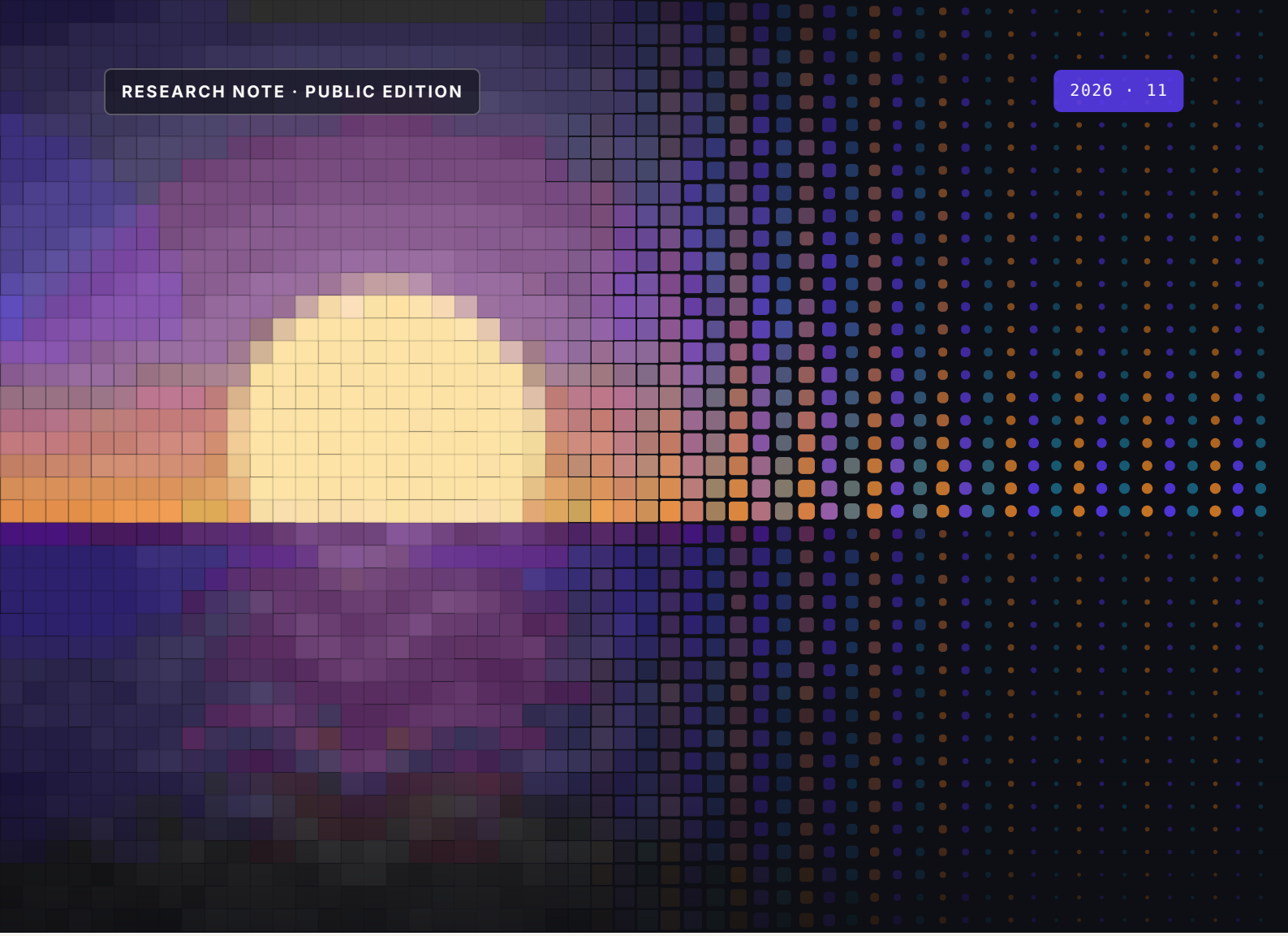


LEARNED COMPRESSION · PLAYBACK × SILICON

Neural Coding in 2026

It ships around the codec, not instead of it

Playback · Photography · Calls · Generative AI · Games · Silicon

> **THE VERDICT**
>
> **Neural networks already work at both ends of the codecs that carry the world's video. End-to-end neural codecs ship only in closed loops, and mainstream screens wait for H.267.**

Tan Shuai

Technologist and founder · AI researcher and developer

tanshuai.com · i@tanshuai.com

Data as of November 12, 2026

Money in US dollars

CONTENTS

Contents



◆ **Executive summary** — Mainstream already, but around the codec — 04

01 **Scope and method** — What “neural coding” means here, and how I looked — 07

02 **Playback and streaming** — Playback and streaming: neural at both ends, standard in the middle — 08

03 **Photography and images** — Photography and images: neural inside the camera, classical on the web — 11

04 **Calls and meetings** — Calls and meetings: where end-to-end neural codecs actually ship — 12

05 **Generative AI** — Generative AI: the largest deployment nobody calls a codec — 14

06 **Games and graphics** — Games and graphics: neural reconstruction is mainstream, neural assets are starting — 15

07 **Specialist links** — Specialist links: satellites, drones and robots — 16

08 **Silicon** — The silicon view: energy, compute, memory and exactness — 17

09 **Standards and patents** — Standards and patents: where neural tools become licensable — 21

10 **Who gets paid** — Who gets paid, and where a company can play — 22

11 **Claims and evidence** — Claims you will hear, and what the evidence says — 24

12 **Signposts** — What to watch, 2027 to 2030 — 25

13 **Open questions** — 26

◆ **About the author** — Tan Shuai — 27

◆ **Sources** — 28

FIGURES

Figures

FIG 01 Where neural coding ships in 2026 06
FIG 02 Three ways a neural network enters a codec 07
FIG 03 From a finished standard to mainstream screens takes most of a decade 10
FIG 04 In calls, a neural codec beats hardware HEVC by 65–76% 13
FIG 05 When one company controls the stack, neural reconstruction spreads fast 15
FIG 06 Energy per decoded pixel: fixed-function decoders are an order of magnitude ahead 18
FIG 07 A neural tool costs tens of thousands of MACs per pixel; a whole neural decoder costs hundreds 19
FIG 08 The split that ships: fixed-function hardware for the bitstream, a neural engine for the pixels 20
FIG 09 Who gets paid from neural coding in 2026 23

Tags in the text

- estimate — the author's arithmetic or inference from cited data
- vendor claim — published by the company that sells the product, not independently verified
- analysis — the author's interpretation of the evidence

Data as of November 12, 2026 · Money in US dollars

EXECUTIVE SUMMARY

Mainstream already, but around the codec

THE VERDICT

Neural coding is already mainstream in 2026, but it works around the standard codecs, not instead of them.

≈75%
of mobile data traffic is video, end of 2025 (Ericsson)

1.26 billion
smartphones shipped in 2025, every one decoding video in fixed-function hardware

≈12×
the energy per pixel of a measured HEVC decoder chip with DRAM, for the only published neural decoder chip design

≈22 TOPS
of sustained compute for a DCVC-RT-class neural decoder at 1080p30 estimate

75.5%
less bitrate than hardware HEVC for MLVC at 360p in calls, on Microsoft's own test set vendor claim

End of 2029
H.267 version 1, the first standard likely to carry neural tools to every screen

Where it ships

Neural networks now run at both ends of the codecs that carry almost all of the world's video. On the way in they shape the picture: AI image signal processors in phone cameras, and a neural downscaler in Netflix's encoding pipeline. On the way out they repair it: YouTube's Super Resolution on TVs, super-resolution at Meta, and TV processors that advertise hundreds of neural networks. In between, the bitstream is still AV1, HEVC, AVC or VVC, decoded by fixed-function hardware. No streaming service sends a neural video format to TVs or phones; where a neural video decoder runs on a device today, it runs inside one company's own app.

Closed loops

End-to-end neural codecs, where both the encoder and the decoder are learned, ship only in closed loops, where one company controls both ends and can fall back to a standard codec. Microsoft has run an AI audio codec in all two-party Teams and Skype calls since 2021, and in July 2026 it open-sourced MLVC, a neural video codec now rolling out in peer-to-peer Teams calls: at 360p it needs about 75% less bitrate than hardware HEVC for the same rated quality. Every speech and video generation model runs on a learned codec, from 12.5 Hz speech tokenizers to video VAEs with a 1:192 compression ratio. Neural texture compression has shipped in one AAA game. These are real deployments, and they share one condition: nobody else has to decode the stream.

Timing

Mass playback of neural bitstreams waits for a standard and a silicon cycle. The first standard likely to carry neural tools, the successor to VVC that I will call H.267, issued its call for proposals in July 2026, evaluates responses in January 2027 and targets a first version at the end of 2029. AV1 needed nearly eight years from finished specification to 30% of Netflix viewing. On that pattern, the first H.267 chips arrive around 2031 and mainstream use comes in the mid-2030s estimate. Meanwhile, the codec that will reach screens next is AV2, finished in May

2026 with no neural decoder in it, and three IP vendors announced AV2 decoder customers within three weeks.

Silicon

The reason is in the silicon. The only published neural video decoder chip design needs 14.6 nJ per pixel, about 12 times a measured HEVC decoder chip counted with its DRAM and about 47 times its core alone. Microsoft's real-time research codec DCVC-RT needs, by my count of its own per-module table, about 175 thousand multiply-accumulates per decoded pixel, or roughly 22 TOPS of sustained compute for 1080p30 estimate : about half the 40 TOPS that defines an AI PC, spent on one video stream. A neural decoder also has to produce bit-identical results on every device, or its entropy decoding falls apart. That is why MLVC trades away compression to decode identically across NPUs and runs at 540p, not 4K.

Business

For a business, the money follows the standards and the platforms, not stand-alone neural codecs. Platforms keep the savings from neural pre- and post-processing. IP vendors license fixed-function decoders for each new standard, plus NPUs for imaging. Patent pools collect per device: $0.50 per phone and $1.00–1.50 per 4K TV for VVC in the US and other Region 1 markets. The three best-known neural codec startups, WaveOne, iSIZE and Deep Render, were bought by Apple, Sony and InterDigital. The open opportunities are in neural pre- and post-processing, in closed-loop products, in silicon that pairs a fixed-function codec with an NPU, and in the evaluation and standards work in between.

Question	Judgement	Key numbers (sources in the chapters)
① **Where does neural coding ship at scale today?**	Around the codec: camera ISPs, encode-side downscaling, playback super-resolution. End to end only in closed loops: calls, game assets, generative-AI tokenizers	Video ≈75% of mobile data; 1.26 billion phones and ≈200 million TVs shipped in 2025, all decoding in fixed-function hardware; DLSS 4 in 250+ games and apps
② **Will neural codecs replace AV1, HEVC and VVC in playback? When?**	Not before a standard and a device cycle. H.267 is the first realistic carrier, and it will likely mix classical and neural tools	H.267 call for proposals July 2026, evaluation January 2027, version 1 end of 2029; mainstream mid-2030s estimate
③ **What does it cost in silicon?**	An order of magnitude more energy per pixel than fixed-function decoding today, plus memory traffic and a bit-exactness requirement	14.6 nJ/pixel (neural) against 1.19 (HEVC, with DRAM); ≈175 kMAC per decoded pixel ≈ 22 TOPS at 1080p30 estimate
④ **What do standards and patents allow?**	Neural post-filters can already be signalled in VVC streams; JPEG AI is an ISO standard; H.267 will decide which neural tools become licensable	ECM (classical) ≈28% over VVC; neural loop filters 6–14% at 10–22+ kMAC/pixel; ByteDance 645 patent families in beyond-VVC work
⑤ **Who gets paid, and where can a company play?**	Platforms, IP vendors, patent holders. Pure neural-codec startups were acquired. Play in pre/post-processing, closed loops, codec-plus-NPU silicon, evaluation	VVC pool $0.50 per phone; AV2 decoder IP licensed within weeks of the spec; three acquisitions, terms undisclosed
⑥ **Which claims should be discounted?**	Gains quoted against old codecs, at low resolution, without decode cost, or from demos rather than shipped products	MLVC: 88% less than H.264 but 65–76% less than hardware HEVC; textures: 85% VRAM saving in a demo, about 30% in the shipped game

FIG 01

Where neural coding ships in 2026

Neural networks already run at both ends of mainstream codecs. End-to-end neural codecs ship only where one company controls both the encoder and the decoder.

