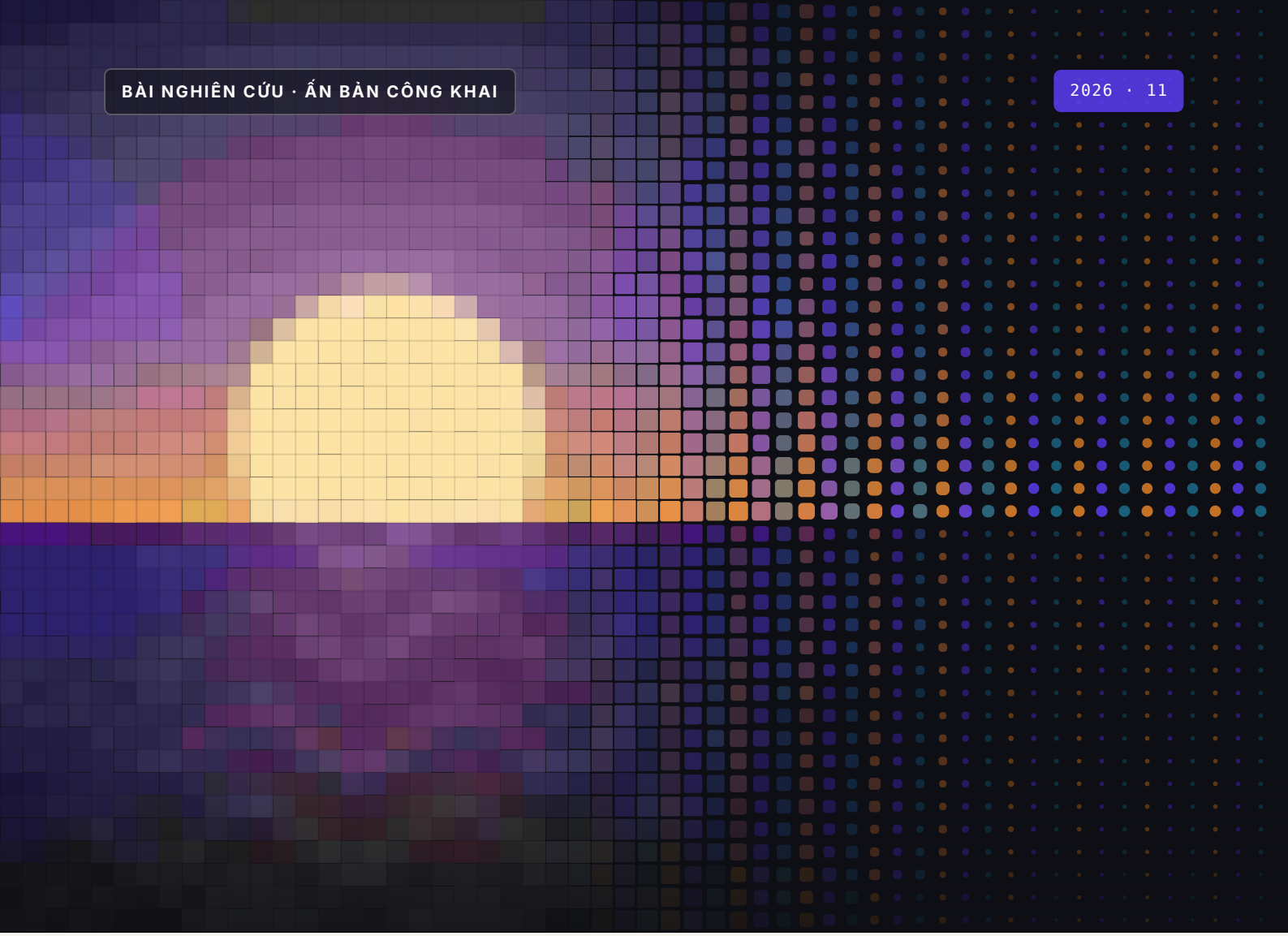


BÀI NGHIÊN CỨU · ẤN BẢN CÔNG KHAI

2026 · 11

NÉN BẰNG HỌC MÁY · PHÁT VIDEO × CHIP

Mã hóa nơ-ron năm 2026

Triển khai quanh codec, chứ không thay cho codec

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

KẾT LUẬN

Mạng nơ-ron đã hoạt động ở cả hai đầu của các codec đang truyền tải video của thế giới. Codec nơ-ron đầu cuối chỉ được triển khai trong các vòng khép kín, còn màn hình đại chúng phải chờ H.267.

