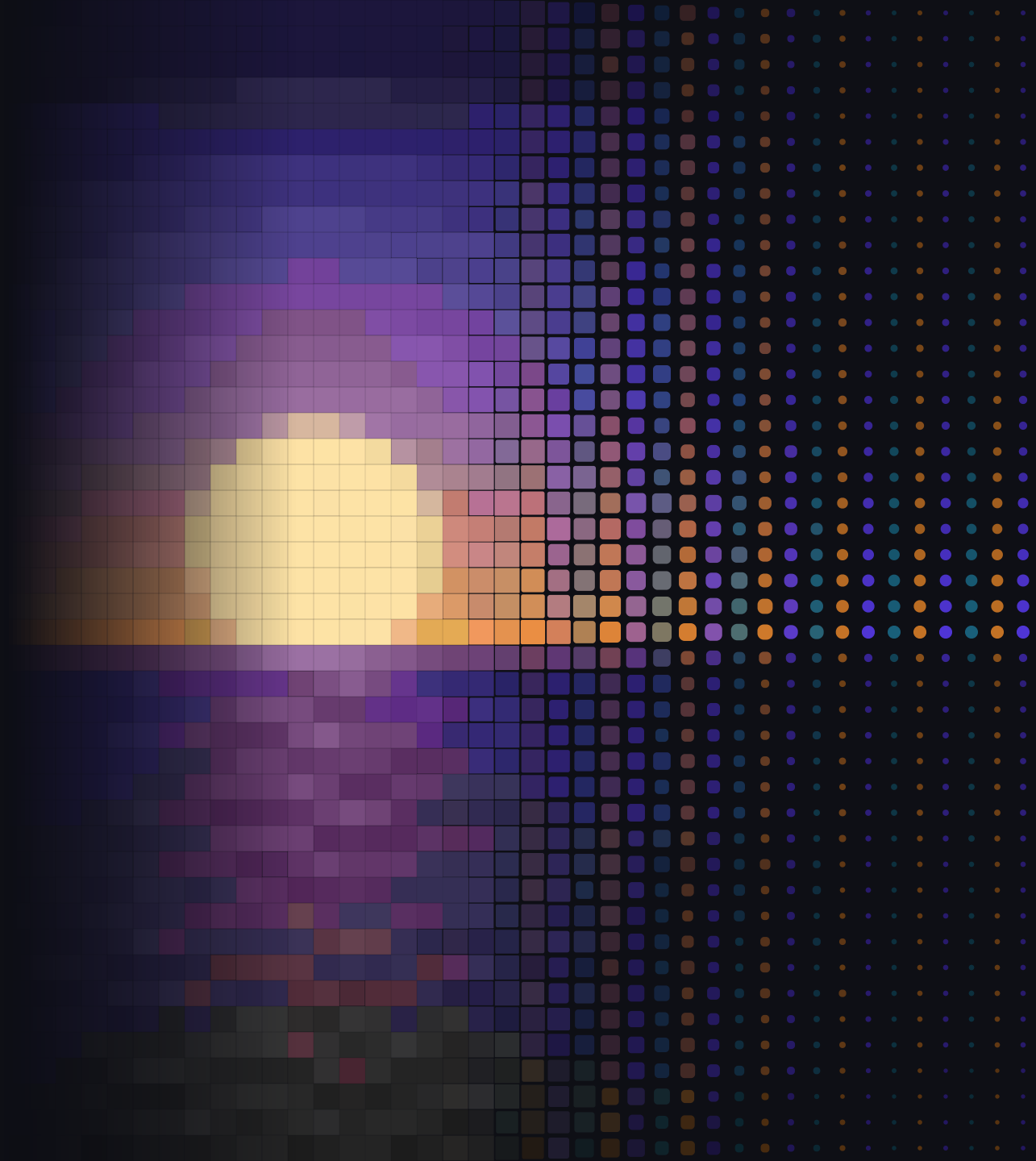
Bài nghiên cứu · Tháng 11/2026

Mã hóa nơ-ron năm 2026

Quanh codec, chứ không thay cho codec

Phát video · Nhiếp ảnh · Cuộc gọi · AI tạo sinh · Trò chơi · Chip

Tan Shuai
tanshuai.com · i@tanshuai.com



KẾT LUẬN

Đã phổ biến, nhưng nằm quanh codec

01 Mạng nơ-ron đã phổ biến ở cả hai đầu của các codec chuẩn

ISP camera, giảm độ phân giải phía mã hóa và siêu phân giải khi phát. Luồng bit ở giữa vẫn là AV1, HEVC, AVC hoặc VVC, do phần cứng chức năng cố định giải mã.

02 Codec nơ-ron đầu cuối chỉ được triển khai trong vòng khép kín

Cuộc gọi, bộ token hóa của AI tạo sinh và một trò chơi AAA: một công ty kiểm soát cả hai đầu và có thể chuyển về codec chuẩn.

03 Luồng bit nơ-ron trên màn hình đại chúng phải chờ H.267

Phiên bản 1 dự kiến ra đời cuối năm 2029; chip đầu tiên khoảng năm 2031 và phổ cập vào giữa thập niên 2030 ước tính.

≈75%
lưu lượng dữ liệu di động là video (cuối năm 2025)

1,26 tỷ
điện thoại thông minh xuất xưởng năm 2025, đều giải mã video bằng phần cứng chức năng cố định

≈12×
năng lượng mỗi pixel so với bộ giải mã HEVC phần cứng, ở thiết kế chip giải mã nơ-ron duy nhất đã công bố

Ba cách mạng nơ-ron đi vào codec



Quanh codec: không cần ai cho phép. Trong tiêu chuẩn: cần một tiêu chuẩn. Thay cho codec: cần kiểm soát cả hai đầu.

