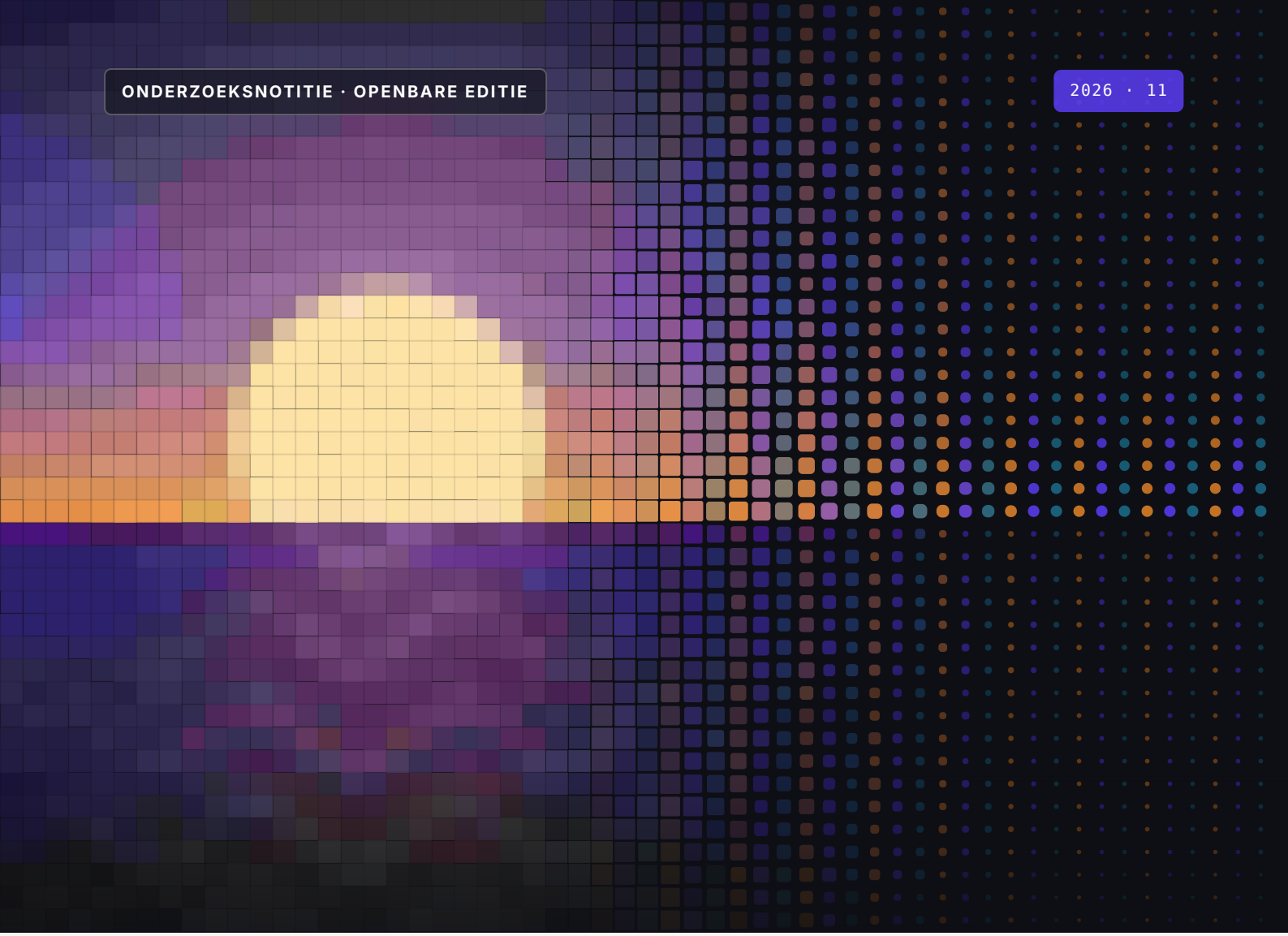


NEURALE COMPRESSIE · AFSPELEN × SILICIUM

Neurale codering in 2026

Ingezet rond de codec, niet in plaats ervan

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

HET OORDEEL

Neurale netwerken werken al aan beide kanten van de codecs die de video van de wereld vervoeren. End-to-end neurale codecs worden alleen in gesloten ketens ingezet, en gangbare schermen wachten op H.267.

Tan Shuai

Technoloog en oprichter · AI-onderzoeker en -ontwikkelaar

tanshuai.com · i@tanshuai.com

Gegevens per 12 november 2026
Bedragen in Amerikaanse dollars

INHOUD

Inhoud



◆ **Managementsamenvatting** 04
AI gemeengoed, maar rond de codec

01 **Afbakening en methode** 07
Wat “neurale codering” hier betekent, en hoe ik heb gekeken

02 **Afspelen en streaming** 09
Afspelen en streaming: neuraal aan beide kanten, standaard in het midden

03 **Fotografie en beeld** 12
Fotografie en beeld: neuraal in de camera, klassiek op het web

04 **Bellen en vergaderen** 13
Bellen en vergaderen: waar end-to-end neurale codecs echt worden ingezet

05 **Generatieve AI** 15
Generatieve AI: de grootste toepassing die niemand een codec noemt

06 **Games en graphics** 16
Games en graphics: neurale reconstructie is gemeengoed, neurale assets komen op gang

07 **Gespecialiseerde verbindingen** 18
Gespecialiseerde verbindingen: satellieten, drones en robots

08 **Silicium** 19
De blik vanuit het silicium: energie, rekenkracht, geheugen en exactheid

09 **Standaarden en patenten** 24
Standaarden en patenten: waar neurale tools licentieerbaar worden

10 **Wie eraan verdient** 25
Wie eraan verdient, en waar een bedrijf kan meedoen

11 **Claims en bewijs** 27
Claims die je zult horen, en wat het bewijs zegt

12 **Wegwijzers** 28
Waar je op moet letten, 2027 tot 2030

13 **Open vragen** 29

◆ **Over de auteur** 30
Tan Shuai

◆ **Bronnen** 31

FIGUREN

Figuren

FIGUUR 01 Waar neurale codering in 2026 wordt ingezet 06

FIGUUR 02 Drie manieren waarop een neuraal netwerk een codec binnenkomt 08

FIGUUR 03 Van voltooide standaard tot gangbare schermen: het grootste deel van een decennium 10

FIGUUR 04 In gesprekken verslaat een neurale codec hardware-HEVC met 65–76% 14

FIGUUR 05 Als één bedrijf de hele stack beheerst, verspreidt neurale reconstructie zich snel 17

FIGUUR 06 Energie per gedecodeerde pixel: fixed-function decoders liggen een ordegrootte voor 20

FIGUUR 07 Een neurale tool kost tienduizenden MAC's per pixel; een complete neurale decoder kost er honderdduizenden 21

FIGUUR 08 De taakverdeling in producten: fixed-function hardware voor de bitstream, een neurale engine voor de pixels 22

FIGUUR 09 Wie er in 2026 aan neurale codering verdient 25

Labels in de tekst

schatting — berekening of afleiding van de auteur op basis van geciteerde gegevens

leveranciersclaim — gepubliceerd door het bedrijf dat het product verkoopt, niet onafhankelijk geverifieerd

analyse — interpretatie van het bewijs door de auteur

Dit is een vertaling van de Engelse editie; waar de twee verschillen, is de Engelse editie leidend.

Gegevens per 12 november 2026 · Bedragen in Amerikaanse dollars

MANAGEMENTSAMENVATTING

AI gemeengoed, maar rond de codec

HET OORDEEL

Neurale codering is in 2026 al gemeengoed, maar werkt rond de standaardcodecs, niet in plaats ervan.

≈75%
van het mobiele dataverkeer is video, eind 2025 (Ericsson)

1,26 miljard
smartphones geleverd in 2025, die stuk voor stuk video decoderen in fixed-function hardware

≈12×
de energie per pixel van een gemeten HEVC-decoderchip met DRAM, voor het enige gepubliceerde ontwerp van een neurale decoderchip

≈22 TOPS
aan continue rekenkracht voor een neurale decoder van DCVC-RT-klasse bij 1080p30 schatting

75,5%
minder bitrate dan hardware-HEVC voor MLVC bij 360p in gesprekken, op Microsofts eigen testset leverancierclaim

Eind 2029
H.267 versie 1, de eerste standaard die waarschijnlijk neurale tools naar elk scherm brengt

Waar het wordt ingezet

Neurale netwerken draaien nu aan beide kanten van de codecs die vrijwel alle video ter wereld vervoeren. Aan de ingang geven ze het beeld vorm: AI-beeldsignaalprocessors in telefooncamera's en een neurale downscaler in de coderingspijplijn van Netflix. Aan de uitgang herstellen ze het: Super Resolution van YouTube op tv's, superresolutie bij Meta en tv-processors die worden aangeprezen met honderden neurale netwerken. Daartussen is de bitstream nog altijd AV1, HEVC, AVC of VVC, gedecodeerd door fixed-function hardware. Geen enkele streamingdienst stuurt een neuraal videoformaat naar tv's of telefoons; waar vandaag een neurale videodecoder op een apparaat draait, draait die binnen de eigen app van één bedrijf.

Gesloten ketens

End-to-end neurale codecs, waarbij zowel de encoder als de decoder getraind is, worden alleen ingezet in gesloten ketens, waar één bedrijf beide kanten beheerst en kan terugvallen op een standaardcodec. Microsoft laat sinds 2021 in alle Teams- en Skype-gesprekken tussen twee deelnemers een AI-audiocodec draaien, en maakte in juli 2026 MLVC open source, een neurale videocodec die nu wordt uitgerold in peer-to-peer-gesprekken in Teams: bij 360p heeft die voor dezelfde beoordeelde kwaliteit ongeveer 75% minder bitrate nodig dan hardware-HEVC. Elk model voor spraak- en videogeneratie draait op een getrainde codec, van spraaktokenizers op 12,5 Hz tot video-VAE's met een compressieverhouding van 1:192. Neurale texturecompressie is verschenen in één AAA-game. Dit zijn echte toepassingen, en ze hebben één voorwaarde gemeen: niemand anders hoeft de stream te decoderen.

Tijdpad

Afspelen van neurale bitstreams op massaschaal wacht op een standaard en een siliciumcyclus. Voor de eerste standaard die waarschijnlijk neurale tools zal bevatten, de opvolger van VVC die ik H.267 noem, ging in juli 2026 de oproep tot voorstellen uit; de inzendingen worden

in januari 2027 beoordeeld en een eerste versie staat gepland voor eind 2029. AV1 had bijna acht jaar nodig van voltooide specificatie tot 30% van de kijktijd bij Netflix. Volgens dat patroon komen de eerste H.267-chips rond 2031 en volgt breed gebruik halverwege de jaren 2030 schatting. Intussen is AV2 de codec die als volgende de schermen bereikt: in mei 2026 voltooid, zonder neurale decoder, en binnen drie weken kondigden drie IP-leveranciers klanten voor AV2-decoders aan.

Silicium

De reden zit in het silicium. Het enige gepubliceerde ontwerp van een neurale videodecoderchip heeft 14,6 nJ per pixel nodig, ongeveer 12 keer zoveel als een gemeten HEVC-decoderchip inclusief DRAM en ongeveer 47 keer zoveel als alleen de kern daarvan. DCVC-RT, de realtime onderzoekscodec van Microsoft, heeft volgens mijn telling van de eigen tabel per module ongeveer 175 duizend multiply-accumulate-bewerkingen per gedecodeerde pixel nodig, ofwel zo'n 22 TOPS aan continue rekenkracht voor 1080p30 schatting: ongeveer de helft van de 40 TOPS waarmee een AI-pc wordt gedefinieerd, besteed aan één videostream. Een neurale decoder moet bovendien op elk apparaat bitidentieke resultaten opleveren, anders valt de entropiedecodering uit elkaar. Daarom levert MLVC compressie in om op alle NPU's identiek te decoderen, en draait het op 540p, niet op 4K.

Zakelijk

Voor een bedrijf volgt het geld de standaarden en de platforms, niet neurale codecs als zelfstandig product. Platforms houden de besparingen van neurale voor- en nabewerking zelf. IP-leveranciers licentiëren fixed-function decoders voor elke nieuwe standaard, plus NPU's voor beeldverwerking. Patentpools innen per apparaat: voor VVC 0,50 dollar per telefoon en 1,00–1,50 dollar per 4K-tv in de VS en andere markten in Regio 1. De drie bekendste startups in neurale codecs, WaveOne, iSIZE en Deep Render, werden gekocht door Apple, Sony en InterDigital. De open kansen liggen in neurale voor- en nabewerking, in producten voor gesloten ketens, in silicium dat een fixed-function codec combineert met een NPU, en in het evaluatie- en standaardisatiewerk daartussen.