Tan Shuai

Chuyên gia công nghệ và nhà sáng lập · Nhà nghiên cứu và phát triển AI

tanshuai.com · i@tanshuai.com

Số liệu tính đến ngày 12/11/2026

Số tiền tính bằng đô la Mỹ

MỤC LỤC

Mục lục

◆ **Tóm tắt tổng quan** — Đã phổ biến, nhưng nằm quanh codec — 04

01 **Phạm vi và phương pháp** — "Mã hóa nơ-ron" ở đây nghĩa là gì, và tôi đã tìm hiểu ra sao — 07

02 **Phát video và phát trực tuyến** — Phát video và phát trực tuyến: nơ-ron ở hai đầu, chuẩn ở giữa — 08

03 **Nhiếp ảnh và hình ảnh** — Nhiếp ảnh và hình ảnh: nơ-ron trong camera, truyền thống trên web — 11

04 **Cuộc gọi và họp trực tuyến** — Cuộc gọi và họp trực tuyến: nơi codec nơ-ron đầu cuối thực sự được triển khai — 12

05 **AI tạo sinh** — AI tạo sinh: nơi triển khai lớn nhất mà không ai gọi là codec — 14

06 **Trò chơi và đồ họa** — Trò chơi và đồ họa: tái tạo nơ-ron đã phổ biến, tài nguyên nơ-ron mới bắt đầu — 15

07 **Liên kết chuyên dụng** — Liên kết chuyên dụng: vệ tinh, drone và robot — 17

08 **Chip** — Góc nhìn từ chip: năng lượng, tính toán, bộ nhớ và độ chính xác — 17

09 **Tiêu chuẩn và bằng sáng chế** — Tiêu chuẩn và bằng sáng chế: nơi công cụ nơ-ron trở thành đối tượng cấp phép — 21

10 **Ai kiếm được tiền** — Ai kiếm được tiền, và doanh nghiệp có thể tham gia ở đâu — 23

11 **Tuyên bố và bằng chứng** — Những tuyên bố bạn sẽ nghe, và bằng chứng nói gì — 25

12 **Các mốc cần theo dõi** — Những gì cần theo dõi, 2027 đến 2030 — 26

13 **Câu hỏi mở** — 27

◆ **Về tác giả** — Tan Shuai — 28

◆ **Nguồn tham khảo** — 29

DANH MỤC HÌNH

Danh mục hình

HÌNH 01	Mã hóa nơ-ron được triển khai ở đâu năm 2026	06
HÌNH 02	Ba cách mạng nơ-ron đi vào codec	07
HÌNH 03	Từ tiêu chuẩn hoàn tất đến màn hình đại chúng mất gần trọn một thập kỷ	10
HÌNH 04	Trong cuộc gọi, codec nơ-ron vượt HEVC phần cứng 65–76%	13
HÌNH 05	Khi một công ty kiểm soát cả hệ thống, tái tạo bằng mạng nơ-ron lan nhanh	15
HÌNH 06	Năng lượng mỗi pixel giải mã: bộ giải mã chức năng cố định dẫn trước một bậc độ lớn	18
HÌNH 07	Một công cụ nơ-ron tốn hàng chục nghìn MAC mỗi pixel; cả một bộ giải mã nơ-ron tốn hàng trăm nghìn	19
HÌNH 08	Phân chia đã vào sản phẩm: phần cứng chức năng cố định cho luồng bit, khối xử lý nơ-ron cho pixel	20
HÌNH 09	Ai kiếm được tiền từ mã hóa nơ-ron năm 2026	23

Nhãn trong văn bản

- **ước tính** — phép tính hoặc suy luận của tác giả từ dữ liệu được trích dẫn
- **theo nhà cung cấp** — do chính công ty bán sản phẩm công bố, chưa được kiểm chứng độc lập
- **phân tích** — cách tác giả diễn giải bằng chứng

Đây là bản dịch từ ấn bản tiếng Anh; nếu hai bản có chỗ khác nhau, ấn bản tiếng Anh được ưu tiên.

Số liệu tính đến ngày 12/11/2026 · Số tiền tính bằng đô la Mỹ

TÓM TẮT TỔNG QUAN

Đã phổ biến, nhưng nằm quanh codec

KẾT LUẬN

Năm 2026, mã hóa nơ-ron đã trở nên phổ biến, nhưng nó hoạt động quanh các codec chuẩn, chứ không thay cho chúng.