Mã hóa nơ-ron được triển khai ở đâu năm 2026

	Quanh một codec chuẩn tiền xử lý, tăng cường, nâng độ phân giải	Codec nơ-ron đầu cuối bộ mã hóa và giải mã học máy	Trong tiêu chuẩn công cụ nơ-ron trong luồng bit hoặc API chuẩn
01 **Phát video và phát trực tuyến**	QUY MÔ LỚN Netflix giảm độ phân giải; YouTube, Meta dùng siêu phân giải; TV có bộ nâng độ phân giải AI	CHƯA TRIỂN KHAI Chưa dịch vụ nào phát định dạng video nơ-ron; TV và điện thoại giải mã codec chuẩn bằng phần cứng	ĐANG CHUẨN HÓA H.267 phiên bản 1 dự kiến cuối năm 2029; công cụ nơ-ron nằm trong số ứng viên
02 **Nhiếp ảnh và xử lý ảnh**	QUY MÔ LỚN ISP AI trong điện thoại cao cấp; ISP nơ-ron hoàn chỉnh ra mắt dạng IP tại CES 2026	DEMO, THÍ ĐIỂM JPEG AI là tiêu chuẩn ISO: có demo trên điện thoại, chưa có sản phẩm	CHƯA TRIỂN KHAI Định dạng đang lan rộng là định dạng truyền thống: JPEG XL bật mặc định trong Chrome 155
03 **Cuộc gọi và họp trực tuyến**	QUY MÔ LỚN Siêu phân giải phía nhận và khử tiếng ồn bằng mạng nơ-ron	QUY MÔ LỚN Âm thanh nơ-ron trong cuộc gọi hai người trên Teams từ 2021; video MLVC đang triển khai trên Teams	TRIỂN KHAI HẠN CHẾ Opus 1.5 thêm công cụ ML tùy chọn, tương thích ngược
04 **AI tạo sinh**	—	QUY MÔ LỚN Mọi mô hình video và giọng nói đều chạy một bộ token hóa học máy, như VAE video 1:192	—
05 **Trò chơi và đồ họa**	QUY MÔ LỚN Nâng độ phân giải và tạo khung hình nơ-ron: riêng DLSS 4 có trong hơn 250 trò chơi, ứng dụng	TRIỂN KHAI HẠN CHẾ Texture nơ-ron đã có trong một trò chơi AAA; SDK của NVIDIA đang ở bản beta	ĐANG CHUẨN HÓA DirectX đang bổ sung phép toán ma trận cho shader (bản xem trước)
06 **Liên kết chuyên dụng: vệ tinh, robot**	DEMO, THÍ ĐIỂM Khôi phục phía máy chủ cho nguồn cấp bitrate thấp	DEMO, THÍ ĐIỂM Codec nơ-ron chịu mất gói đang được nghiên cứu và thí điểm	—

Các hàng được sắp theo số người mà mỗi lĩnh vực tiếp cận.

Nơ-ron ở hai đầu, chuẩn ở giữa

- **Ở bộ mã hóa**

  Bộ giảm độ phân giải nơ-ron của Netflix chạy từ năm 2022 mà không đổi gì trên thiết bị; khoảng 77% người xem ưa chọn nó trong thử nghiệm của Netflix **theo nhà cung cấp**.

- **Ở màn hình**

  YouTube Super Resolution trên TV từ tháng 10/2025; Meta chạy siêu phân giải trên máy chủ và trên điện thoại.

- **Trong TV**

  Bộ xử lý NQ8 AI Gen3 của Samsung quảng cáo 768 mạng nơ-ron để nâng độ phân giải.

LUỒNG BIT VẪN THEO CHUẨN

30%

lượng xem Netflix bằng AV1 (tháng 12/2025)

Trên 90%

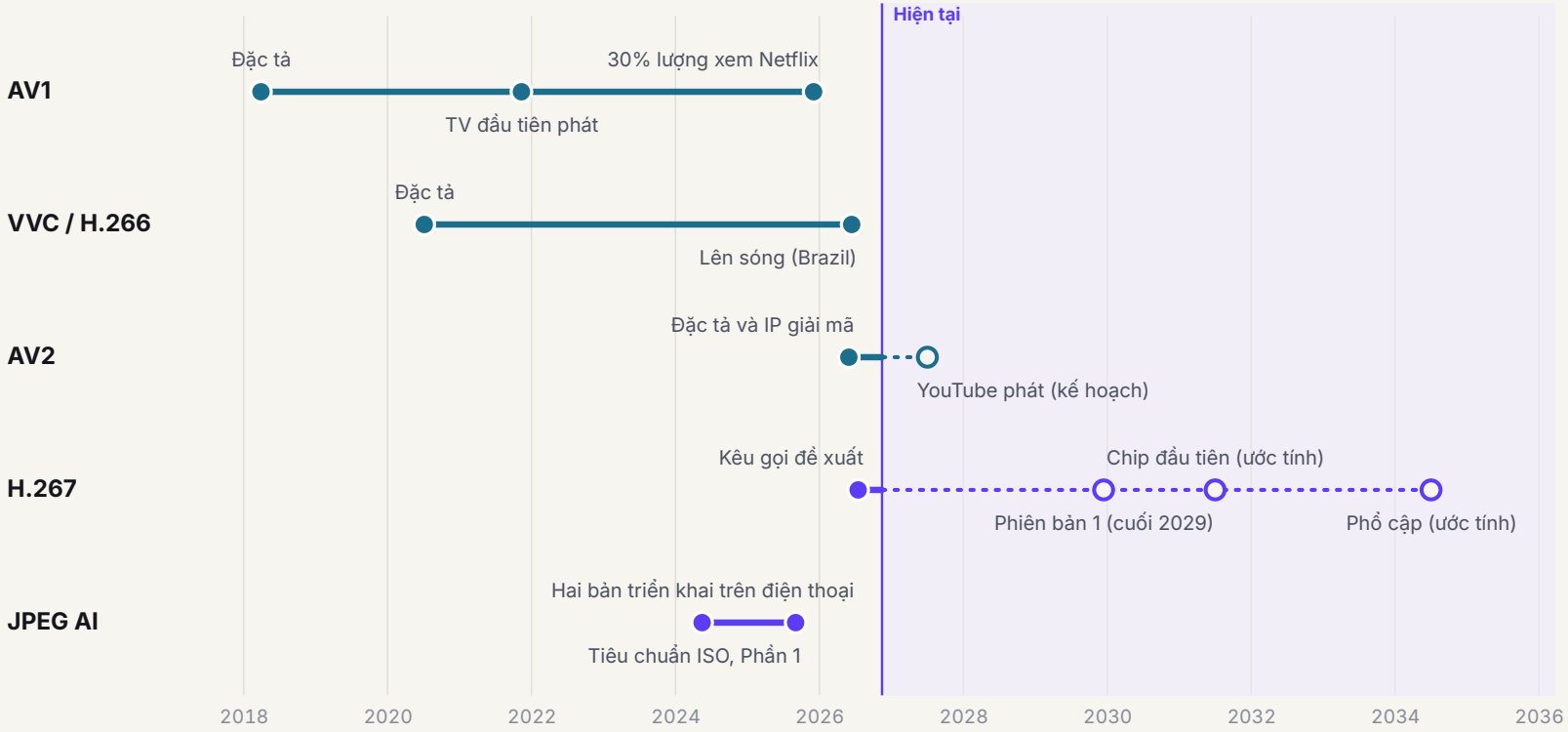
kho nội dung YouTube, tính theo thời lượng xem, có sẵn ở AV1 (tháng 3/2026)