	Around a standard codec pre-processing, enhancement, upscaling	End-to-end neural codec learned encoder and decoder	Inside a standard neural tools in a standard bitstream or API
01 **Playback and streaming**	AT SCALE Netflix downscaler; YouTube and Meta super-resolution; TV AI upscalers	NOT DEPLOYED No streaming service sends a neural video format; TVs and phones decode standard codecs in hardware	STANDARDIZING H.267 version 1 is due at the end of 2029; neural tools are among the candidates
02 **Photography and imaging**	AT SCALE AI ISPs in flagship phones; a full neural ISP shown as IP at CES 2026	DEMOS, PILOTS JPEG AI is an ISO standard: phone demos, no product	NOT DEPLOYED The format gaining ground is classical: JPEG XL on by default in Chrome 155
03 **Calls and meetings**	AT SCALE Receiver-side super-resolution and neural noise suppression	AT SCALE Neural audio in Teams two-party calls since 2021; MLVC video rolling out in Teams	SHIPPING, LIMITED Opus 1.5 adds optional, backward-compatible ML tools
04 **Generative AI**	—	AT SCALE Every video and voice model runs a learned tokenizer, e.g. a 1:192 video VAE	—
05 **Games and graphics**	AT SCALE Neural upscaling and frame generation: DLSS 4 alone in 250+ games and apps	SHIPPING, LIMITED Neural textures shipped in one AAA game; NVIDIA's SDK is in beta	STANDARDIZING DirectX is adding matrix operations for shaders (preview)
06 **Specialist links: satellite, robots**	DEMOS, PILOTS Server-side restoration of low-bitrate feeds	DEMOS, PILOTS Loss-resilient neural codecs in research and pilots	—

Source: Company, standards and developer documents cited in the report; status as of November 12, 2026. Rows are ordered by how many people each domain reaches.

CHAPTER 01

01 SCOPE AND METHOD
What “neural coding” means here, and how I looked

Three ways a neural network enters a codec

“Neural coding” covers three different things, and most arguments about it come from mixing them up.

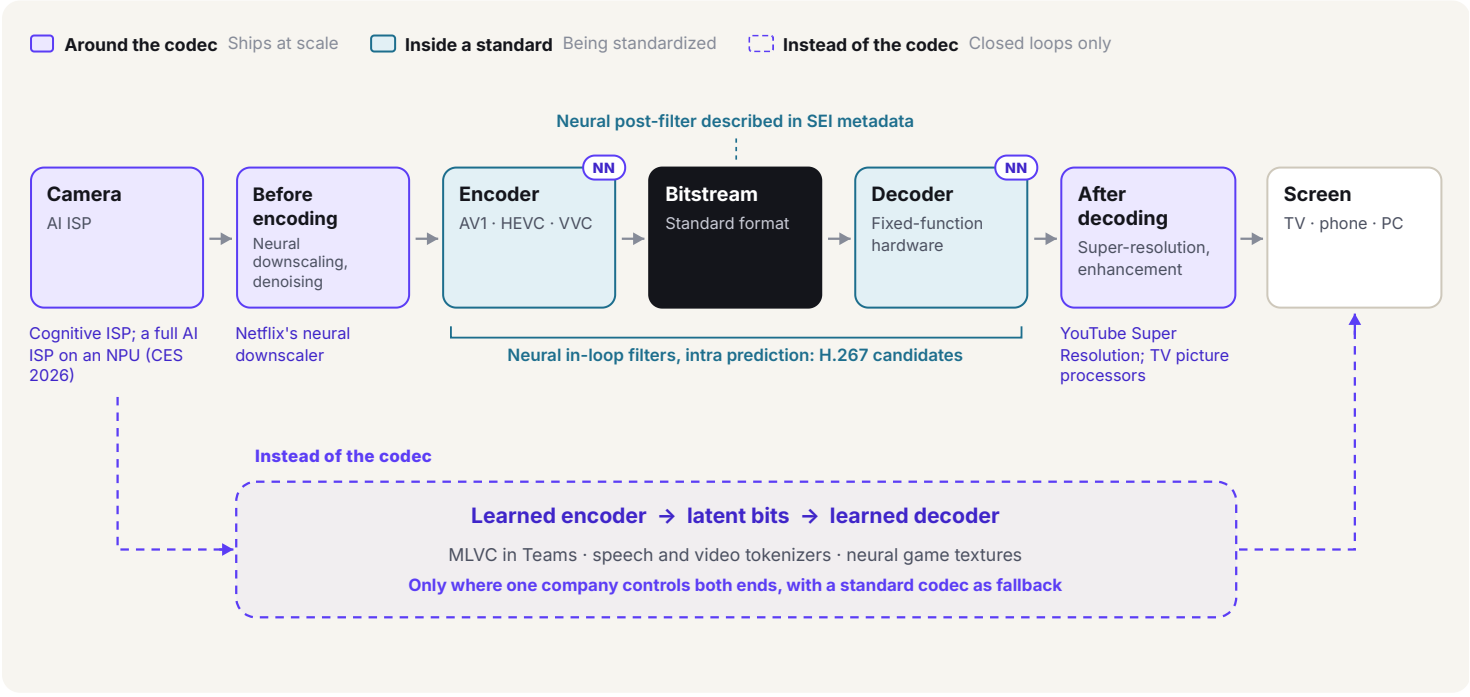
01 **Around the codec.** A network processes pixels before encoding or after decoding, and the bitstream stays standard. Examples: an AI image signal processor (ISP) in a camera, Netflix’s neural downscaler before its encoder, super-resolution on a TV. Any device can play the stream; the network only changes what goes in or what comes out.

02 **Inside a standard.** A network becomes a tool in a standard codec, for example a neural in-loop filter or intra predictor that every decoder must run, so it has to be specified exactly, typically in integer arithmetic. A lighter form is a neural post-filter that the bitstream describes in metadata and decoders may choose to apply.

03 **Instead of the codec.** Both the encoder and the decoder are learned end to end. The bitstream is a set of quantized latent values that only the matching network can decode. Examples: Microsoft’s MLVC in Teams, speech tokenizers in voice models, neural textures in a game.

FIG 02 **Three ways a neural network enters a codec**

Most of what ships in 2026 works around a standard codec. Neural tools inside a standard are being specified. End-to-end neural codecs replace the codec only where one company controls both ends.



Source: The author's synthesis of the sources in this report; status as of November 12, 2026.

The first kind needs no one's permission. The second needs a standard. The third needs control of both ends. That is the order in which they reach people, and it explains most of what follows.

Ordered by the real world, not by the research agenda

Research papers mostly study the third kind, because that is where the compression records are. This report orders the domains by how many people they reach, because that is what decides where money and engineering go: playback and streaming first, then photography, calls, generative AI, games, and the specialist links (satellite, drones, robots) last. Video is about 75% of all mobile data traffic, by Ericsson's count at the end of 2025. In 2025 the industry shipped about 1.26 billion smartphones (IDC) and roughly 196–209 million TVs (TrendForce and Omdia estimates). Every one of them decodes video in fixed-function hardware.

How I worked

This is desk research, done the way I would review a chip or a product plan. I read standards documents and meeting reports (JVET, MPEG, JPEG, AOMedia, IETF), company engineering blogs and release notes, chip papers, patent-pool rate cards, export-control texts and the press. Every number carries its date. Where I derive a number myself, such as compute per pixel from a paper's tables, I say so and mark it estimate. Claims I could not check against a primary source are marked vendor claim. Money is in US dollars. Status is as of November 12, 2026.

I compare like with like where I can, and say so where I cannot. Compression gains mean little without four things attached: the anchor (gain over what), the metric (PSNR, VMAF or a panel of viewers), the content and resolution, and what it costs to decode. Most of the disagreement about neural codecs disappears once those four are stated.

CHAPTER 02

PLAYBACK AND STREAMING

02 Playback and streaming: neural at both ends, standard in the middle

What viewers receive is still a standard bitstream

The largest video services finished one codec transition and are starting the next, and neither involves a neural decoder. In December 2025 Netflix reported that AV1 carried about 30% of its viewing, its second most-used codec after AVC. AV1 reached TVs only when hardware decoders did: Netflix added AV1 to its device certification in 2019, streamed it to TVs from late 2021, and says 88% of the large-screen devices submitted for certification between 2021 and 2025 supported it, almost all since 2023. YouTube said that as of March 19, 2026, more than 90% of its catalogue, weighted by watch time, was available in AV1.

AV2, AV1's successor from the Alliance for Open Media, reached version 1.0.0 on May 28, 2026, with about 30% better compression than AV1 by AOMedia's objective measurements. YouTube plans AV2 streams on the web "as soon as early 2027". The silicon followed at once: Chips&Media completed an AV2 hardware decoder in May and licensed it to an unnamed North American big-tech company in June; Allegro DVT had delivered its AV2 decoder IP to three customers by June 16; VeriSilicon had released an AV2-capable video IP to several customers by June

10. AV2 is a classical codec: its gains come from better partitioning, prediction, transforms and in-loop filters (one of them a "data-driven" guided detail filter), not from an end-to-end network.

VVC, the current ITU/ISO standard, found its first broadcast home in Brazil, whose DTV+ (TV 3.0) system moved into commercial operation in June 2026, timed for the World Cup. As of July 2026 no TV on sale in Brazil had a built-in DTV+ receiver; the first are expected between late 2026 and early 2027.

At the encoder: Netflix's neural downscaler

The cleanest example of neural coding in playback has been running since 2022 and changes nothing a viewer's device has to do. Netflix encodes each title at several resolutions, and the downscaling step that produces 1080p, 720p and 540p from a 4K master traditionally used a fixed filter such as Lanczos. Netflix replaced it with a small neural network, a preprocessing block followed by a resizing block, trained to improve the quality seen after a standard upscaler on the device. Netflix describes it as a drop-in solution: no other change on the encoding side or on the client side, so "millions of devices that support Netflix streaming automatically benefit". In Netflix's preference tests about 77% of viewers chose the neural downscaler's output vendor claim. To keep the cost acceptable at Netflix's scale, "just a few neural network layers were sufficient".

That is the pattern to remember. The network runs once per title in the data centre, where compute is cheap and can be scheduled, and its output is an ordinary AV1 or HEVC stream.

