Energy-Efficient AI Video Analytics with MPEG-5 LCEVC: GPU-Optimised Results

Guendalina Cobianchi
V-Nova Ltd
London, UK
guendalina@v-nova.com

Manzoor Razaak
V-Nova Ltd
London, UK
manzoor.razaak@v-nova.com

Ben Cope
Intel Corporation
London, UK
ben.cope@intel.com

Palanivel Guruvareddiar
Intel Corporation
Chandler, USA
palanivel.guruvareddiar@intel.com

***Abstract*— The scaling of visual Artificial Intelligence (AI) for applications like smart cities and retail analytics is increasingly constrained by data pipeline inefficiencies rather than model computation. Conventional video codecs necessitate full-resolution decoding and extensive data movement, resulting in unnecessary power expenditure and GPU under-utilization. This paper evaluates MPEG-5 Low Complexity Enhancement Video Coding (LCEVC) [1] as a hierarchical solution that enables "base-first" operation, allowing AI inference to process only the low-resolution base stream. Using Intel's Deep Learning Streamer framework on an integrated GPU architecture (Intel Core Ultra 7 258V), we quantify LCEVC's impact in a high-density, eight Ultra High Definition (UHD) parallel AI inference stream environment compared to native AVC. Results confirm LCEVC-enhanced AVC achieves up to 3.4x pipeline speed-up, 71% lower average power consumption per channel, and a crucial 83% reduction in memory bandwidth utilization. These findings position LCEVC as a practical enabler for energy-efficient, high-density edge AI systems.**

I. Introduction

The escalating global deployment of intelligent video analytics necessitates a change in basic assumptions towards sustainable computing practices. As AI workloads, particularly in physical AI, expand, the dominant computational cost is shifting from arithmetic operations (FLOPS) to data orchestration, specifically the energy and time budgets consumed by video decoding, memory transfer, and pre-processing. This reliance on legacy video formats (e.g., AVC, HEVC) results in a system-wide bottleneck. These formats, designed for full-frame human consumption, compel the system to decode and move high volumes of pixel data, even when the AI model requires only a low-resolution input (e.g., 448x256 pixels). This overhead consumes excessive energy and contributes to "GPU starvation", where high-performance accelerators sit idle waiting for data movement across the memory fabric. Studies [2] indicate that I/O operations and data loading can consume up to 70% of overall model training time, highlighting data movement as the core inefficiency. Hierarchical, compute-aware data formats offer an architectural solution by enabling selective data access. This paper quantifies the resource-efficient properties of MPEG-5 LCEVC, a standardized layering technology that decouples high-resolution fidelity from low-resolution perception requirements. We evaluate LCEVC's ability to reduce power consumption and increase channel density on modern GPU hardware in a constrained edge computing environment.

Prior work [3] by the authors is the first known to show the performance benefits of scalable CODECs for AI pipelines for both SMPTE VC-6 and MPEG-5 LC-EVC. That work did not cover the key data orchestration benefits of the impact on memory bandwidth and power consumption. This paper extends the analysis to the impact on memory bandwidth where contributions include showing energy efficiency.

II. LCEVC: Architecture for Resource Efficiency

LCEVC is a codec-agnostic enhancement technology that achieves efficiency through a multi-layered structure. It transmits a low-resolution base stream, encoded by any base codec (e.g., AVC), together with a lightweight enhancement layer containing the residual information required for full-fidelity reconstruction.

This architecture enables base-first operation, which is highly advantageous for resource-efficient AI pipelines. Since tasks such as object detection typically require far lower resolution (448×256 in our tests), the system can decode and process only the LCEVC base stream, bypassing the more intensive enhancement layer.

Selective decoding alleviates several pipeline bottlenecks:

1. **Reduced decode load**: The GPU's fixed-function media engine processes a lower-resolution stream, lowering computational demand.
2. **Lower memory bandwidth**: Pixel traffic across the memory bus is significantly reduced. The base layer represents approximately 75–80% of total bitrate, yielding 20–25% transfer and storage savings when the enhancement layer is omitted. These intrinsic savings complement LCEVC's compression efficiency, previously validated by MPEG [4] and other studies [5].
3. **Lighter Pre-processing:** The computational cost of scaling the base layer down to the model's required input resolution is lower than scaling from full resolution.

While this paper focuses on LCEVC enhancing AVC due to its near-term deployment relevance, the same approach applies to next-generation codecs such as VVC. For tasks that occasionally require higher spatial detail (e.g., fine-grained recognition), the enhancement layer can be decoded on demand as a second pass, enabling a controlled balance between efficiency and fidelity. This selective, two-step inference mode will be investigated in future work.