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

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

≈12×
năng lượng mỗi pixel so với một chip giải mã HEVC đã đo (tính cả DRAM), ở thiết kế chip giải mã nơ-ron duy nhất đã công bố

≈22 TOPS
năng lực tính toán liên tục cho một bộ giải mã nơ-ron cỡ DCVC-RT ở 1080p30
ước tính

75,5%
bitrate ít hơn HEVC phần cứng: MLVC ở 360p trong cuộc gọi, trên bộ thử nghiệm của chính Microsoft
theo nhà cung cấp

Cuối năm 2029
H.267 phiên bản 1, tiêu chuẩn đầu tiên có khả năng đưa công cụ nơ-ron đến mọi màn hình

Triển khai ở đâu

Mạng nơ-ron hiện chạy ở cả hai đầu của những codec đang truyền tải gần như toàn bộ video trên thế giới. Ở chiều vào, chúng định hình hình ảnh: bộ xử lý tín hiệu hình ảnh AI trong camera điện thoại, và một bộ giảm độ phân giải nơ-ron trong quy trình mã hóa của Netflix. Ở chiều ra, chúng sửa lại hình ảnh: Super Resolution của YouTube trên TV, siêu phân giải tại Meta, và các bộ xử lý TV quảng cáo hàng trăm mạng nơ-ron. Ở giữa, luồng bit vẫn là AV1, HEVC, AVC hoặc VVC, do phần cứng chức năng cố định giải mã. Không dịch vụ phát trực tuyến nào gửi định dạng video nơ-ron đến TV hay điện thoại; ở đâu có bộ giải mã video nơ-ron đang chạy trên thiết bị hôm nay, nó chạy bên trong ứng dụng riêng của một công ty.

Vòng khép kín

Codec nơ-ron đầu cuối, trong đó cả bộ mã hóa lẫn bộ giải mã đều được huấn luyện, chỉ được triển khai trong các vòng khép kín, nơi một công ty kiểm soát cả hai đầu và có thể chuyển về một codec chuẩn. Microsoft đã chạy một codec âm thanh AI trong mọi cuộc gọi hai người trên Teams và Skype từ năm 2021, và vào tháng 7/2026 đã mở mã nguồn MLVC, một codec video nơ-ron đang được triển khai trong các cuộc gọi ngang hàng trên Teams: ở 360p, nó cần ít hơn khoảng 75% bitrate so với HEVC phần cứng để đạt cùng mức chất lượng được chấm điểm. Mọi mô hình tạo giọng nói và video đều chạy trên một codec học máy, từ các bộ token hóa giọng nói 12,5 Hz đến các VAE video có tỷ lệ nén 1:192. Nén texture nơ-ron đã có mặt trong một trò chơi AAA. Đây là những đợt triển khai thực sự, và chúng có chung một điều kiện: không ai khác phải giải mã luồng dữ liệu.

Thời điểm

Phát video đại chúng bằng luồng bit nơ-ron phải chờ một tiêu chuẩn và một chu kỳ chip. Tiêu chuẩn đầu tiên có khả năng mang công cụ nơ-ron, thế hệ kế nhiệm VVC mà tôi sẽ gọi là H.267, đã phát hành lời kêu gọi đề xuất vào tháng 7/2026, sẽ đánh giá các hồ sơ phản hồi vào tháng

1/2027 và đặt mục tiêu ra phiên bản đầu tiên vào cuối năm 2029. AV1 cần gần tám năm từ khi hoàn tất đặc tả đến lúc chiếm 30% lượng xem của Netflix. Theo khuôn mẫu đó, những con chip H.267 đầu tiên sẽ xuất hiện vào khoảng năm 2031 và phổ cập vào giữa thập niên 2030 ước tính. Trong khi đó, codec sẽ đến với màn hình tiếp theo là AV2, hoàn tất vào tháng 5/2026 và không có bộ giải mã nơ-ron nào bên trong, và ba nhà cung cấp IP đã công bố khách hàng cho bộ giải mã AV2 trong vòng ba tuần.

Chip

Lý do nằm ở chip. Thiết kế chip giải mã video nơ-ron duy nhất đã công bố cần 14,6 nJ mỗi pixel, gấp khoảng 12 lần một chip giải mã HEVC đã đo khi tính cả DRAM, và khoảng 47 lần riêng phần lõi của chip đó. Codec nghiên cứu thời gian thực DCVC-RT của Microsoft, theo cách tôi cộng từ chính bảng số liệu theo từng mô-đun của nó, cần khoảng 175 nghìn phép nhân-cộng dồn (MAC) cho mỗi pixel giải mã, tức chừng 22 TOPS năng lực tính toán liên tục cho 1080p30 ước tính: khoảng một nửa trong số 40 TOPS dùng để định nghĩa một PC AI, dành cho một luồng video. Bộ giải mã nơ-ron còn phải cho ra kết quả giống hệt nhau đến từng bit trên mọi thiết bị, nếu không phần giải mã entropy sẽ đổ vỡ. Đó là lý do MLVC đánh đổi khả năng nén để giải mã giống hệt nhau trên các NPU, và chạy ở 540p chứ không phải 4K.