At the screen: super-resolution in the app and in the TV

At the other end, networks repair what the codec and the source lost. On October 29, 2025 YouTube announced Super Resolution for its TV app: videos uploaded below 1080p, from 240p to 720p, are upscaled to HD with machine learning, with 4K as the stated goal. The upscaled versions are labelled, viewers can switch back to the original, and creators can opt out. Meta described in October 2025 how it runs video super-resolution both on its servers at upload, on standard x86 CPUs in partnership with Intel, and on the client at playback, and found that the two complement each other. TV makers do the same in silicon: Samsung's flagship 8K processor, the NQ8 AI Gen3, advertises 768 neural networks for upscaling.

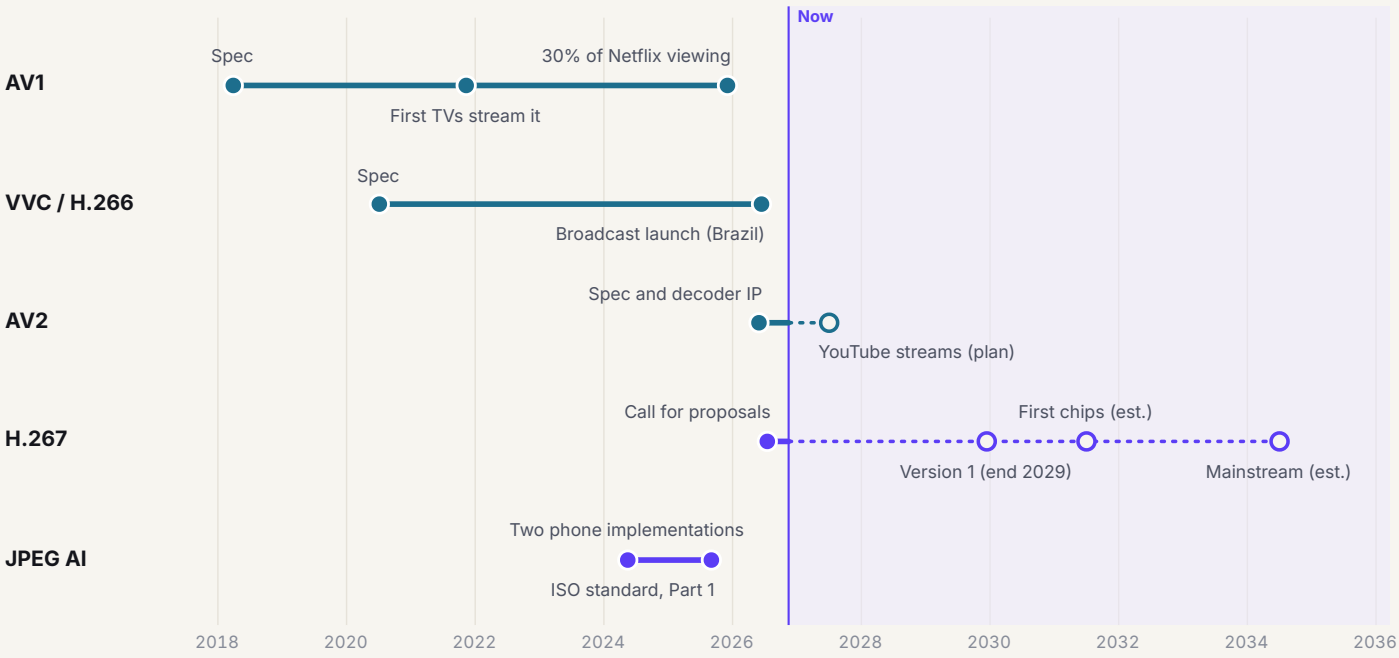
The economics differ by where the network runs. Server-side processing costs the platform compute for every video it touches. Device-side processing costs the platform nothing; it uses an NPU the buyer has already paid for. That is why the largest gains in 2026 come from both.

Why no TV decodes a neural bitstream

A playback format must decode on every device that will ever play it, and those devices last years. That makes the path from a finished standard to mainstream screens slow even for classical codecs. AV1's specification was finished in March 2018; the dav1d software decoder followed within months; TVs streamed it from late 2021; it reached about 30% of Netflix viewing in December 2025, nearly eight years after the specification. VVC was finished in July 2020 and reached its first broadcast service in June 2026.

FIG 03 **From a finished standard to mainstream screens takes most of a decade**

Playback needs a hardware decoder in the device. AV1 took nearly eight years from finished spec to 30% of Netflix viewing; the first standard likely to carry neural tools is due in late 2029.



Source: Netflix Tech Blog (Dec 2025); AOMedia; Allegro DVT; TV Technology; Fraunhofer IIS on Brazil's DTV+; MPEG-155; JPEG committee; ISO. Hollow marks are targets or the author's estimates.

A neural bitstream adds two problems that classical codecs solved long ago: energy per pixel and bit-exact decoding. Both are covered in Chapter 8. Together they mean that no TV or phone maker will put a neural video decoder in every device before a standard tells them exactly what to build.

What would change it: H.267

The standard that could carry neural tools to every screen is the successor to VVC, developed jointly by ITU-T and ISO/IEC in the Joint Video Experts Team (JVET). Its name is not final; I will call it H.267, as much of the industry does.

In October 2025 JVET evaluated responses to a call for evidence: five responses from 16 companies and research labs, with 20 codec configurations tested. The committee found evidence of better compression than VVC and judged both improved conventional technology and neural-network technology "interesting". Neural approaches showed clear perceptual gains, particularly on 8K HDR content, and smaller gains on gaming content. InterDigital, one respondent, submitted one design on the conventional Enhanced Compression Model (ECM) and one that overlaid AI tools on VVC.

The formal call for proposals was issued at the July 2026 meeting in Geneva. The main submissions were due at the end of October 2026, a supplemental package is due in mid-December, and the responses will be evaluated at the JVET meeting in Brisbane in January 2027. JVET expects a first test model in the second half of 2027 and a first version of the standard by the end of 2029, and calls the effort "probably a larger effort than ever done before to kick off a new standard".

The classical track sets a high bar. ECM, built with conventional signal processing, saves roughly 28% over VVC in random-access configuration. JVET's neural-network exploration reports in-loop filters that save about 6% at 10 thousand multiply-accumulates per pixel, 8% at 22 thousand, and 14% at a high-complexity operating point.

My reading is that H.267 will most likely be a classical core with a few neural tools where the bits saved justify the decoder cost, and that the January 2027 evaluation will show how many analysis.

There is already a bridge. VVC and HEVC streams can carry supplemental enhancement information (SEI) that describes a neural post-filter for the decoder side to apply, and the new editions of VVC and of its SEI standard (VSEI), finished in October 2025, extended those messages. A stream can say which network to run after decoding without making every decoder run it.

What this means for businesses in playback

Sell what works around the codec now: encode-side preprocessing for services, and post-processing blocks and tuning for TV and set-top chips. Do not build a playback business on a neural bitstream that has to decode on third-party devices before the 2030s. If you make silicon, the near-term revenue is AV2 and VVC decoders; the neural part is an NPU path for enhancement beside them (Chapter 8).

CHAPTER 03

PHOTOGRAPHY AND IMAGES

03 Photography and images: neural inside the camera, classical on the web

The camera pipeline is already neural

Every photo from a recent flagship phone has passed through neural networks before it is compressed. Qualcomm's Snapdragon 8 Gen 2, in phones from 2023, introduced a "Cognitive ISP" that runs real-time semantic segmentation during photo and video capture, adjusting faces, hair, skies and backgrounds separately; Samsung's Galaxy S23 Ultra was the first phone to use it on the selfie camera. In January 2026 Chips&Media and Visionary.ai showed at CES what they call the first full AI-based ISP: an end-to-end imaging pipeline that runs entirely on an NPU, Chips&Media's WAVE-N, in 16-bit floating point, without a fixed-function ISP, and can be updated over the air. "Existing pipelines bolt neural blocks onto a fixed-function ISP. We replace the conventional ISP entirely," Visionary.ai's CEO said. It was offered to manufacturers for evaluation after the show.

Then the result is saved as JPEG or HEIF, classical formats any device can open. The pattern is the same as in playback: neural inside one company's pipeline, standard at the boundary.

JPEG AI: a finished neural standard with no product yet

JPEG AI is the first learned image coding standard from ISO/IEC and ITU-T (ISO/IEC 6048, ITU-T T.840). Part 1, the core coding system, was published in 2025. The JPEG committee reports compression gains of 12.5% to 27.9% over VVC intra coding, across a range of decoder complexity from 7 to 216 thousand multiply-accumulates per pixel. Its design has two encoders and three decoders, trained jointly, so a receiver can implement only one decoder and still read every stream. By May 2024 the committee reported two smartphone implementations. In parallel, the IEEE published IEEE 1857.11-2024, a standard for neural-network-based image coding designed for NPUs, in December 2024.

As of November 2026 I found no phone camera, photo app or browser that writes or reads JPEG AI by default.

What the web adopted instead: JPEG XL

The image format that did gain ground in 2026 is classical. On October 6, 2026 Google enabled JPEG XL decoding by default in Chrome 155, using a new memory-safe decoder written in Rust, four years after it removed experimental support. Safari has decoded JPEG XL since 2023, and Mozilla has been moving the same way. Google's case for it is better compression than JPEG, lossless recompression of existing JPEG files, HDR and progressive decoding.

The lesson generalizes. A format that must open everywhere wins with a fast, deterministic decoder that runs on any CPU and can be made memory-safe. A neural image decoder needs an NPU or GPU and exact arithmetic across devices. Neural coding wins inside a device's own pipeline and in systems where the same company encodes and decodes, such as a cloud photo service that never hands the compressed file to anyone else.

What this means for businesses in imaging

The money is in the capture pipeline: AI ISP IP, NPUs sized for imaging, and the tuning work that camera makers buy. For storage and delivery, a learned image codec is a closed-system tool today, not a format to bet on.

CHAPTER 04

CALLS AND MEETINGS

04 Calls and meetings: where end-to-end neural codecs actually ship

Audio went first

Microsoft introduced Satin, an AI-powered speech codec, in February 2021, and said it was already used for all two-party calls in Teams and Skype. Satin delivers super-wideband speech, up to 16 kHz audio bandwidth, at 6 kbps, where Skype's earlier Silk codec was narrowband at that rate. The bitrate it saves is spent on redundancy, so calls survive packet loss better.

