A 1970s Vibe to Energy-Conserving AI Monitoring

Originally published at https://www.unite.ai/a-1970s-vibe-to-energy-conserving-ai-monitoring/

New research shows most video AI does not need color at all, switching it on only at key moments and cutting data use by over 90% with little loss in accuracy.

Remote streaming cameras and other untethered, battery-driven video devices demand tightly-optimized monitoring setups, since they may rely on unstable power sources – such as solar – or require periodic recharging, or other forms of human intervention, in situations where, ideally, no one should need to be present.

In tandem with this line of research, interest in camera-equipped wearables has also grown (even though such devices were already tightly constrained by power and compute limits), because edge AI now promises to make them significantly more useful.

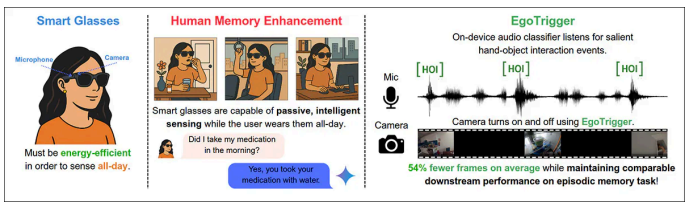
Beyond these considerations, the long-term impetus to reduce edge AI and monitoring costs (particularly in cases where such savings don't need to be passed on to the customer) make a compelling case for innovation in energy conservation approaches for 'edge' use-cases.

Sound Off

In the field of streaming video-sensing, resource-deprived edge monitoring devices must use the least possible energy, whilst still spending enough power to monitor for 'interesting' events – at which point, it will be worth spending more resources.

Effectively, this is a similar use case as movement-driven lights, which provide illumination only when low-energy-drain sensors determine there is someone there to appreciate it.

Since audio-monitoring and compression is notably less resource-intensive than video, several approaches in recent years have attempted to use sound-driven cues to 'switch on' attention in constrained systems; frameworks such as *Listen to Look* and *Egotrigger*:



In the Egotrigger system, audio-driven triggering selectively activates image capture from hand-object interaction cues, reducing redundant frames while preservin episodic memory performance in resource-constrained smart-glasses systems. Source

Clearly audio is not the ideal medium in which to search out visual events, since many essential such events may have no associated audio cue, or may occur out of range of edge microphones.

Light Sleeper

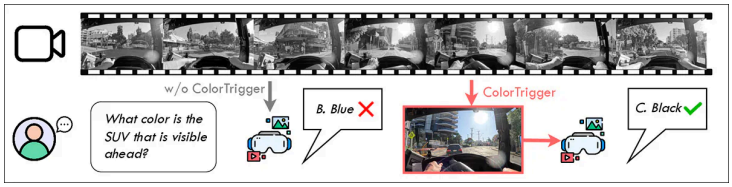
What might be better, one new paper suggests, is a video stream that can work together with AI to increase resources as soon as a watched-for event occurs. The simulation below* gives a general idea of the concept – low-resolution monitoring is maintained at the minimum signal-level level necessary for object-detection frameworks to operate, and to tell the system to increase resolution due to the triggering of an event:

CLICK TO PLAY VIDEO ON SITE

A simulation of the desired behavior – that streaming and analysis operate at its lowest level of resource consumption by default; just enough to trigger higher resource consumption when 'interesting' or sought events are detected in the grayscale stream. The black-and-white surveillance style may be rather 'retro', but it could be a sign of things to come. This video was created by the author purely for illustrative purposes in relation to the core ideas of the new paper. Source:

The new work, an academic collaboration between various UK institutions and Huawei, proposes a training-free, AI-facilitated, *grayscale-always, color-on-demand* schema for edge monitoring – designed to operate at low token usage when no 'key events' are taking place, and to ramp up consumption only for the duration of the event.

