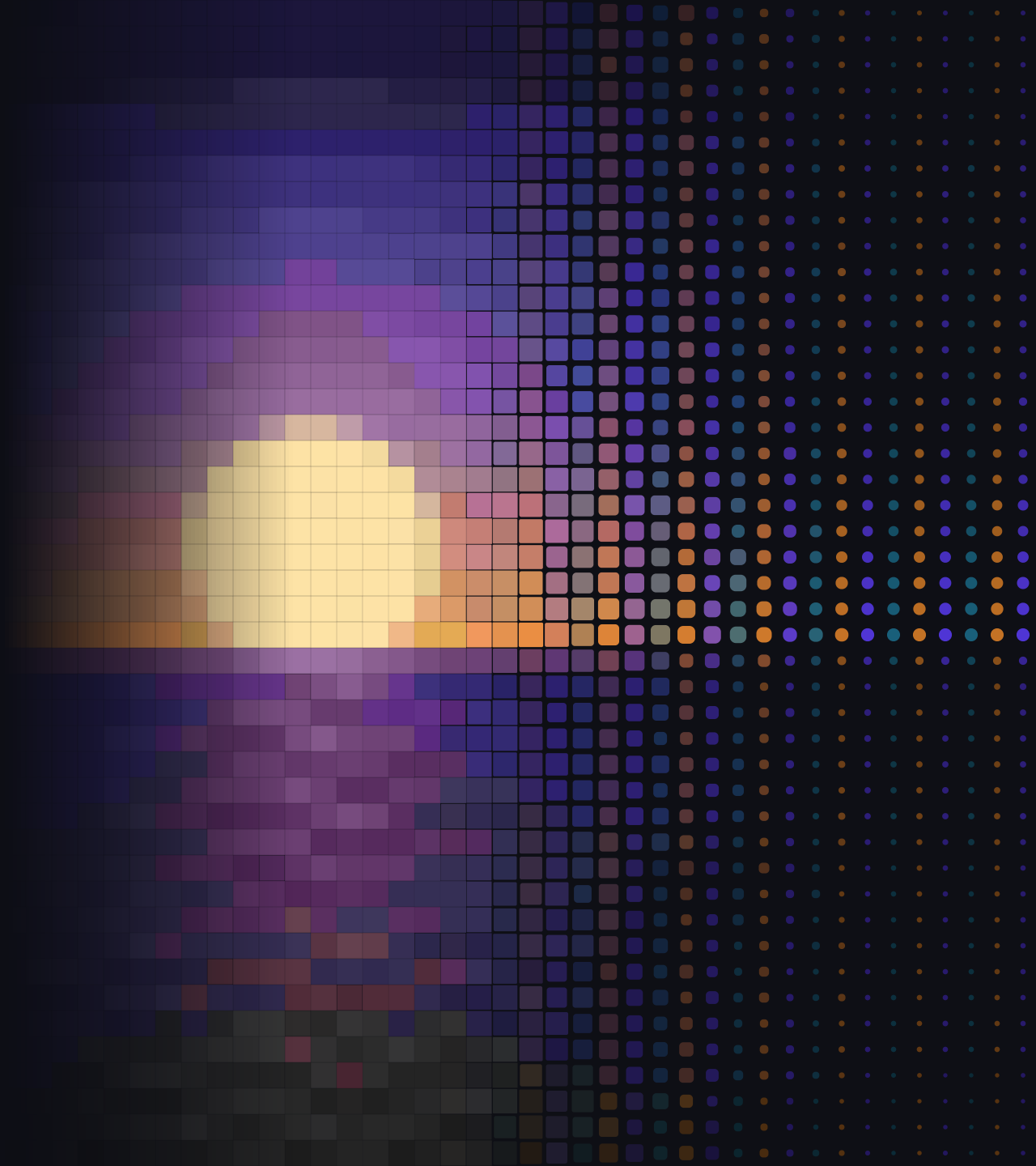
Onderzoeksnotitie · november 2026

Neurale codering in 2026

Rond de codec, niet in plaats ervan

Afspelen · Fotografie · Bellen · Generatieve AI · Games · Silicium

Tan Shuai
tanshuai.com · i@tanshuai.com



AI gemeengoed, maar rond de codec

01 Neurale netwerken zijn gemeengoed aan beide kanten van standaardcodecs

Camera-ISP's, downscaling aan de encoderkant en superresolutie bij het afspelen. De bitstream daartussen is nog altijd AV1, HEVC, AVC of VVC, gedecodeerd door fixed-function hardware.

02 End-to-end neurale codecs worden alleen in gesloten ketens ingezet

Gesprekken, tokenizers voor generatieve AI en één AAA-game: één bedrijf beheerst beide kanten en kan terugvallen op een standaardcodec.

03 Neurale bitstreams op gangbare schermen wachten op H.267

Versie 1 wordt eind 2029 verwacht; eerste chips rond 2031 en breed gebruik halverwege de jaren 2030 schatting.

≈75%

van het mobiele dataverkeer is video (eind 2025)

1,26 miljard

smartphones geleverd in 2025, die allemaal video decoderen in fixed-function hardware

≈12×

de energie per pixel van een HEVC-hardwaredecoder, voor het enige gepubliceerde ontwerp van een neurale decoderchip

DEFINITIES

Drie manieren waarop een neuraal netwerk een codec binnenkomt



> Rond de codec: niemands toestemming nodig. Binnen een standaard: een standaard nodig. In plaats van de codec: controle over beide kanten nodig.