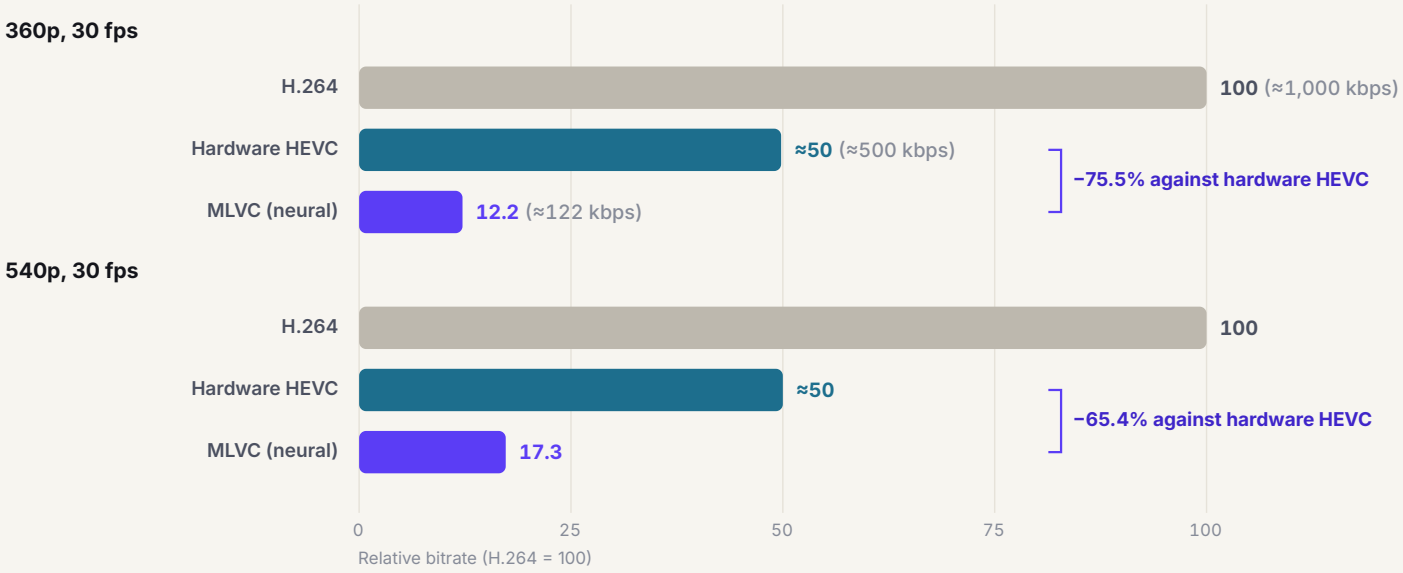
MLVC: a neural video codec in production calls

On July 24, 2026 Microsoft released MLVC, its machine-learning video codec, as open source under the MIT licence, with the model code, trained weights, training scripts and the conversion tools that target different NPUs. MLVC is the product version of the DCVC family Microsoft Research has published since 2021. Microsoft says it is rolling MLVC out in Teams for peer-to-peer calls, with live telemetry, A/B testing and the conventional codecs as fallback.

MLVC runs at 540p and 360p at 30 frames per second and is engineered to use no more than half of the NPU on Apple's Neural Engine and on Qualcomm and Intel NPUs, splitting the work between CPU and NPU. In subjective tests to ITU-T P.910 on Microsoft's own Video Conferencing Dataset, it needed 87.8% less bitrate than H.264 and 75.5% less than hardware HEVC at 360p, and 82.7% and 65.4% less at 540p vendor claim. At 360p that is roughly 122 kbps where H.264 needs about 1 Mbps.

FIG 04 **In calls, a neural codec beats hardware HEVC by 65–76%**

Bitrate needed for the same rated quality on Microsoft's video-conferencing test set, H.264 = 100. MLVC runs on laptop and phone NPUs at up to 540p30, using at most half of the NPU.



Source: Microsoft, MLVC release (July 24, 2026): ITU-T P.910 subjective tests on its Video Conferencing Dataset. HEVC bars derived from the reported savings. Vendor claims, not independent measurements.

Two things temper the headline. The 88% is against H.264; against the hardware HEVC that the same laptops already contain, it is 65–76%, still large. And MLVC gave up compression to be deployable: an independent analysis notes that a perceptually trained DCVC-RT scores 81.9% on the same measure where MLVC scores 75.5%. The price of decoding identically on hardware Microsoft does not own was about six points.

Why calls are the natural home

A call is a closed loop. Both endpoints run the same application, negotiate what they can do at the start of the call, and can fall back to a standard codec in the middle of it. Resolutions are modest, networks are often poor, and the laptops on both ends now ship with NPUs, as phones have for years. Microsoft sets 40 TOPS of NPU as the bar for its AI PCs, and Apple's M4 Neural Engine is rated at 38 TOPS. Every condition that blocks a neural codec in playback is absent here.

The counterexamples: Opus and MLow

Not every calling platform made the same choice, and the reasons are instructive.

Opus, the open audio codec in WebRTC and most calling apps, added deep learning in version 1.5 (March 2024), its first use of networks to process or generate the signal itself. But it added them as options: decoder-side enhancement (LACE and NoLACE) that needs no bitstream change, a neural packet-loss concealment, and a deep redundancy scheme (DRED); all are off by default so old devices are not burdened. Opus 1.6 added a neural bandwidth extension and a DRED model three times smaller, and stays fully compatible with the 2012 standard. The approach is "inside a standard, optional and backward compatible".

Meta went further back. In June 2024 it launched MLow, a low-bitrate speech codec built on classical signal processing rather than a network, now in all Instagram and Messenger calls and rolling out on WhatsApp. Meta's reasons were its devices: more than 20% of its calls are made on ARMv7 phones, and tens of millions of WhatsApp calls a day run on phones more than ten years old. Neural codecs, including Meta's own EnCodec, run reliably

only on expensive handsets, Meta wrote. MLow delivers about twice Opus's quality at 6 kbps while using 10% less compute.

For a service whose users span flagship laptops and decade-old phones, the complexity budget decides. Neural codecs ship first where the installed base is new and the vendor controls both ends.

What this means for businesses in communications

Enterprise conferencing, telepresence, remote assistance, contribution links and private networks are the places where an end-to-end neural codec is a product today, provided both ends are yours. Plan for three things from the start: a fallback codec, decode tests on every NPU you support, and a quality evaluation on your own content rather than a vendor's dataset.

CHAPTER 05



GENERATIVE AI

05 Generative AI: the largest deployment nobody calls a codec

Every voice and video model runs on a learned codec

The biggest deployment of end-to-end neural compression is not in any codec standard. It is inside generative AI. A speech model does not listen to waveforms or produce them directly: it reads and writes the discrete tokens of a learned audio codec. Kyutai's Mimi, released in 2024, turns speech into tokens at 12.5 frames per second and 1.1 kbps, lower than any telephony codec, combining semantic and acoustic information in one stream. Meta's EnCodec, published in 2022, is another; Meta's own engineers later described such codecs as excellent at low bitrates but too heavy for older phones (Chapter 4).

Video generation works the same way. A diffusion or transformer model generates in the latent space of a learned video autoencoder (VAE), and the VAE's decoder turns the result into pixels. LTX-Video's VAE compresses video 1:192: each latent token stands for a block of 32×32 pixels over 8 frames, with 128 channels. Its decoder also performs the final denoising step, so it partly generates detail rather than only reconstructing it.

Why these are not distribution formats

These codecs run on every request to a voice or video model, but none of them is a format anyone else decodes. Each model has its own latent space; there is no rate control, conformance test or guarantee that a different decoder would produce the same picture. They are designed for generation quality, not for fidelity at a given bitrate, and they decode on data-centre GPUs. Their success shows that learned compression works at scale when one company owns the encoder, the decoder and the hardware.

What this means for businesses

Two practical points. First, tokenizer efficiency is a direct cost line for anyone serving generative video or voice: for a transformer that attends over all tokens, compute grows faster than the number of tokens, so a more compressive latent cuts generation cost more than proportionally analysis. Codec engineers who understand latent design, entropy models and decoder efficiency are valuable to generative-AI teams. Second, the boundary is moving: standards bodies have started work on coding for machines rather than people. MPEG's Video Coding

for Machines (ISO/IEC 23888-2) reached draft international standard in early 2026 and is planned to be final in late 2026, and JPEG AI was designed from the start to serve computer-vision tasks as well as human viewing.

CHAPTER 06

GAMES AND GRAPHICS

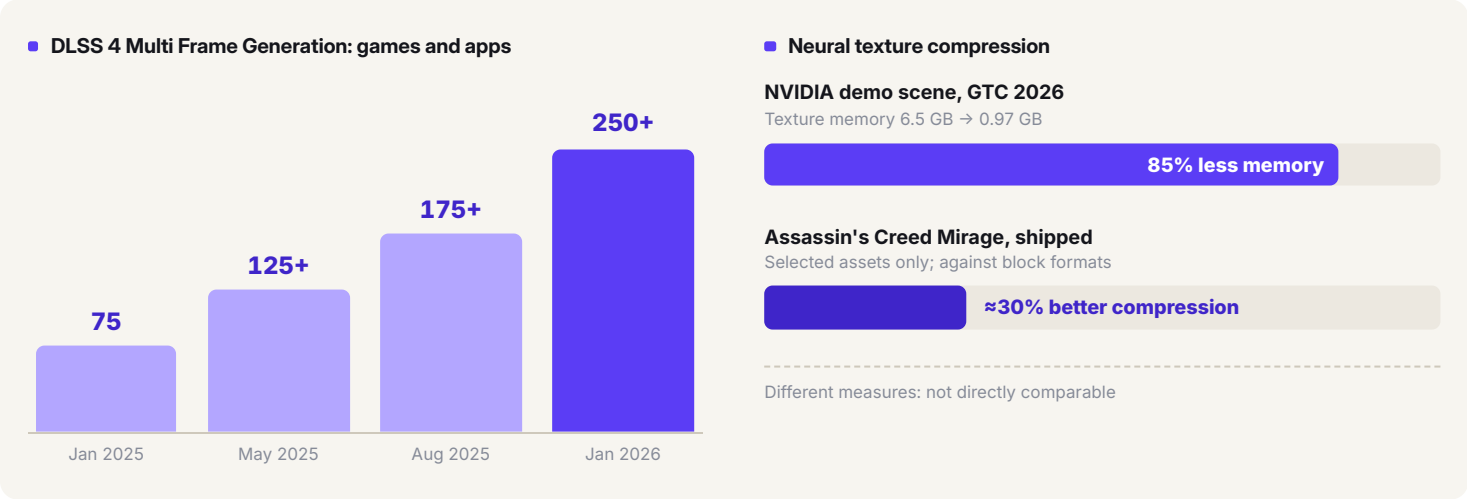
06 Games and graphics: neural reconstruction is mainstream, neural assets are starting

Rendering fewer pixels and letting a network fill in

In real-time graphics a network already reconstructs most of what players see. NVIDIA's DLSS renders fewer pixels and fewer frames and uses networks to produce the rest. DLSS 4 with Multi Frame Generation launched in January 2025 with support in 75 games and apps; NVIDIA counted more than 125 by May 2025, more than 175 by August and more than 250 by January 2026, when it announced DLSS 4.5 with a second-generation transformer model for super-resolution. That is a decoder in all but name: the "bitstream" is a low-resolution frame plus motion data, and the network rebuilds the full image.

FIG 05 **When one company controls the stack, neural reconstruction spreads fast**

Left: games and apps supporting NVIDIA's DLSS 4 Multi Frame Generation. Right: neural texture compression in a demo scene and in a shipped game, which used it only on selected assets because of its runtime cost.



Source: NVIDIA (January 2025 to January 2026; GTC 2026 demo); Ubisoft La Forge (February 2026). Vendor claims; the two texture figures measure different things.

Neural textures have shipped, selectively

Textures are compressed in fixed block formats (BCn) that GPUs decode in hardware. Neural texture compression replaces a set of material textures with a compact latent representation and a small network that reconstructs texels when they are needed. In February 2026 Ubisoft's La Forge lab described shipping it in Assassin's Creed Mirage, a 2023 game, with about 30% higher effective compression than the conventional formats at the same visual quality vendor claim. It was not used everywhere: "usage was constrained by runtime performance", so Ubisoft applied it to a subset of assets with high instance counts and heavy memory pressure.

