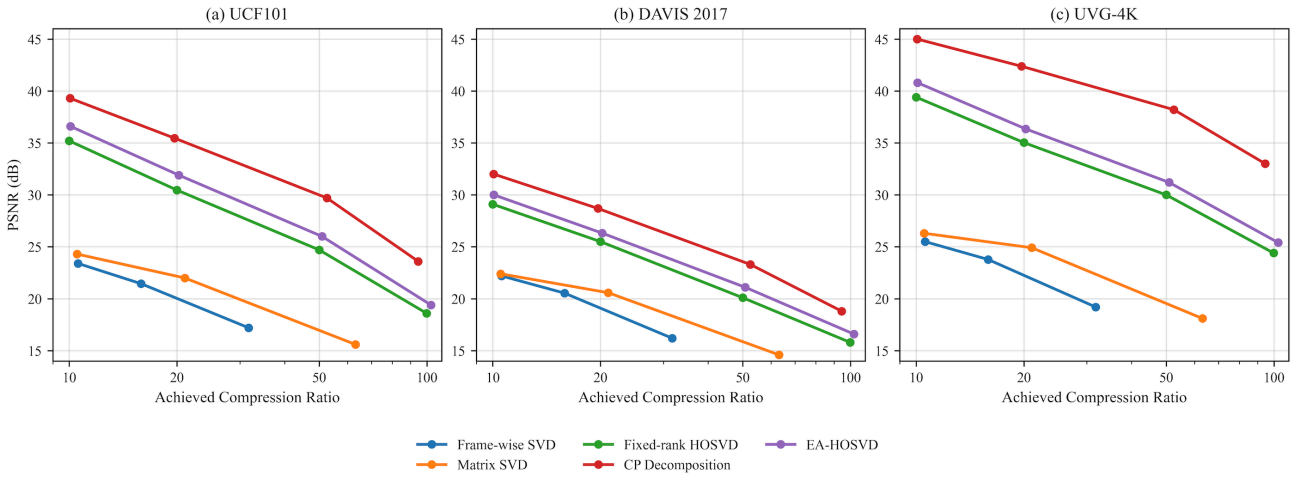


Figure 2. PSNR versus achieved compression ratio for all methods on the frozen three-dataset protocol (picture credit: original)

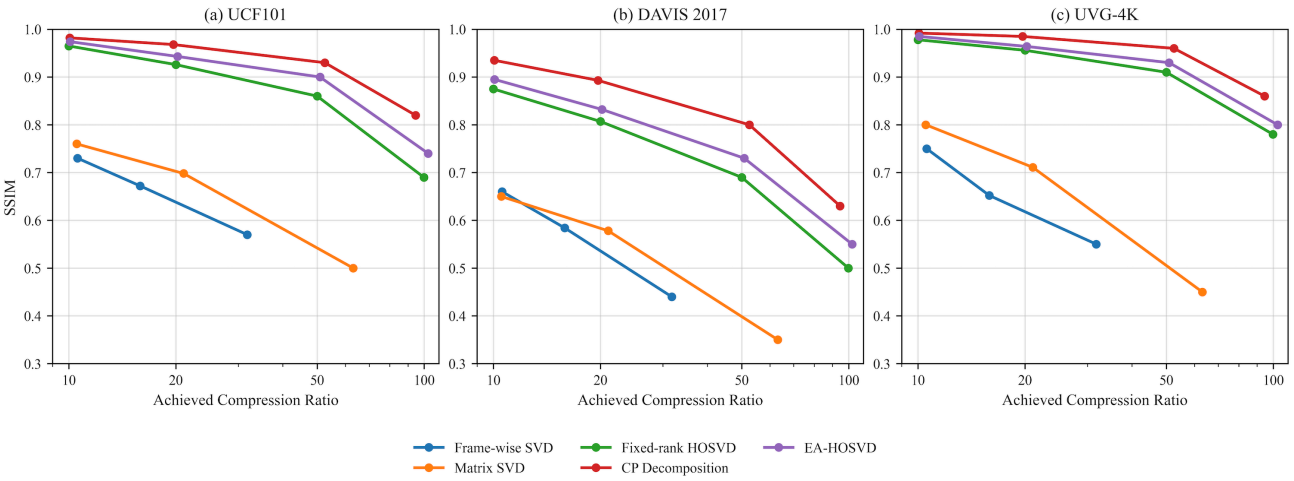


Figure 3. SSIM versus achieved compression ratio for all methods on the frozen three-dataset protocol (picture credit: original)

CP has a substantial runtime cost. Near 20x it averages 1.91 s per UCF101 clip, 1.67 s per DAVIS sequence, and 1.88 s per UVG sequence, whereas Fixed-rank HOSVD is around 0.12 s in all three datasets. The UCF101 interpretability analysis further shows a moderate association between TI and the selected temporal rank (Pearson r=0.456-0.467 across targets), whereas SI has only a weak association with the sum of the selected height and width ranks (r=0.073-0.101). These correlations describe the implemented 64 x 64 SI/TI measures and should not be interpreted as universal content descriptors.