III. Method: GPU-Centric Sustainable Evaluation

A. Hardware and Platform

Experiments were conducted on an edge-optimized platform featuring an integrated Intel Core Ultra 7 258V processor using the Intel Deep Learning Streamer framework [6]. The unified memory architecture emphasizes the cost associated with memory bandwidth, making it suitable for measuring real-world I/O efficiency gains. All measurements are GPU-only. The baseline was native AVC (H.264), compared against LCEVC-AVC.

B. Test Scenario and Pipeline

The scenario models a high-density, resource-constrained edge scenario dedicated to AI inference, where the objective is to maximize concurrent video channels processed within a

fixed platform Thermal Design Power (TDP) budget. This is characteristic of headless surveillance or smart-city nodes.

The AI task was object detection using the MobileNetV2-SSD model [7], along with object tracking, on high-resolution traffic clips recorded from UAV and available on open-source datasets, including the VisDrone dataset [8]. The pipeline ran eight asynchronous inferences on UHD streams in parallel to stress the system's resource limits. The LCEVC pipeline was configured for base-first operation, decoding only the low-resolution base stream for inference. The end-to-end pipeline is shown in Attachment 1. In a traditional pipeline the video decoding step decodes the full resolution video which not only means a higher video decode requirement it also has follow on impact on more complexity in the pre-processing stage.

C. Sustainability Evaluation Criteria

The evaluation focused on metrics relevant to sustainable operation under TDP constraints:

- **End-to-End Pipeline Performance:** Improved performance, lower energy per processed frame.
- **Average Power Consumption per Channel:** Electrical power savings per sustained active stream.
- **Average Memory Bandwidth Utilization per Channel:** Reduction in data movement, serving as a primary proxy for mitigating GPU starvation and associated power savings at the DRAM level.

IV. Results and Quantified Sustainability Impact

A. Efficiency Gains

The evaluation confirmed that LCEVC maintains consistent inference accuracy, with detection quality comparable between native AVC and LCEVC-AVC pipelines (see Attachment 2). The GPU-based evaluation, of a resource-constrained eight-stream scenario, demonstrates a substantial gain in operational efficiency compared to native AVC.

TABLE I. AI inference comparison for AVC vs LCEVC-AVC

	Total fps (8 streams)		Avg per-stream FPS		Fps gain
Video Number	**AVC**	**LCEVC AVC**	**AVC**	**LCEVC AVC**	
Video 1	242.46	824.2	30.31	103.03	***340%***
Video 2	243.4	692.74	30.42	86.59	***285%***
Video 3	243.1	821.1	30.39	102.64	***338%***
Video 4	220.93	693.82	27.62	86.73	***314%***

TABLE II. GPU Perf. comparison for AVC vs LCEVC-AVC

Video Number	**Avg Power per stream (index)**		***LCEVC Power Savings (%)***	**Total Memory Bandwidth per stream (MB/s)**		***Memory Bandwidth Savings (%)***
	AVC	**LCEVC AVC**		**AVC**	**LCEVC AVC**	
Video 1	1.00	0.29	**-71%**	1673.02	290.75	**-83%**
Video 2	1.02	0.35	**-66%**	1571.45	329.9	**-79%**
Video 3	1.03	0.31	**-70%**	1731.15	306.79	**-82%**
Video 4	1.10	0.35	**-68%**	1699.04	352.22	**-79%**

Table I reports the end-to-end inference throughput across eight parallel 4K channels, showing that LCEVC-AVC consistently achieves a 2.85× to 3.40× increase in both total and per-stream FPS relative to native AVC. These gains indicate that the reduced decoding complexity of the LCEVC-enhanced bitstreams directly translates into higher effective AI processing capacity under identical hardware constraints.

Table II shows that across all test sequences, LCEVC-AVC reduces average per-stream power consumption by 66–71% and lowers memory bandwidth utilization by 79–83%. Such reductions in operational load demonstrate that LCEVC not only increases throughput but also markedly improves power and memory efficiency that are two critical limiting factors in multi-stream edge-AI deployments. Our results confirm that LCEVC delivers substantial computational and energy savings while preserving inference accuracy.

B. Analysis of Sustainability Implications

The performance gains translate into sustainable compute.

The **2.9–3.4x pipeline speed-up** indicates that the hardware completes the same amount of work in reduced wall-clock time, minimising the energy required per frame and increasing the number of supported AI streams.

The **66%-71% power savings per channel** is a consequence of base-first decoding. By avoiding full-resolution processing in the fixed-function and scaling units, the system reduces operational power draw per stream.