NVIDIA's RTX Neural Texture Compression SDK, in beta on GitHub, shows the ceiling. In a demonstration at GTC 2026, a villa scene that used 6.5 GB of texture memory with BCn needed 970 MB with neural textures, 85% less vendor claim. The SDK offers two paths: decode at load time into BCn, which saves disk space and download size but not video memory, or decode at sampling time, which saves the memory but costs shader time every frame.

The API is arriving

Graphics APIs are adding the matrix operations that neural decoding needs inside shaders. Microsoft shipped Shader Model 6.9 in February 2026; the cooperative-vector feature it previewed in 2025 was folded into a unified linear-algebra design for a later shader model, with DirectX Linear Algebra entering public preview in April 2026. When that is in every driver, a neural texture decoder can run on any vendor's GPU from one code path.

What this means for businesses

Games ship neural decoding because the conditions of a closed loop hold: one engine, one asset pipeline, tensor hardware in the GPU and a fallback format. The open work is in tools, asset pipelines that encode neural materials, cross-vendor runtimes and, above all, mobile, where the energy cost per decoded texel decides.

CHAPTER 07

07 SPECIALIST LINKS
Specialist links: satellites, drones and robots

These are the cases where end-to-end neural coding fits best on paper. One owner controls both ends of the link, bandwidth is scarce and lossy, the content is narrow enough to train for, and a perceptually or even semantically faithful picture is often enough. Loss-resilient neural codecs for such links are an active research area, and server-side restoration of low-bitrate feeds is used.

They are also a small and increasingly regulated market compared with consumer video analysis. In China, Announcement No. 49 of 2026 from the Ministry of Commerce and two other bodies, issued on October 19 and effective November 18, 2026, revises export controls on 13 categories of drone-related items. Among the technology entries, one covers unmanned ground platforms defined partly by thresholds such as 75% semantic image compression and image reconstruction within 150 milliseconds, as reported in the trade press. Semantic compression itself has become an export-control parameter. In the United States, the Federal Communications Commission added foreign-made drones and their critical components to its Covered List on December 22, 2025, so new foreign models can no longer receive equipment authorization.

The verdict: a sound niche for a specialist company with industrial or government customers and the capacity to manage export compliance on both sides, but not a mainstream market, and not the place to learn whether neural coding works at scale.

CHAPTER 08

SILICON

08 The silicon view: energy, compute, memory and exactness

This is the chapter where my own background matters most. I have many years of fabless chip R&D and management behind me, including full-custom ASIC datapath design and multimedia camera-chip projects such as video codec accelerators, and the questions a chip team asks about a new codec are not about BD-rate. They are: how many joules per pixel, how many bytes of DRAM traffic per pixel, how many gates, and can every decoder in the world compute exactly the same bits? On all four, a neural decoder in 2026 is where hardware HEVC was before its first chips: promising in software, expensive in silicon.

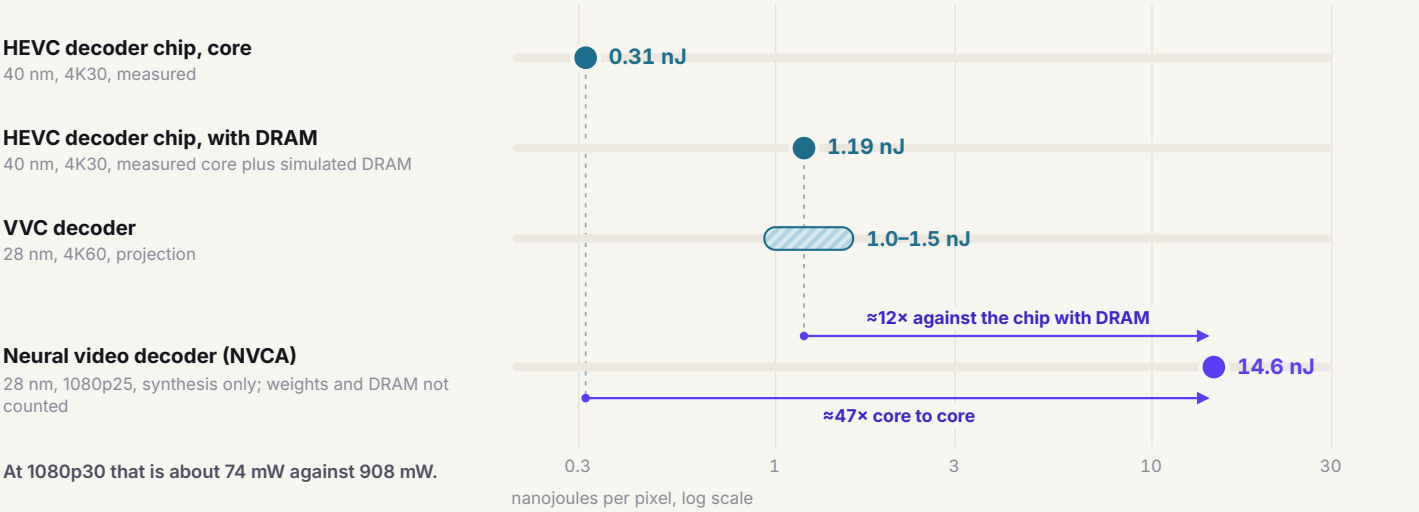
Energy per pixel: an order of magnitude apart

Few neural decoders have been designed as chips, so the comparison rests on a handful of published designs. The anchor on the classical side is MIT's HEVC decoder chip (Tikekar et al., ISSCC 2013 and IEEE JSSC 2014): 40 nm, 1.77 mm², decoding 4K at 30 frames per second at 200 MHz. Its core consumes 78 mW, or 0.31 nJ per pixel; counted with simulated DRAM power, the system needs 1.19 nJ per pixel. For VVC, production figures are proprietary; an analysis in Electronic Design in 2026, scaling the MIT chip to 28 nm and applying VVC's higher complexity, projects 1–1.5 nJ per pixel for a 4K60 decoder of about one million gates.

On the neural side, the only published end-to-end neural video decoder ASIC design, as Electronic Design counts it, is NVCA (Zhang et al., DATE 2024), synthesized in TSMC 28 nm with 5.01 million gates, not counting its weights, which sit off chip. It decodes 1080p at 25 frames per second at 760 mW, or 14.6 nJ per pixel.

FIG 06 **Energy per decoded pixel: fixed-function decoders are an order of magnitude ahead**

The only published neural video decoder chip design needs about 12 times the energy per pixel of a measured HEVC decoder chip counted with its DRAM, and about 47 times its core alone. The neural design's own weight and DRAM traffic is not counted.



Source: Tikekar et al., IEEE JSSC 2014 (MIT); NVCA, Zhang et al., DATE 2024 (arXiv 2312.10716); the VVC range is Electronic Design's projection (2026). The HEVC chip is on an older node (40 nm against 28 nm), which flatters the neural design.

That is about 12 times the HEVC chip counted with its DRAM, and about 47 times its core alone. The comparison flatters NVCA three ways: the HEVC chip is on an older node, NVCA's figure leaves out its weight and DRAM traffic, and NVCA is a synthesis result rather than measured silicon. At 1080p30 the difference is about 74 mW against roughly 900 mW: a decoder nobody notices, against one that shows in a phone's battery and thermal budget.

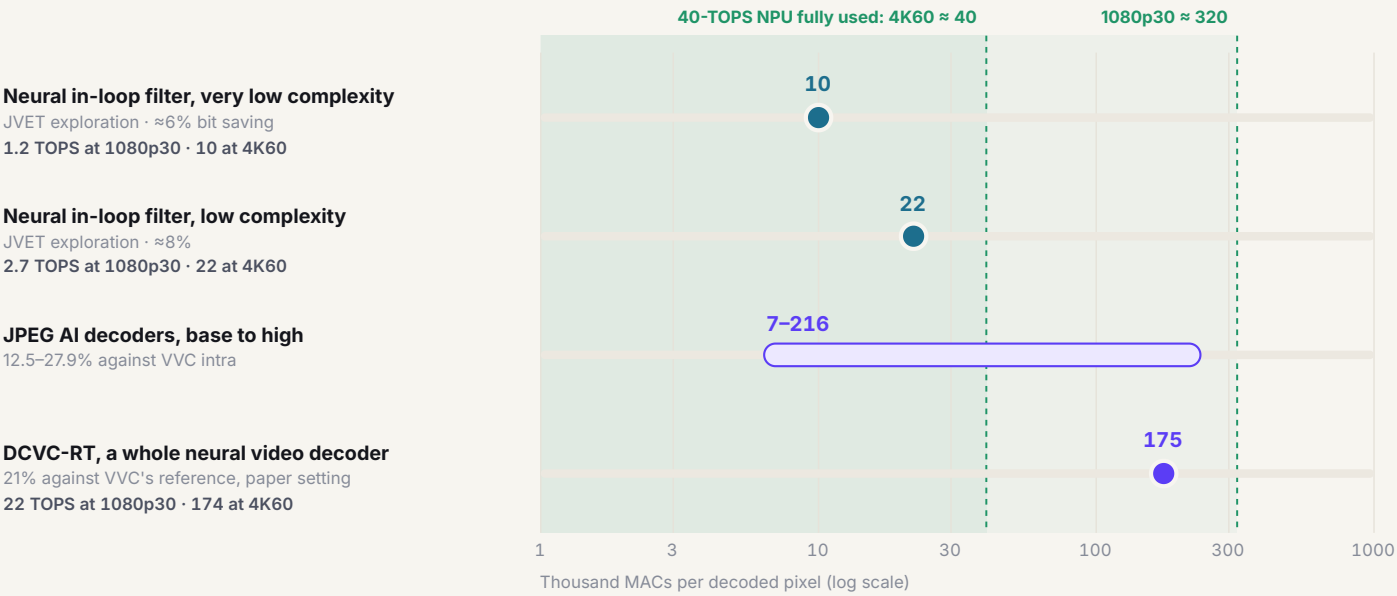
Compute per pixel: what a neural decoder asks of an NPU

The fastest published neural video codec I found is Microsoft Research's DCVC-RT (Jia et al., CVPR 2025). It encodes and decodes 1080p at 125 and 113 frames per second on an NVIDIA A100 GPU while saving 21% bitrate against VVC's reference encoder in the paper's test setting, and it uses integer arithmetic so that different devices decode consistently. Its supplementary material lists the complexity of each module. Adding the four modules a decoder must run (decoder, feature extractor, entropy model and reconstruction), I get about 175 thousand multiply-accumulates per decoded pixel estimate.

At 1080p30 that is about 11 trillion multiply-accumulates per second, or about 22 TOPS at two operations per MAC, before any loss of utilization; 4K60 needs eight times as much estimate. For scale, Microsoft defines its AI PCs by an NPU of at least 40 TOPS, and Apple rates the M4's Neural Engine at 38 TOPS. Those are peak figures, shared with everything else the device does.

FIG 07 **A neural tool costs tens of thousands of MACs per pixel; a whole neural decoder costs hundreds**

Decoder multiply-accumulates per pixel, log scale, against what a fully used 40-TOPS NPU could afford at 4K60 and at 1080p30. Gains are against different anchors and are not directly comparable.



Source: JVET via the SC 29 chair (April 2026); JPEG committee (2024); DCVC-RT, CVPR 2025 supplementary material, the author's sum of decoder modules; Microsoft's Copilot+ PC definition. The author's arithmetic: two operations per MAC, full utilization.