Vraag	Oordeel	Kerncijfers (bronnen in de hoofdstukken)
1 **Waar wordt neurale codering vandaag op grote schaal ingezet?**	Rond de codec: camera-ISP's, downscaling aan de encoderkant, superresolutie bij het afspelen. End-to-end alleen in gesloten ketens: gesprekken, game-assets, tokenizers voor generatieve AI	Video ≈75% van het mobiele dataverkeer; 1,26 miljard telefoons en ≈200 miljoen tv's geleverd in 2025, die allemaal in fixed-function hardware decoderen; DLSS 4 in 250+ games en apps
2 **Gaan neurale codecs AV1, HEVC en VVC vervangen bij het afspelen? Wanneer?**	Niet vóór een standaard en een apparaatcyclus. H.267 is de eerste realistische drager, en die zal waarschijnlijk klassieke en neurale tools combineren	Oproep tot voorstellen voor H.267 juli 2026, beoordeling januari 2027, versie 1 eind 2029; breed gebruik halverwege de jaren 2030 schatting
3 **Wat kost het in silicium?**	Vandaag een ordegrootte meer energie per pixel dan fixed-function decodering, plus geheugenverkeer en een eis van bitexactheid	14,6 nJ/pixel (neuraal) tegenover 1,19 (HEVC, met DRAM); ≈175 kMAC per gedecodeerde pixel ≈ 22 TOPS bij 1080p30 schatting
4 **Wat staan standaarden en patenten toe?**	Neurale postfilters kunnen al in VVC-streams worden gesignaleerd; JPEG AI is een ISO-standaard; H.267 beslist welke neurale tools licentieerbaar worden	ECM (klassiek) ≈28% beter dan VVC; neurale in-loopfilters 6–14% bij 10–22+ kMAC/pixel; ByteDance 645 patentfamilies in werk voorbij VVC
5 **Wie verdient eraan, en waar kan een bedrijf meedoen?**	Platforms, IP-leveranciers, patenthouders. Pure startups in neurale codecs zijn overgenomen. Kansen in voor- en nabewerking, gesloten ketens, silicium met codec plus NPU, evaluatie	VVC-pool 0,50 dollar per telefoon; AV2-decoder-IP binnen enkele weken na de specificatie in licentie gegeven; drie overnames, voorwaarden niet bekendgemaakt

Vraag	Oordeel	Kerncijfers (bronnen in de hoofdstukken)
⑥ **Welke claims moet je met een korrel zout nemen?**	Winst die wordt opgegeven tegenover oude codecs, bij lage resolutie, zonder decodeerkosten, of uit demo's in plaats van uitgebrachte producten	MLVC: 88% minder dan H.264 maar 65–76% minder dan hardware-HEVC; textures: 85% VRAM-besparing in een demo, ongeveer 30% in de uitgebrachte game

FIGUUR 01

Waar neurale codering in 2026 wordt ingezet

Neurale netwerken draaien al aan beide kanten van gangbare codecs. End-to-end neurale codecs worden alleen ingezet waar één bedrijf zowel de encoder als de decoder beheerst.

	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	—

Bron: Documenten van bedrijven, standaardisatieorganen en ontwikkelaars die in het rapport worden geciteerd; stand per 12 november 2026. De rijen zijn geordend naar het aantal mensen dat elk domein bereikt.

AFBAKENING EN METHODE

01 Wat “neurale codering” hier betekent, en hoe ik heb gekeken

Drie manieren waarop een neuraal netwerk een codec binnenkomt

“Neurale codering” omvat drie verschillende dingen, en de meeste discussies erover ontstaan doordat ze door elkaar worden gehaald.

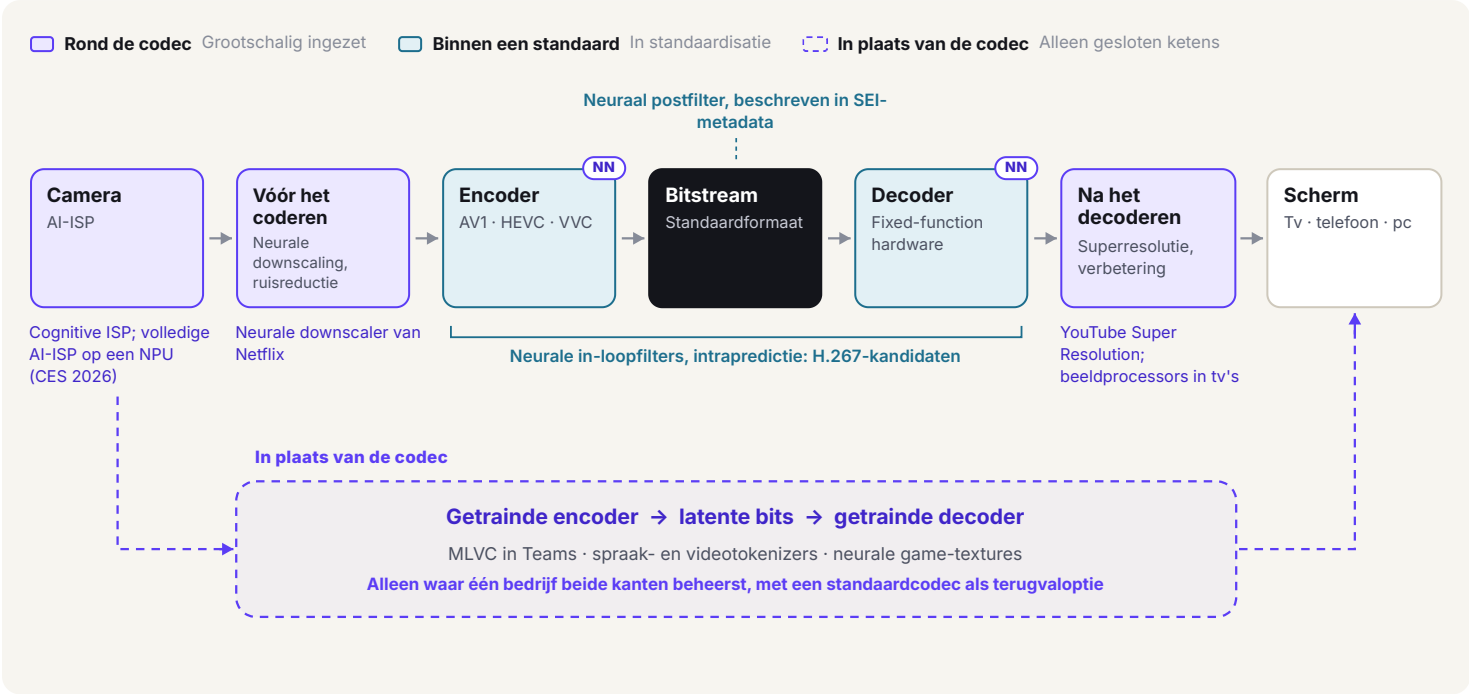
01 **Rond de codec.** Een netwerk bewerkt pixels vóór het coderen of na het decoderen, en de bitstream blijft standaard. Voorbeelden: een AI-beeldsignaalprocessor (ISP) in een camera, de neurale downscaler die Netflix vóór zijn encoder gebruikt, superresolutie op een tv. Elk apparaat kan de stream afspelen; het netwerk verandert alleen wat erin gaat of wat eruit komt.

02 **Binnen een standaard.** Een netwerk wordt een tool in een standaardcodec, bijvoorbeeld een neuraal in-loopfilter of een neurale intrapredictor die elke decoder moet uitvoeren, zodat het exact gespecificeerd moet worden, meestal in integerrekenkunde. Een lichtere vorm is een neuraal postfilter dat de bitstream in metadata beschrijft en dat decoders naar keuze kunnen toepassen.

03 **In plaats van de codec.** Zowel de encoder als de decoder is end-to-end getraind. De bitstream is een verzameling gekwantiseerde latente waarden die alleen het bijbehorende netwerk kan decoderen. Voorbeelden: MLVC van Microsoft in Teams, spraaktokenizers in spraakmodellen, neurale textures in een game.

FIGUUR 02 **Drie manieren waarop een neuraal netwerk een codec binnenkomt**

Het meeste wat in 2026 wordt ingezet, werkt rond een standaardcodec. Neurale tools binnen een standaard worden gespecificeerd. End-to-end neurale codecs vervangen de codec alleen waar één bedrijf beide kanten beheerst.



Bron: Synthese door de auteur van de bronnen in dit rapport; stand per 12 november 2026.

De eerste soort heeft niemands toestemming nodig. De tweede heeft een standaard nodig. De derde vereist controle over beide kanten. In die volgorde bereiken ze mensen, en dat verklaart het meeste van wat volgt.

Geordend naar de praktijk, niet naar de onderzoeksagenda

Onderzoekspapers bestuderen vooral de derde soort, omdat daar de compressierecords liggen. Dit rapport ordent de domeinen naar het aantal mensen dat ze bereiken, omdat dat bepaalt waar geld en engineering naartoe gaan: eerst afspelen en streaming, dan fotografie, bellen, generatieve AI, games, en als laatste de gespecialiseerde verbindingen (satelliet, drones, robots). Video is goed voor ongeveer 75% van al het mobiele dataverkeer, volgens de telling van Ericsson eind 2025. In 2025 leverde de industrie ongeveer 1,26 miljard smartphones (IDC) en zo'n 196–209 miljoen tv's (schattingen van TrendForce en Omdia). Elk daarvan decodeert video in fixed-function hardware.

Hoe ik heb gewerkt

Dit is bureauonderzoek, gedaan zoals ik een chip of een productplan zou beoordelen. Ik las standaardisatiedocumenten en vergaderverslagen (JVET, MPEG, JPEG, AOMedia, IETF), technische blogs en release notes van bedrijven, chippapers, tarievenlijsten van patentpools, teksten over exportcontrole en de pers. Bij elk cijfer staat de datum. Waar ik zelf een cijfer afleid, zoals rekenwerk per pixel uit de tabellen van een paper, zeg ik dat en markeer ik het met `schatting`. Claims die ik niet aan een primaire bron kon toetsen, zijn gemarkeerd met `leverancierclaim`. Bedragen zijn in Amerikaanse dollars. De stand van zaken is die van 12 november 2026.

Ik vergelijk waar dat kan gelijk met gelijk, en zeg het wanneer dat niet kan. Compressiewinst zegt weinig zonder vier dingen erbij: de referentie (winst ten opzichte van wat), de maatstaf (PSNR, VMAF of een panel van kijkers), de inhoud en de resolutie, en wat het decoderen kost. Het meeste meningsverschil over neurale codecs verdwijnt zodra die vier zijn vermeld.

HOOFDSTUK 02

AFSPELEN EN STREAMING

02 Afspelen en streaming: neuraal aan beide kanten, standaard in het midden

Wat kijkers ontvangen, is nog altijd een standaardbitstream

De grootste videodiensten hebben één codecovergang afgerond en beginnen aan de volgende, en bij geen van beide komt een neurale decoder kijken. In december 2025 meldde Netflix dat AV1 ongeveer 30% van zijn kijktijd voor zijn rekening nam, als op één na meest gebruikte codec, na AVC. AV1 bereikte tv's pas toen hardwaredecoders dat deden: Netflix nam AV1 in 2019 op in zijn apparaatcertificering, streamde het vanaf eind 2021 naar tv's, en zegt dat 88% van de apparaten met een groot scherm die tussen 2021 en 2025 ter certificering werden aangeboden het ondersteunden, sinds 2023 bijna allemaal. YouTube zei dat per 19 maart 2026 meer dan 90% van zijn catalogus, gewogen naar kijktijd, in AV1 beschikbaar was.

AV2, de opvolger van AV1 van de Alliance for Open Media, bereikte op 28 mei 2026 versie 1.0.0, met ongeveer 30% betere compressie dan AV1 volgens de objectieve metingen van AOMedia. YouTube plant AV2-streams op het web “al begin 2027”. Het silicium volgde meteen: Chips&Media voltooide in mei een AV2-hardwaredecoder en gaf die in juni in licentie aan een niet bij naam genoemd groot Noord-Amerikaans techbedrijf; Allegro DVT had zijn AV2-decoder-IP op 16 juni al aan drie klanten geleverd; VeriSilicon had op 10 juni al een video-IP met AV2-ondersteuning vrijgegeven aan meerdere klanten. AV2 is een klassieke codec: de winst komt van betere partitionering, predictie, transformaties en in-loopfilters (waarvan één een “datagedreven” guided detail filter is), niet van een end-to-end netwerk.

VVC, de huidige ITU/ISO-standaard, vond zijn eerste thuis in de omroep in Brazilië, waar het DTV+-systeem (TV 3.0) in juni 2026 commercieel in gebruik ging, op tijd voor het WK. In juli 2026 was in Brazilië nog geen tv te koop met een ingebouwde DTV+-ontvanger; de eerste worden tussen eind 2026 en begin 2027 verwacht.

Bij de encoder: de neurale downscaler van Netflix

Het zuiverste voorbeeld van neurale codering bij het afspelen draait al sinds 2022 en verandert niets aan wat het apparaat van een kijker moet doen. Netflix codeert elke titel in meerdere resoluties, en de downscalingstap die uit een 4K-master 1080p, 720p en 540p maakt, gebruikte traditioneel een vast filter zoals Lanczos. Netflix verving dat door een klein neuraal netwerk, een voorbewerkingsblok gevolgd door een schalingsblok, getraind om de kwaliteit te verbeteren die je ziet na een standaard-upscaler op het apparaat. Netflix noemt het een drop-in-oplossing: geen enkele andere wijziging aan de coderingskant of aan de clientkant, zodat “miljoenen apparaten die Netflix-streaming ondersteunen er automatisch van profiteren”. In de voorkeurstests van Netflix koos ongeveer 77% van de kijkers de uitvoer van de neurale downscaler leveranciersclaim. Om de kosten op de schaal van Netflix aanvaardbaar te houden, “waren slechts een paar lagen van een neuraal netwerk voldoende”.