In streaming video understanding benchmarks, the new system, dubbed *ColorTrigger*, was able to achieve 91.6% of full-color baseline performance while using only 8.1% of the RGB frames in those standards:

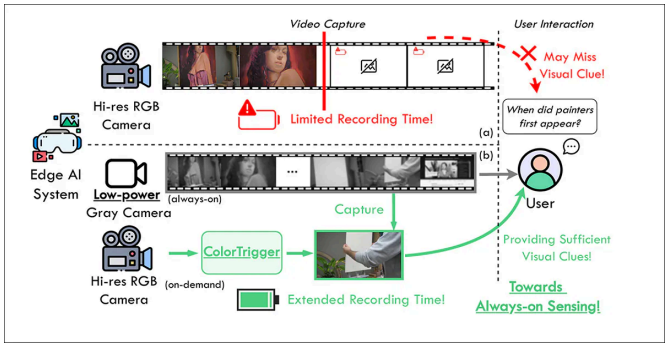


When the model only sees grayscale video, it confuses key details and gives wrong answers; but triggering color at the right moments disambiguates the image ai mistakes triggered by tasks that depend on color. Source

The new paper is titled *Color When It Counts: Grayscale-Guided Online Triggering for Always-On Streaming Video Sensing*, and comes from eight researchers across Queen Mary University of London, Durham University, Imperial College London, and Huawei Noah's Ark Lab. The paper also has an accompanying project page.

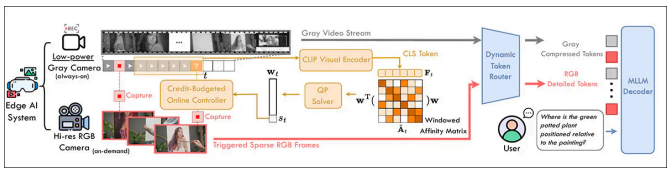
Method

To preserve temporal structure in the new system, ColorTrigger maintains constant low-bandwidth grayscale surveillance. A causal online trigger analyzes a sliding window (i.e., a flexible plus-minus range of frames around a particular time, such as the sensing of an event trigger) of the low-resolution stream:



Continuous high-resolution RGB capture rapidly drains power, so recording stops early and key moments can be missed. Conversely, ColorTrigger keeps a low-power grayscale stream running at all times, and only activates the RGB camera at selected moments – extending recording time, while still capturing the visual details needed to answer later queries. Source

While the system is in ‘passive’ mode (i.e., it has not identified a trigger event yet), its dynamic token router allocates limited capacity to an asymmetric decoder, always looking for redundancy, and for events indicating novelty, at which point the token flow re-prioritizes capacity over compression:



Schema for ColorTrigger. The system monitors a sliding-window analysis of recent frames to detect redundancy and change, triggering high-resolution RGB capture only when needed, under a credit-based budget. A dynamic token router allocates fewer tokens to grayscale inputs and more to selected RGB frames, preserving temporal order for downstream Multimodal Large Language Model (MLLM) processing.

On a frame-by-frame basis, the system needs to decide whether the current moment contains new information worth the cost of capturing color. The short recent history of grayscale frames in the sliding window allows ColorTrigger to compare the current frame against its *immediate* past. Each frame is converted into a compact feature representation, and these features are compared with one another to measure how similar or different their host frames are.

This comparison process is organized into a structure that summarizes *how much* each frame overlaps with the others, effectively capturing whether the scene is repeating or changing. A lightweight optimization step assigns an importance score to each frame in the window, favoring novelty.

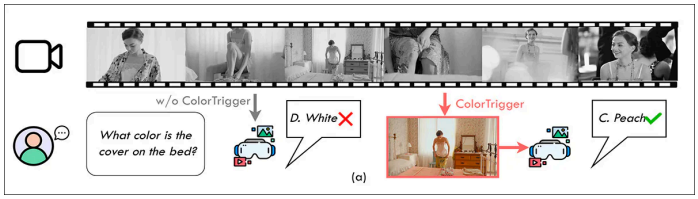
Color Balance

To prevent excessive use of color, a simple ‘credit system’ limits how often color can be triggered over time. Credits accumulate gradually, and are spent when color is requested, ensuring that bursts of activity are allowed, but overall usage remains controlled. A frame is only ‘upgraded’ to color if it is both informative, and if there are enough credits available.

The Dynamic Token Router controls how much detail each frame receives, instead of processing every frame at full quality. When nothing important is detected, the grayscale frame is kept low-resolution and turned into a small, compressed set of tokens. When an important moment is detected, the system switches to color and processes that frame at higher resolution, offering a richer and more detailed representation.

Both types of frames go through the same model, but grayscale frames are handled in a lighter way, while selected color frames are given more attention. The outputs are then combined in their original order and sent to the model as a continuous stream.