The ladder in Figure 7 is the useful way to read the field. Neural tools that cost tens of thousands of MACs per pixel, such as JVET's in-loop filters (about 6% saving at 10 thousand, 8% at 22 thousand), are plausible inside a future standard decoder. JPEG AI spans the range, from 7 to 216 thousand MACs per pixel for 12.5–27.9% over VVC intra. A whole neural video decoder at a couple of hundred thousand MACs per pixel belongs on an NPU, at modest resolution, in a closed loop. That is exactly where MLVC sits: 540p30, using at most half of the NPU.

Memory traffic, not MACs, is often the real limit

The DCVC-RT authors found that what limited speed was not arithmetic but "operational cost": memory input and output and the number of function calls. They removed explicit motion modules and multi-scale processing largely for that reason. NVCA's designers built a special dataflow to cut off-chip memory traffic from motion and residual features. A chip designer will recognize the pattern. In the MIT HEVC chip, DRAM accounts for about three-quarters of system energy (0.88 of 1.19 nJ per pixel), even after a cache that cut DRAM bandwidth by 67% analysis. A neural decoder that keeps large multi-channel feature maps between frames makes the memory problem worse, not better.

Exactness: every decoder must compute the same bits

Entropy coding works only if the decoder's probability model matches the encoder's exactly. In a classical codec that model is a table or a small set of integer rules. In a neural codec it is a network, and a floating-point difference between two devices, from a different kernel, a different accumulation order or a different NPU, can make the decoder read the wrong symbols and lose the stream from that point on.

Everyone who has shipped or standardized a neural codec has had to engineer this away. DCVC-RT converts its model to integers for consistent cross-device coding. MLVC uses a shared-scale hyperprior design to survive cross-platform mismatch, which is likely part of why it gives up compression against DCVC-RT analysis. JVET's neural-network exploration runs every network through SADL, a small library with 16-bit integer inference. JPEG

AI defines conformance points in the latent, residual and pixel domains. For a chip team this is the same discipline as HEVC's integer transforms: the standard has to define the arithmetic exactly, which constrains the network design and rules out vendor-specific shortcuts.

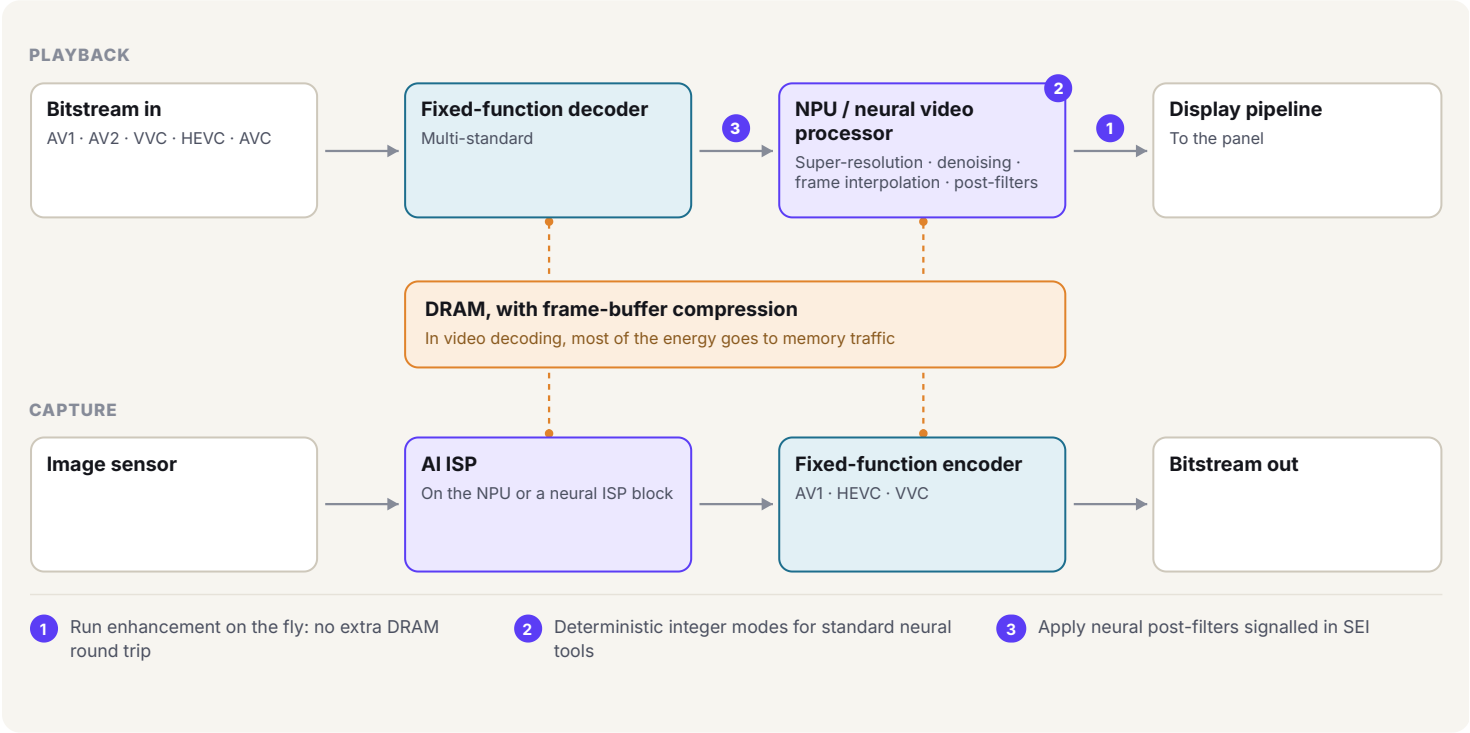
Area is not the obstacle; verification and lifetime are

NVCA's 5 million gates against about one million for a projected VVC decoder look large, but on a modern phone or TV SoC that already carries a multi-TOPS NPU, a few million gates is not what decides the design. What costs real money is verification against a conformance suite, keeping a programmable decoder correct across software updates for a device's lifetime, and power at 4K60 HDR. Those costs only make sense once a standard fixes what has to be built.

The architecture that wins now: a fixed-function codec plus an NPU

FIG 08 **The split that ships: fixed-function hardware for the bitstream, a neural engine for the pixels**

A 2026 media SoC. The codec block decodes every standard at a fraction of a nanojoule per pixel; the NPU does what networks are good at and standards do not constrain. DRAM traffic is the cost that matters.



Source: The author's architecture view, drawing on product disclosures from Tencent (Canghai V2), Chips&Media, Allegro DVT and Samsung (2026).

The silicon that sells in 2026 pairs a fixed-function multi-standard codec with a neural engine beside it. The codec block handles AV1, AV2, VVC, HEVC and AVC at a fraction of a nanojoule per pixel. The neural engine handles what networks are good at and standards do not constrain: denoising and the ISP on capture, super-resolution, frame interpolation and enhancement on display. The two share frame-buffer compression and the DRAM schedule, because memory traffic is the cost that matters.

The market shows the split. Tencent's Canghai (沧海) video codec ASIC, developed in-house since 2019, has more than 100,000 first-generation chips deployed across Tencent Cloud and its own live, on-demand and cloud-gaming services; the second generation, announced in May 2026 as powered on and entering volume production for service in the second half of the year, supports H.265 and H.266 in one architecture with more than 10% and 30% better compression than its predecessor, about twice the throughput, and on-chip picture enhancement such as super-resolution. Chips&Media sells codec IP and a video-oriented NPU, WAVE-N, side by side. Allegro

DVT sells AV2 decoders and has added its first neural video processing IP. Samsung's TV processors pair standard decoders with hundreds of upscaling networks.

What I would build, by role

SoC and TV-chip teams. Keep fixed-function decoders for every standard you must play, including AV2 now and VVC where broadcast requires it. Size the NPU for post-processing at display resolution and frame rate, and design the path from decoder to NPU to display so that enhancement does not add a round trip through DRAM. Support neural post-filter SEI. Provide deterministic integer modes in the NPU. Treat H.267 as a 2029-or-later design start, and keep some programmability for its neural tools.

IP vendors. Sell AV2 and VVC decoders now, plus the pieces neural processing needs around them: post-filter accelerators, frame-buffer compression, and conformance tools.

Video platforms. Neural preprocessing at the encoder pays back first, because it runs once per title and needs nothing from devices. Use client-side super-resolution where devices allow it. Keep end-to-end neural codecs for closed loops you control.

CHAPTER 09

STANDARDS AND PATENTS

09 Standards and patents: where neural tools become licensable

The standards pipeline

Standard	Body	Neural content	Status, November 2026
H.267 (working name; beyond VVC)	JVET (ITU-T and ISO/IEC)	Open. Classical tools (ECM) and neural tools (in-loop filters, intra prediction, super-resolution) are both candidates	Call for proposals July 2026; responses evaluated January 2027; first test model second half of 2027; version 1 end of 2029
VVC and HEVC, with VSEI	JVET	A neural post-filter can be described in SEI metadata; applying it is optional for the decoder	In force; messages extended in the October 2025 editions
AV2	AOMedia	Classical tools, royalty-free	Version 1.0.0 on May 28, 2026; decoder IP licensed from June 2026; YouTube web streams planned for early 2027
JPEG AI	JPEG (ISO/IEC and ITU-T)	End-to-end learned image codec, for human and machine vision	Part 1 published 2025; a second version under discussion; no default support in products found
IEEE 1857.11	IEEE	Neural-network image coding designed for NPUs	Published December 2024
Video Coding for Machines	MPEG	Coding for machine analysis rather than viewing	Draft international standard early 2026; final planned for late 2026

Standard	Body	Neural content	Status, November 2026
Opus ML extensions	IETF	Optional deep redundancy, decoder-side enhancement, bandwidth extension	In libopus 1.5 and 1.6; deep redundancy being standardized

Two things stand out. First, neural content enters standards at the edges before the core: as optional post-filters, optional enhancement and image coding, all of which leave the main decoding path unchanged. Second, the one standard that could put neural tools in the decoding path of every screen, H.267, is three years from a first version.

Who is positioned in the patents

Standard tools become licensable patents, so the H.267 process is also a patent race, and the neural tools are part of it. GreyB's 2025 landscape of 3,916 beyond-VVC patents puts ByteDance in front, with 645 patent families and 251 JVET contributions, and notes university–industry partnerships such as Xiaomi and Xidian University with more than 20 families. MLex reported in March 2026 that Tencent, Alibaba and ByteDance are emerging as active participants in video codec patent pools. In May 2026 Sauvegarder Investment Management bought 126 video-codec patent assets from Alibaba covering AV1, AV2 and Video Coding for Machines. InterDigital's purchase of Deep Render in October 2025 added Deep Render's AI video coding patents to its portfolio.

The money per device is easy to see in the pool rate cards. Access Advance's rates for VVC (as of July 1, 2026, for licensees in compliance and using its trademark) are $0.50 per mobile phone, tablet or laptop in Region 1 markets such as the US, EU and Japan, and $0.25 in Region 2; for 4K TVs, $1.00 to $1.50 depending on price in Region 1; with an annual enterprise cap of $75 million. AV1 and AV2 are royalty-free from AOMedia's members. Whatever H.267 contains, its tools will be priced the same way, which is why companies with neural-coding patents are contributing to JVET now analysis.

CHAPTER 10

10

WHO GETS PAID

Who gets paid, and where a company can play

The money in 2026

FIG 09 **Who gets paid from neural coding in 2026**

The money follows the standards and the platforms. I found no stand-alone company earning its living by licensing an end-to-end neural video codec.