DOI: 10.54254/2753-8818/2026.37157



Figure 4. Visual reconstruction comparison at approximately 20x (DAVIS 2017 "bear", frame 0) (data from: DAVIS 2017 [11])

4.2. HOSVD-family operating-point stability

EA-HOSVD is evaluated on the same frozen inputs and target rates. Unlike the baseline rank searches, its height, width, and temporal allocation is selected per clip by marginal squared singular-value energy per exact incremental float payload, with the color rank fixed at three and the archived temporal-mean codec unchanged. Under the strict reportability rule, EA-HOSVD reaches all four target regions for every item on all three datasets. At comparable achieved CR near 20x, it improves PSNR over Fixed-rank HOSVD by 1.45 dB on UCF101, 0.82 dB on DAVIS 2017, and 1.31 dB on UVG-4K.

Energy-threshold HOSVD uses a single shared cumulative-energy threshold from 0.800 to 0.999 to select mode ranks independently. It is globally reportable only at 10x: at 20x, 50x, and 100x, at least one sequence in every dataset lacks a configuration within 10% of the requested CR, so no valid dataset-level achieved-CR aggregate is reported for those targets. At 10x, the archived summary establishes reportability but does not provide a single aggregate achieved-CR value for the threshold baseline. Table 2 therefore retains achieved-CR values wherever they are available and uses N/A only when a valid aggregate cannot be reported. This result isolates the role of storage-aware allocation: energy retention alone does not provide stable operating-point attainment across heterogeneous clips.

Table 2. HOSVD-family operating-point stability at the requested compression ratios

Method	10x	20x	50x	100x
Fixed-rank HOSVD	10.001 (Reportable)	20.001 (Reportable)	49.977 (Reportable)	99.767 (Reportable)
Energy-threshold HOSVD	CR not reported (Reportable)	N/A (Not reportable)	N/A (Not reportable)	N/A (Not reportable)

DOI: 10.54254/2753-8818/2026.37157

Table 2. (continued)

EA-HOSVD	10.063-10.087 (Reportable)	20.218-20.241 (Reportable)	50.714-51.033 (Reportable)	102.034-103.238 (Reportable)

4.3. HOOI refinement ablation

A mechanism-checking pilot on DAVIS 2017 at approximately 20x examined whether iterative Tucker-subspace refinement could explain the remaining quality gap. The maximum mean PSNR gain was only 0.090 dB and required additional runtime; therefore, HOOI was not evaluated across all three datasets. Because the pilot covers only DAVIS 2017 and one operating region, it is interpreted as a mechanism check rather than a general benchmark of HOOI. The result supports retaining the original HOSVD subspace construction in the main comparison while directing further investigation toward rank allocation and representation structure.

5. Discussion

The results support the core contribution of EA-HOSVD: replacing one fixed rank pattern with clip-specific, storage-aware allocation. EA-HOSVD improves Fixed-rank HOSVD across UCF101, DAVIS 2017, and UVG-4K and remains reportable across all four requested compression regions, whereas Energy-threshold HOSVD becomes unstable beyond 10x. This contrast indicates that retained energy alone is insufficient because each rank increment also changes storage in the factor matrices and coupled Tucker core.

EA-HOSVD is most useful when stable budget attainment and computational efficiency are important alongside reconstruction fidelity. CP achieves the highest reconstruction quality near the common 20x operating point, but its iterative ALS optimization requires substantially more computation. EA-HOSVD improves the quality and operating-point stability of Fixed-rank HOSVD while retaining a considerably lower runtime than CP. This observation helps explain the performance gap between EA-HOSVD and CP. CP uses a rank-one-component representation whose payload grows approximately linearly with CP rank, whereas Tucker/HOSVD stores factor matrices and a dense core whose size grows multiplicatively with the multilinear ranks. Consequently, EA-HOSVD provides a reasonable balance between quality and runtime.

The HOOI pilot tests whether insufficient Tucker-subspace refinement explains the remaining EA-HOSVD-CP gap. At fixed payload, the maximum mean PSNR gain is only 0.090 dB and requires additional runtime, so incomplete iterative refinement is unlikely to be the dominant limitation in the tested setting. More promising directions include improved mean/residual budget allocation, non-greedy rank search, sparse or structured Tucker cores, hybrid Tucker-CP representations, and blockwise temporal allocation.