Kinh doanh

Với doanh nghiệp, tiền đi theo các tiêu chuẩn và nền tảng, chứ không theo các codec nơ-ron độc lập. Các nền tảng giữ lại phần tiết kiệm từ tiền xử lý và hậu xử lý nơ-ron. Nhà cung cấp IP cấp phép bộ giải mã chức năng cố định cho mỗi tiêu chuẩn mới, kèm NPU cho xử lý ảnh. Các nhóm bằng sáng chế thu phí theo thiết bị: 0,50 USD mỗi điện thoại và 1,00–1,50 USD mỗi TV 4K cho VVC tại Mỹ và các thị trường Khu vực 1 khác. Ba startup codec nơ-ron nổi tiếng nhất, WaveOne, iSIZE và Deep Render, đã được Apple, Sony và InterDigital mua lại. Các cơ hội còn để ngỏ nằm ở tiền xử lý và hậu xử lý nơ-ron, ở các sản phẩm vòng khép kín, ở những con chip kết hợp codec chức năng cố định với NPU, và ở công việc đánh giá và tiêu chuẩn nằm giữa các mảng đó.

Câu hỏi	Nhận định	Số liệu chính (nguồn trong các chương)
① **Mã hóa nơ-ron hiện được triển khai ở quy mô lớn ở đâu?**	Quanh codec: ISP camera, giảm độ phân giải phía mã hóa, siêu phân giải khi phát. Đầu cuối thì chỉ trong vòng khép kín: cuộc gọi, tài nguyên trò chơi, bộ token hóa của AI tạo sinh	Video chiếm ≈75% dữ liệu di động; 1,26 tỷ điện thoại và ≈200 triệu TV xuất xưởng năm 2025, tất cả đều giải mã bằng phần cứng chức năng cố định; DLSS 4 có trong hơn 250 trò chơi và ứng dụng
② **Codec nơ-ron có thay thế AV1, HEVC và VVC trong phát video không? Khi nào?**	Không trước khi có một tiêu chuẩn và một chu kỳ thiết bị. H.267 là phương tiện thực tế đầu tiên, và nhiều khả năng sẽ kết hợp công cụ truyền thống với công cụ nơ-ron	H.267 kêu gọi đề xuất tháng 7/2026, đánh giá tháng 1/2027, phiên bản 1 cuối năm 2029; phổ cập giữa thập niên 2030 ước tính
③ **Chi phí trên chip là bao nhiêu?**	Năng lượng mỗi pixel cao hơn một bậc độ lớn so với giải mã bằng phần cứng chức năng cố định hiện nay, cộng thêm lưu lượng bộ nhớ và yêu cầu chính xác đến từng bit	14,6 nJ/pixel (nơ-ron) so với 1,19 (HEVC, kèm DRAM); ≈175 kMAC mỗi pixel giải mã ≈ 22 TOPS ở 1080p30 ước tính
④ **Tiêu chuẩn và bằng sáng chế cho phép những gì?**	Bộ lọc hậu xử lý nơ-ron đã có thể được báo hiệu trong luồng VVC; JPEG AI là tiêu chuẩn ISO; H.267 sẽ quyết định công cụ nơ-ron nào trở thành đối tượng cấp phép	ECM (truyền thống) ≈28% so với VVC; bộ lọc nơ-ron trong vòng lặp 6–14% ở mức 10–22+ kMAC/pixel; ByteDance có 645 họ bằng sáng chế trong công việc thế hệ sau VVC
⑤ **Ai kiếm được tiền, và doanh nghiệp có thể tham gia ở đâu?**	Nền tảng, nhà cung cấp IP, chủ sở hữu bằng sáng chế. Các startup thuần codec nơ-ron đã được mua lại. Tham gia ở tiền/hậu xử lý, vòng khép kín, chip codec cộng NPU, đánh giá	Nhóm bằng sáng chế VVC thu 0,50 USD mỗi điện thoại; IP giải mã AV2 được cấp phép vài tuần sau khi có đặc tả; ba thương vụ mua lại, không công bố điều khoản

Câu hỏi	Nhận định	Số liệu chính (nguồn trong các chương)
⑥ **Những tuyên bố nào cần dè dặt?**	Mức cải thiện tính so với codec cũ, ở độ phân giải thấp, không kèm chi phí giải mã, hoặc từ bản demo thay vì sản phẩm đã phát hành	MLVC: ít hơn H.264 88% nhưng ít hơn HEVC phần cứng 65–76%; texture: tiết kiệm 85% VRAM trong demo, khoảng 30% trong trò chơi đã phát hành

HÌNH 01

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

Mạng nơ-ron đã chạy ở cả hai đầu của các codec chủ đạo. Codec nơ-ron đầu cuối chỉ được triển khai ở nơi một công ty kiểm soát cả bộ mã hóa lẫn bộ giải mã.