3 tuần

từ AV2 phiên bản 1.0.0 (28/5/2026) đến khi ba nhà cung cấp có khách hàng IP giải mã

THỜI ĐIỂM

Từ tiêu chuẩn hoàn tất đến màn hình đại chúng mất gần trọn một thập kỷ



AV1 mất gần tám năm từ đặc tả (2018) đến 30% lượng xem Netflix (tháng 12/2025). Các mốc rỗng là mục tiêu hoặc ước tính của tôi.

H.267 quyết định công cụ nơ-ron có đến mọi màn hình không

Tháng 10/2025
Kêu gọi bằng chứng: năm hồ sơ từ 16 công ty và phòng thí nghiệm; lợi ích nơ-ron rõ nhất ở 8K HDR

Tháng 7/2026
Phát hành kêu gọi đề xuất tại Geneva

Tháng 1/2027
Đánh giá hồ sơ phản hồi tại Brisbane

Nửa cuối năm 2027
Mô hình thử nghiệm đầu tiên

Cuối năm 2029
Phiên bản 1 của tiêu chuẩn

≈28%
tiết kiệm so với VVC nhờ hướng ECM truyền thống

6–14%
tiết kiệm nhờ bộ lọc nơ-ron trong vòng lặp, với 10 đến hơn 22 nghìn MAC mỗi pixel

SEI
luồng VVC đã có thể báo hiệu một bộ lọc hậu xử lý nơ-ron để bộ giải mã áp dụng

Nhận định của tôi: một lõi truyền thống cùng vài công cụ nơ-ron ở nơi lượng bit tiết kiệm xứng với chi phí giải mã phân tích.

Nơ-ron trong camera, truyền thống trên web

- **Camera**

  Cognitive ISP của Qualcomm chạy phân đoạn thời gian thực trên điện thoại từ năm 2023; một ISP AI hoàn chỉnh chạy trên NPU đã ra mắt tại CES 2026.

- **JPEG AI**

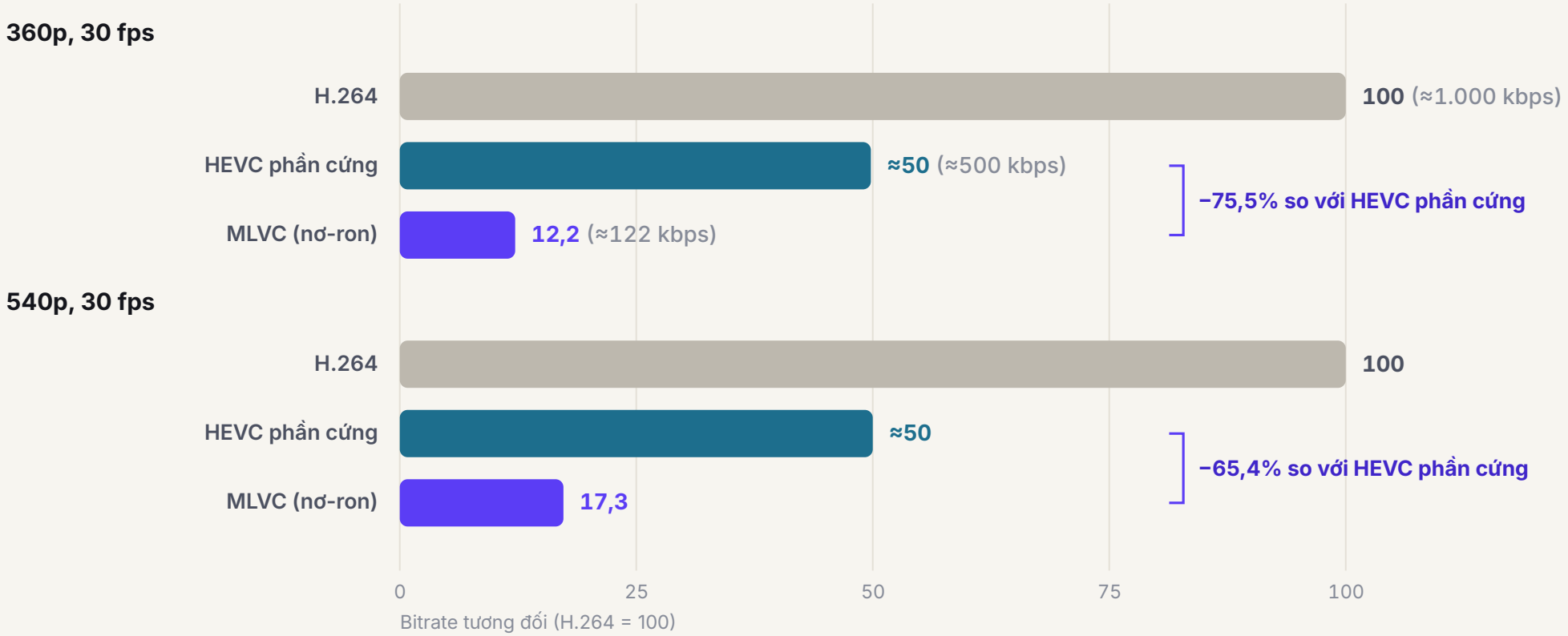
  Ban hành năm 2025: hơn VVC nội khung 12,5–27,9%, với 7 đến 216 nghìn MAC mỗi pixel. Tôi không thấy sản phẩm nào dùng nó mặc định.