EARNING NOW

Platforms

Keep the savings

Netflix's neural downscaler; YouTube and Meta super-resolution; Microsoft open-sourced MLVC under MIT

EARNING NOW

Device and chip makers

Sell differentiation

Camera ISPs; TV upscaling processors; Tencent's in-house Canghai codec chips

EARNING NOW

IP vendors

License each new standard

AV2 decoder customers at Chips&Media, Allegro DVT and VeriSilicon within three weeks of the spec

EARNING NOW

GPU vendors

Build platform lock-in

DLSS 4 in 250+ games and apps; neural texture SDK in beta

PER DEVICE

Patent holders and pools

Collect per device

VVC: $0.50 per phone, $1.00–1.50 per 4K TV; H.267 tools after 2029

EXIT BY ACQUISITION

Neural-codec startups

Acquired

WaveOne → Apple (2023); iSIZE → Sony (2023); Deep Render → InterDigital (2025)

Source: Company and patent-pool disclosures cited in the report; status as of November 12, 2026. Pool rates: Access Advance, VVC, Region 1, in compliance with trademark, as of July 1, 2026.

Platforms keep the savings. Netflix, YouTube, Meta, Microsoft and Tencent build neural processing in-house and keep the benefit as lower bandwidth, lower storage and better quality per bit. When it helps adoption they give the code away: Microsoft released MLVC under the MIT licence.

Device and chip makers sell differentiation. Camera ISPs and TV picture processors are where neural networks touch every user, and they are sold as features: Qualcomm's Cognitive ISP, Samsung's upscaling processors.

IP vendors sell decoders for every new standard, plus NPUs. Within three weeks of AV2's specification, Chips&Media, Allegro DVT and VeriSilicon all announced AV2 decoder customers. The same vendors now sell neural processing blocks beside the codecs.

GPU vendors build platforms. DLSS and neural texture compression make NVIDIA's tensor hardware part of how games are made.

Patent holders collect per device, through pools for VVC and HEVC and, after 2029, H.267.

Neural-codec startups have been acquired. WaveOne went to Apple in 2023, iSIZE to Sony Interactive Entertainment in 2023, and Deep Render to InterDigital in 2025, all on undisclosed terms. In each case a platform or a licensor bought the team and the patents. I found no stand-alone company in 2026 earning its living by licensing an end-to-end neural video codec.

Where a company can play

Play	Who buys	Why now	Main risk
Neural preprocessing at the encoder (downscaling, denoising, per-title tuning)	Streaming and user-video platforms, broadcasters	No device change needed; proven at Netflix	Large platforms build it themselves
Display-side enhancement IP and tuning (super-resolution, frame interpolation, picture quality)	TV, set-top and phone chip makers	NPUs are already in the devices; visible differentiation	Commoditization; chip makers in-house
Closed-loop neural codecs (conferencing, remote operations, private links)	Enterprise communications, industrial and telepresence vendors	MLVC proves feasibility, and its code is open	You must control both ends; NPU fragmentation
Codec plus NPU silicon or IP (AV2, VVC and post-processing)	SoC makers	The AV2 cycle is starting	Strong incumbents
Evaluation, conformance and deterministic-inference tooling	Anyone shipping neural tools	Bit-exactness is a hard requirement	Small market unless bundled
Standards contributions to H.267	Licensors, large platforms	Responses are evaluated in January 2027	Long, costly, uncertain

For most companies the first two rows are where revenue is available in 2027. The third is the most interesting for a focused product company. The last needs patience and a licensing strategy.

CHAPTER 11

11 CLAIMS AND EVIDENCE
Claims you will hear, and what the evidence says

Claim	What the evidence says
"Neural codecs already beat VVC."	In the lab, yes. DCVC-RT saves 21% against VVC's reference encoder at 1080p in its paper's setting; JPEG AI saves 12.5–27.9% against VVC intra. The decoder costs about 175 thousand MACs per pixel, or the energy of an NPU, against a fraction of a nanojoule in fixed-function hardware.
"AI codecs will replace H.26x and AV1 within a few years."	Not in playback. The first standard able to carry neural tools targets the end of 2029, chips follow around 2031 and mainstream use in the mid-2030s estimate. The codec arriving now, AV2, is classical.
"MLVC cuts bandwidth by 88%."	Against H.264 at 360p, on Microsoft's own test set. Against hardware HEVC it is 75.5% at 360p and 65.4% at 540p, in calls where Microsoft controls both ends.
"Neural textures cut video memory by 85–96%."	In demonstration scenes. The shipped game gained about 30% effective compression and used neural textures only on selected assets because of runtime cost.

Claim	What the evidence says
"JPEG AI is an ISO standard, so cameras and browsers will adopt it."	I found no default support in products. The format the web adopted in 2026, JPEG XL in Chrome 155, is classical.
"Neural decoding is too expensive for phones."	Partly true. MLVC decodes 540p30 within half an NPU; 4K60 with a DCVC-RT-class model would need around 175 TOPS of sustained compute estimate.
"Satellite and drone links are the big market for neural video."	Technically the best fit, but small and increasingly export-controlled on both sides of the Pacific.
"Big platforms will license a neural codec from a startup."	They built their own (Netflix, Microsoft, Tencent) or bought the startup (Apple, Sony, InterDigital).

A simple test cuts through most pitches: ask for the anchor, the metric, the resolution and the decode cost. A claim that cannot state all four is a research result, not a product.

CHAPTER 12

12

SIGNPOSTS

What to watch, 2027 to 2030

When	Signpost	Why it matters
January 2027	JVET evaluates the H.267 responses in Brisbane	How many top responses rely on neural tools, and at what cost per pixel
Early 2027	YouTube's first AV2 web streams; AV2 decoders in announced SoCs	The next classical cycle sets device budgets until about 2030
Second half of 2027	H.267's first test model	Whether neural tools are in the core design or only in exploration
2027	MLVC beyond peer-to-peer Teams calls; other vendors adopting the open code	The first test of an open neural codec outside its owner
2027–2028	A first product that writes or reads JPEG AI by default	The first neural format used across company boundaries
2027–2028	DirectX Linear Algebra in shipping drivers; neural textures on more games and on mobile	Cross-vendor neural decoding inside graphics
Ongoing	NPU budgets in phones and PCs	At around 100 TOPS sustained, a DCVC-RT-class 1080p30 decoder would take about a fifth of the NPU instead of half analysis
End of 2029	H.267 version 1	The start of the silicon cycle for neural tools in mainstream playback

CHAPTER 13

13 OPEN QUESTIONS Open questions

Several questions in this report have no public answer yet, and they are the ones I find most interesting technically.

Exactness at scale. How close can a neural decoder get to bit-exact, integer-only arithmetic across NPUs from different vendors before it gives up most of its compression advantage? MLVC's six-point trade is one data point; a standard needs a general answer.

Memory before MACs. What does neural post-processing cost in DRAM traffic per pixel when it runs on the fly between a hardware decoder and the display, and how small can its working set be made?

The price of a neural tool in H.267. At how many MACs and nanojoules per pixel does a neural in-loop filter stop paying for the bits it saves in a 4K60 decoder?

Closed loops beyond calls. Which other closed systems, from cloud photo storage and game streaming to robots and private contribution links, can adopt end-to-end neural codecs before H.267 arrives?

Decoders that generate. When a decoder generates detail rather than reconstructing it, as video VAEs already do, how should fidelity be measured and specified?

These are areas we can explore together in the years ahead, through technical exchange and technical collaboration: i@tanshuai.com.

Limitations of this report

This is desk research based on public documents up to November 12, 2026. I did no primary interviews and no independent measurements for this edition. Several key numbers are vendor or author claims on their own test sets, such as MLVC's bitrate savings and neural texture compression ratios; they are labelled as such. My compute estimates for DCVC-RT add up the per-module figures its authors publish and ignore utilization losses, so real NPU load will be higher. The energy comparison rests on a few published designs on different process nodes; production VVC and AV2 decoder figures are proprietary. Standards timelines are the committees' own targets. The H.267 dates after 2029 are my estimates from the AV1 and VVC histories.

About the author

Tan Shuai is a technologist and founder with more than twenty years of work across chips, hardware and system software, and 20 granted patents. He has many years of fabless chip R&D and management behind him, including full-custom ASIC datapath design and multimedia camera-chip projects such as video codec accelerators, which is the lens of this report: energy per pixel, memory traffic, and what a standard lets a decoder fix in silicon. His work has taken him through Silicon Valley, New York, Chicago, Taiwan, the Greater Bay Area (Shenzhen and Hong Kong), Vietnam and Malaysia. Today he works as an AI researcher and developer.

Technical discussion: i@tanshuai.com · tanshuai.com

Notes and disclaimer

This report is independent desk research based on public documents available up to November 12, 2026. It is provided for information only and is not investment, legal or engineering advice. Products, standards timelines, patent-pool rates and export rules change; verify anything you rely on against the primary source. Figures marked vendor claim, estimate or analysis should be read as such. Company and product names belong to their owners and are used for identification only; no company reviewed or sponsored this report.

© 2026 Tan Shuai. You may share this report unchanged with attribution.

Sources

All accessed between November 3 and November 12, 2026. Grouped by chapter.

Traffic, devices and scope

- Ericsson, *Mobility Report*, June 2026 (video about 75% of mobile data traffic at the end of 2025): https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/june-2026 ; summary: https://telecomlead.com/5g/ericsson-mobility-report-2026-global-mobile-data-traffic-to-reach-515-eb-by-2031-as-india-leads-smartphone-data-consumption-126290
- IDC, worldwide smartphone shipments, Q4 and full year 2025 (1.26 billion), January 13, 2026: https://www.idc.com/resource-center/press-releases/q425mobilephonetop5/
- TrendForce, 2025 TV shipment forecast (195.71 million): https://iconnect007.com/article/146406/early-shipments-by-tv-brands-push-global-1h25-volume-up-2-but-raise-concerns-for-secondhalf-peak-season/146403/ein ; Omdia, 2025 TV forecast (208.7 million): https://iconnect007.com/article/145954/omdia-forecasts-flat-tv-shipments-in-2025-but-mini-led-poised-for-explosive-growth-at-829/145951/ein

Playback and streaming

- Netflix Technology Blog, "AV1 — Now Powering 30% of Netflix Streaming", December 2025: https://netflixtechblog.com/av1-now-powering-30-of-netflix-streaming-02f592242d80
- Netflix Technology Blog, "For your eyes only: improving Netflix video quality with neural networks", November 2022: https://netflixtechblog.com/for-your-eyes-only-improving-netflix-video-quality-with-neural-networks-5b8d032da09c
- AOMedia, AV2 Bitstream and Decoding Process Specification v1.0.0, May 28, 2026: https://av2.aomedia.org/v1.0.0/index.html ; "AV2: What the Final Specification Means for Implementers": https://aomedia.org/blog%20posts/AV2-What-the-Final-Specification-Means-for-Implementers/
- TV Technology, "YouTube to Commence AV2 Streaming in 2027" (AV1 share of YouTube's catalogue): https://www.tvtechnology.com/platform/streaming/youtube-to-commence-av2-streaming-in-2027
- Google Research, "The in-loop filtering pipeline in AV2", 2026: https://research.google/pubs/the-in-loop-filtering-pipeline-in-av2/
- Fraunhofer IIS, "The Future of Broadcast in Brazil" (DTV+ commercial operation, June 2026): https://mpegh.com/brazil/ ; CartaCapital on receivers in TVs: https://www.cartacapital.com.br/?p=628337 ; Olhar Digital on VVC in TV 3.0: https://olhardigital.com.br/2026/05/01/pro/em-breve-a-tv-aberta-vai-mudar-no-brasil/
- 9to5Google, "YouTube adds AI-upscaled 'Super Resolution' for low-quality videos", October 29, 2025: https://9to5google.com/2025/10/29/youtube-adds-ai-upscaled-super-resolution-for-low-quality-videos/
- Meta, "How Meta Deployed Super Resolution at Scale to Transform Video Quality", @Scale, October 16, 2025: https://atscaleconference.com/how-meta-deployed-super-resolution-at-scale-to-transform-video-quality/
- Samsung, 8K TV highlights (NQ8 AI Gen3, 768 neural networks): https://www.samsung.com/us/tvs/8k-tv/highlights/