Waar neurale codering in 2026 wordt ingezet

	Rond een standaardcodec voorbewerking, verbetering, upscaling	End-to-end neurale codec getrainde encoder en decoder	Binnen een standaard neurale tools in een standaardbitstream of -API
01 **Afspelen en streaming**	**GROOTSCHALIG** Downscaler van Netflix; superresolutie bij YouTube en Meta; AI-upscalers in tv's	**NIET INGEZET** Geen streamingdienst verstuurt een neuraal videoformaat; tv's en telefoons decoderen standaardcodecs in hardware	**IN STANDAARDISATIE** H.267 versie 1 wordt eind 2029 verwacht; neurale tools behoren tot de kandidaten
02 **Fotografie en beeldverwerking**	**GROOTSCHALIG** AI-ISP's in vlaggenschiptelefoons; een volledig neurale ISP als IP getoond op CES 2026	**DEMO'S, PILOTS** JPEG AI is een ISO-standaard: telefoondemo's, geen product	**NIET INGEZET** Het formaat dat terrein wint, is klassiek: JPEG XL standaard aan in Chrome 155
03 **Bellen en vergaderen**	**GROOTSCHALIG** Superresolutie aan ontvangerszijde en neurale ruisonderdrukking	**GROOTSCHALIG** Neurale audio in Teams-gesprekken met twee deelnemers sinds 2021; MLVC-video in uitrol in Teams	**BEPERKT INGEZET** Opus 1.5 voegt optionele, achterwaarts compatibele ML-tools toe
04 **Generatieve AI**	—	**GROOTSCHALIG** Elk video- en spraakmodel draait een getrainde tokenizer, bijv. een video-VAE van 1:192	—
05 **Games en graphics**	**GROOTSCHALIG** Neurale upscaling en framegeneratie: alleen al DLSS 4 in 250+ games en apps	**BEPERKT INGEZET** Neurale textures in één uitgebrachte AAA-game; de SDK van NVIDIA is in bèta	**IN STANDAARDISATIE** DirectX krijgt matrixbewerkingen voor shaders (preview)
06 **Gespecialiseerde verbindingen: satelliet, robots**	**DEMO'S, PILOTS** Herstel aan serverzijde van feeds met lage bitrate	**DEMO'S, PILOTS** Verliesbestendige neurale codecs in onderzoek en pilots	—

De rijen zijn geordend naar het aantal mensen dat elk domein bereikt.

Neuraal aan beide kanten, standaard in het midden

- **Bij de encoder**

  De neurale downscaler van Netflix draait sinds 2022 zonder enige apparaatwijziging; in tests van Netflix verkoos zo'n 77% van de kijkers het resultaat leverancierclaim.

- **Bij het scherm**

  YouTube Super Resolution op tv's sinds oktober 2025; Meta past superresolutie toe op zijn servers en op telefoons.

- **In de tv**

  Samsung prijst zijn NQ8 AI Gen3-processor aan met 768 neurale netwerken voor upscaling.

DE BITSTREAM BLIJFT STANDAARD

30%

van de kijktijd bij Netflix in AV1 (december 2025)

90%+

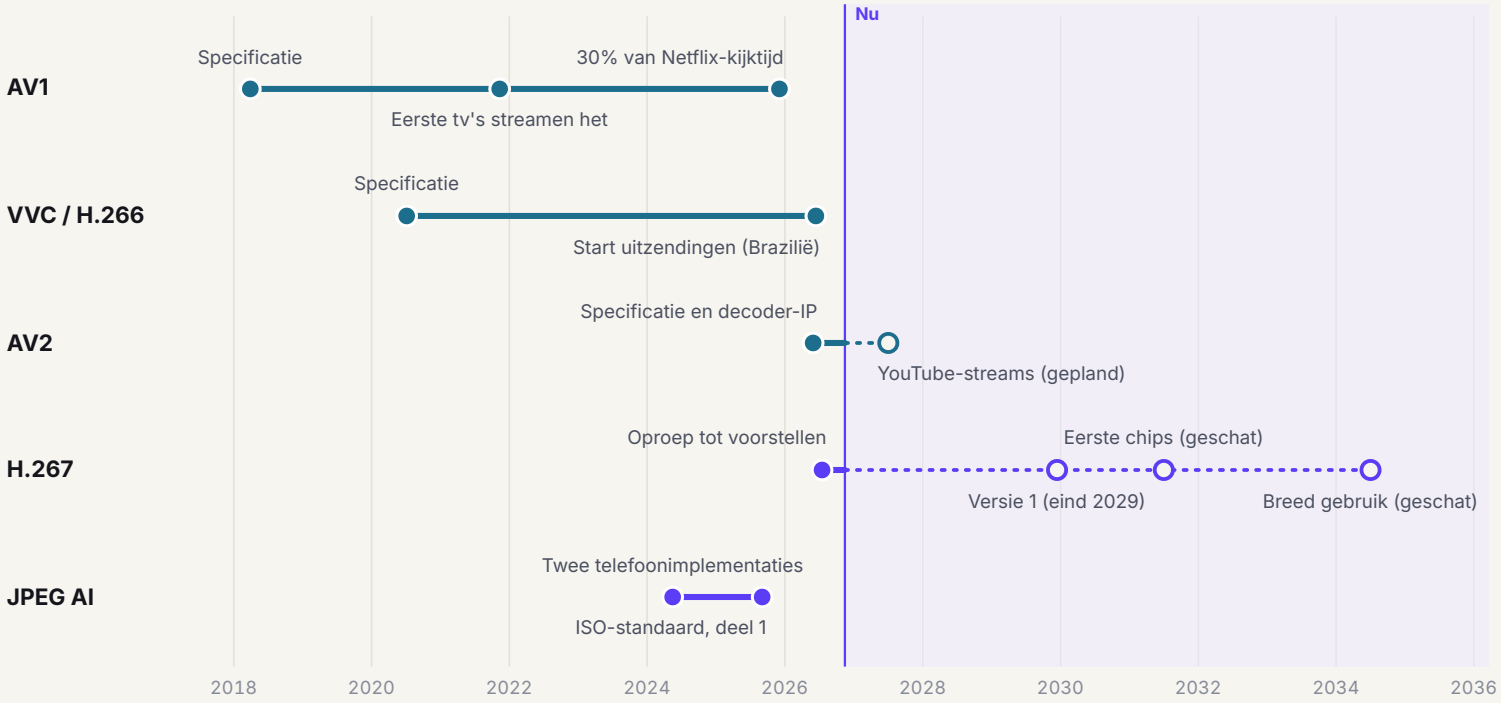
van de YouTube-catalogus, gewogen naar kijktijd, beschikbaar in AV1 (maart 2026)