- **Web**

  Chrome 155 bật JPEG XL mặc định từ ngày 6/10/2026, với bộ giải mã an toàn bộ nhớ viết bằng Rust.

> **Định dạng phải mở được ở mọi nơi vẫn là truyền thống; mã hóa nơ-ron thắng thế bên trong quy trình của một công ty.**

Cuộc gọi là nơi codec nơ-ron đầu cuối được triển khai



Satin chạy trong mọi cuộc gọi hai người trên Teams và Skype từ năm 2021. MLVC, mã nguồn mở theo giấy phép MIT từ ngày 24/7/2026, đang triển khai trong cuộc gọi ngang hàng trên Teams, tối đa 540p30, dùng không quá một nửa NPU **theo nhà cung cấp**.

Không phải nền tảng nào cũng chọn nơ-ron, và lý do rất đáng học hỏi

TRONG TIÊU CHUẨN, TÙY CHỌN

Opus 1.5 và 1.6

Tăng cường nơ-ron, che giấu mất gói, dư thừa sâu và mở rộng băng thông, tất cả đều tắt mặc định và tương thích hoàn toàn với tiêu chuẩn năm 2012.

TRUYỀN THỐNG, CÓ CHỦ ĐÍCH

Meta MLow (2024)

Hơn 20% cuộc gọi của Meta chạy trên điện thoại ARMv7, và mỗi ngày có hàng chục triệu cuộc gọi WhatsApp trên điện thoại hơn mười năm tuổi. MLow cho chất lượng gấp khoảng đôi Opus ở 6 kbps, với ít hơn 10% tính toán.

Ngân sách độ phức tạp của lượng thiết bị đang dùng là yếu tố quyết định. Codec nơ-ron đi trước ở nơi thiết bị còn mới và một nhà cung cấp kiểm soát cả hai đầu.

Nơi triển khai lớn nhất mà không ai gọi là codec

- **Codec học máy ở khắp nơi**

  Mọi mô hình giọng nói và video đều đọc và ghi token hoặc biến tiềm ẩn của một codec học máy.

- **Không phải định dạng phân phối**

  Mỗi mô hình có không gian tiềm ẩn riêng: không kiểm soát bitrate, không kiểm tra tuân thủ, không bộ giải mã thứ hai.

- **Một dòng chi phí**

  Với ai cung cấp video tạo sinh, biểu diễn tiềm ẩn nén mạnh hơn sẽ giảm chi phí tạo sinh nhiều hơn mức tương ứng phân tích.

HAI VÍ DỤ

12,5 Hz · 1,1 kbps

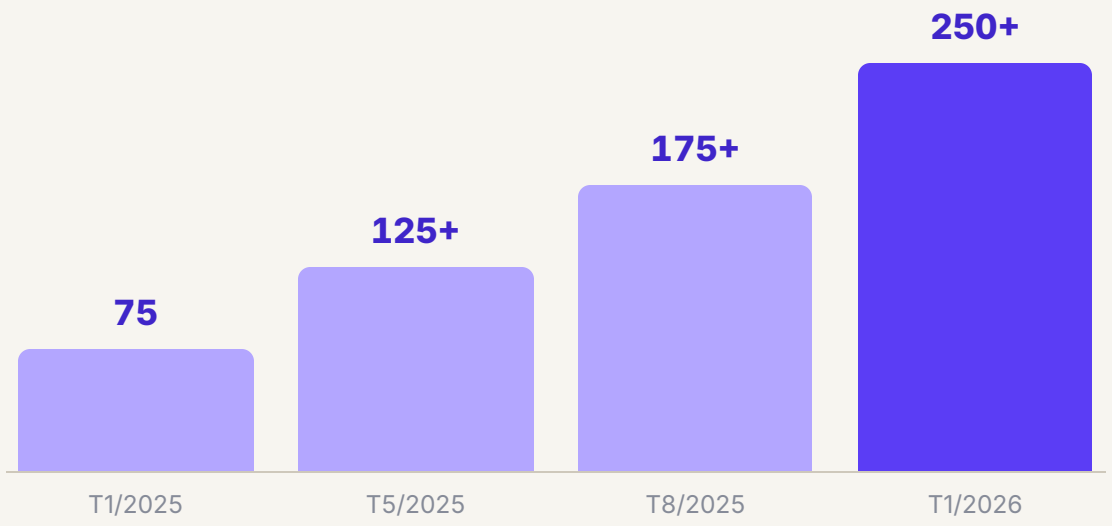
token giọng nói Mimi của Kyutai, thấp hơn mọi codec điện thoại

1:192

VAE video của LTX-Video: một token tiềm ẩn cho 32×32 pixel qua 8 khung hình

Tái tạo nơ-ron đã phổ biến; tài nguyên nơ-ron mới bắt đầu

DLSS 4 Multi Frame Generation: trò chơi và ứng dụng



Nén texture nơ-ron

Cảnh demo của NVIDIA, GTC 2026
Bộ nhớ texture 6,5 GB → 0,97 GB

Giảm 85% bộ nhớ

Assassin's Creed Mirage, đã phát hành
Chỉ một số tài nguyên; so với định dạng khối

Nén tốt hơn ≈30%

Thước đo khác nhau: không so sánh trực tiếp được

DirectX Linear Algebra ra bản xem trước công khai vào tháng 4/2026: một đường mã duy nhất để giải mã nơ-ron trong shader trên GPU của mọi hãng.