Dat is het patroon om te onthouden. Het netwerk draait één keer per titel in het datacenter, waar rekenkracht goedkoop is en kan worden ingepland, en de uitvoer is een gewone AV1- of HEVC-stream.

Bij het scherm: superresolutie in de app en in de tv

Aan de andere kant herstellen netwerken wat de codec en de bron hebben verloren. Op 29 oktober 2025 kondigde YouTube Super Resolution aan voor zijn tv-app: video's die onder 1080p zijn geüpload, van 240p tot 720p, worden met machine learning naar HD opgeschaald, met 4K als uitgesproken doel. De opgeschaalde versies zijn gelabeld, kijkers kunnen terugschakelen naar het origineel en makers kunnen zich ervoor afmelden. Meta beschreef in oktober 2025 hoe het videosuperresolutie toepast, zowel op zijn servers bij het uploaden, op standaard x86-CPU's in samenwerking met Intel, als op de client bij het afspelen, en ontdekte dat de twee elkaar aanvullen. Tv-makers doen hetzelfde in silicium: Samsung prijst zijn vlaggenschipprocessor voor 8K, de NQ8 AI Gen3, aan met 768 neurale netwerken voor upscaling.

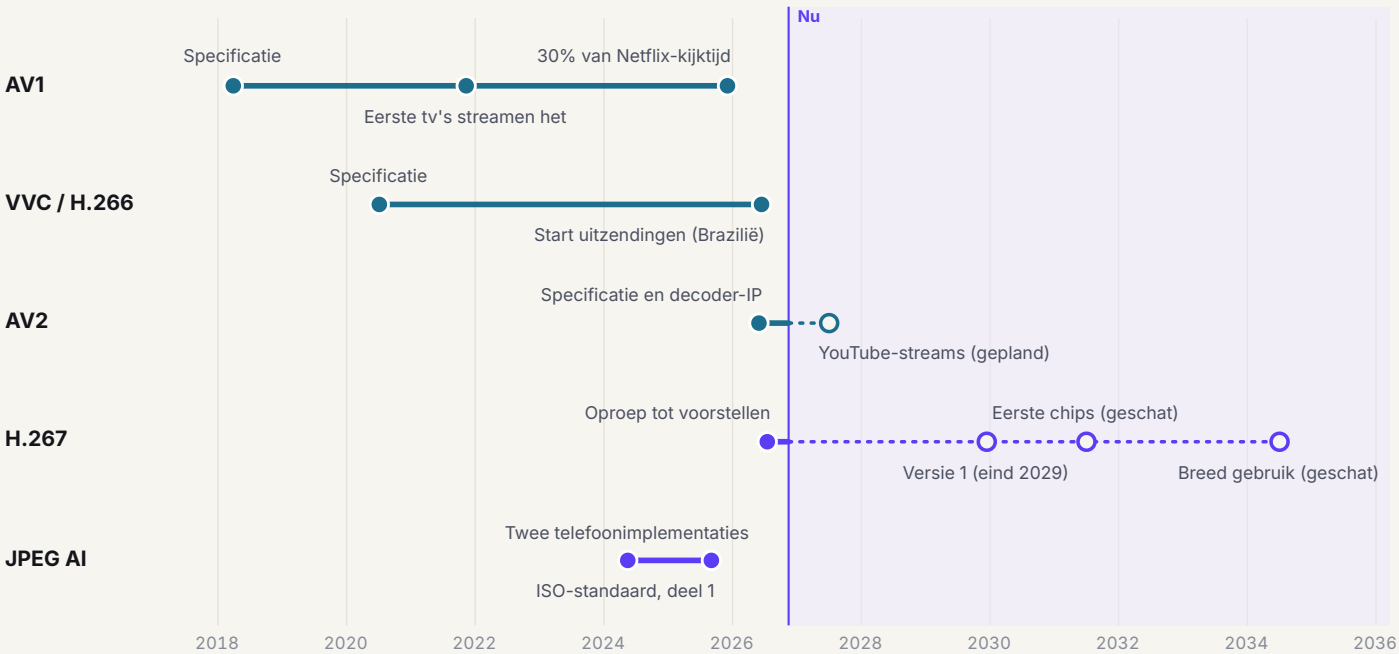
De rekensom verschilt per plek waar het netwerk draait. Bewerking aan serverzijde kost het platform rekenkracht voor elke video die het aanraakt. Bewerking op het apparaat kost het platform niets; die gebruikt een NPU waarvoor de koper al heeft betaald. Daarom komt de grootste winst in 2026 van beide.

Waarom geen enkele tv een neurale bitstream decodeert

Een afspeelformaat moet decoderen op elk apparaat dat het ooit zal afspelen, en die apparaten gaan jaren mee. Daardoor is de weg van een voltooide standaard naar gangbare schermen zelfs voor klassieke codecs traag. De specificatie van AV1 was in maart 2018 klaar; de softwaredecoder dav1d volgde binnen enkele maanden; tv's streamden het vanaf eind 2021; in december 2025 bereikte het ongeveer 30% van de kijktijd bij Netflix, bijna acht jaar na de specificatie. VVC was in juli 2020 klaar en bereikte in juni 2026 zijn eerste omroepdienst.

FIGUUR 03 **Van voltooide standaard tot gangbare schermen: het grootste deel van een decennium**

Afspelen vereist een hardwaredecoder in het apparaat. AV1 had bijna acht jaar nodig van voltooide specificatie tot 30% van de kijktijd bij Netflix; de eerste standaard die waarschijnlijk neurale tools bevat, wordt eind 2029 verwacht.



Bron: Netflix Tech Blog (dec. 2025); AOMedia; Allegro DVT; TV Technology; Fraunhofer IIS over DTV+ in Brazilië; MPEG-155; JPEG-comité; ISO. Open markeringen zijn doelen of schattingen van de auteur.

Een neurale bitstream voegt twee problemen toe die klassieke codecs lang geleden hebben opgelost: energie per pixel en bitexacte decodering. Beide komen aan bod in hoofdstuk 8. Samen betekenen ze dat geen tv- of tele-

foonmaker een neurale videodecoder in elk apparaat stopt voordat een standaard precies voorschrijft wat er gebouwd moet worden.

Wat het zou veranderen: H.267

De standaard die neurale tools naar elk scherm zou kunnen brengen, is de opvolger van VVC, die ITU-T en ISO/IEC samen ontwikkelen in het Joint Video Experts Team (JVET). De naam ligt nog niet vast; ik noem hem H.267, zoals een groot deel van de industrie doet.

In oktober 2025 beoordeelde JVET de inzendingen op een oproep om bewijs: vijf inzendingen van 16 bedrijven en onderzoekslabs, met 20 geteste codecconfiguraties. Het comité vond bewijs voor betere compressie dan VVC en noemde zowel verbeterde conventionele technologie als technologie op basis van neurale netwerken “interessant”. Neurale benaderingen lieten duidelijke perceptuele winst zien, vooral bij 8K-HDR-materiaal, en kleinere winst bij gamecontent. InterDigital, een van de inzenders, diende één ontwerp in op basis van het conventionele Enhanced Compression Model (ECM) en één dat AI-tools bovenop VVC legde.

De formele oproep tot voorstellen ging uit tijdens de vergadering van juli 2026 in Genève. De belangrijkste inzendingen moesten eind oktober 2026 binnen zijn, een aanvullend pakket volgt half december, en de inzendingen worden beoordeeld op de JVET-vergadering in Brisbane in januari 2027. JVET verwacht een eerste testmodel in de tweede helft van 2027 en een eerste versie van de standaard tegen eind 2029, en noemt het geheel “waarschijnlijk een grotere inspanning dan ooit is geleverd om een nieuwe standaard op gang te brengen”.

Het klassieke spoor legt de lat hoog. ECM, gebouwd met conventionele signaalverwerking, bespaart in random-accessconfiguratie zo'n 28% ten opzichte van VVC. De verkenning van JVET naar neurale netwerken rapporteert in-loopfilters die ongeveer 6% besparen bij 10 duizend multiply-accumulate-bewerkingen per pixel, 8% bij 22 duizend en 14% in een werkpunt met hoge complexiteit. Mijn interpretatie is dat H.267 hoogstwaarschijnlijk een klassieke kern wordt met enkele neurale tools waar de bespaarde bits de decoderkosten rechtvaardigen, en dat de beoordeling in januari 2027 zal laten zien hoeveel analyse.

Er is al een brug. VVC- en HEVC-streams kunnen supplemental enhancement information (SEI) bevatten die een neuraal postfilter beschrijft dat de decoderkant kan toepassen, en de nieuwe edities van VVC en van zijn SEI-standaard (VSEI), voltooid in oktober 2025, hebben die berichten uitgebreid. Een stream kan aangeven welk netwerk na het decoderen moet draaien zonder elke decoder te dwingen het uit te voeren.

Wat dit betekent voor bedrijven in de afspeelmarkt

Verkoop nu wat rond de codec werkt: voorbewerking aan de encoderkant voor diensten, en nabewerkingsblokken en afstemming voor chips in tv's en settopboxen. Bouw geen afspeelbedrijf op een neurale bitstream die vóór de jaren 2030 op apparaten van derden moet decoderen. Als je silicium maakt, zit de omzet op korte termijn in AV2- en VVC-decoders; het neurale deel is een NPU-pad voor verbetering daarnaast (hoofdstuk 8).

HOOFDSTUK 03

FOTOGRAFIE EN BEELD

03 Fotografie en beeld: neuraal in de camera, klassiek op het web

De camerapijplijn is al neuraal

Elke foto van een recente vlaggenschiptelefoon is door neurale netwerken gegaan voordat hij wordt gecomprimeerd. De Snapdragon 8 Gen 2 van Qualcomm, in telefoons uit 2023, introduceerde een “Cognitive ISP” die tijdens foto- en video-opnamen realtime semantische segmentatie uitvoert en gezichten, haar, luchten en achtergronden afzonderlijk bijstelt; de Galaxy S23 Ultra van Samsung was de eerste telefoon die hem voor de selfiecamera gebruikte. In januari 2026 toonden Chips&Media en Visionary.ai op CES wat zij de eerste volledig op AI gebaseerde ISP noemen: een end-to-end beeldpijplijn die volledig op een NPU draait, WAVE-N van Chips&Media, in 16-bits floating point, zonder fixed-function ISP, en die draadloos kan worden bijgewerkt. “Bestaande pijplijnen schroeven neurale blokken vast aan een fixed-function ISP. Wij vervangen de conventionele ISP volledig”, zei de CEO van Visionary.ai. Na de beurs werd het fabrikanten ter evaluatie aangeboden.

Daarna wordt het resultaat opgeslagen als JPEG of HEIF, klassieke formaten die elk apparaat kan openen. Het patroon is hetzelfde als bij afspelen: neuraal binnen de pijplijn van één bedrijf, standaard aan de grens.

JPEG AI: een voltooide neurale standaard, nog zonder product

JPEG AI is de eerste standaard van ISO/IEC en ITU-T voor beeldcodering met getrainde netwerken (ISO/IEC 6048, ITU-T T.840). Deel 1, het kerncoderingssysteem, verscheen in 2025. Het JPEG-comité rapporteert compressiewinst van 12,5% tot 27,9% ten opzichte van VVC-intracodering, over een bereik aan decodercomplexiteit van 7 tot 216 duizend multiply-accumulate-bewerkingen per pixel. Het ontwerp heeft twee encoders en drie decoders, samen getraind, zodat een ontvanger maar één decoder hoeft te implementeren en toch elke stream kan lezen. In mei 2024 meldde het comité twee implementaties op smartphones. Daarnaast publiceerde de IEEE in december 2024 IEEE 1857.11-2024, een standaard voor beeldcodering op basis van neurale netwerken die is ontworpen voor NPU's.

Per november 2026 vond ik geen telefooncamera, foto-app of browser die JPEG AI standaard schrijft of leest.

Wat het web in plaats daarvan koos: JPEG XL

Het beeldformaat dat in 2026 wél terrein won, is klassiek. Op 6 oktober 2026 schakelde Google in Chrome 155 de decodering van JPEG XL standaard in, met een nieuwe geheugenveilige decoder die in Rust is geschreven, vier jaar nadat het de experimentele ondersteuning had verwijderd. Safari decodeert JPEG XL sinds 2023, en Mozilla gaat al een tijd dezelfde kant op. Google voert als argumenten aan: betere compressie dan JPEG, verliesloze hercompressie van bestaande JPEG-bestanden, HDR en progressieve decodering.

De les geldt breder. Een formaat dat overal moet openen, wint met een snelle, deterministische decoder die op elke CPU draait en geheugenveilig kan worden gemaakt. Een neurale beelddecoder heeft een NPU of GPU nodig, en exacte rekenkunde op alle apparaten. Neurale codering wint binnen de eigen pijplijn van een apparaat en in systemen waarin hetzelfde bedrijf codeert en decodeert, zoals een fotodienst in de cloud die het gecomprimeerde bestand nooit aan iemand anders overdraagt.