Because most frames stay lightweight and only a few are upgraded, the system saves a large amount of computation while still capturing the key details when they matter:



From the paper, another example where the system is required to temporarily increase resources in order to distinguish a color.

Data and Tests

To test the system, the researchers evaluated against the StreamingBench and OVO-Bench video benchmarks, avoiding the processing of *future* content (which is a potential hazard in offline tests).

The frozen Multimodal Large Language Model (MLLM) used was InternVL3.5-8B-Instruct, with the causal trigger implemented via CLIP ViT-B/16.

The grayscale stream was limited to the luminance channel in the CIELAB color space, in accordance with prior work, with the resulting grayscale frames resized to 224x224px prior to patchification (the splitting of an image into small fixed-size blocks, so that each block can be processed as a separate unit by the model).

The RGB frames, conversely, enjoyed a higher bitrate, and were processed at 448x448px, producing 256 tokens, in contrast to the 64 tokens produced for the grayscale frames.

Common optimization tools were used to make the system’s decisions: CVXPY (a Python library for setting up optimization problems), and OSQP Solver (a fast algorithm that calculates when to trigger color).

Video was processed at 1fps, with a cap of 128 frames per clip, to keep computation low.

Proprietary systems tested were Gemini 1.5 Pro; GPT-4o; and Claude 3.5 Sonnet. Open source video MLLMs tested were LLaVA-OneVision-7B; Video-LLaMA2-7B; and Qwen2.5-VL-7B.

Streaming MLLMs tested were Flash-VStream-7B; VideoLLM-online-8B; Dispider-7B; and TimeChat-Online-7B.

InternVL-3.5-8B and Qwen3-VL-8B were tested in various configurations, detailed in the first results table below, regarding StreamingBench:

Model	#Frames	RGB (%)	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	All
Human	-	-	89.47	92.00	93.60	91.47	95.65	92.52	88.00	88.75	89.74	91.30	91.46
Proprietary MLLMs													
Gemini 1.5 Pro [42]	1 fps	100	79.02	80.47	83.54	79.67	80.00	84.74	77.78	64.23	71.95	48.70	75.69
GPT-4o [18]	64	100	77.11	80.47	83.91	76.47	70.19	83.80	66.67	62.19	69.12	49.22	73.28
Claude 3.5 Sonnet [1]	20	100	80.49	77.34	82.02	81.73	72.33	75.39	61.11	61.79	69.32	43.09	72.44
Open-Source Video MLLMs													
LLaVA-OneVision-7B [20]	32	100	80.38	74.22	76.03	80.72	72.67	71.65	67.59	65.45	65.72	45.08	71.12
Video-LLaMA2-7B [7]	32	100	55.86	55.47	57.41	58.17	52.80	43.61	39.81	42.68	45.61	35.23	49.52
Qwen2.5-VL-7B [2]	1 fps	100	78.32	80.47	78.86	80.45	76.73	78.50	79.63	63.41	66.19	53.19	73.68
Streaming MLLMs													
Flash-VStream-7B [49]	-	100	25.89	43.57	24.91	23.87	27.33	13.08	18.52	25.20	23.87	48.70	23.23
VideoLLM-online-8B [6]	2 fps	100	39.07	40.06	34.49	31.05	45.96	32.40	31.48	34.16	42.49	27.89	35.99
Dispider-7B [30]	1 fps	100	74.92	75.53	74.10	73.08	74.44	59.92	76.14	62.91	62.16	45.80	67.63
TimeChat-Online-7B [47]	1 fps	100	80.22	82.03	79.50	83.33	76.10	78.50	78.70	64.63	69.60	57.98	75.36
InternVL-3.5-8B	128	100	83.47	82.03	82.65	84.62	75.47	80.06	81.48	67.89	70.45	59.04	77.20
InternVL-3.5-8B grayscale	128	0	60.98	78.91	60.88	55.45	65.41	63.24	75.00	60.16	62.22	55.85	62.08
ColorTrigger (Ours, 8B)	128	8.1 (↓91.9%)	75.07	77.34	72.56	75.64	71.70	70.40	76.85	65.04	66.48	57.98	70.72 (↑8.64%)
ColorTrigger (Ours, 8B)	128	34.3 (↓65.7%)	81.57	78.91	77.29	84.94	76.10	77.57	81.48	67.48	67.61	56.91	75.24 (↑13.16%)

Performance on StreamingBench for real-time visual understanding tasks, comparing proprietary, open-source, and streaming MLLMs under different color budgets. RGB (%) indicates the proportion of frames kept in color after triggering, where 100 denotes full color and 0 denotes grayscale-only input. ColorTrigger is evaluated at two operating points, retaining 8.1% and 34.3% color frames, and demonstrates improved overall accuracy over the grayscale InternVL-3.5-8B baseline while substantially reducing color usage relative to the full-color setting.