Vệ tinh, drone và robot: phù hợp nhất, thị trường nhỏ

- **Vì sao phù hợp**

  Một chủ sở hữu kiểm soát cả hai đầu, băng thông khan hiếm, và nội dung đủ hẹp để huấn luyện riêng.

- **Trung Quốc**

  Thông báo số 49 năm 2026, hiệu lực từ ngày 18/11, sửa đổi quy định kiểm soát xuất khẩu drone; một mục dùng nén ảnh ngữ nghĩa làm ngưỡng.

- **Mỹ**

  Ngày 22/12/2025, FCC đưa drone sản xuất ở nước ngoài và linh kiện quan trọng của chúng vào Covered List.

> **Một ngách vững cho doanh nghiệp chuyên biệt có năng lực tuân thủ xuất khẩu, không phải thị trường đại chúng.**

Năng lượng mỗi pixel cao hơn một bậc độ lớn

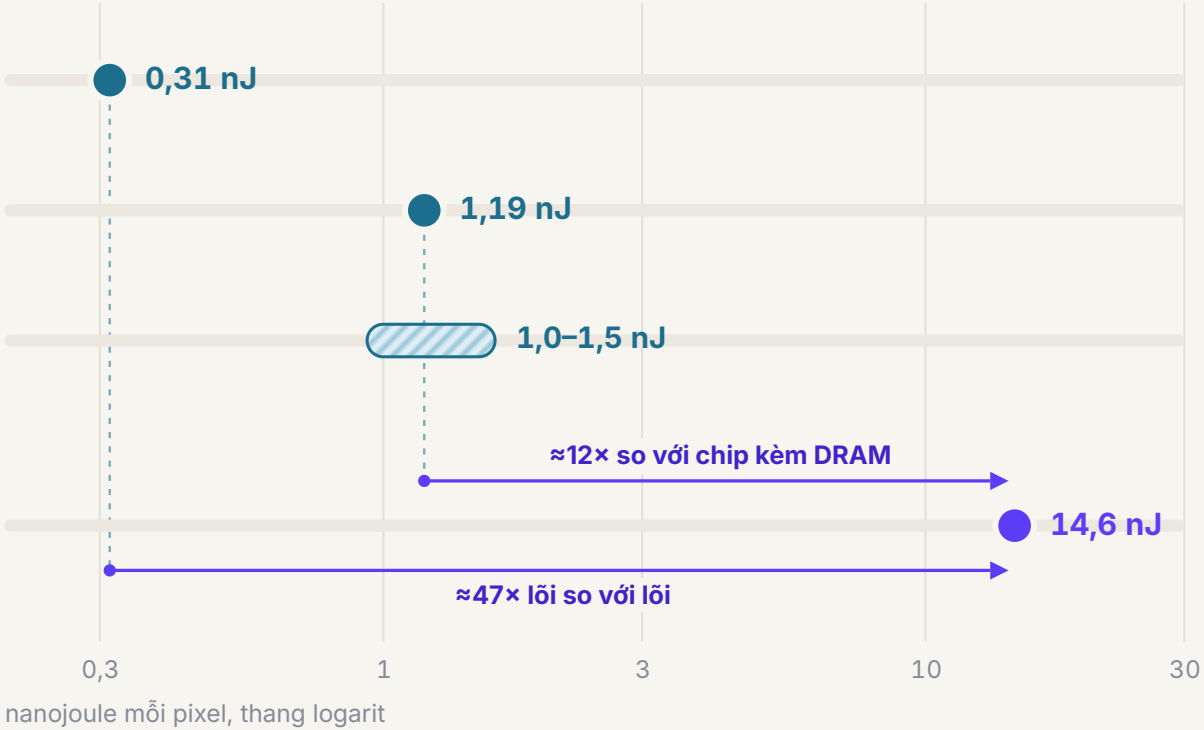
Chip giải mã HEVC, lõi
40 nm, 4K30, đo thực tế

Chip giải mã HEVC, kèm DRAM
40 nm, 4K30, lõi đo thực tế cộng DRAM mô phỏng

Bộ giải mã VVC
28 nm, 4K60, dự phóng

Bộ giải mã video nơ-ron (NVCA)
28 nm, 1080p25, chỉ tổng hợp; chưa tính trọng số và DRAM

Ở 1080p30: khoảng 74 mW so với 908 mW.



Ở 1080p30: khoảng 74 mW cho chip HEVC kèm DRAM, so với chừng 900 mW cho NVCA. Chip HEVC dùng tiến trình cũ hơn, NVCA chưa tính trọng số và DRAM: phép so sánh này có lợi cho thiết kế nơ-ron.