3 weken

van AV2 versie 1.0.0 (28 mei 2026) tot klanten voor decoder-IP bij drie leveranciers

TIJDPAD

Van voltooide standaard tot gangbare schermen: het grootste deel van een decennium



AV1 had bijna acht jaar nodig van zijn specificatie (2018) tot 30% van de kijktijd bij Netflix (december 2025). Open markeringen zijn doelen of mijn schattingen.

H.267 beslist of neurale tools elk scherm bereiken

Oktober 2025
Oproep om bewijs: vijf inzendingen van 16 bedrijven en labs; neurale winst het duidelijkst bij 8K HDR

Juli 2026
Oproep tot voorstellen uitgebracht in Genève

Januari 2027
Inzendingen beoordeeld in Brisbane

Tweede helft van 2027
Eerste testmodel

Eind 2029
Versie 1 van de standaard

≈28%
besparing t.o.v. VVC door het klassieke ECM-spoor

6–14%
besparing door neurale in-loopfilters bij 10 tot 22+ duizend MAC's per pixel

SEI
VVC-streams kunnen al een neuraal postfilter signaleren dat decoders mogen toepassen

Mijn interpretatie: een klassieke kern met enkele neurale tools waar de bespaarde bits de decoderkosten rechtvaardigen analyse.

Neuraal in de camera, klassiek op het web

- **Camera**

  De Cognitive ISP van Qualcomm voert sinds 2023 realtime segmentatie uit in telefoons; op CES 2026 werd een volledige AI-ISP getoond die op een NPU draait.

- **JPEG AI**

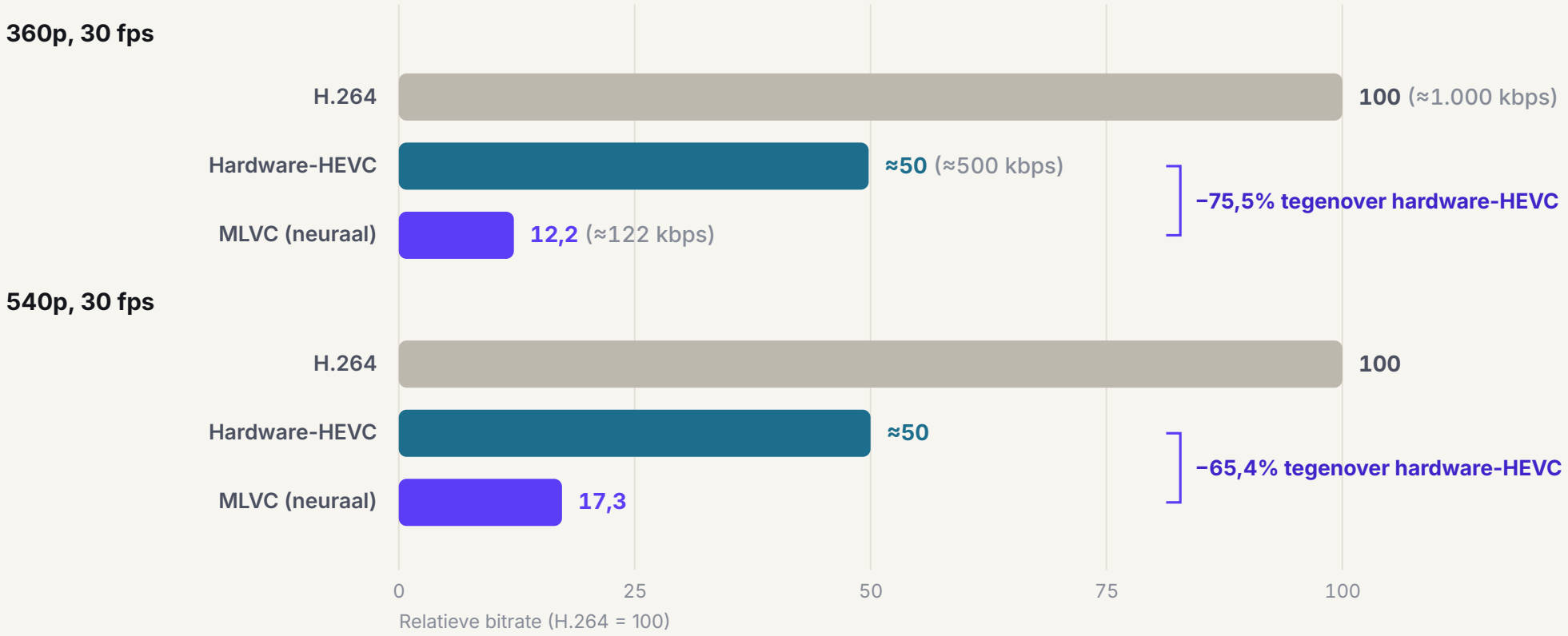
  Gepubliceerd in 2025: 12,5–27,9% beter dan VVC intra bij 7 tot 216 duizend MAC's per pixel. Ik vond geen product dat het standaard gebruikt.