The **79%-83% memory bandwidth reduction** addresses one of the main bottlenecks in visual AI pipelines: I/O and memory access. Lower pixel traffic across the memory bus decreases the energy consumed by memory subsystems and helps maintain stable GPU utilisation. This improvement in memory efficiency reduces the overall system power in addition to the reduced SOC processing power.

V. Discussion, Conclusion, and Future Work

The results confirm that LCEVC is an effective method for improving the resource efficiency of AI video analytics. By allowing pipelines to decode only the low-resolution base layer, LCEVC reduces reliance on high-cost operations such as full-resolution decoding and heavy memory movement. These per-stream reductions accumulate at system level, increasing thermal headroom and enabling more concurrent channels within a fixed TDP budget – a key requirement for sustainable, high-density edge deployments.

LCEVC therefore provides a practical means of lowering per-frame energy consumption and improving the overall energy-to-throughput ratio. By aligning the data layer with the model's actual input needs, it mitigates fundamental compute and I/O bottlenecks of traditional pipelines. The measured gains – up to 3.4× end-to-end speed-up and 83% memory-bandwidth savings – offer a credible and deployable route to building energy-efficient video analytics systems at scale.

Future work will focus on: (1) completing comparative tests with VVC as the base codec, where similar or greater benefits are expected; (2) integrating live power and thermal telemetry to provide finer-grained sustainability metrics, including energy per processed frame; and (3) investigating selective enhancement-layer decoding for multi-stage inference tasks that occasionally require higher spatial detail.

Whilst this work shows accuracy is on-par (see attachment 2) a proper quantified comparison is future work for authors.

This work is a proof point that AI pipeline optimisation beyond model optimisation, such as adjusting the CODEC and/or pre-processing stages to improve data orchestration is an exciting area of research for performance, quality of result and energy efficiency benefits. The work in [3] further supports this from a performance perspective for SMPTE VC-6 and the authors plan to investigate how video super resolution would compare in bringing equivalent benefits.

REFERENCES

[1] G. Meardi, S. Ferrara, L. Ciccarelli, G. Cobianchi, S. Poularakis, F. Maurer, S. Battista, and A. Byagowi, "MPEG-5 part 2: Low Complexity enhancement Video Coding Overview and performance evaluation," Applications of Digital Image Processing XLIII, vol. 11510, pp. 238–257, 2020.

[2] Hyperbolic AI, "How to Identify and Resolve GPU Bottleneck Issues for AI Workloads (Citing Google and Microsoft Research)," 2025, https://www.hyperbolic.ai/blog/how-to-identify-and-resolve-gpu-bottleneck-issues-for-ai-workloads

[3] G. Cobianchi, M. Razaak, B. Cope and P. Guruvareddiar, "Enabling Resource-Efficient AI Video Analytics with Hierarchical Codecs," *to appear* ACM Mile High Video (MHV), 2026.

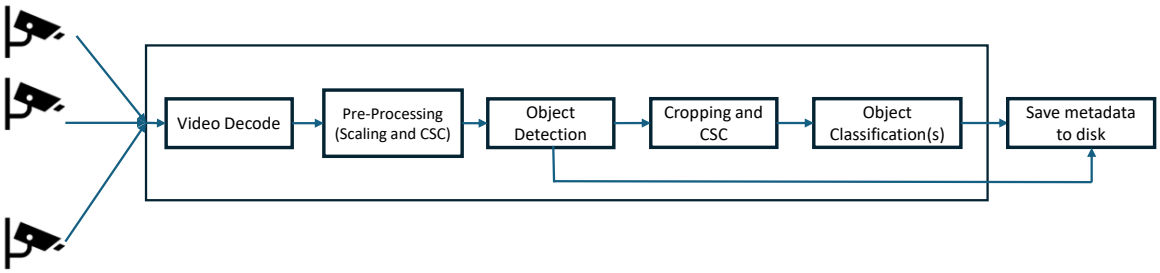
[4] Intel Corporation, "Intel Deep Learning Streamer (DL Streamer)," https://dlstreamer.github.io/, 2025, accessed: 2025-07-21.

[5] MPEG, "Verification Test Report on the Compression Performance of Low Complexity Enhancement Video Coding", April 2021, Meeting MPEG-134, available https://www.lcevc.org/wp-content/uploads/MPEG-Verification-Test-Report-on-the-Compression-Performance-of-LCEVC-Meeting-MPEG-134-May-2021.pdf.