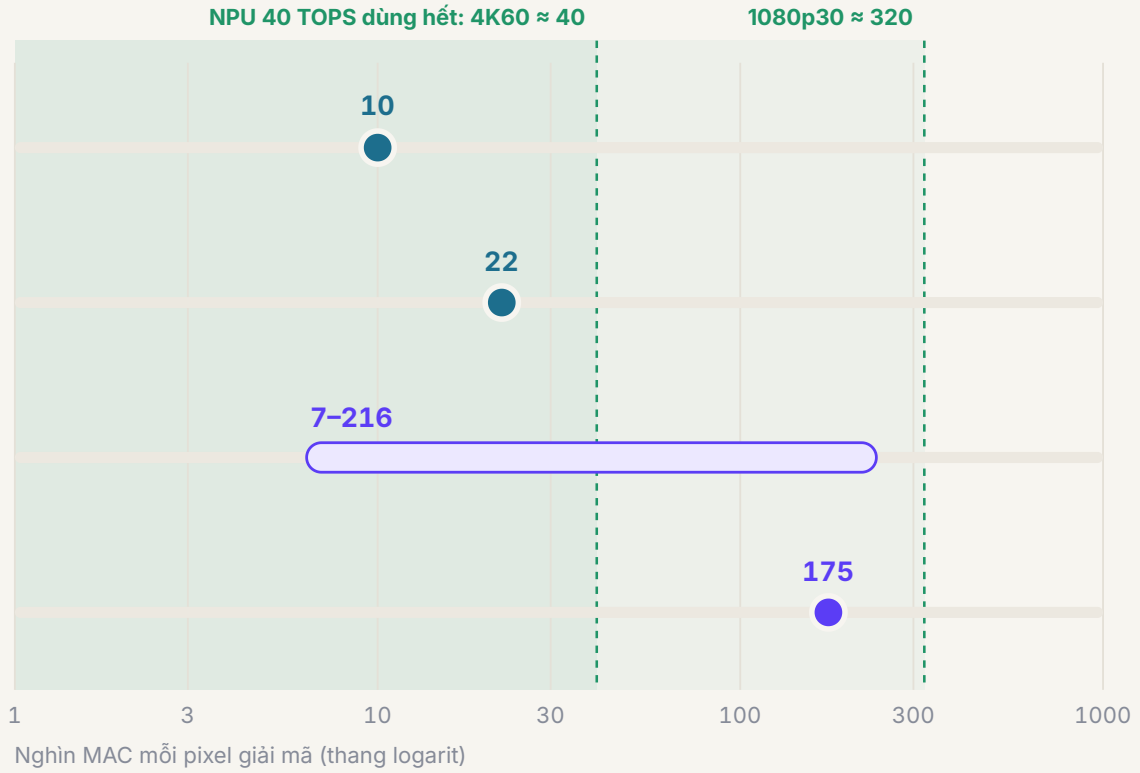
Bộ giải mã nơ-ron đòi hỏi gì ở NPU

Bộ lọc nơ-ron trong vòng lặp, độ phức tạp rất thấp
JVET thăm dò · tiết kiệm ≈6% bit
1,2 TOPS ở 1080p30 · 10 ở 4K60

Bộ lọc nơ-ron trong vòng lặp, độ phức tạp thấp
JVET thăm dò · ≈8%
2,7 TOPS ở 1080p30 · 22 ở 4K60

Bộ giải mã JPEG AI, từ cơ bản đến cao
12,5–27,9% so với VVC nội khung

DCVC-RT, cả một bộ giải mã video nơ-ron
21% so với tham chiếu VVC, thiết lập bài báo
22 TOPS ở 1080p30 · 174 ở 4K60



DCVC-RT cần khoảng 175 nghìn MAC mỗi pixel: khoảng 22 TOPS ở 1080p30 và 174 ở 4K60, khi dùng hết công suất **ước tính**. Một NPU 40 TOPS dùng hết công suất đáp ứng được khoảng 40 nghìn ở 4K60.

Lưu lượng bộ nhớ và độ chính xác từng bit là giới hạn cứng

DRAM CHIẾM PHẦN LỚN

Bộ nhớ trước MAC

Trong chip HEVC của MIT, DRAM chiếm khoảng ba phần tư năng lượng hệ thống: 0,88 trên 1,19 nJ mỗi pixel. Bộ giải mã nơ-ron mang bản đồ đặc trưng lớn giữa các khung hình, và nhóm tác giả DCVC-RT thấy chính I/O bộ nhớ, chứ không phải tính toán, giới hạn tốc độ.