Here the authors comment:

> *'ColorTrigger achieves competitive performance on the Real-time Visual Understanding subtask of StreamingBench.*
>
> *'Our model with 34.3% RGB frames scores 75.24, outperforming recent online model Dispider-7B and close to TimeChat-Online-7B, while being comparable to proprietary models such as Gemini 1.5 Pro (75.69) and surpassing GPT-4o (73.28) and Claude 3.5 Sonnet (72.44).'*

InternVL-3.5-8B scored 77.20 using full color, while ColorTrigger reached 75.24 using 65.7% fewer RGB frames – and even with just 8.1% color frames, it scored 70.72, outperforming the grayscale baseline of 62.08 by 8.64%, and remaining competitive with other streaming models.

Next, OVO-Bench was tested:

Model	#Frames	RGB (%)	Real-Time Visual Perception						Backward Tracing				Forward Active Responding				Overall
			OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	REC	SSR	CRR	Avg.	Avg.
Human Agents	-	-	94.0	92.6	94.8	92.7	91.1	94.0	93.2	92.6	93.0	91.4	92.3	95.5	89.7	93.6	92.9	92.8
Proprietary MLLMs																		
Gemini 1.5 Pro [42]	1 fps	100	87.3	67.0	80.2	54.5	68.3	67.4	70.8	68.6	75.7	52.7	62.3	35.5	74.2	61.7	57.2	65.3
GPT-4o [18]	64	100	69.1	65.1	65.5	50.0	68.3	63.7	63.6	49.8	71.0	55.4	58.7	27.6	73.2	59.4	53.4	58.6
Open-Source Video MLLMs																		
LLaVA-NeXT-Video-7B [53]	64	100	69.8	59.6	66.4	50.6	72.3	61.4	63.3	51.2	64.2	9.7	41.7	34.1	67.6	60.8	54.2	53.1
LLaVA-OneVision-7B [20]	64	100	67.1	58.7	69.8	49.4	71.3	60.3	62.8	52.5	58.8	23.7	45.0	24.8	66.9	60.8	50.9	52.9
Qwen2-VL-7B [43]	64	100	69.1	53.2	63.8	50.6	66.3	60.9	60.7	44.4	66.9	34.4	48.6	30.1	65.7	50.8	48.9	52.7
LongVU-7B [38]	1 fps	100	55.7	49.5	59.5	48.3	68.3	63.0	57.4	43.1	66.2	9.1	39.5	16.6	69.0	60.0	48.5	48.5
Streaming MLLMs																		
Flash-VStream-7B [49]	1 fps	100	25.5	32.1	29.3	33.7	29.7	28.8	29.9	36.4	33.8	5.9	25.4	5.4	67.3	60.0	44.2	33.2
VideoLLM-online-8B [6]	2 fps	100	8.1	23.9	12.1	14.0	45.5	21.2	20.8	22.2	18.8	12.2	17.7	-	-	-	-	-
TimeChat-Online-7B [47]	1 fps	100	75.2	46.8	70.7	47.8	69.3	61.4	61.9	55.9	59.5	9.7	41.7	31.6	38.5	40.0	36.7	46.7
Qwen2.5-VL-7B [2]	1 fps	100	73.8	56.0	68.1	46.6	71.3	60.3	62.7	48.2	64.9	26.9	46.6	36.2	41.8	47.1	41.7	50.3
InternVL-3.5-8B	128	100	77.9	64.2	75.9	56.7	75.3	70.7	70.1	56.6	66.2	34.4	52.4	48.2	62.7	41.3	50.7	57.7
InternVL-3.5-8B grayscale	128	0	61.1	46.8	53.5	47.8	56.4	57.1	53.8	46.1	54.7	37.1	46.0	35.3	54.3	42.1	43.9	47.9
ColorTrigger (Ours, 8B)	128	7.1 (↓92.9%)	65.1	52.3	61.2	50.6	66.3	63.6	59.9 (↑6.1%)	50.5	58.1	32.3	47.0	32.5	59.3	41.3	44.3	50.4 (↑2.5%)
ColorTrigger (Ours, 8B)	128	33.1 (↓66.9%)	77.9	56.0	69.8	55.1	67.3	65.2	65.2 (↑11.4%)	50.2	62.8	30.7	41.3	32.7	59.4	40.8	44.3	52.5 (↑4.6%)