- **Web**

  Chrome 155 schakelde JPEG XL op 6 oktober 2026 standaard in, met een geheugenveilige decoder geschreven in Rust.

Formaten die overal moeten openen, blijven klassiek; neurale codering wint binnen de pijplijn van één bedrijf.

Bellen: hier worden end-to-end neurale codecs ingezet



Satin draait sinds 2021 in alle Teams- en Skype-gesprekken tussen twee deelnemers. MLVC, sinds 24 juli 2026 open source onder MIT, wordt uitgerold in peer-to-peer-gesprekken in Teams tot 540p30, met hooguit de helft van de NPU leverancierclaim.

Niet elk platform koos voor neuraal, en de redenen zijn leerzaam

BINNEN EEN STANDAARD, OPTIONEEL

Opus 1.5 en 1.6

Neurale verbetering, maskering van pakketverlies, diepe redundantie en bandbreedte-uitbreiding, allemaal uit in de basisinstelling en volledig compatibel met de standaard uit 2012.

KLASSIEK, BEWUST GEKOZEN

Meta MLow (2024)

Ruim 20% van de gesprekken van Meta loopt via ARMv7-telefoons, en dagelijks tientallen miljoenen WhatsApp-gesprekken via telefoons ouder dan tien jaar. MLow haalt bij 6 kbps zo'n twee keer de kwaliteit van Opus, met 10% minder rekenwerk.

Het complexiteitsbudget van de geïnstalleerde basis beslist. Neurale codecs komen het eerst waar apparaten nieuw zijn en één leverancier beide kanten beheerst.

GENERATIEVE AI

De grootste toepassing die niemand een codec noemt

- **Overal getrainde codecs**

  Elk spraak- en videomodel leest en schrijft de tokens of latente waarden van een getrainde codec.

- **Geen distributieformaten**

  Elk model heeft zijn eigen latente ruimte: geen bitratesturing, geen conformiteitstest, geen tweede decoder.

- **Een kostenpost**

  Voor wie generatieve video aanbiedt, verlaagt een sterker comprimerende latente representatie de generatiekosten meer dan evenredig analyse.

TWEE VOORBEELDEN

12,5 Hz · 1,1 kbps

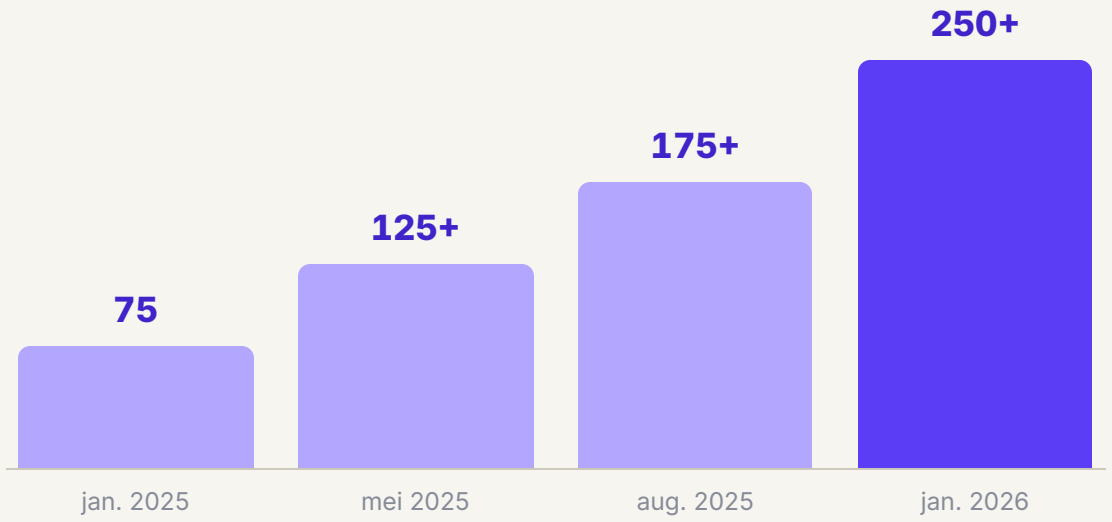
spraaktokens van Mimi (Kyutai), lager dan elke telefoniecodec