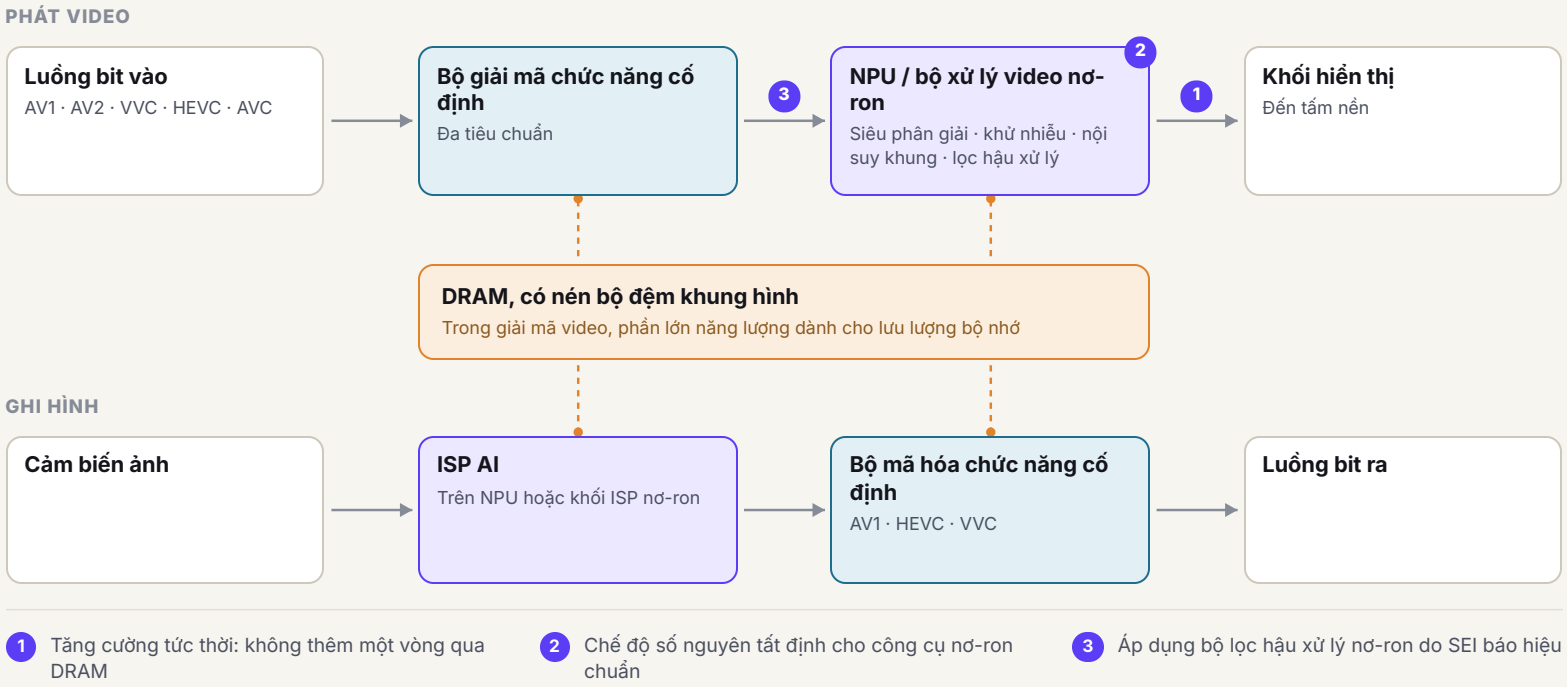
ĐỘ CHÍNH XÁC

Mọi bộ giải mã phải tính ra cùng các bit

Giải mã entropy thất bại nếu hai thiết bị tính một xác suất khác nhau. Cách khắc phục: mô hình số nguyên (DCVC-RT), hyperprior dùng chung thang đo (MLVC), suy luận số nguyên 16 bit (SADL của JVET), điểm tuân thủ (JPEG AI).

Với đội thiết kế chip, đây lại là kỷ luật biến đổi số nguyên của HEVC: tiêu chuẩn phải định nghĩa chính xác phép tính.

Phân chia đã vào sản phẩm: codec chức năng cố định cộng NPU



> Canghai của Tencent: hơn 100.000 chip thế hệ đầu đã triển khai; thế hệ thứ hai bổ sung H.266 và siêu phân giải trên chip. Chips&Media và Allegro DVT bán song song IP codec và IP video nơ-ron.

Tôi sẽ xây dựng gì, theo từng vai trò

Đội SoC và chip TV

- Bộ giải mã chức năng cố định cho mọi tiêu chuẩn phải phát, AV2 ngay từ bây giờ
- NPU đủ sức tăng cường ở độ phân giải hiển thị, không thêm vòng qua DRAM
- SEI cho bộ lọc hậu xử lý nơ-ron và chế độ số nguyên tất định
- H.267 là điểm khởi đầu thiết kế từ năm 2029 trở đi

Nhà cung cấp IP

- Bộ giải mã AV2 và VVC ngay bây giờ
- Bộ tăng tốc lọc hậu xử lý và nén bộ đệm khung hình
- Công cụ kiểm tra tuân thủ cho xử lý nơ-ron

Nền tảng video

- Tiền xử lý nơ-ron ở bộ mã hóa hoàn vốn trước tiên
- Siêu phân giải phía máy khách ở nơi thiết bị cho phép
- Codec nơ-ron đầu cuối chỉ trong vòng khép kín do bạn kiểm soát

Nơi công cụ nơ-ron trở thành đối tượng cấp phép

Tiêu chuẩn	Thành phần nơ-ron	Tình trạng, tháng 11/2026
H.267 (tên tạm)	Còn mở: cả công cụ truyền thống (ECM) và nơ-ron đều là ứng viên	Đánh giá hồ sơ tháng 1/2027; phiên bản 1 cuối năm 2029
VVC và HEVC, với VSEI	Bộ lọc hậu xử lý nơ-ron mô tả trong SEI; tùy chọn với bộ giải mã	Đang có hiệu lực; mở rộng vào tháng 10/2025
AV2	Truyền thống, miễn phí bản quyền	Phiên bản 1.0.0 ngày 28/5/2026; YouTube dự kiến phát trên web đầu năm 2027
JPEG AI	Codec ảnh học máy đầu cuối	Phần 1 ban hành năm 2025; chưa thấy sản phẩm hỗ trợ mặc định
IEEE 1857.11	Mã hóa ảnh nơ-ron thiết kế cho NPU	Ban hành tháng 12/2024
Video Coding for Machines	Mã hóa cho máy phân tích thay vì để xem	Dự thảo tiêu chuẩn quốc tế đầu năm 2026
Phần mở rộng ML của Opus	Tùy chọn và tương thích ngược	Có trong libopus 1.5 và 1.6

645

họ bằng sáng chế thế hệ sau VVC do ByteDance nắm giữ (GreyB, 2025)

0,50 · 1,00–1,50 USD

phí nhóm bằng sáng chế VVC cho mỗi điện thoại và mỗi TV 4K, Khu vực 1

0 USD

phí bản quyền cho AV1 và AV2 từ các thành viên AOMedia

Tiền đi theo các tiêu chuẩn và nền tảng

ĐANG CÓ DOANH THU

Nền tảng

Giữ lại phần tiết kiệm

Bộ giảm độ phân giải nơ-ron của Netflix; siêu phân giải của YouTube và Meta; Microsoft mở mã nguồn MLVC theo giấy phép MIT