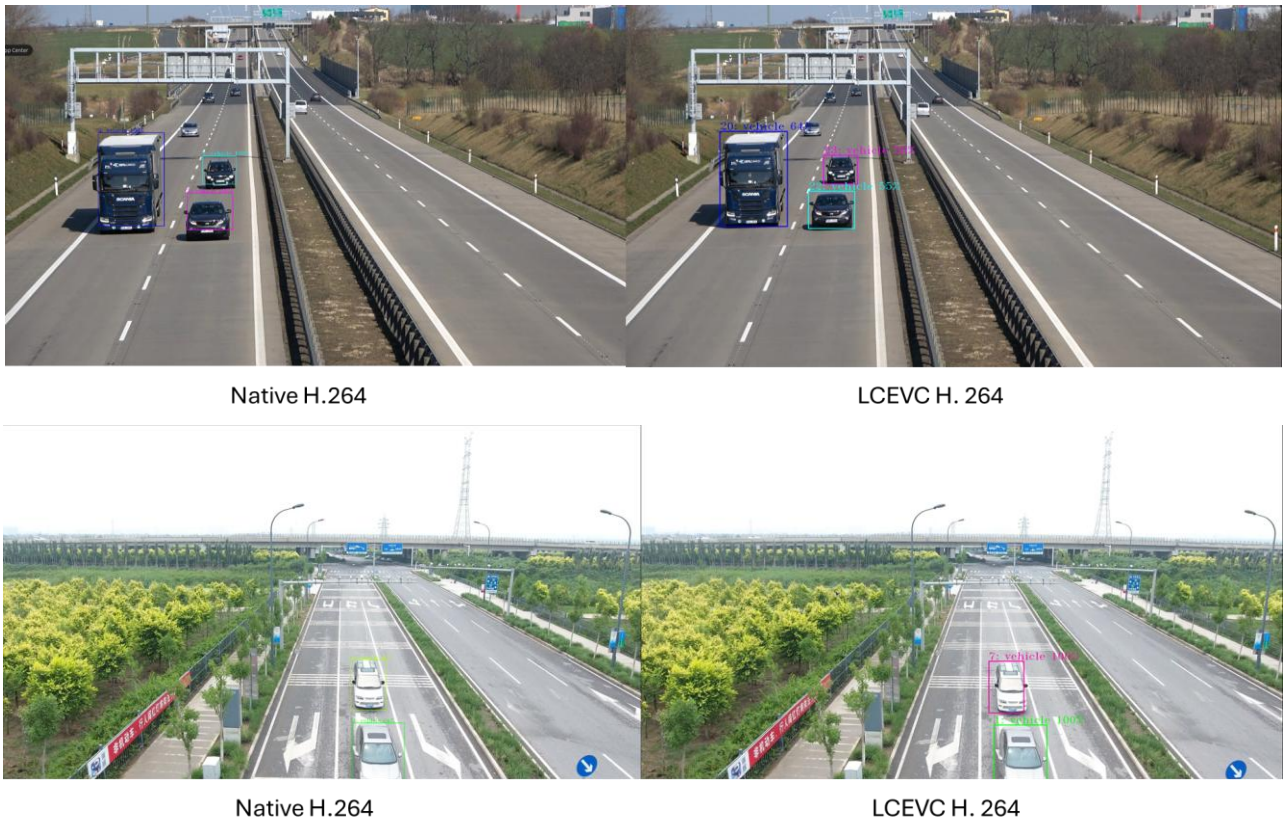
[6] G. Cobianchi, G. Meardi, S. Poularakis, A. Walisiewicz, O, Abdelkafi, F. Ben Amara, F. Kossentini, C. Stejerean, H. Tmar. "Enhancing SVT-AV1 with LCEVC to improve quality-cycles trade-offs and enhance sustainability of VOD transcoding", October 2022, available https://www.spiedigitallibrary.org/conference-proceedings-of-spie/12226/2633882/Enhancing-SVT-AV1-with-LCEVC-to-improve-quality-cycles-trade/10.1117/12.2633882.full?tab=ArticleLink.

[7] S. Saunack, "MobileNetV2-SSD: Implementation of Single Shot MultiBox Detector with MobileNetV2 as feature extractor," GitHub repository, https://github.com/saunack/MobileNetv2-SSD, accessed: Sep. 30, 2025.

[8] D. Du, P. Zhu, L. Wen, X. Bian, H. Lin, Q. Hu, T. Peng, J. Zheng, X. Wang, Y. Zhang, and others, "The VisDrone-DET2018 dataset: A large-scale benchmark for object detection in video captured by drones," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops, 2018, pp. 214–215. Available: http://aiskyeye.com/ D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, arXiv:1312.6114. [Online]. Available: https://arxiv.org/abs/1312.6114



Attachment 1: Diagram of end to end pipeline for the tests in this paper



Attachment 2: Examples of comparative frames