Wat dit betekent voor bedrijven in de beeldverwerking

Het geld zit in de opnamepijplijn: IP voor AI-ISP's, NPU's die op beeldverwerking zijn gedimensioneerd, en het afstemmingswerk dat cameramakers inkopen. Voor opslag en distributie is een getrainde beeldcodec vandaag een hulpmiddel voor gesloten systemen, geen formaat om op te wedden.

HOOFDSTUK 04



BELLEN EN VERGADEREN

04 Bellen en vergaderen: waar end-to-end neurale codecs echt worden ingezet

Audio ging voorop

Microsoft introduceerde Satin, een door AI aangedreven spraakcodec, in februari 2021 en zei dat die al werd gebruikt voor alle gesprekken tussen twee deelnemers in Teams en Skype. Satin levert superwideband-spraak, met een audiobandbreedte tot 16 kHz, bij 6 kbps, waar de eerdere Silk-codec van Skype bij die bitrate smalband was. De bespaarde bitrate gaat naar redundantie, zodat gesprekken beter bestand zijn tegen pakketverlies.

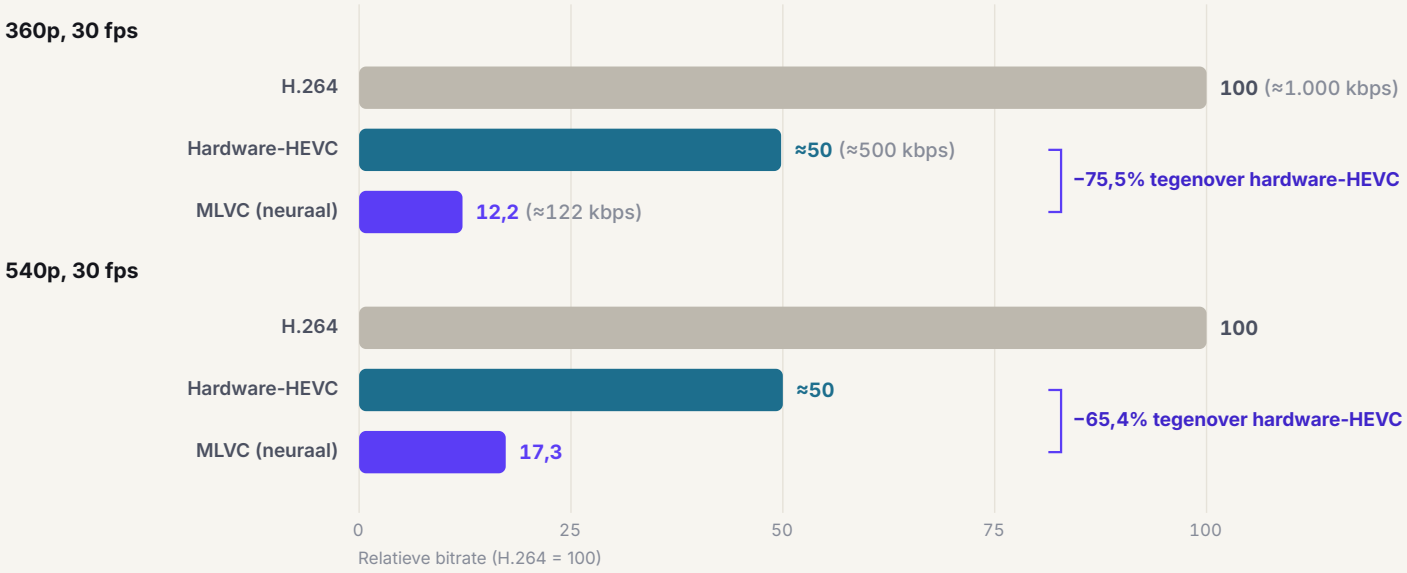
MLVC: een neurale videocodec in echte gesprekken

Op 24 juli 2026 bracht Microsoft MLVC, zijn videocodec op basis van machine learning, uit als open source onder de MIT-licentie, met de modelcode, getrainde gewichten, trainingsscripts en de conversietools voor verschillende NPU's. MLVC is de productversie van de DCVC-familie die Microsoft Research sinds 2021 publiceert. Microsoft zegt MLVC in Teams uit te rollen voor peer-to-peer-gesprekken, met live telemetrie, A/B-tests en de conventionele codecs als terugvaloptie.

MLVC draait op 540p en 360p met 30 frames per seconde en is zo ontworpen dat het niet meer dan de helft van de NPU gebruikt op de Neural Engine van Apple en op NPU's van Qualcomm en Intel, waarbij het werk wordt verdeeld tussen CPU en NPU. In subjectieve tests volgens ITU-T P.910 op Microsofts eigen Video Conferencing Dataset had het bij 360p 87,8% minder bitrate nodig dan H.264 en 75,5% minder dan hardware-HEVC, en bij 540p 82,7% en 65,4% minder leverancierclaim. Bij 360p is dat zo'n 122 kbps, waar H.264 ongeveer 1 Mbps nodig heeft.

FIGUUR 04 **In gesprekken verslaat een neurale codec hardware-HEVC met 65–76%**

Benodigde bitrate voor dezelfde beoordeelde kwaliteit op Microsofts testset voor videovergaderen, H.264 = 100. MLVC draait op NPU's van laptops en telefoons tot 540p30 en gebruikt daarbij hooguit de helft van de NPU.



Bron: Microsoft, release van MLVC (24 juli 2026): subjectieve tests volgens ITU-T P.910 op de eigen Video Conferencing Dataset. HEVC-balken afgeleid uit de gerapporteerde besparingen. Claims van de leverancier, geen onafhankelijke metingen.

Twee dingen relativeren het kopcijfer. De 88% geldt tegenover H.264; tegenover de hardware-HEVC die dezelfde laptops al bevatten, is het 65–76%, nog altijd veel. En MLVC leverde compressie in om inzetbaar te zijn: een onafhankelijke analyse merkt op dat een perceptueel getrainde DCVC-RT op dezelfde maat 81,9% scoort, waar MLVC op 75,5% uitkomt. De prijs van identiek decoderen op hardware die niet van Microsoft is, was ongeveer zes procentpunten.

Waarom gesprekken de natuurlijke plek zijn

Een gesprek is een gesloten keten. Beide eindpunten draaien dezelfde applicatie, stemmen aan het begin van het gesprek af wat ze kunnen, en kunnen halverwege terugvallen op een standaardcodec. De resoluties zijn bescheiden, de netwerken zijn vaak slecht, en de laptops aan beide kanten worden nu met NPU's geleverd, zoals telefoons al jaren. Microsoft legt de lat voor zijn AI-pc's op 40 TOPS aan NPU-rekenkracht, en de Neural Engine van Apples M4 is goed voor 38 TOPS. Elke voorwaarde die een neurale codec bij het afspelen blokkeert, ontbreekt hier.

De tegenvoorbeelden: Opus en MLow

Niet elk belplatform maakte dezelfde keuze, en de redenen zijn leerzaam.

Opus, de open audiocodec in WebRTC en de meeste belapps, voegde in versie 1.5 (maart 2024) deep learning toe: voor het eerst gebruikte het netwerken om het signaal zelf te bewerken of te genereren. Maar het voegde ze toe als opties: verbetering aan decoderzijde (LACE en NoLACE) waarvoor de bitstream niet hoeft te veranderen, een neurale maskering van pakketverlies en een schema voor diepe redundantie (DRED); ze staan allemaal in de basisinstelling uit, zodat oude apparaten niet worden belast. Opus 1.6 voegde een neurale bandbreedte-uitbreiding toe en een DRED-model dat drie keer kleiner is, en blijft volledig compatibel met de standaard uit 2012. De aanpak is “binnen een standaard, optioneel en achterwaarts compatibel”.

Meta ging nog een stap terug. In juni 2024 lanceerde het MLow, een spraakcodec voor lage bitrates die is gebouwd op klassieke signaalverwerking in plaats van op een netwerk, nu in alle gesprekken op Instagram en Messenger en in uitrol op WhatsApp. Meta's redenen waren zijn apparaten: meer dan 20% van zijn gesprekken wordt gevoerd op ARMv7-telefoons, en dagelijks lopen tientallen miljoenen WhatsApp-gesprekken via telefoons van meer dan tien jaar oud. Neurale codecs, waaronder Meta's eigen EnCodec, draaien alleen betrouwbaar op dure toestellen, schreef Meta. MLow levert bij 6 kbps ongeveer twee keer de kwaliteit van Opus en gebruikt daarbij 10% minder rekenkracht.

Voor een dienst met gebruikers van vlaggenschiplaptops tot tien jaar oude telefoons beslist het complexiteitsbudget. Neurale codecs worden het eerst ingezet waar de geïnstalleerde basis nieuw is en de leverancier beide kanten beheerst.

Wat dit betekent voor bedrijven in de communicatie

Zakelijk videovergaderen, telepresence, hulp op afstand, contributieverbindingen en privénetwerken zijn de plekken waar een end-to-end neurale codec vandaag een product is, mits beide kanten van jou zijn. Houd vanaf het begin rekening met drie dingen: een terugvalcodec, decodeertests op elke NPU die je ondersteunt, en een kwaliteitsevaluatie op je eigen content in plaats van op de dataset van een leverancier.

HOOFDSTUK 05

GENERATIEVE AI

05 Generatieve AI: de grootste toepassing die niemand een codec noemt

Elk spraak- en videomodel draait op een getrainde codec

De grootste toepassing van end-to-end neurale compressie zit in geen enkele codecstandaard. Ze zit in generatieve AI. Een spraakmodel luistert niet rechtstreeks naar golfvormen en produceert ze ook niet rechtstreeks: het leest en schrijft de discrete tokens van een getrainde audiocodec. Mimi van Kyutai, uitgebracht in 2024, zet spraak om in tokens met 12,5 frames per seconde en 1,1 kbps, lager dan welke telefoniecodec ook, en combineert semantische en akoestische informatie in één stream. EnCodec van Meta, gepubliceerd in 2022, is een ander voorbeeld; Meta's eigen engineers beschreven zulke codecs later als uitstekend bij lage bitrates, maar te zwaar voor oudere telefoons (hoofdstuk 4).

Videogeneratie werkt op dezelfde manier. Een diffusie- of transformermodel genereert in de latente ruimte van een getrainde video-autoencoder (VAE), en de decoder van de VAE zet het resultaat om in pixels. De VAE van LTX-Video comprimeert video 1:192: elk latent token staat voor een blok van 32×32 pixels over 8 frames, met 128 kanalen. De decoder voert ook de laatste ruisreductiestap uit, zodat hij deels details genereert in plaats van ze alleen te reconstrueren.

Waarom dit geen distributieformaten zijn

Deze codecs draaien bij elk verzoek aan een spraak- of videomodel, maar geen ervan is een formaat dat iemand anders decodeert. Elk model heeft zijn eigen latente ruimte; er is geen bitratesturing, geen conformiteitstest en geen garantie dat een andere decoder hetzelfde beeld zou produceren. Ze zijn ontworpen met het oog op generatiekwaliteit, niet op getrouwheid bij een gegeven bitrate, en ze decoderen op GPU's in datacenters. Hun succes

laat zien dat neurale compressie op grote schaal werkt als één bedrijf de encoder, de decoder en de hardware bezit.

Wat dit betekent voor bedrijven

Twee praktische punten. Ten eerste is de efficiëntie van de tokenizer een directe kostenpost voor iedereen die generatieve video of spraak aanbiedt: voor een transformer met attention over alle tokens groeit de rekenlast sneller dan het aantal tokens, zodat een sterker comprimerende latente representatie de generatiekosten meer dan evenredig verlaagt analyse. Codec-engineers die verstand hebben van het ontwerp van latente representaties, entropiemodellen en decoderefficiëntie, zijn waardevol voor teams in generatieve AI. Ten tweede schuift de grens op: standaardisatieorganen zijn begonnen aan codering voor machines in plaats van mensen. Video Coding for Machines van MPEG (ISO/IEC 23888-2) bereikte begin 2026 de status van internationale ontwerpstandaard en moet eind 2026 definitief worden, en JPEG AI is vanaf het begin ontworpen om naast menselijke kijkers ook computervisietaken te bedienen.