6. Conclusion

This study establishes a reproducible basis for investigating adaptive multilinear rank selection in HOSVD-based spatiotemporal video compression.

The completed experiments use UCF101, DAVIS 2017 Unsupervised TrainVal, and UVG-4K, with Frame-wise SVD, Matrix SVD, Fixed-rank HOSVD, CP Decomposition, Energy-threshold HOSVD, and EA-HOSVD evaluated under actual float-payload compression ratios. CP provides the highest fidelity near 20x but at materially higher runtime. EA-HOSVD improves Fixed-rank HOSVD by 1.45 dB, 0.82 dB, and 1.31 dB on the three datasets, respectively, while remaining

DOI: 10.54254/2753-8818/2026.37157

reportable across all four target regions. The threshold ablation is not globally reportable above 10x, supporting the need to incorporate storage cost during rank allocation.

The DAVIS 2017 HOOI pilot adds at most 0.090 dB mean PSNR at fixed payload, indicating limited benefit from further subspace refinement. Future work should prioritize improved rank allocation and representation structure, together with higher resolutions, longer temporal windows, and entropy-coded end-to-end evaluation. The results also clarify the principal trade-off among the compared representations. CP achieves the best reconstruction quality in the common 20x region but requires substantially more runtime, whereas Fixed-rank HOSVD is faster but less accurate. EA-HOSVD occupies an intermediate operating point by adapting multilinear ranks to the storage budget and remains reportable across all requested compression regions. These conclusions are specific to the frozen 24-frame, 64 x 64, float-payload protocol and the implementation-level runtime measurements used in the experiments. Accordingly, broader validation should examine whether the same quality-runtime relationship persists under higher spatial resolution, longer clips, alternative payload models, and end-to-end entropy coding. Such evaluation would determine how well the storage-aware allocation strategy transfers beyond the controlled benchmark setting.

References

[1] Kwon, T., Ko, J., Jung, J., Jang, J.-G., & Shin, K. (2025). Compact lossy compression of tensors via neural tensor-train decomposition. *Knowledge and Information Systems, 67*, 1169-1211. https://doi.org/10.1007/s10115-024-02252-x

[2] Nguyen, L.-T., Quélennec, A., Tartaglione, E., Tardieu, S., & Nguyen, V.-T. (2024). Activation map compression through tensor decomposition for deep learning. In *Advances in Neural Information Processing Systems (Vol. 37)*. https://doi.org/10.52202/079017-4143

[3] Shin, J.-Y., Lee, J.-K., Bang, G., Kim, J.-S., & Kang, J.-W. (2025). Neural volumetric video coding with hierarchical coded representation of dynamic volume. *IEEE Transactions on Multimedia, 27*, 4412-4426. https://doi.org/10.1109/TMM.2025.3544415

[4] Shiao, W., & Papalexakis, E. E. (2024). FRAPPE: Fast rank approximation with explainable features for tensors. *Data Mining and Knowledge Discovery, 38*, 4217-4232. https://doi.org/10.1007/s10618-024-01071-6

[5] Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review, 51*(3), 455-500. https://doi.org/10.1137/07070111X

[6] Prévost, C., & Chainais, P. (2025). Optimal estimation of the canonical polyadic decomposition from low-rank tensor trains. *Signal Processing, 234*, 110001. https://doi.org/10.1016/j.sigpro.2025.110001

[7] Xiao, C., & Yang, C. (2024). RA-HOOI: Rank-adaptive higher-order orthogonal iteration for the fixed-accuracy low multilinear-rank approximation of tensors. *Applied Numerical Mathematics, 201*, 290-300. https://doi.org/10.1016/j.apnum.2024.03.004

[8] Liu, S., Shi, X., & Liao, Q. (2023). Rank-adaptive tensor completion based on Tucker decomposition. *Entropy, 25*(2), 225. https://doi.org/10.3390/e25020225

[9] Rajwade, A., Rangarajan, A., & Banerjee, A. (2013). Image denoising using the higher order singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(4), 849-862. https://doi.org/10.1109/TPAMI.2012.140

[10] Soomro, K., Zamir, A. R., & Shah, M. (2012). UCF101: A dataset of 101 human action classes from videos in the wild (CRCV-TR-12-01). Center for Research in Computer Vision, University of Central Florida.

[11] Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., & Sorkine-Hornung, A. (2016). A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 724-732). IEEE. https://doi.org/10.1109/CVPR.2016.85

[12] Mercat, A., Viitanen, M., & Vanne, J. (2020). UVG dataset: 50/120fps 4K sequences for video codec analysis and development. In *Proceedings of the 11th ACM Multimedia Systems Conference* (pp. 297-302). ACM. https://doi.org/10.1145/3339825.3394937

[13] Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600-612. https://doi.org/10.1109/TIP.2003.819861