1:192

video-VAE van LTX-Video: één latent token voor 32×32 pixels over 8 frames

Neurale reconstructie is gemeengoed; neurale assets komen op gang

DLSS 4 Multi Frame Generation: games en apps



Neurale texturecompressie

Demoscène van NVIDIA, GTC 2026
Texturegeheugen 6,5 GB → 0,97 GB

85% minder geheugen

Assassin's Creed Mirage, uitgebracht
Alleen geselecteerde assets; tegenover blokformaten

≈30% betere compressie

Verschillende maatstaven: niet direct vergelijkbaar

> DirectX Linear Algebra ging in april 2026 in openbare preview: één codepad voor neurale decodering in shaders op GPU's van elke leverancier.

Satellieten, drones en robots: de beste match, een kleine markt

- **Waarom het past**

  Eén eigenaar beheerst beide kanten, bandbreedte is schaars en de inhoud is specifiek genoeg om op te trainen.

- **China**

  Bekendmaking nr. 49 van 2026, van kracht op 18 november, herziet de exportcontrole op drones; één post gebruikt semantische beeldcompressie als drempelwaarde.

- **Verenigde Staten**

  De FCC zette op 22 december 2025 in het buitenland gemaakte drones en hun kritieke onderdelen op haar Covered List.

> **Een gezonde niche voor een specialist die exportcompliance aankan, geen massamarkt.**

Een ordegrootte meer energie per pixel

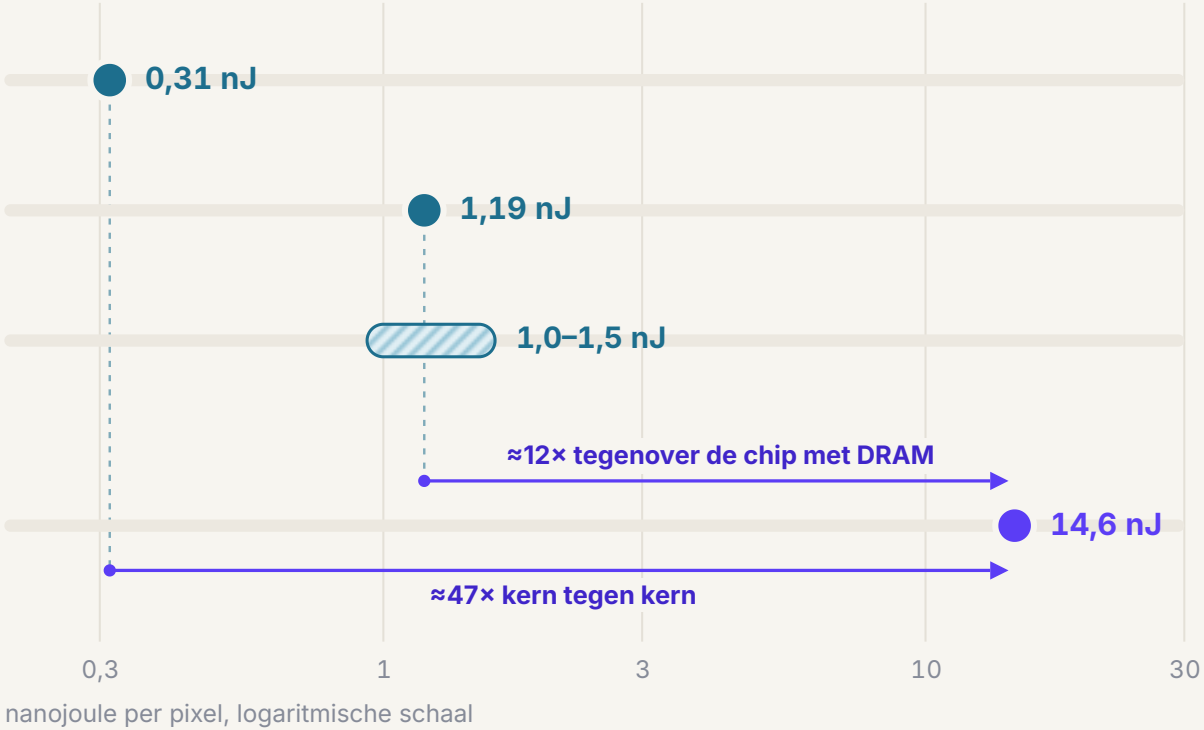
HEVC-decoderchip, kern
40 nm, 4K30, gemeten

HEVC-decoderchip, met DRAM
40 nm, 4K30, gemeten kern plus gesimuleerd DRAM

VVC-decoder
28 nm, 4K60, projectie

Neurale videodecoder (NVCA)
28 nm, 1080p25, alleen synthese; gewichten en DRAM niet meegeteld

Bij 1080p30 is dat ongeveer 74 mW tegen 908 mW.



Bij 1080p30: ongeveer 74 mW voor de HEVC-chip met DRAM tegen zo'n 900 mW voor NVCA. Een ouder procedé voor de HEVC-chip, gewichten en DRAM niet meegeteld voor NVCA: de vergelijking flatteert het neurale ontwerp.