	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	—

Nguồn: Tài liệu của các công ty, tổ chức tiêu chuẩn và nhà phát triển được trích dẫn trong báo cáo; tình trạng tính đến ngày 12/11/2026. Các hàng được sắp theo số người mà mỗi lĩnh vực tiếp cận.

PHẠM VI VÀ PHƯƠNG PHÁP

01 “Mã hóa nơ-ron” ở đây nghĩa là gì, và tôi đã tìm hiểu ra sao

Ba cách mạng nơ-ron đi vào codec

“Mã hóa nơ-ron” bao hàm ba thứ khác nhau, và phần lớn tranh luận về nó bắt nguồn từ việc lẫn lộn chúng.

01 **Quanh codec.** Một mạng nơ-ron xử lý pixel trước khi mã hóa hoặc sau khi giải mã, còn luồng bit vẫn theo chuẩn. Ví dụ: bộ xử lý tín hiệu hình ảnh (ISP) AI trong camera, bộ giảm độ phân giải nơ-ron của Netflix đặt trước bộ mã hóa, siêu phân giải trên TV. Mọi thiết bị đều phát được luồng này; mạng nơ-ron chỉ thay đổi những gì đi vào hoặc đi ra.

02 **Trong tiêu chuẩn.** Mạng nơ-ron trở thành một công cụ trong codec chuẩn, ví dụ một bộ lọc nơ-ron trong vòng lặp hoặc bộ dự đoán nội khung mà mọi bộ giải mã phải chạy, nên nó phải được đặc tả chính xác, thường bằng số học số nguyên. Một dạng nhẹ hơn là bộ lọc hậu xử lý nơ-ron được luồng bit mô tả trong siêu dữ liệu, và bộ giải mã có thể chọn áp dụng.

03 **Thay cho codec.** Cả bộ mã hóa lẫn bộ giải mã đều được huấn luyện từ đầu đến cuối. Luồng bit là một tập giá trị tiềm ẩn đã lượng tử hóa mà chỉ mạng tương ứng mới giải mã được. Ví dụ: MLVC của Microsoft trong Teams, bộ token hóa giọng nói trong các mô hình giọng nói, texture nơ-ron trong một trò chơi.

HÌNH 02 **Ba cách mạng nơ-ron đi vào codec**

Phần lớn những gì được triển khai năm 2026 hoạt động quanh một codec chuẩn. Công cụ nơ-ron trong tiêu chuẩn đang được đặc tả. Codec nơ-ron đầu cuối chỉ thay cho codec ở nơi một công ty kiểm soát cả hai đầu.



Nguồn: Tác giả tổng hợp từ các nguồn trong báo cáo này; tình trạng tính đến ngày 12/11/2026.

Loại thứ nhất không cần ai cho phép. Loại thứ hai cần một tiêu chuẩn. Loại thứ ba cần kiểm soát cả hai đầu. Đó là thứ tự chúng đến với người dùng, và nó giải thích phần lớn những gì trình bày tiếp theo.

Sắp xếp theo thế giới thực, không theo chương trình nghiên cứu

Các bài báo nghiên cứu chủ yếu tìm hiểu loại thứ ba, vì đó là nơi có các kỷ lục nén. Báo cáo này sắp xếp các lĩnh vực theo số người mà chúng tiếp cận, vì đó là điều quyết định tiền và công sức kỹ thuật đổ vào đâu: phát video và phát trực tuyến trước tiên, sau đó là nhiếp ảnh, cuộc gọi, AI tạo sinh, trò chơi, và cuối cùng là các liên kết chuyên dụng (vệ tinh, drone, robot). Theo số liệu của Ericsson vào cuối năm 2025, video chiếm khoảng 75% toàn bộ lưu lượng dữ liệu di động. Năm 2025, ngành công nghiệp đã xuất xưởng khoảng 1,26 tỷ điện thoại thông minh (IDC) và chừng 196–209 triệu TV (ước tính của TrendForce và Omdia). Tất cả chúng đều giải mã video bằng phần cứng chức năng cố định.