HOOFDSTUK 06

GAMES EN GRAPHICS

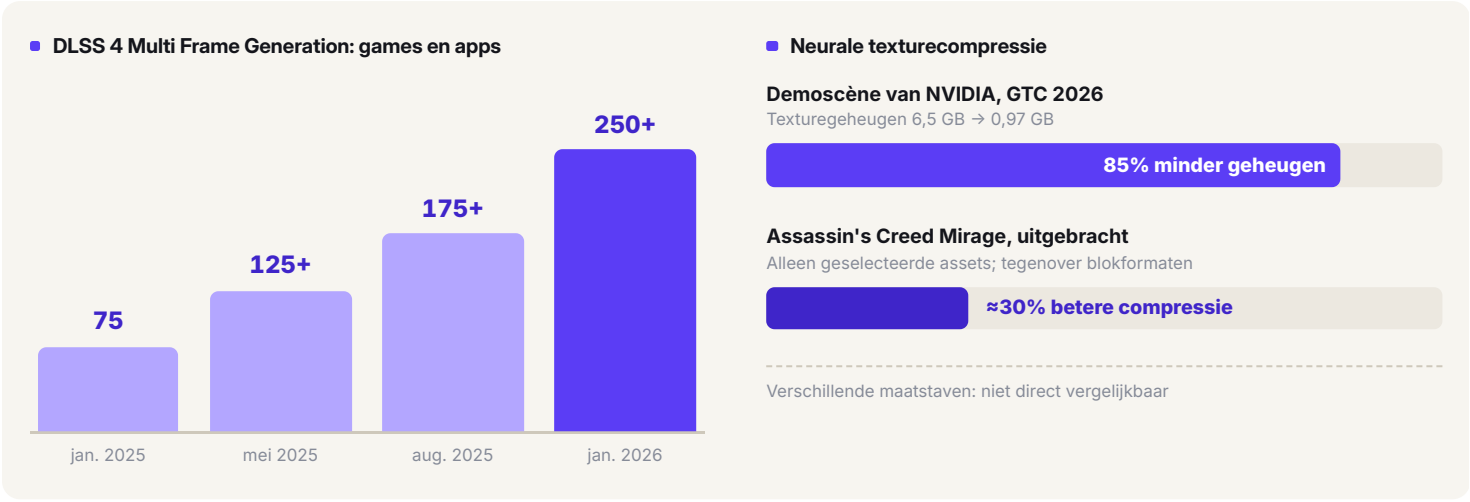
06 Games en graphics: neurale reconstructie is gemeengoed, neurale assets komen op gang

Minder pixels renderen en een netwerk de rest laten invullen

In realtime graphics reconstrueert een netwerk al het grootste deel van wat spelers zien. DLSS van NVIDIA rendert minder pixels en minder frames en gebruikt netwerken om de rest te produceren. DLSS 4 met Multi Frame Generation verscheen in januari 2025 met ondersteuning in 75 games en apps; NVIDIA telde er in mei 2025 meer dan 125, in augustus meer dan 175 en in januari 2026 meer dan 250, toen het DLSS 4.5 aankondigde met een transformermodel van de tweede generatie voor superresolutie. Dat is in alles behalve de naam een decoder: de “bitstream” is een frame met lage resolutie plus bewegingsdata, en het netwerk reconstrueert het volledige beeld.

FIGUUR 05 **Als één bedrijf de hele stack beheerst, verspreidt neurale reconstructie zich snel**

Links: games en apps met ondersteuning voor DLSS 4 Multi Frame Generation van NVIDIA. Rechts: neurale texturecompressie in een demoscène en in een uitgebrachte game, die de techniek vanwege de kosten tijdens runtime alleen voor geselecteerde assets gebruikte.



Bron: NVIDIA (januari 2025 tot januari 2026; demo op GTC 2026); Ubisoft La Forge (februari 2026). Claims van de leveranciers; de twee texturecijfers meten verschillende dingen.

Neurale textures zijn verschenen, selectief

Textures worden gecomprimeerd in vaste blokformaten (BCn) die GPU's in hardware decoderen. Neurale texturecompressie vervangt een set materiaaltextures door een compacte latente representatie en een klein netwerk dat texels reconstrueert wanneer ze nodig zijn. In februari 2026 beschreef het La Forge-lab van Ubisoft hoe het de techniek had meegeleverd in Assassin's Creed Mirage, een game uit 2023, met ongeveer 30% hogere effectieve compressie dan de conventionele formaten bij dezelfde visuele kwaliteit leveranciersclaim. De techniek werd niet overal gebruikt: “het gebruik werd beperkt door de runtimeprestaties”, dus paste Ubisoft die toe op een deel van de assets met veel instanties en een zware geheugendruk.

De RTX Neural Texture Compression SDK van NVIDIA, in bèta op GitHub, laat het plafond zien. In een demonstratie op GTC 2026 had een villascène die met BCn 6,5 GB texturegeheugen gebruikte, met neurale textures 970 MB nodig, 85% minder leveranciersclaim. De SDK biedt twee routes: decoderen bij het laden naar BCn, wat schijfruimte en downloadgrootte bespaart maar geen videogeheugen, of decoderen bij het samplen, wat het geheugen bespaart maar elk frame shadertijd kost.

De API komt eraan

Grafische API's krijgen de matrixbewerkingen die neurale decodering binnen shaders nodig heeft. Microsoft bracht in februari 2026 Shader Model 6.9 uit; de cooperative-vectorfunctie die het in 2025 als preview had getoond, werd opgenomen in een uniform ontwerp voor lineaire algebra voor een later shadermodel, en DirectX Linear Algebra ging in april 2026 in openbare preview. Als dat in elke driver zit, kan een neurale texturedecoder via één codepad op de GPU van elke leverancier draaien.

Wat dit betekent voor bedrijven

Games zetten neurale decodering in omdat de voorwaarden van een gesloten keten gelden: één engine, één assetpijplijn, tensorhardware in de GPU en een terugvalformaat. Het open werk zit in tools, in assetpijplijnen die neurale materialen coderen, in runtimes voor GPU's van verschillende leveranciers en vooral in mobiel, waar de energiekosten per gedecodeerde texel de doorslag geven.

HOOFDSTUK 07

GESPECIALISEERDE VERBINDINGEN

07 Gespecialiseerde verbindingen: satellieten, drones en robots

Dit zijn de gevallen waarin end-to-end neurale codering op papier het best past. Eén eigenaar beheerst beide kanten van de verbinding, bandbreedte is schaars en verliesgevoelig, de inhoud is specifiek genoeg om op te trainen, en een perceptueel of zelfs semantisch getrouw beeld is vaak genoeg. Verliesbestendige neurale codecs voor zulke verbindingen zijn een actief onderzoeksgebied, en herstel aan serverzijde van feeds met lage bitrate wordt toegepast.

Ze vormen ook een kleine en steeds sterker gereguleerde markt vergeleken met consumentenvideo analyse. In China herziet Bekendmaking nr. 49 van 2026 van het ministerie van Handel en twee andere instanties, uitgevaardigd op 19 oktober en van kracht vanaf 18 november 2026, de exportcontrole op 13 categorieën dronegerelateerde artikelen. Onder de technologieposten is er één voor onbemande grondplatforms, die deels worden gedefinieerd aan de hand van drempels zoals 75% semantische beeldcompressie en beeldreconstructie binnen 150 milliseconden, zoals de vakpers meldt. Semantische compressie zelf is een parameter voor exportcontrole geworden. In de Verenigde Staten zette de Federal Communications Commission op 22 december 2025 in het buitenland gemaakte drones en hun kritieke onderdelen op haar Covered List, zodat nieuwe buitenlandse modellen geen goedkeuring voor apparatuur meer kunnen krijgen.

Het oordeel: een gezonde niche voor een gespecialiseerd bedrijf met industriële of overheidsklanten en de capaciteit om exportcompliance aan beide kanten te beheren, maar geen massamarkt, en niet de plek om te leren of neurale codering op grote schaal werkt.

HOOFDSTUK 08

SILICIUM

08 De blik vanuit het silicium: energie, rekenkracht, geheugen en exactheid

Dit is het hoofdstuk waarin mijn eigen achtergrond het zwaarst telt. Ik heb vele jaren R&D en management voor fabless chips achter de rug, waaronder full-custom ontwerp van ASIC-datapaden en multimediaprojecten voor camerachips, zoals videocodecversnellers, en de vragen die een chipteam over een nieuwe codec stelt, gaan niet over BD-rate. Ze luiden: hoeveel joule per pixel, hoeveel bytes DRAM-verkeer per pixel, hoeveel gates, en kan elke decoder ter wereld precies dezelfde bits berekenen? Op alle vier staat een neurale decoder in 2026 waar hardware-HEVC stond vóór de eerste chips: veelbelovend in software, duur in silicium.

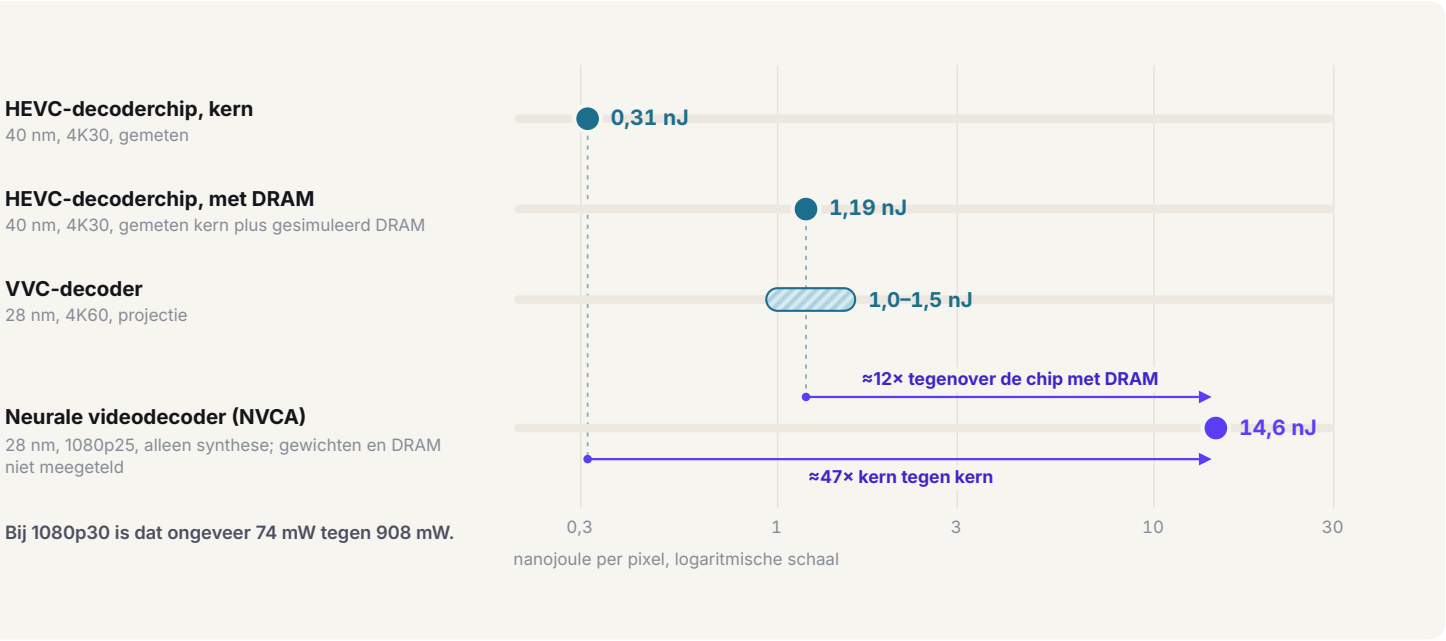
Energie per pixel: een ordegrootte verschil

Weinig neurale decoders zijn als chip ontworpen, dus de vergelijking rust op een handvol gepubliceerde ontwerpen. Het ijkpunt aan de klassieke kant is de HEVC-decoderchip van MIT (Tikekar et al., ISSCC 2013 en IEEE JSSC 2014): 40 nm, 1,77 mm², die 4K decodeert met 30 frames per seconde op 200 MHz. De kern verbruikt 78 mW, ofwel 0,31 nJ per pixel; inclusief gesimuleerd DRAM-vermogen heeft het systeem 1,19 nJ per pixel nodig. Voor VVC zijn productiecijfers niet openbaar; een analyse in Electronic Design uit 2026, die de MIT-chip naar 28 nm schaalt en de hogere complexiteit van VVC verrekent, raamt 1–1,5 nJ per pixel voor een 4K60-decoder van ongeveer één miljoen gates.

Aan de neurale kant is het enige gepubliceerde ASIC-ontwerp van een end-to-end neurale videodecoder, volgens de telling van Electronic Design, NVCA (Zhang et al., DATE 2024), gesynthetiseerd in TSMC 28 nm met 5,01 miljoen gates, exclusief de gewichten, die buiten de chip zijn opgeslagen. Het decodeert 1080p met 25 frames per seconde bij 760 mW, ofwel 14,6 nJ per pixel.

FIGUUR 06 **Energie per gedecodeerde pixel: fixed-function decoders liggen een ordegrootte voor**

Het enige gepubliceerde ontwerp van een neurale videodecoderchip heeft per pixel ongeveer 12 keer zoveel energie nodig als een gemeten HEVC-decoderchip inclusief DRAM, en ongeveer 47 keer zoveel als alleen de kern daarvan. Het eigen verkeer voor gewichten en DRAM van het neurale ontwerp is niet meegeteld.



Bron: Tikekar et al., IEEE JSSC 2014 (MIT); NVCA, Zhang et al., DATE 2024 (arXiv 2312.10716); het VVC-bereik is een projectie van Electronic Design (2026). De HEVC-chip zit op een ouder procedé (40 nm tegen 28 nm), wat het neurale ontwerp flatteert.

Dat is ongeveer 12 keer de HEVC-chip inclusief DRAM, en ongeveer 47 keer alleen de kern daarvan. De vergelijking flatteert NVCA op drie manieren: de HEVC-chip zit op een ouder procedé, het cijfer van NVCA laat het verkeer voor gewichten en DRAM weg, en NVCA is een syntheseresultaat en geen gemeten silicium. Bij 1080p30 is het verschil ongeveer 74 mW tegen zo'n 900 mW: een decoder die niemand opmerkt, tegenover een die je terugziet in het batterij- en warmtebudget van een telefoon.