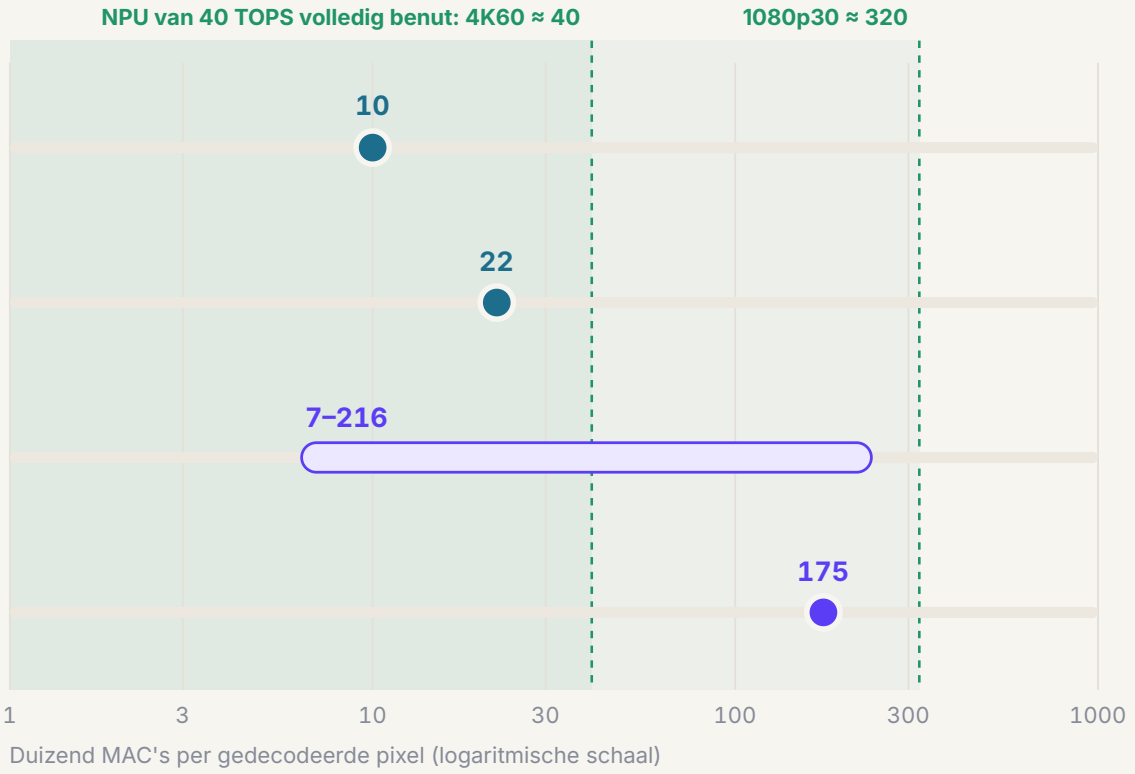
Wat een neurale decoder van een NPU vraagt

Neuraal in-loopfilter, zeer lage complexiteit
JVET-verkenning · ≈6% bitbesparing
1,2 TOPS bij 1080p30 · 10 bij 4K60

Neuraal in-loopfilter, lage complexiteit
JVET-verkenning · ≈8%
2,7 TOPS bij 1080p30 · 22 bij 4K60

JPEG AI-decoders, basis tot hoog
12,5–27,9% tegenover VVC intra

DCVC-RT, een complete neurale videodecoder
21% tegenover VVC-referentie, opzet van de paper
22 TOPS bij 1080p30 · 174 bij 4K60



DCVC-RT heeft ongeveer 175 duizend MAC's per pixel nodig: ongeveer 22 TOPS bij 1080p30 en 174 bij 4K60, bij volledige benutting **schatting**. Een volledig benutte NPU van 40 TOPS kan bij 4K60 ongeveer 40 duizend aan.

Geheugenverkeer en bitexactheid zijn de harde grenzen

DRAM DOMINEERT

Geheugen vóór MAC's

In de HEVC-chip van MIT is DRAM goed voor ongeveer driekwart van de systeemenergie: 0,88 van 1,19 nJ per pixel. Neurale decoders slepen grote feature maps mee tussen frames, en de auteurs van DCVC-RT stelden vast dat geheugen-I/O, niet rekenwerk, de snelheid beperkte.