Cách tôi làm việc

Đây là nghiên cứu tài liệu, được thực hiện theo cách tôi thẩm định một con chip hay một kế hoạch sản phẩm. Tôi đã đọc tài liệu tiêu chuẩn và báo cáo cuộc họp (JVET, MPEG, JPEG, AOMedia, IETF), blog kỹ thuật và ghi chú phát hành của các công ty, bài báo về chip, bảng phí của các nhóm bằng sáng chế, văn bản kiểm soát xuất khẩu và báo chí. Mọi con số đều kèm thời điểm. Khi tự suy ra một con số, chẳng hạn khối lượng tính toán mỗi pixel từ các bảng của một bài báo, tôi nói rõ và gắn nhãn **ước tính**. Những tuyên bố tôi không thể đối chiếu với nguồn gốc được gắn nhãn **theo nhà cung cấp**. Số tiền tính bằng đô la Mỹ. Tình trạng tính đến ngày 12/11/2026.

Khi có thể, tôi so sánh những thứ tương đương nhau, và nói rõ khi không thể. Mức cải thiện về nén ít có ý nghĩa nếu không kèm bốn điều: mốc so sánh (cải thiện so với cái gì), thước đo (PSNR, VMAF hay một nhóm người xem đánh giá), nội dung và độ phân giải, và chi phí giải mã. Phần lớn bất đồng về codec nơ-ron biến mất khi bốn điều đó được nêu rõ.

CHƯƠNG 02

PHÁT VIDEO VÀ PHÁT TRỰC TUYẾN

02 Phát video và phát trực tuyến: nơ-ron ở hai đầu, chuẩn ở giữa

Những gì người xem nhận được vẫn là luồng bit chuẩn

Các dịch vụ video lớn nhất đã hoàn tất một lần chuyển đổi codec và đang bắt đầu lần tiếp theo, và cả hai đều không liên quan đến bộ giải mã nơ-ron. Tháng 12/2025, Netflix cho biết AV1 chiếm khoảng 30% lượng xem của họ, là codec được dùng nhiều thứ hai sau AVC. AV1 chỉ đến được TV khi bộ giải mã phần cứng đến: Netflix đưa AV1 vào quy trình chứng nhận thiết bị năm 2019, phát nó đến TV từ cuối năm 2021, và cho biết 88% thiết bị màn hình lớn được gửi chứng nhận trong giai đoạn 2021–2025 hỗ trợ nó, gần như tất cả là từ năm 2023. YouTube cho biết tính đến ngày 19/3/2026, hơn 90% kho nội dung của họ, tính theo thời lượng xem, đã có sẵn ở AV1.

AV2, thế hệ kế nhiệm AV1 của Alliance for Open Media, đạt phiên bản 1.0.0 ngày 28/5/2026, với khả năng nén tốt hơn AV1 khoảng 30% theo các phép đo khách quan của AOMedia. YouTube dự kiến phát luồng AV2 trên web "sớm nhất là đầu năm 2027". Chip đi theo ngay lập tức: Chips&Media hoàn thành bộ giải mã phần cứng AV2 vào tháng 5 và cấp phép cho một công ty công nghệ lớn giấu tên ở Bắc Mỹ vào tháng 6; Allegro DVT đã giao IP giải mã AV2 cho ba khách hàng trước ngày 16/6; VeriSilicon đã phát hành một IP video hỗ trợ AV2 cho một số khách

hàng trước ngày 10/6. AV2 là codec truyền thống: mức cải thiện của nó đến từ phân vùng, dự đoán, phép biến đổi và bộ lọc trong vòng lặp tốt hơn (một trong số đó là bộ lọc chi tiết có hướng dẫn "dựa trên dữ liệu"), chứ không phải từ một mạng nơ-ron đầu cuối.

VVC, tiêu chuẩn ITU/ISO hiện hành, có nơi phát sóng đầu tiên tại Brazil, nơi hệ thống DTV+ (TV 3.0) đi vào vận hành thương mại vào tháng 6/2026, đúng dịp World Cup. Tính đến tháng 7/2026, chưa có TV nào bán tại Brazil tích hợp sẵn đầu thu DTV+; những chiếc đầu tiên dự kiến xuất hiện trong khoảng từ cuối năm 2026 đến đầu năm 2027.