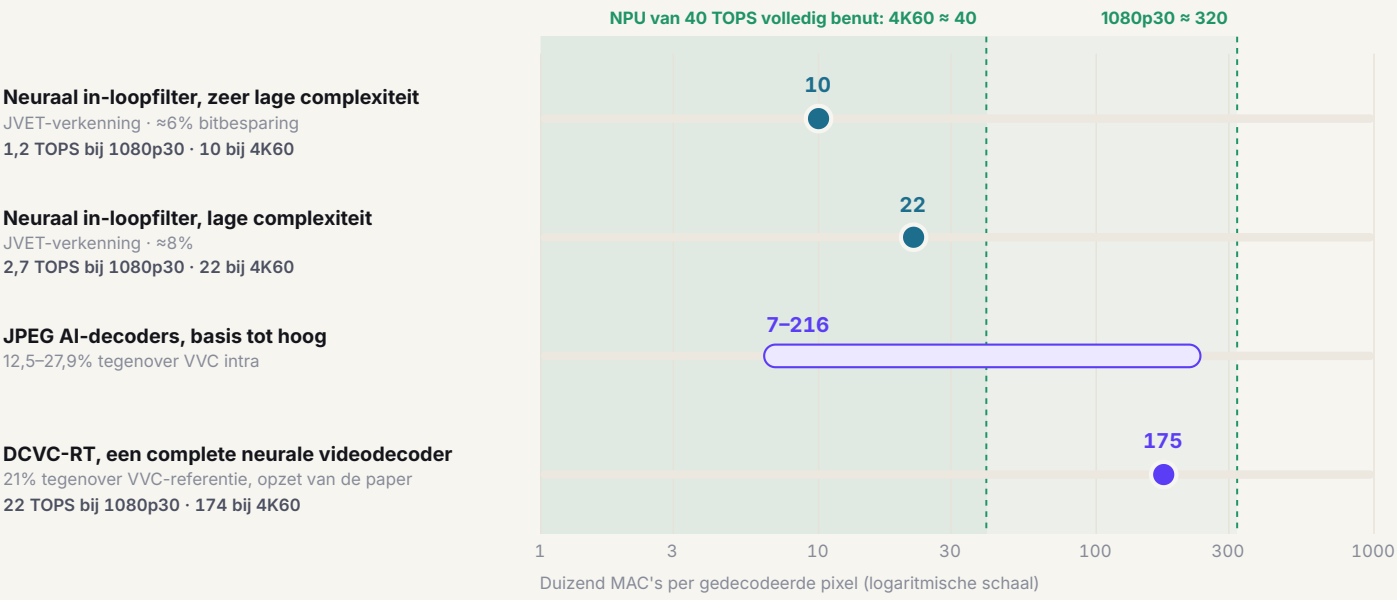
Rekenwerk per pixel: wat een neurale decoder van een NPU vraagt

De snelste gepubliceerde neurale videocodec die ik vond, is DCVC-RT van Microsoft Research (Jia et al., CVPR 2025). Hij codeert en decodeert 1080p met 125 en 113 frames per seconde op een NVIDIA A100-GPU en bespaart daarbij 21% bitrate tegenover de referentie-encoder van VVC in de testopzet van de paper, en hij gebruikt integerrekenkunde zodat verschillende apparaten consistent decoderen. Het aanvullende materiaal vermeldt de complexiteit van elke module. Als ik de vier modules optel die een decoder moet uitvoeren (decoder, feature extractor, entropiemodel en reconstructie), kom ik op ongeveer 175 duizend multiply-accumulate-bewerkingen per gedecodeerde pixel schatting.

Bij 1080p30 is dat ongeveer 11 biljoen multiply-accumulate-bewerkingen per seconde, ofwel ongeveer 22 TOPS bij twee bewerkingen per MAC, vóór enig verlies aan benutting; 4K60 vraagt acht keer zoveel schatting. Ter vergelijking: Microsoft definieert zijn AI-pc's aan de hand van een NPU van minstens 40 TOPS, en Apple geeft de Neural Engine van de M4 op met 38 TOPS. Dat zijn piekwaarden, die worden gedeeld met al het andere wat het apparaat doet.

FIGUUR 07 **Een neurale tool kost tienduizenden MAC's per pixel; een complete neurale decoder kost er honderdduizenden**

Multiply-accumulate-bewerkingen van de decoder per pixel, logaritmische schaal, tegenover wat een volledig benutte NPU van 40 TOPS aankan bij 4K60 en bij 1080p30. De winst is gemeten tegen verschillende referenties en niet direct vergelijkbaar.



Bron: JVET via de voorzitter van SC 29 (april 2026); JPEG-comité (2024); DCVC-RT, aanvullend materiaal bij CVPR 2025, som van de decodermodules door de auteur; Microsofts definitie van een Copilot+ pc. Berekening van de auteur: twee bewerkingen per MAC, volledige benutting.

Het veld laat zich het best lezen aan de hand van de ladder in figuur 7. Neurale tools die tienduizenden MAC's per pixel kosten, zoals de in-loopfilters van JVET (ongeveer 6% besparing bij 10 duizend, 8% bij 22 duizend), zijn plausibel binnen een toekomstige standaarddecoder. JPEG AI bestrijkt het hele bereik, van 7 tot 216 duizend MAC's per pixel voor 12,5–27,9% ten opzichte van VVC intra. Een complete neurale videodecoder van een paar honderdduizend MAC's per pixel hoort op een NPU, bij een bescheiden resolutie, in een gesloten keten. Dat is precies waar MLVC zit: 540p30, met hooguit de helft van de NPU.

Geheugenverkeer, niet MAC's, is vaak de echte grens

De auteurs van DCVC-RT stelden vast dat niet het rekenwerk de snelheid beperkte, maar de “operationele kosten”: geheugeninvoer en -uitvoer en het aantal functieaanroepen. Grotendeels om die reden verwijderden ze expliciete bewegingsmodules en verwerking op meerdere schalen. De ontwerpers van NVCA bouwden een speciale dataflow om het geheugenverkeer buiten de chip voor bewegings- en residufeatures te beperken. Een chipontwerper herkent het patroon. In de HEVC-chip van MIT is DRAM goed voor ongeveer driekwart van de systeemenergie (0,88 van 1,19 nJ per pixel), zelfs na een cache die de DRAM-bandbreedte met 67% verlaagde analyse. Een neurale decoder die grote feature maps met veel kanalen tussen frames bewaart, maakt het geheugenprobleem erger, niet kleiner.

Exactheid: elke decoder moet dezelfde bits berekenen

Entropiecodering werkt alleen als het kansmodel van de decoder exact overeenkomt met dat van de encoder. In een klassieke codec is dat model een tabel of een kleine set integerregels. In een neurale codec is het een netwerk, en een verschil in floating-pointberekening tussen twee apparaten, door een andere kernel, een andere ac-

cumulatievolgorde of een andere NPU, kan ertoe leiden dat de decoder de verkeerde symbolen leest en de stream vanaf dat punt kwijtraakt.

Iedereen die een neurale codec heeft uitgebracht of gestandaardiseerd, heeft dit moeten wegwerken. DCVC-RT zet zijn model om naar integers voor consistente codering op verschillende apparaten. MLVC gebruikt een hyperprior-ontwerp met gedeelde schaal om afwijkingen tussen platforms te overleven, wat waarschijnlijk deels verklaart waarom het compressie inlevert tegenover DCVC-RT analyse. De verkenning van JVET naar neurale netwerken laat elk netwerk draaien via SADL, een kleine bibliotheek met 16-bits integerinferentie. JPEG AI definieert conformiteitspunten in het latente, het residu- en het pixeldomein. Voor een chipteam is dit dezelfde discipline als bij de integertransformaties van HEVC: de standaard moet de rekenkunde exact vastleggen, wat het netwerkontwerp beperkt en leveranciersspecifieke kortere wegen uitsluit.

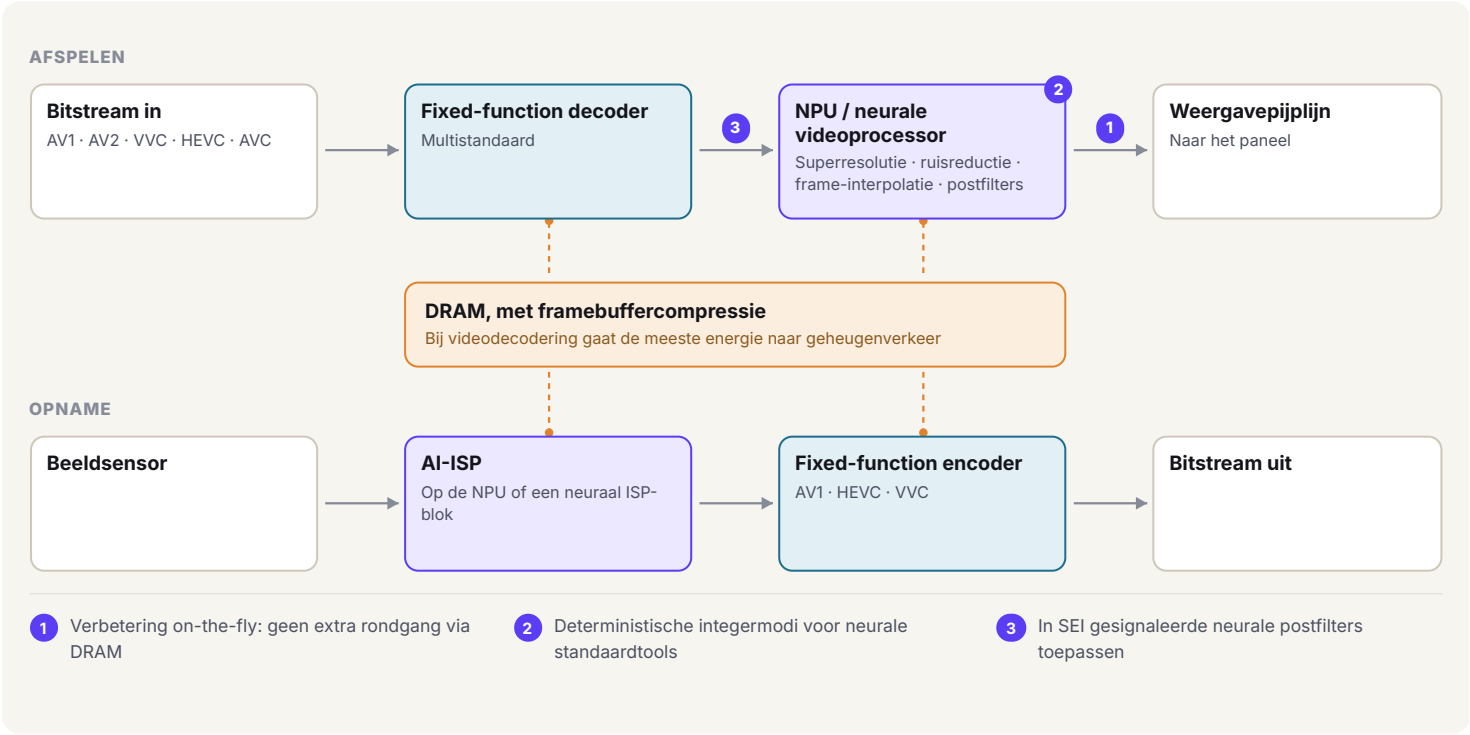
Chipoppervlak is niet het obstakel; verificatie en levensduur wel

De 5 miljoen gates van NVCA tegenover ongeveer één miljoen voor een geprojecteerde VVC-decoder lijken veel, maar op een moderne SoC voor telefoon of tv die al een NPU van meerdere TOPS bevat, zijn een paar miljoen gates niet wat het ontwerp bepaalt. Wat echt geld kost, is verificatie tegen een conformiteitssuite, een programmeerbare decoder gedurende de hele levensduur van een apparaat correct houden over software-updates heen, en het vermogen bij 4K60 HDR. Die kosten hebben pas zin als een standaard vastlegt wat er gebouwd moet worden.

De architectuur die nu wint: een fixed-function codec plus een NPU

FIGUUR 08 **De taakverdeling in producten: fixed-function hardware voor de bitstream, een neurale engine voor de pixels**

Een media-SoC uit 2026. Het codecblok decodeert elke standaard voor een fractie van een nanojoule per pixel; de NPU doet waar netwerken goed in zijn en waar standaarden niets over voorschrijven. DRAM-verkeer is de kostenpost die telt.



Bron: Architectuurbeeld van de auteur, op basis van productinformatie van Tencent (Canghai V2), Chips&Media, Allegro DVT en Samsung (2026).

Het silicium dat in 2026 verkoopt, combineert een fixed-function multistandaardcodec met een neurale engine ernaast. Het codecblok verwerkt AV1, AV2, VVC, HEVC en AVC voor een fractie van een nanojoule per pixel. De neurale engine doet waar netwerken goed in zijn en waar standaarden niets over voorschrijven: ruisreductie en de ISP bij opname, superresolutie, frame-interpolatie en verbetering bij weergave. De twee delen framebuffer-compressie en de DRAM-planning, omdat geheugenverkeer de kostenpost is die telt.