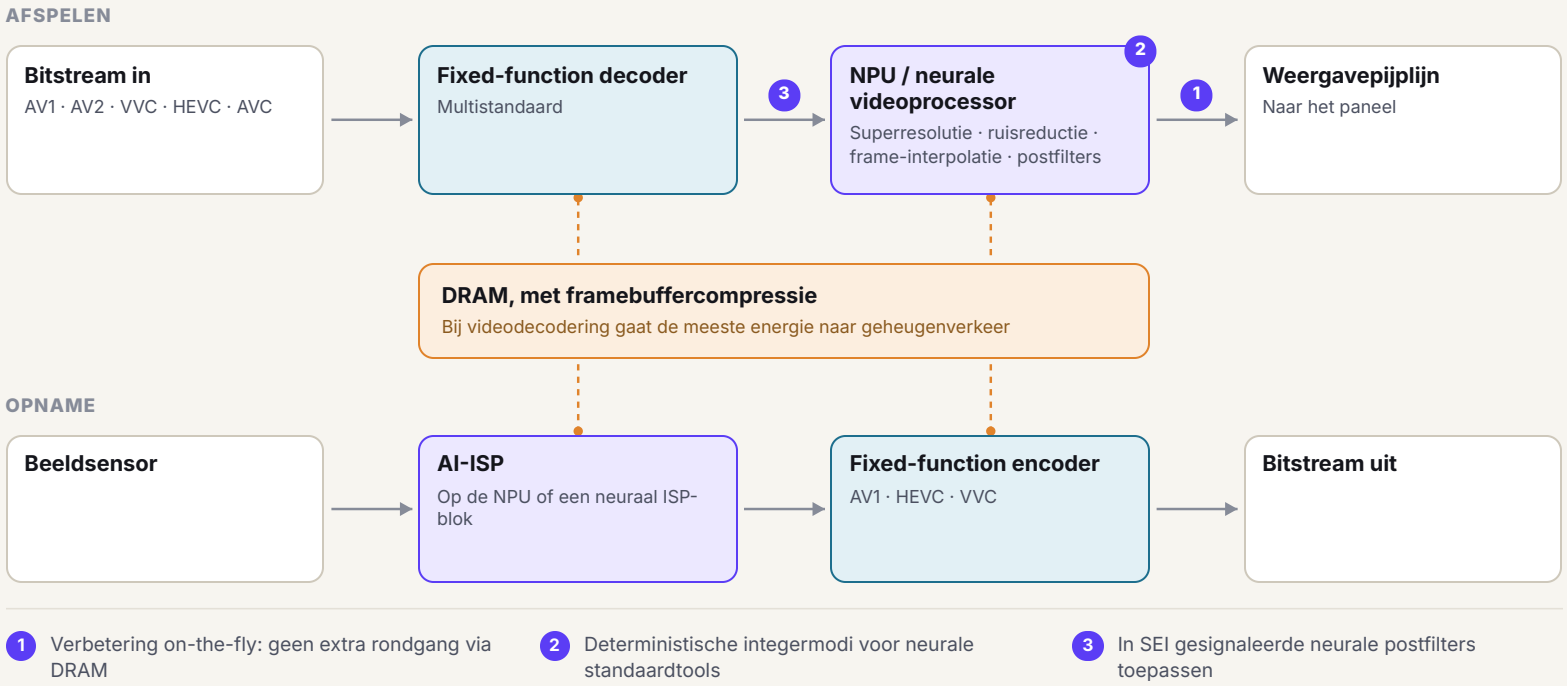
EXACTHEID

Elke decoder moet dezelfde bits berekenen

Entropiedecodering faalt als twee apparaten een kans verschillend berekenen. De oplossingen: integermodellen (DCVC-RT), een hyperprior met gedeelde schaal (MLVC), 16-bits integerinferentie (SADL van JVET), conformiteitspunten (JPEG AI).

Voor een chipteam is het opnieuw de discipline van de integertransformaties van HEVC: de standaard moet de rekenkunde exact vastleggen.

De taakverdeling in producten: een fixed-function codec plus een NPU



> Canghai van Tencent: meer dan 100.000 chips van de eerste generatie in gebruik; de tweede generatie voegt H.266 en superresolutie op de chip toe. Chips&Media en Allegro DVT verkopen codec-IP en neurale video-IP naast elkaar.

Wat ik zou bouwen, per rol

SoC- en tv-chipteams

- Fixed-function decoders voor elke standaard die je moet afspelen, nu AV2
- Een NPU gedimensioneerd voor verbetering op schermresolutie, zonder extra rondgang via DRAM
- SEI voor neurale postfilters en deterministische integermodi
- H.267 als ontwerpstart in 2029 of later

IP-leveranciers

- Nu AV2- en VVC-decoders
- Postfilterversnellers en framebuffercompressie
- Conformiteitstools voor neurale bewerking

Videoplatforms

- Neurale voorbewerking bij de encoder verdient zich het eerst terug
- Superresolutie aan clientzijde waar apparaten het toelaten
- End-to-end neurale codecs alleen in je eigen gesloten ketens

Waar neurale tools licentieerbaar worden

Standaard	Neurale inhoud	Status, november 2026
H.267 (werknaam)	Open: klassieke (ECM) en neurale tools zijn allebei kandidaat	Inzendingen beoordeeld in januari 2027; versie 1 eind 2029
VVC en HEVC, met VSEI	Neuraal postfilter beschreven in SEI; optioneel voor decoders	Van kracht; uitgebreid in oktober 2025
AV2	Klassiek, royaltyvrij	Versie 1.0.0 op 28 mei 2026; webstreams van YouTube gepland voor begin 2027
JPEG AI	End-to-end getrainde beeldcodec	Deel 1 gepubliceerd in 2025; geen product gevonden dat het standaard ondersteunt
IEEE 1857.11	Neurale beeldcodering ontworpen voor NPU's	Gepubliceerd december 2024
Video Coding for Machines	Codering voor analyse door machines in plaats van kijken	Internationale ontwerpstandaard begin 2026
ML-uitbreidingen van Opus	Optioneel en achterwaarts compatibel	In libopus 1.5 en 1.6