ĐANG CÓ DOANH THU

Hãng thiết bị và chip

Bán sự khác biệt

ISP camera; chip nâng độ phân giải cho TV; chip codec Canghai tự phát triển của Tencent

ĐANG CÓ DOANH THU

Nhà cung cấp IP

Cấp phép mỗi tiêu chuẩn mới

Chips&Media, Allegro DVT và VeriSilicon có khách hàng IP giải mã AV2 trong vòng ba tuần sau khi có đặc tả

ĐANG CÓ DOANH THU

Hãng GPU

Khóa chân bằng nền tảng

DLSS 4 trong hơn 250 trò chơi và ứng dụng; SDK texture nơ-ron ở bản beta

THEO THIẾT BỊ

Chủ sở hữu và nhóm bằng sáng chế

Thu phí theo thiết bị

VVC: 0,50 USD mỗi điện thoại, 1,00–1,50 USD mỗi TV 4K; công cụ H.267 sau năm 2029

THOÁI VỐN QUA MUA LẠI

Startup codec nơ-ron

Được mua lại

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

Ba startup codec nơ-ron nổi tiếng nhất, WaveOne, iSIZE và Deep Render, đã được Apple, Sony và InterDigital mua lại.

Sáu hướng đi, và rủi ro của từng hướng

Hướng đi	Ai mua	Vì sao là lúc này	Rủi ro chính
Tiền xử lý nơ-ron ở bộ mã hóa	Nền tảng phát trực tuyến và video người dùng, đài truyền hình	Không cần đổi thiết bị; đã chứng minh ở Netflix	Các nền tảng lớn tự xây dựng
IP và tinh chỉnh tăng cường phía hiển thị	Hãng chip cho TV, đầu thu và điện thoại	NPU đã có sẵn trong thiết bị	Bị hàng hóa hóa; đội ngũ nội bộ
Codec nơ-ron vòng khép kín	Hãng giải pháp liên lạc doanh nghiệp, công nghiệp và hiện diện từ xa	MLVC chứng minh là khả thi, và mã nguồn đã mở	Phải kiểm soát cả hai đầu; NPU phân mảnh
Chip hoặc IP codec cộng NPU	Hãng SoC	Chu kỳ AV2 đang bắt đầu	Các hãng đương nhiệm mạnh
Đánh giá, kiểm tra tuân thủ và suy luận tất định	Bất kỳ ai triển khai công cụ nơ-ron	Chính xác đến từng bit là yêu cầu bắt buộc	Thị trường nhỏ nếu không bán kèm
Đóng góp tiêu chuẩn cho H.267	Bên cấp phép, nền tảng lớn	Hồ sơ phản hồi được đánh giá vào tháng 1/2027	Lâu dài, tốn kém, bất định

NHỮNG TUYÊN BỐ BẠN SẼ NGHE

Hãy dè dặt với tuyên bố chưa nêu đủ bốn điều

Tuyên bố	Bằng chứng cho thấy gì
"Codec nơ-ron đã vượt VVC."	Trong phòng thí nghiệm. DCVC-RT tiết kiệm 21% so với bộ mã hóa tham chiếu của VVC; bộ giải mã của nó tốn khoảng 175 nghìn MAC mỗi pixel.
"Codec AI sẽ thay thế H.26x và AV1 trong vài năm tới."	Không phải trong phát video: H.267 phiên bản 1 cuối năm 2029, chip khoảng năm 2031 ước tính. Codec đang đến, AV2, là codec truyền thống.
"MLVC giảm 88% băng thông."	So với H.264 ở 360p. So với HEVC phần cứng: 75,5% ở 360p và 65,4% ở 540p, trên bộ thử nghiệm của chính Microsoft.
"Texture nơ-ron giảm 85–96% bộ nhớ video."	Trong cảnh demo. Trò chơi đã phát hành cải thiện khoảng 30%, chỉ trên tài nguyên chọn lọc.
"JPEG AI là tiêu chuẩn ISO, nên camera và trình duyệt sẽ áp dụng."	Chưa thấy sản phẩm nào hỗ trợ mặc định. Web đã chọn JPEG XL, một định dạng truyền thống.

Hãy hỏi mốc so sánh, thước đo, độ phân giải và chi phí giải mã. Tuyên bố nào không nêu được đủ bốn điều thì là kết quả nghiên cứu, không phải sản phẩm.

CẦN THEO DÕI GÌ

Những gì cần theo dõi, và các câu hỏi mở

CÁC MỐC

Tháng 1/2027	Đánh giá hồ sơ H.267: bao nhiêu hồ sơ dựa vào công cụ nơ-ron, với chi phí mỗi pixel bao nhiêu
Đầu năm 2027	YouTube phát AV2 trên web lần đầu; bộ giải mã AV2 trong các SoC đã công bố
Nửa cuối năm 2027	Mô hình thử nghiệm đầu tiên của H.267
2027	MLVC ra ngoài cuộc gọi ngang hàng trên Teams
2027–2028	Sản phẩm đầu tiên dùng JPEG AI mặc định
Cuối năm 2029	H.267 phiên bản 1 khởi động chu kỳ chip

CÂU HỎI MỞ

01 Độ chính xác ở quy mô lớn: giải mã số nguyên chính xác từng bit trên NPU của các hãng, mất bao nhiêu khả năng nén?

02 Bộ nhớ trước MAC: lưu lượng DRAM mỗi pixel của hậu xử lý nơ-ron

03 Chi phí mỗi pixel mà tại đó một công cụ nơ-ron hết đáng giá trong H.267

04 Vòng khép kín nào ngoài cuộc gọi sẽ áp dụng codec nơ-ron đầu cuối trước

05 Đo độ trung thực thế nào khi bộ giải mã tự tạo chi tiết

Đây là những lĩnh vực chúng ta có thể cùng khám phá, qua trao đổi kỹ thuật và hợp tác kỹ thuật.

i@tanshuai.com · tanshuai.com