De markt laat de taakverdeling zien. Van Tencents videocodec-ASIC Canghai (沧海), sinds 2019 in eigen huis ontwikkeld, zijn meer dan 100.000 chips van de eerste generatie in gebruik bij Tencent Cloud en bij Tencents eigen live-, on-demand- en cloudgamingdiensten; de tweede generatie, waarvan in mei 2026 werd gemeld dat ze was opgestart en in volumeproductie ging voor gebruik in de tweede helft van het jaar, ondersteunt H.265 en H.266 in één architectuur, met meer dan 10% en 30% betere compressie dan de voorganger, ongeveer twee keer de doorvoer en beeldverbetering op de chip zoals superresolutie. Chips&Media verkoopt codec-IP en een op video gerichte NPU, WAVE-N, naast elkaar. Allegro DVT verkoopt AV2-decoders en heeft zijn eerste IP voor neurale videobewerking toegevoegd. De tv-processors van Samsung combineren standaarddecoders met honderden upscalingnetwerken.

Wat ik zou bouwen, per rol

SoC- en tv-chipteams. Behoud fixed-function decoders voor elke standaard die je moet afspelen, inclusief AV2 nu en VVC waar omroepen het vereisen. Dimensioneer de NPU voor nabewerking op de resolutie en framerate van het scherm, en ontwerp het pad van decoder via NPU naar scherm zo dat verbetering geen rondgang via DRAM toevoegt. Ondersteun SEI voor neurale postfilters. Bied deterministische integermodi in de NPU. Behandel H.267 als een ontwerpstart in 2029 of later, en houd enige programmeerbaarheid achter de hand voor de neurale tools ervan.

IP-leveranciers. Verkoop nu AV2- en VVC-decoders, plus de onderdelen die neurale bewerking eromheen nodig heeft: postfilterversnellers, framebuffercompressie en conformiteitstools.

Videoplatforms. Neurale voorbewerking bij de encoder verdient zich het eerst terug, omdat ze één keer per titel draait en niets van apparaten vraagt. Gebruik superresolutie aan clientzijde waar apparaten het toelaten. Bewaar end-to-end neurale codecs voor gesloten ketens die je zelf beheerst.

HOOFDSTUK 09

09

STANDAARDEN EN PATENTEN

Standaarden en patenten: waar neurale tools licentieerbaar worden

De standaardisatiepijplijn

Standaard	Instantie	Neurale inhoud	Status, november 2026
H.267 (werknaam; voorbij VVC)	JVET (ITU-T en ISO/IEC)	Open. Klassieke tools (ECM) en neurale tools (in-loopfilters, intrapredictie, superresolutie) zijn allebei kandidaat	Oproep tot voorstellen juli 2026; inzendingen beoordeeld januari 2027; eerste testmodel tweede helft van 2027; versie 1 eind 2029
VVC en HEVC, met VSEI	JVET	Een neuraal postfilter kan in SEI-metadata worden beschreven; toepassen is optioneel voor de decoder	Van kracht; berichten uitgebreid in de edities van oktober 2025
AV2	AOMedia	Klassieke tools, royaltyvrij	Versie 1.0.0 op 28 mei 2026; decoder-IP in licentie gegeven vanaf juni 2026; webstreams van YouTube gepland voor begin 2027
JPEG AI	JPEG (ISO/IEC en ITU-T)	End-to-end getrainde beeldcodec, voor menselijk en machinaal zien	Deel 1 gepubliceerd in 2025; een tweede versie in bespreking; geen product gevonden dat het standaard ondersteunt
IEEE 1857.11	IEEE	Beeldcodering met neurale netwerken, ontworpen voor NPU's	Gepubliceerd december 2024
Video Coding for Machines	MPEG	Codering voor analyse door machines in plaats van kijken	Internationale ontwerpstandaard begin 2026; definitief gepland voor eind 2026
ML-uitbreidingen van Opus	IETF	Optionele diepe redundantie, verbetering aan decoderzijde, bandbreedte-uitbreiding	In libopus 1.5 en 1.6; diepe redundantie in standaardisatie

Er vallen twee dingen op. Ten eerste komt neurale inhoud standaarden binnen via de randen, vóór de kern: als optionele postfilters, optionele verbetering en beeldcodering, die het hoofdpad van de decodering allemaal ongemoeid laten. Ten tweede ligt de enige standaard die neurale tools in het decodeerpad van elk scherm zou kunnen brengen, H.267, nog drie jaar van een eerste versie af.

Wie zich positioneert in het patentlandschap

Standaardtools worden licentieerbare patenten, dus het H.267-proces is ook een patentrace, en de neurale tools maken daar deel van uit. Het landschapsonderzoek van GreyB uit 2025 naar 3.916 patenten voor technologie voorbij VVC zet ByteDance aan kop, met 645 patentfamilies en 251 JVET-bijdragen, en wijst op samenwerkingen tussen universiteiten en industrie, zoals Xiaomi en de Xidian-universiteit, met meer dan 20 families. MLex meldde in maart 2026 dat Tencent, Alibaba en ByteDance opkomen als actieve deelnemers in patentpools voor videoco-

decs. In mei 2026 kocht Sauvegarder Investment Management van Alibaba 126 patentactiva voor videocodecs, op het gebied van AV1, AV2 en Video Coding for Machines. Met de overname van Deep Render in oktober 2025 voegde InterDigital de patenten van Deep Render op AI-videocodering aan zijn portefeuille toe.

Het geld per apparaat is eenvoudig af te lezen aan de tarievenlijsten van de pools. De tarieven van Access Advance voor VVC (per 1 juli 2026, voor licentienemers die aan hun verplichtingen voldoen en het merk gebruiken) zijn 0,50 dollar per mobiele telefoon, tablet of laptop in markten van Regio 1 zoals de VS, de EU en Japan, en 0,25 dollar in Regio 2; voor 4K-tv's 1,00 tot 1,50 dollar in Regio 1, afhankelijk van de prijs; met een jaarlijks plafond per onderneming van 75 miljoen dollar. Voor AV1 en AV2 vragen de leden van AOMedia geen royalty's. Wat H.267 ook bevat, de tools ervan zullen op dezelfde manier worden geprijsd, en daarom dragen bedrijven met patenten op neurale codering nu bij aan JVET analyse.

HOOFDSTUK 10

10 WIE ERAAN VERDIENT
Wie eraan verdient, en waar een bedrijf kan meedoen

Het geld in 2026

FIGUUR 09 **Wie er in 2026 aan neurale codering verdient**

Het geld volgt de standaarden en de platforms. Ik vond geen zelfstandig bedrijf dat zijn geld verdient met het in licentie geven van een end-to-end neurale videocodec.

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)

Bron: Openbare informatie van bedrijven en patentpools die in het rapport wordt geciteerd; stand per 12 november 2026. Pooltarieven: Access Advance, VVC, Regio 1, bij naleving en met gebruik van het merk, per 1 juli 2026.

Platforms houden de besparingen zelf. Netflix, YouTube, Meta, Microsoft en Tencent ontwikkelen neurale bewerking in eigen huis en houden het voordeel zelf, in de vorm van minder bandbreedte, minder opslag en een betere kwaliteit per bit. Als het de adoptie helpt, geven ze de code weg: Microsoft bracht MLVC uit onder de MIT-licentie.

Apparaat- en chipmakers verkopen onderscheidend vermogen. Camera-ISP's en beeldprocessors in tv's zijn de plekken waar neurale netwerken elke gebruiker raken, en ze worden als functies verkocht: de Cognitive ISP

van Qualcomm, de upscalingprocessors van Samsung.

IP-leveranciers verkopen decoders voor elke nieuwe standaard, plus NPU's. Binnen drie weken na de specificatie van AV2 kondigden Chips&Media, Allegro DVT en VeriSilicon alle drie klanten voor AV2-decoders aan. Dezelfde leveranciers verkopen nu blokken voor neurale bewerking naast de codecs.

GPU-leveranciers bouwen platforms. DLSS en neurale texturecompressie maken de tensorhardware van NVIDIA tot een onderdeel van hoe games worden gemaakt.

Patenthouders innen per apparaat, via pools voor VVC en HEVC en, na 2029, voor H.267.

Startups in neurale codecs zijn overgenomen. WaveOne ging in 2023 naar Apple, iSIZE in 2023 naar Sony Interactive Entertainment en Deep Render in 2025 naar InterDigital, alle drie tegen niet bekendgemaakte voorwaarden. Telkens kocht een platform of een licentiegever het team en de patenten. Ik vond in 2026 geen zelfstandig bedrijf dat zijn geld verdient met het in licentie geven van een end-to-end neurale videocodec.

Waar een bedrijf kan meedoen

Kans	Wie koopt	Waarom nu	Grootste risico
Neurale voorbewerking bij de encoder (downscaling, ruisreductie, afstemming per titel)	Streaming- en gebruikersvideoplatforms, omroepen	Geen wijziging aan apparaten nodig; bewezen bij Netflix	Grote platforms bouwen het zelf
IP en afstemming voor verbetering aan schermzijde (superresolutie, frame-interpolatie, beeldkwaliteit)	Chipmakers voor tv's, settopboxen en telefoons	NPU's zitten al in de apparaten; zichtbaar onderscheid	Commoditisering; chipmakers doen het zelf
Neurale codecs in gesloten ketens (videovergaderen, bediening op afstand, privéverbindingen)	Leveranciers van bedrijfscommunicatie, industriële systemen en telepresence	MLVC bewijst de haalbaarheid, en de code is open	Je moet beide kanten beheersen; versnippering van NPU's
Codec plus NPU, als silicium of IP (AV2, VVC en nabewerking)	SoC-makers	De AV2-cyclus begint	Sterke gevestigde spelers
Tooling voor 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	De inzendingen worden in januari 2027 beoordeeld	Lang, duur, onzeker

Voor de meeste bedrijven is in 2027 omzet te halen in de eerste twee rijen. De derde is het interessantst voor een gericht productbedrijf. De laatste vraagt geduld en een licentiestrategie.

HOOFDSTUK 11

11 CLAIMS EN BEWIJS
Claims die je zult horen, en wat het bewijs zegt

Claim	Wat het bewijs zegt
"Neurale codecs verslaan VVC al."	In het lab wel. DCVC-RT bespaart 21% tegenover de referentie-encoder van VVC bij 1080p in de opzet van de paper; JPEG AI bespaart 12,5–27,9% tegenover VVC intra. De decoder kost ongeveer 175 duizend MAC's per pixel, ofwel het energieverbruik van een NPU, tegenover een fractie van een nanojoule in fixed-function hardware.
"AI-codecs vervangen H.26x en AV1 binnen een paar jaar."	Niet bij het afspelen. De eerste standaard die neurale tools kan bevatten, mikt op eind 2029, chips volgen rond 2031 en breed gebruik halverwege de jaren 2030 schatting. De codec die nu komt, AV2, is klassiek.
"MLVC bespaart 88% bandbreedte."	Tegenover H.264 bij 360p, op Microsofts eigen testset. Tegenover hardware-HEVC is het 75,5% bij 360p en 65,4% bij 540p, in gesprekken waarin Microsoft beide kanten beheerst.
"Neurale textures besparen 85–96% videogeheugen."	In demonstratiescènes. De uitgebrachte game won ongeveer 30% effectieve compressie en gebruikte neurale textures vanwege de kosten tijdens runtime alleen voor geselecteerde assets.
"JPEG AI is een ISO-standaard, dus camera's en browsers gaan het gebruiken."	Ik vond geen product dat het standaard ondersteunt. Het formaat dat het web in 2026 koos, JPEG XL in Chrome 155, is klassiek.
"Neurale decodering is te duur voor telefoons."	Deels waar. MLVC decodeert 540p30 binnen een halve NPU; 4K60 met een model van DCVC-RT-klasse zou zo'n 175 TOPS aan continue rekenkracht vragen schatting.
"Satelliet- en droneverbindingen zijn de grote markt voor neurale video."	Technisch de beste match, maar klein en aan beide kanten van de Stille Oceaan steeds sterker onder exportcontrole.
"Grote platforms gaan een neurale codec van een startup in licentie nemen."	Ze bouwden hun eigen (Netflix, Microsoft, Tencent) of kochten de startup (Apple, Sony, InterDigital).

Een eenvoudige test prikt door de meeste pitches heen: vraag naar de referentie, de maatstaf, de resolutie en de decodeerkosten. Een claim die niet alle vier kan noemen, is een onderzoeksresultaat, geen product.