645

patentfamilies van ByteDance voor technologie voorbij VVC (GreyB, 2025)

$ 0,50 · $ 1,00–1,50

VVC-pooltarief per telefoon en per 4K-tv, Regio 1

$ 0

royalty's voor AV1 en AV2 van de leden van AOMedia

WIE ERAAN VERDIENT

Het geld volgt de standaarden en de platforms

VERDIENEN NU

Platforms

Houden de besparingen zelf

Neurale downscaler van Netflix; superresolutie bij YouTube en Meta; Microsoft maakte MLVC open source onder MIT

VERDIENEN NU

Apparaat- en chipmakers

Verkopen onderscheidend vermogen

Camera-ISP's; upscalingprocessors voor tv's; Tencents eigen Canghai-codecchips

VERDIENEN NU

IP-leveranciers

Licentiëren elke nieuwe standaard

Klanten voor AV2-decoders bij Chips&Media, Allegro DVT en VeriSilicon binnen drie weken na de specificatie

VERDIENEN NU

GPU-leveranciers

Bouwen aan platformbinding

DLSS 4 in 250+ games en apps; SDK voor neurale textures in bèta

PER APPARAAT

Patenthouders en pools

Innen per apparaat

VVC: 0,50 dollar per telefoon, 1,00–1,50 dollar per 4K-tv; H.267-tools na 2029

EXIT VIA OVERNAME

Startups in neurale codecs

Overgenomen

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

De drie bekendste startups in neurale codecs, WaveOne, iSIZE en Deep Render, werden gekocht door Apple, Sony en InterDigital.

Zes kansen, en het risico van elk

Kans	Wie koopt	Waarom nu	Grootste risico
Neurale voorbewerking bij de encoder	Streaming- en gebruikersvideoplatforms, omroepen	Geen apparaatwijziging; bewezen bij Netflix	Grote platforms bouwen het zelf
IP en afstemming voor verbetering aan schermzijde	Chipmakers voor tv's, settopboxen en telefoons	NPU's zitten al in de apparaten	Commoditisering; eigen teams
Neurale codecs in gesloten ketens	Leveranciers van bedrijfscommunicatie, industriële systemen en telepresence	MLVC bewijst dat het werkt, en de code is open	Je moet beide kanten beheersen; versnippering van NPU's
Codec plus NPU, als silicium of IP	SoC-makers	De AV2-cyclus begint	Sterke gevestigde spelers
Evaluatie, conformiteit en deterministische inferentie	Iedereen die neurale tools levert	Bitexactheid is een harde eis	Kleine markt, tenzij gebundeld
Bijdragen aan de standaard H.267	Licentiegevers, grote platforms	Inzendingen worden in januari 2027 beoordeeld	Lang, duur, onzeker

CLAIMS DIE JE ZULT HOREN

Neem een claim pas serieus als hij vier dingen vermeldt

Claim	Wat het bewijs zegt
“Neurale codecs verslaan VVC al.”	In het lab. DCVC-RT bespaart 21% tegenover de referentie-encoder van VVC; de decoder kost ongeveer 175 duizend MAC's per pixel.
“AI-codecs vervangen H.26x en AV1 binnen een paar jaar.”	Niet bij het afspelen: H.267 versie 1 eind 2029, chips rond 2031 schatting. De codec die nu komt, AV2, is klassiek.
“MLVC bespaart 88% bandbreedte.”	Tegenover H.264 bij 360p. Tegenover hardware-HEVC: 75,5% bij 360p en 65,4% bij 540p, op Microsofts eigen testset.
“Neurale textures besparen 85–96% videogeheugen.”	In demoscènes. De uitgebrachte game won ongeveer 30%, alleen op geselecteerde assets.
“JPEG AI is een ISO-standaard, dus camera's en browsers gaan het gebruiken.”	Geen product gevonden dat het standaard ondersteunt. Het web koos JPEG XL, een klassiek formaat.

Vraag naar de referentie, de maatstaf, de resolutie en de decodeerkosten. Een claim die niet alle vier kan noemen, is een onderzoeksresultaat, geen product.

Waar je op moet letten, en open vragen

WEGWIJZERS

Januari 2027	Inzendingen voor H.267 beoordeeld: hoeveel leunen op neurale tools, tegen welke kosten per pixel
Begin 2027	Eerste AV2-webstreams van YouTube; AV2-decoders in aangekondigde SoC's
Tweede helft van 2027	Eerste testmodel van H.267
2027	MLVC buiten peer-to-peer-gesprekken in Teams
2027–2028	Een eerste product dat JPEG AI standaard gebruikt
Eind 2029	H.267 versie 1 start de siliciumcyclus

OPEN VRAGEN

01 Exactheid op schaal: bitexacte integerdecodering op NPU's van verschillende leveranciers, tegen welk compressieverlies?

02 Geheugen vóór MAC's: het DRAM-verkeer per pixel van neurale nabewerking

03 De kosten per pixel waarboven een neurale tool zich in H.267 niet meer terugverdient

04 Welke gesloten ketens buiten gesprekken als eerste end-to-end neurale codecs gebruiken

05 Hoe je getrouwheid meet als een decoder details genereert

Dit zijn gebieden die we samen kunnen verkennen, via technische uitwisseling en technische samenwerking.

i@tanshuai.com · tanshuai.com