Ở bộ mã hóa: bộ giảm độ phân giải nơ-ron của Netflix

Ví dụ rõ ràng nhất về mã hóa nơ-ron trong phát video đã chạy từ năm 2022 và không thay đổi gì những việc thiết bị của người xem phải làm. Netflix mã hóa mỗi nội dung ở nhiều độ phân giải, và bước giảm độ phân giải tạo ra các bản 1080p, 720p và 540p từ bản gốc 4K trước đây dùng một bộ lọc cố định như Lanczos. Netflix thay nó bằng một mạng nơ-ron nhỏ, gồm một khối tiền xử lý tiếp theo là một khối đổi kích thước, được huấn luyện để cải thiện chất lượng hình ảnh sau khi đi qua một bộ nâng độ phân giải chuẩn trên thiết bị. Netflix mô tả đây là giải pháp thay thế trực tiếp (drop-in): không cần thay đổi gì khác ở phía mã hóa hay phía máy khách, nên "hàng triệu thiết bị hỗ trợ phát trực tuyến Netflix tự động được hưởng lợi". Trong các thử nghiệm về mức độ ưa thích do Netflix thực hiện, khoảng 77% người xem chọn kết quả của bộ giảm độ phân giải nơ-ron **theo nhà cung cấp**. Để giữ chi phí ở mức chấp nhận được với quy mô của Netflix, "chỉ vài lớp mạng nơ-ron là đủ".

Đó là khuôn mẫu cần ghi nhớ. Mạng nơ-ron chạy một lần cho mỗi nội dung trong trung tâm dữ liệu, nơi năng lực tính toán rẻ và có thể lên lịch, và đầu ra của nó là một luồng AV1 hoặc HEVC bình thường.

Ở màn hình: siêu phân giải trong ứng dụng và trong TV

Ở đầu bên kia, mạng nơ-ron sửa lại những gì codec và nguồn đã làm mất. Ngày 29/10/2025, YouTube công bố Super Resolution cho ứng dụng TV: video tải lên dưới 1080p, từ 240p đến 720p, được nâng lên HD bằng học máy, với mục tiêu đã nêu là 4K. Các phiên bản được nâng độ phân giải có gắn nhãn, người xem có thể chuyển về bản gốc, và nhà sáng tạo có thể chọn không tham gia. Tháng 10/2025, Meta mô tả cách hãng chạy siêu phân giải video cả trên máy chủ khi tải lên, trên CPU x86 tiêu chuẩn với sự hợp tác của Intel, lẫn trên máy khách khi phát, và nhận thấy hai cách này bổ sung cho nhau. Các hãng TV làm điều tương tự bằng chip: bộ xử lý 8K hàng đầu của Samsung, NQ8 AI Gen3, quảng cáo 768 mạng nơ-ron để nâng độ phân giải.

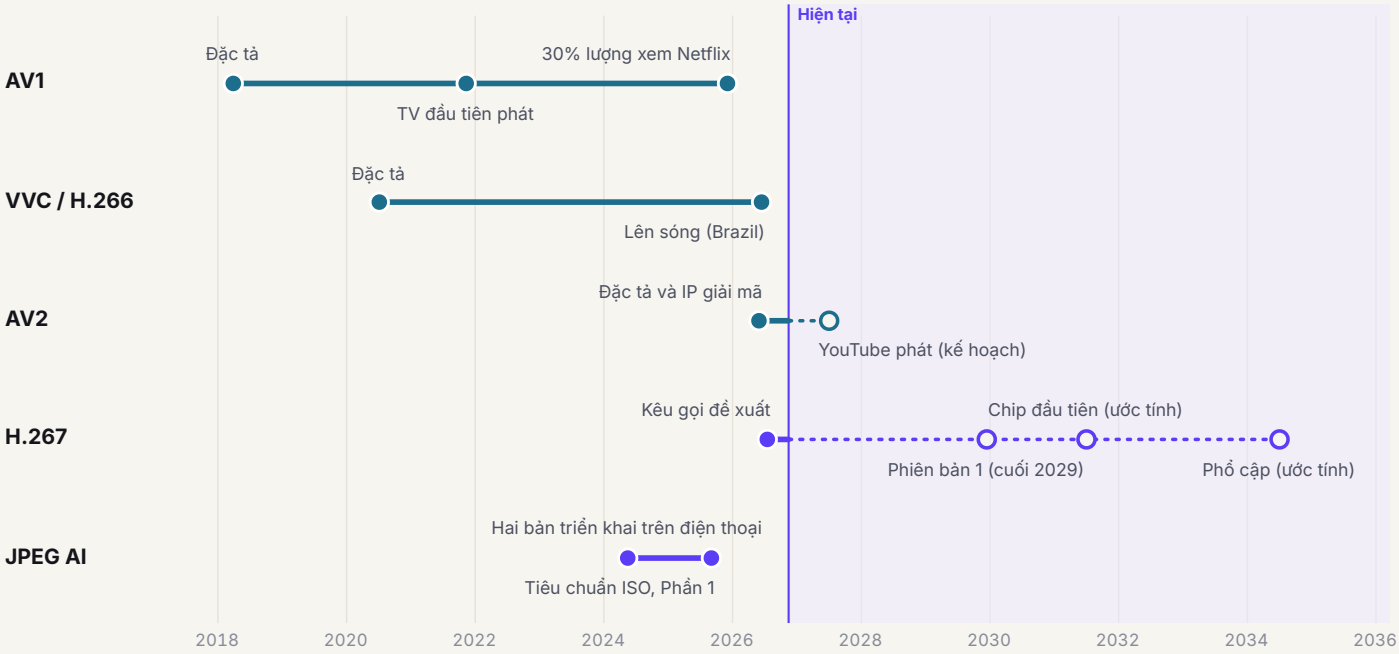
Bài toán kinh tế khác nhau tùy nơi mạng nơ-ron chạy. Xử lý phía máy chủ khiến nền tảng tốn năng lực tính toán cho mọi video mà nó xử lý. Xử lý phía thiết bị không tốn gì của nền tảng; nó dùng NPU mà người mua đã trả tiền. Đó là lý do những mức cải thiện lớn nhất trong năm 2026 đến từ cả hai.