HOOFDSTUK 12

12 WEGWIJZERS
Waar je op moet letten, 2027 tot 2030

Wanneer	Wegwijzer	Waarom het ertoe doet
Januari 2027	JVET beoordeelt in Brisbane de inzendingen voor H.267	Hoeveel topinzendingen leunen op neurale tools, en tegen welke kosten per pixel
Begin 2027	Eerste AV2-webstreams van YouTube; AV2-decoders in aangekondigde SoC's	De volgende klassieke cyclus bepaalt de apparaatbudgetten tot ongeveer 2030
Tweede helft van 2027	Eerste testmodel van H.267	Of neurale tools in het kernontwerp zitten of alleen in de verkenning
2027	MLVC buiten peer-to-peer-gesprekken in Teams; andere leveranciers die de open code overnemen	De eerste test van een open neurale codec buiten zijn eigenaar
2027–2028	Een eerste product dat JPEG AI standaard schrijft of leest	Het eerste neurale formaat dat over bedrijfsgrenzen heen wordt gebruikt
2027–2028	DirectX Linear Algebra in uitgebrachte drivers; neurale textures in meer games en op mobiel	Neurale decodering binnen graphics, over leveranciers heen
Doorlopend	NPU-budgetten in telefoons en pc's	Bij ongeveer 100 TOPS continu zou een 1080p30-decoder van DCVC-RT-klasse ongeveer een vijfde van de NPU innemen in plaats van de helft analyse
Eind 2029	H.267 versie 1	Het begin van de siliciumcyclus voor neurale tools bij het afspelen op massaschaal

HOOFDSTUK 13

13 OPEN VRAGEN Open vragen

Verschillende vragen in dit rapport hebben nog geen publiek antwoord, en technisch vind ik juist die het interessantst.

> **Exactheid op schaal.** Hoe dicht kan een neurale decoder bij bitexacte rekenkunde met alleen integers komen op NPU's van verschillende leveranciers, voordat hij het grootste deel van zijn compressievoordeel kwijt is? De zes procentpunten die MLVC inlevert, zijn één datapunt; een standaard heeft een algemeen antwoord nodig.

> **Geheugen vóór MAC's.** Wat kost neurale nabewerking aan DRAM-verkeer per pixel als ze on-the-fly draait tussen een hardwaredecoder en het scherm, en hoe klein kan de werkset worden gemaakt?

> **De prijs van een neurale tool in H.267.** Bij hoeveel MAC's en nanojoule per pixel verdient een neuraal inloopfilter in een 4K60-decoder de bits die het bespaart niet meer terug?

> **Gesloten ketens buiten gesprekken.** Welke andere gesloten systemen, van foto-opslag in de cloud en gamestreaming tot robots en privécontributieverbindingen, kunnen end-to-end neurale codecs gebruiken voordat H.267 er is?

> **Decoders die genereren.** Als een decoder details genereert in plaats van ze te reconstrueren, zoals video-VAE's al doen, hoe moet getrouwheid dan worden gemeten en gespecificeerd?

Dit zijn gebieden die we de komende jaren samen kunnen verkennen, via technische uitwisseling en technische samenwerking: i@tanshuai.com.

Beperkingen van dit rapport

Dit is bureauonderzoek op basis van openbare documenten tot en met 12 november 2026. Voor deze editie heb ik geen primaire interviews afgenomen en geen onafhankelijke metingen gedaan. Verschillende kerncijfers zijn claims van leveranciers of auteurs op hun eigen testsets, zoals de bitratebesparingen van MLVC en de compressieverhoudingen van neurale textures; ze zijn als zodanig gemarkeerd. Mijn schattingen van het rekenwerk voor DCVC-RT tellen de cijfers per module op die de auteurs publiceren en negeren benuttingsverliezen, dus de werkelijke belasting van de NPU zal hoger zijn. De energievergelijking rust op een paar gepubliceerde ontwerpen op verschillende procedés; productiecijfers voor VVC- en AV2-decoders zijn niet openbaar. De tijdpaden van standaarden zijn de eigen doelen van de comités. De jaartallen voor H.267 na 2029 zijn mijn schattingen op basis van de geschiedenis van AV1 en VVC.

Over de auteur

Tan Shuai is technoloog en oprichter, met meer dan twintig jaar werk in chips, hardware en systeemsoftware, en 20 verleende patenten. Hij heeft vele jaren R&D en management voor fabless chips achter de rug, waaronder full-custom ontwerp van ASIC-datapaden en multimediaprojecten voor camerachips, zoals videocodecversnellers, en dat is de lens van dit rapport: energie per pixel, geheugenverkeer, en wat een standaard een decoder in silicium laat vastleggen. Zijn werk bracht hem naar Silicon Valley, New York, Chicago, Taiwan, de Greater Bay Area (Shenzhen en Hongkong), Vietnam en Maleisië. Tegenwoordig werkt hij als AI-onderzoeker en -ontwikkelaar.

Technische discussie: i@tanshuai.com · tanshuai.com

Opmerkingen en disclaimer

Dit rapport is onafhankelijk bureauonderzoek op basis van openbare documenten die tot en met 12 november 2026 beschikbaar waren. Het is uitsluitend bedoeld ter informatie en is geen beleggings-, juridisch of technisch advies. Producten, tijdpaden van standaarden, tarieven van patentpools en exportregels veranderen; controleer alles waarop je vertrouwt aan de hand van de primaire bron. Cijfers die zijn gemarkeerd met leverancierclaim, schatting of analyse, moeten als zodanig worden gelezen. Bedrijfs- en productnamen behoren toe aan hun respectieve eigenaars en worden alleen ter identificatie gebruikt; geen enkel bedrijf heeft dit rapport beoordeeld of gesponsord.

© 2026 Tan Shuai. Je mag dit rapport ongewijzigd en met bronvermelding delen.

Bronnen

Alle geraadpleegd tussen 3 en 12 november 2026. Gegroepeerd per hoofdstuk.

Verkeer, apparaten en afbakening

- Ericsson, *Mobility Report*, juni 2026 (video ongeveer 75% van het mobiele dataverkeer eind 2025): https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/june-2026 ; samenvatting: 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, wereldwijde smartphoneleveringen, Q4 en heel 2025 (1,26 miljard), 13 januari 2026: https://www.idc.com/resource-center/press-releases/q425mobilephonetop5/
- TrendForce, prognose van de tv-leveringen in 2025 (195,71 miljoen): 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, tv-prognose voor 2025 (208,7 miljoen): https://iconnect007.com/article/145954/omdia-forecasts-flat-tv-shipments-in-2025-but-mini-led-poised-for-explosive-growth-at-829/145951/ein

Afspelen en 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, 28 mei 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” (aandeel van AV1 in de catalogus van YouTube): 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” (commerciële exploitatie van DTV+, juni 2026): https://mpegh.com/brazil/ ; CartaCapital over ontvangers in tv's: https://www.cartacapital.com.br/?p=628337 ; Olhar Digital over 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”, 29 oktober 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, 16 oktober 2025: https://atscaleconference.com/how-meta-deployed-super-resolution-at-scale-to-transform-video-quality/
- Samsung, overzicht van de 8K-tv's (NQ8 AI Gen3, 768 neurale netwerken): https://www.samsung.com/us/tvs/8k-tv/highlights/

Fotografie en beeld

- 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 en Visionary.ai, volledig op AI gebaseerde ISP, 5 januari 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/ ; interview in PetaPixel: https://petapixel.com/2026/01/18/a-look-at-the-worlds-first-full-ai-based-image-signal-processor
- JPEG-comité, persbericht van de 103e vergadering (winst, complexiteit en smartphone-implementaties van JPEG AI): https://jpeg.org/items/20240507_press.html
- G. J. Sullivan, inleiding tot ISO/IEC JTC 1/SC 29, april 2026 (status van JPEG AI; cijfers voor ECM en NNVC): 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/
- Werkgroep IEEE 1857.11 (beeldcodering op basis van neurale netwerken): https://sagroups.ieee.org/fvc
- Chrome 155 en JPEG XL, 6 oktober 2026: https://xenospectrum.com/en/chrome-155-jpeg-xl-rust/

Bellen en vergaderen

- Microsoft, “Satin: Microsoft's latest AI-powered audio codec for real-time communications”, 17 februari 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)”, 24 juli 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 ; analyse van iReplay: https://ireplay.tv/blog/mlvc-microsoft-neural-video-codec-open-source-streaming/

- Releasenotes van Opus 1.5: 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”, 13 juni 2024: https://engineering.fb.com/2024/06/13/web/mlow-metas-low-bitrate-audio-codec/
- Microsoft, Copilot+ pc's (NPU van 40+ TOPS): https://www.microsoft.com/en-us/windows/copilot-plus-pcs ; Apple, aankondiging van de M4 (Neural Engine van 38 TOPS): https://www.apple.com/newsroom/2024/05/apple-introduces-m4-chip/

Generatieve AI

- Kyutai, modelkaart van Mimi (12,5 Hz, 1,1 kbps): https://huggingface.co/kyutai/mimi
- Lightricks, “LTX-Video: Realtime Video Latent Diffusion” (video-VAE van 1:192): https://arxiv.org/abs/2501.00103
- Persbericht van de 153e MPEG-vergadering (Video Coding for Machines in het DIS-stadium): https://www.mpeg.org/wp-content/uploads/2026/03/press_release_153.pdf

Games en graphics

- NVIDIA, adoptie van DLSS 4: lancering en Computex 2025 (75 en 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”, februari 2026: https://www.ubisoft.com/en-us/studio/laforge/news/415slmB3ZGzv8d2LhO9QT/shipping-neural-texture-compression-in-assassins-creed-mirage
- TweakTown over de demo van neurale textures van NVIDIA op GTC 2026 (6,5 GB naar 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, 26 februari 2026: https://devblogs.microsoft.com/directx/shader-model-6-9-retail-and-more/ ; over cooperative vector: https://devblogs.microsoft.com/directx/shader-model-6-9-and-the-future-of-cooperative-vector/ ; plannen voor GDC 2026: https://windowsreport.com/microsoft-showcases-neural-rendering-and-dxr-2-0-plans-for-directx-at-gdc-2026/

Gespecialiseerde verbindingen

- Global Trade Alert over de Chinese Bekendmaking nr. 49 van 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 (drempels voor onbemande grondplatforms): 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 zet in het buitenland gemaakte drones en onderdelen op de Covered List, 22 december 2025: https://dronelife.com/2025/12/22/fcc-adds-foreign-made-drones-and-components-to-covered-list-citing-national-security-risks/

Silicium

- 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 ; chippagina van MIT (78 mW bij 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 ; aanvullend materiaal (complexiteit per module): https://openaccess.thecvf.com/content/CVPR2025/supplemental/Jia_Towards_Practical_Real-Time_CVPR_2025_supplemental.pdf
- NNVC en de SADL-bibliotheek voor integerinferentie: https://arxiv.org/abs/2309.05846
- Tencent Canghai V2, 27 mei 2026 (Shanghai Securities News via 10jqka): https://stock.10jqka.com.cn/20260527/c677016020.shtml ; samenvatting van de functies: 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, 12 mei 2026: https://www.design-reuse.com/news/202530501-chips-media-completes-development-of-next-gen-av2-hw-decoder-ip/ ; AV2-licentie, 17 juni 2026: https://www.edge-ai-vision.com/?p=58270
- Allegro DVT, AV2-decoder-IP, 16 juni 2026: https://www.edge-ai-vision.com/2026/06/allegro-dvts-pulsar-decoder-ip-adds-support-for-av2-video-codec/
- VeriSilicon, VC9800D met AV2, 10 juni 2026: https://www.businesswirechina.com/en/news/63176.html

Standaarden en patenten

- Persbericht van de 155e MPEG-vergadering (gezamenlijke oproep tot voorstellen voorbij VVC), 5 augustus 2026: https://www.mpeg.org/wp-content/uploads/2026/08/ag3w254SN26474Press-Release-MPEG155.pdf
- Persbericht van de 152e MPEG-vergadering (resultaten van de oproep om bewijs; edities van VVC en VSEI), 30 oktober 2025: https://www.mpeg.org/wp-content/uploads/2025/11/Press-Release-MPEG152.pdf ; MPEG-column van ACM SIGMM: 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, patentlandschap voorbij VVC (H.267): https://greyb.com/blog/exclusive-video-codec-landscape/ ; over AV2-licenties en de verkoop van de Alibaba-portefeuille: https://greyb.com/blog/av2-licensing/
- MLex, “Fragmented codec licensing exposes need for clearer rules for Chinese tech majors”, 27 maart 2026: https://www.mlex.com/mlex/articles/2458402/fragmented-codec-licensing-exposes-need-for-clearer-rules-for-chinese-tech-majors
- Access Advance, royaltytarieven van MCBA en VVC Advance per 1 juli 2026: https://accessadvance.com/wp-content/uploads/2025/07/MCBA-Program-Overview-September-2026.pdf

Geld en bedrijven

- InterDigital neemt Deep Render over, 30 oktober 2025: https://www.advanced-television.com/2025/10/30/interdigital-acquires-deep-render
- TechCrunch, “Apple acquired a startup using AI to compress videos”, 27 maart 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/

NEURALE COMPRESSIE · AFSPELEN × SILICIUM

Neurale codering in 2026

Ingezet rond de codec, niet in plaats ervan

Elk kerncijfer in dit rapport staat vermeld met bron en datum. Waar de auteur een cijfer zelf afleidt, is het gemarkeerd als schatting; waar alleen de verkoper het rapporteert, is het gemarkeerd als leverancierssclaim.

Tan Shuai

Technoloog en oprichter · AI-onderzoeker en -ontwikkelaar

tanshuai.com · i@tanshuai.com

OPENBARE EDITIE · Gegevens per 12 november 2026