Photography and images

- Qualcomm, "Meet the Snapdragon 8 Gen 2 Super Cameras" (Cognitive ISP): https://www.qualcomm.com/news/onq/2023/07/meet-the-snapdragon-8-gen-2-super-cameras
- Chips&Media and Visionary.ai, full AI-based ISP, January 5, 2026: https://www.edge-ai-vision.com/2026/01/chipsmedia-and-visionary-ai-unveil-the-worlds-first-ai-based-full-image-signal-processor-redefining-the-future-of-image-quality/ ; PetaPixel interview: https://petapixel.com/2026/01/18/a-look-at-the-worlds-first-full-ai-based-image-signal-processor
- JPEG committee, 103rd meeting press release (JPEG AI gains, complexity, smartphone implementations): https://jpeg.org/items/20240507_press.html
- G. J. Sullivan, introduction to ISO/IEC JTC 1/SC 29, April 2026 (JPEG AI status; ECM and NNVC figures): https://www.mpeg.org/introduction-to-iso-iec-jtc-1-sc-29-coding-of-audio-picture-multimedia-and-hypermedia-information-the-world-of-jpeg-and-mpeg/
- IEEE 1857.11 working group (neural-network-based image coding): https://sagroups.ieee.org/fvc
- Chrome 155 and JPEG XL, October 6, 2026: https://xenospectrum.com/en/chrome-155-jpeg-xl-rust/

Calls and meetings

- Microsoft, "Satin: Microsoft's latest AI-powered audio codec for real-time communications", February 17, 2021: https://techcommunity.microsoft.com/t5/microsoft-teams-blog/satin-microsoft-s-latest-ai-powered-audio-codec-for-real-time/ba-p/2141382
- Microsoft, "Announcing the Open-Source Release of ML Video Codec (MLVC)", July 24, 2026: https://techcommunity.microsoft.com/blog/linuxandopensourceblog/announcing-the-open-source-release-of-ml-video-codec-mlvc/4539875 ; code: https://github.com/microsoft/mlvc
- Streaming Learning Center, "Microsoft MLVC: Interoperability, Power Realities, and the Neural Compression Patent Landscape": https://streaminglearningcenter.com/articles/microsoft_mlvc.html ; iReplay analysis: https://ireplay.tv/blog/mlvc-microsoft-neural-video-codec-open-source-streaming/

- Opus 1.5 release notes: https://opus-codec.org/demo/opus-1.5 ; Opus 1.6: https://opus-codec.org/demo/opus-1.6
- Meta Engineering, "MLow: Meta's low bitrate audio codec", June 13, 2024: https://engineering.fb.com/2024/06/13/web/mlow-metas-low-bitrate-audio-codec/
- Microsoft, Copilot+ PCs (40+ TOPS NPU): https://www.microsoft.com/en-us/windows/copilot-plus-pcs ; Apple, M4 announcement (38 TOPS Neural Engine): https://www.apple.com/newsroom/2024/05/apple-introduces-m4-chip/

Generative AI

- Kyutai, Mimi model card (12.5 Hz, 1.1 kbps): https://huggingface.co/kyutai/mimi
- Lightricks, "LTX-Video: Realtime Video Latent Diffusion" (1:192 video VAE): https://arxiv.org/abs/2501.00103
- MPEG 153rd meeting press release (Video Coding for Machines at DIS): https://www.mpeg.org/wp-content/uploads/2026/03/press_release_153.pdf

Games and graphics

- NVIDIA, DLSS 4 adoption: launch and Computex 2025 (75 and 125+): https://www.nvidia.com/en-us/geforce/news/125-dlss-4-multi-frame-gen-games-more-announced-computex-2025/ ; Gamescom 2025 (175+): https://www.nvidia.com/en-us/geforce/news/gamescom-2025-dlss-4-full-ray-tracing-game-announcements/ ; CES 2026 (250+, DLSS 4.5): https://www.nvidia.com/en-us/geforce/news/dlss-4-rtx-path-tracing-game-announcements-ces-2026/
- Ubisoft La Forge, "Shipping Neural Texture Compression in Assassin's Creed Mirage", February 2026: https://www.ubisoft.com/en-us/studio/laforge/news/415slmB3ZGzv8d2LhO9QT/shipping-neural-texture-compression-in-assassins-creed-mirage
- TweakTown on NVIDIA's GTC 2026 neural texture demo (6.5 GB to 970 MB): https://www.tweaktown.com/news/110886/nvidias-neural-texture-compression-cuts-vram-usage-from-6-5gb-down-to-970mb/index.html
- Microsoft DirectX blog, Shader Model 6.9 retail, February 26, 2026: https://devblogs.microsoft.com/directx/shader-model-6-9-retail-and-more/ ; on cooperative vector: https://devblogs.microsoft.com/directx/shader-model-6-9-and-the-future-of-cooperative-vector/ ; GDC 2026 plans: https://windowsreport.com/microsoft-showcases-neural-rendering-and-dxr-2-0-plans-for-directx-at-gdc-2026/

Specialist links

- Global Trade Alert on China's Announcement No. 49 of 2026: https://globaltradealert.org/intervention/161396-china-government-revises-export-controls-on-drone-related-items ; DroneXL: https://dronexl.co/2026/10/19/china-drone-export-controls-motors-props-november-18/ ; Army Recognition (unmanned-ground-platform thresholds): https://www.armyrecognition.com/news/army-news/2026/china-tightens-drone-export-controls-raising-supply-chain-risks-for-the-us-europe-and-ukraine
- DroneLife, FCC adds foreign-made drones and components to the Covered List, December 22, 2025: https://dronelife.com/2025/12/22/fcc-adds-foreign-made-drones-and-components-to-covered-list-citing-national-security-risks/

Silicon

- M. Tikekar et al., "A 249-Mpixel/s HEVC Video-Decoder Chip for 4K Ultra-HD Applications", IEEE JSSC 2014: https://dspace.mit.edu/handle/1721.1/93876 ; MIT chip page (78 mW at 4K30): https://eecsg.mit.edu/chips/hevc-video-decoder-for-4k-ultra-hd-applications/
- S. Zhang et al., NVCA, DATE 2024: https://arxiv.org/abs/2312.10716
- Electronic Design, "Why Traditional VVC Will Outlive the Neural-Codec Hype on Silicon", 2026: https://www.electronicdesign.com/home/article/55414955/why-traditional-vvc-will-outlive-the-neural-codec-hype-on-silicon
- Z. Jia et al., "Towards Practical Real-Time Neural Video Compression" (DCVC-RT), CVPR 2025: https://openaccess.thecvf.com/content/CVPR2025/html/Jia_Towards_Practical_Real-Time_Neural_Video_Compression_CVPR_2025_paper.html ; supplementary (per-module complexity): https://openaccess.thecvf.com/content/CVPR2025/supplemental/Jia_Towards_Practical_Real-Time_CVPR_2025_supplemental.pdf
- NNVC and the SADL integer inference library: https://arxiv.org/abs/2309.05846
- Tencent Canghai V2, May 27, 2026 (Shanghai Securities News via 10jqka): https://stock.10jqka.com.cn/20260527/c677016020.shtml ; feature summary: https://www.besthub.dev/articles/tencent-s-canghai-v2-chip-enters-mass-production-what-the-new-capabilities-mean-for-video-encoding-bcfc18dbfa29
- Chips&Media AV2 decoder IP, May 12, 2026: https://www.design-reuse.com/news/202530501-chips-media-completes-development-of-next-gen-av2-hw-decoder-ip/ ; AV2 licence, June 17, 2026: https://www.edge-ai-vision.com/?p=58270
- Allegro DVT, AV2 decoder IP, June 16, 2026: https://www.edge-ai-vision.com/2026/06/allegro-dvts-pulsar-decoder-ip-adds-support-for-av2-video-codec/
- VeriSilicon, VC9800D with AV2, June 10, 2026: https://www.businesswirechina.com/en/news/63176.html

Standards and patents

- MPEG 155th meeting press release (joint call for proposals beyond VVC), August 5, 2026: https://www.mpeg.org/wp-content/uploads/2026/08/ag3w254SN26474Press-Release-MPEG155.pdf
- MPEG 152nd meeting press release (call for evidence results; VVC and VSEI editions), October 30, 2025: https://www.mpeg.org/wp-content/uploads/2025/11/Press-Release-MPEG152.pdf ; ACM SIGMM MPEG column: https://records.sigmm.org/2025/12/09/mpeg-column-152nd-mpeg-meeting/
- Streaming Media, "AI Helps InterDigital Reach Beyond VVC in Race to Develop Next-Gen Codec": https://www.streamingmedia.com/Articles/News/Online-Video-News/AI-Helps-InterDigital-Reach-Beyond-VVC-in-Race-to-Develop-Next-Gen-Codec-171971.aspx

- GreyB, beyond-VVC (H.267) patent landscape: https://greyb.com/blog/exclusive-video-codec-landscape/ ; on AV2 licensing and the Alibaba portfolio sale: https://greyb.com/blog/av2-licensing/
- MLex, "Fragmented codec licensing exposes need for clearer rules for Chinese tech majors", March 27, 2026: https://www.mlex.com/mlex/articles/2458402/fragmented-codec-licensing-exposes-need-for-clearer-rules-for-chinese-tech-majors
- Access Advance, MCBA and VVC Advance royalty rates as of July 1, 2026: https://accessadvance.com/wp-content/uploads/2025/07/MCBA-Program-Overview-September-2026.pdf

Money and companies

- InterDigital acquires Deep Render, October 30, 2025: https://www.advanced-television.com/2025/10/30/interdigital-acquires-deep-render
- TechCrunch, "Apple acquired a startup using AI to compress videos", March 27, 2023: https://techcrunch.com/2023/03/27/apple-acquired-a-startup-using-ai-to-compress-videos
- Neowin, "Sony acquires AI video startup iSIZE", November 2023: https://www.neowin.net/news/sony-acquires-ai-video-startup-isize-to-enhance-playstation-streaming/

LEARNED COMPRESSION · PLAYBACK × SILICON

Neural Coding in 2026

It ships around the codec, not instead of it

Every key number in this report carries its source and date. Where the author derives a number, it is marked as an estimate; where only the seller reports it, it is marked as a vendor claim.

Tan Shuai

Technologist and founder · AI researcher and developer

tanshuai.com · i@tanshuai.com

PUBLIC EDITION · Data as of November 12, 2026