Performance on OVO-Bench across three categories: Real-Time Visual Perception, Backward Tracing, and Forward Active Responding, comparing proprietary, open-source, and streaming MLLMs under different color budgets. RGB (%) indicates the proportion of frames kept in color after triggering, where 100 denotes full color and 0 denotes grayscale-only input. ColorTrigger is evaluated at two operating points, retaining 7.1% and 33.1% color frames, and shows improved overall accuracy over the grayscale InternVL-3.5-8B baseline while substantially reducing color usage relative to the full-color setting.

Of these results, the authors state:

> *'Our model with 33.1% RGB frames achieves an overall score of 52.5, outperforming almost all existing open-source online MLLMs. Compared to the base model InternVL-3.5-8B with full RGB input (57.7), ColorTrigger scores 52.5 while reducing RGB frame usage by 66.9%, representing only a 5.2-point drop in overall performance.*
>
> *'This modest degradation is accompanied by substantial gains in efficiency, demonstrating the effectiveness of our adaptive routing strategy.'*

Real-Time Visual Perception reached 65.2 – an 11.4-point gain over the grayscale-only baseline of 53.8. Even when limited to just 7.1% RGB frames (a 92.9% reduction), ColorTrigger maintained an overall score of 50.4, improving on the grayscale setting by 2.5 points.

Finally the researchers conducted a test against an offline video task (an analytical task not designed to test latency or other 'live' environmental conditions, using the Video-MME long-form video understanding benchmark:

Method	#Frames	RGB (%)	Video-MME			
			short	medium	long	overall
VideoChat2-7B [21]	16	100	48.3	37.0	33.2	39.5
LongVA-7B [52]	128	100	61.1	50.4	46.2	52.6
Kangaroo-7B [24]	64	100	66.1	55.3	46.6	56.0
Video-CCAM-14B [11]	96	100	62.2	50.6	46.7	53.2
VideoXL-7B [39]	128	100	64.0	53.2	49.2	55.5
Dispider-7B [30]	1 fps	100	-	-	-	57.2
VideoChat-Online-4B [17]	2 fps	100	-	-	47.1	54.4
TimeChat-Online-7B [47]	1 fps	100	-	-	48.4	62.4
InternVL-3.5-8B	128	100	76.7	65.3	54.7	65.6
InternVL-3.5-8B grayscale	128	0	63.4	57.9	50.6	57.3
ColorTrigger (Ours, 8B)	128	9.1 (↓90.9%)	71.9	63.0	53.6	62.8 (↑5.5%)
ColorTrigger (Ours, 8B)	128	37.6 (↓62.4%)	76.8	66.3	55.1	66.1 (↑8.8%)

Performance comparison of the trialed systems on the Video-MME benchmark.

In this test, the model achieved an overall score of 66.1, while using 37.6% RGB frames, surpassing the full-color InternVL-3.5-8B baseline score of 65.6, despite using 62.4% fewer color frames.

The authors comment:

‘This demonstrates that our adaptive triggering mechanism not only reduces computational cost but can actually improve performance by focusing RGB capacity on semantically critical moments.

‘Notably, ColorTrigger outperforms all existing online MLLMs including TimeChat-Online-7B at 62.4 and Dispider-7B at 57.2, confirming the effectiveness of combining continuous grayscale context with selective RGB acquisition for long-form video understanding.’

Conclusion

I always enjoy seeing innovations of this type, not least because AI’s high and ever-growing need for (electrical) power has been producing dismal headlines for a long time, and it’s good to see research that at least indirectly addresses the issue.

It’s cynically comforting to know that the power-savings made in such forays are motivated by commercial considerations, since these are less likely to be affected by short-term political decisions than the nobler, but more assailable worries over energy conservation and global warming. Thankfully, the same end is achieved, for different reasons.

* *Created by me, just to encapsulate the paper’s idea for the reader.*

First published Thursday, March 26, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More