Vì sao không TV nào giải mã luồng bit nơ-ron

Một định dạng phát video phải giải mã được trên mọi thiết bị sẽ phát nó, và những thiết bị đó dùng nhiều năm. Điều này khiến con đường từ một tiêu chuẩn hoàn tất đến màn hình đại chúng diễn ra chậm, ngay cả với codec truyền thống. Đặc tả AV1 hoàn tất vào tháng 3/2018; bộ giải mã phần mềm dav1d ra đời trong vòng vài tháng; TV phát nó từ cuối năm 2021; nó chiếm khoảng 30% lượng xem của Netflix vào tháng 12/2025, gần tám năm sau khi có đặc tả. VVC hoàn tất vào tháng 7/2020 và có dịch vụ phát sóng đầu tiên vào tháng 6/2026.

HÌNH 03 **Từ tiêu chuẩn hoàn tất đến màn hình đại chúng mất gần trọn một thập kỷ**

Phát video cần bộ giải mã phần cứng trong thiết bị. AV1 mất gần tám năm từ khi hoàn tất đặc tả đến lúc chiếm 30% lượng xem Netflix; tiêu chuẩn đầu tiên có khả năng mang công cụ nơ-ron dự kiến ra đời cuối năm 2029.



Nguồn: Netflix Tech Blog (tháng 12/2025); AOMedia; Allegro DVT; TV Technology; Fraunhofer IIS về DTV+ của Brazil; MPEG-155; ủy ban JPEG; ISO. Các mốc rỗng là mục tiêu hoặc ước tính của tác giả.

Luồng bit nơ-ron thêm vào hai vấn đề mà codec truyền thống đã giải quyết từ lâu: năng lượng mỗi pixel và giải mã chính xác đến từng bit. Cả hai được trình bày ở Chương 8. Gộp lại, chúng có nghĩa là sẽ không hãng TV hay điện thoại nào đưa bộ giải mã video nơ-ron vào mọi thiết bị trước khi một tiêu chuẩn chỉ rõ chính xác họ phải xây dựng cái gì.

Điều gì có thể thay đổi cục diện: H.267

Tiêu chuẩn có thể đưa công cụ nơ-ron đến mọi màn hình là thế hệ kế nhiệm VVC, do ITU-T và ISO/IEC cùng phát triển trong Nhóm Chuyên gia Video Chung (JVET). Tên của nó chưa chốt; tôi sẽ gọi nó là H.267, như phần lớn ngành vẫn gọi.

Tháng 10/2025, JVET đánh giá các hồ sơ phản hồi lời kêu gọi bằng chứng: năm hồ sơ từ 16 công ty và phòng thí nghiệm nghiên cứu, với 20 cấu hình codec được thử nghiệm. Ủy ban tìm thấy bằng chứng về khả năng nén tốt hơn VVC và đánh giá cả công nghệ truyền thống cải tiến lẫn công nghệ mạng nơ-ron đều "đáng quan tâm". Các cách tiếp cận nơ-ron cho thấy cải thiện rõ rệt về cảm nhận thị giác, đặc biệt với nội dung 8K HDR, và cải thiện nhỏ hơn với nội dung trò chơi. InterDigital, một bên gửi hồ sơ, đã nộp một thiết kế dựa trên Mô hình Nén Nâng cao (ECM) truyền thống và một thiết kế bổ sung các công cụ AI lên VVC.

Lời kêu gọi đề xuất chính thức được phát hành tại cuộc họp tháng 7/2026 ở Geneva. Hạn nộp hồ sơ chính là cuối tháng 10/2026, một gói bổ sung đến hạn vào giữa tháng 12, và các hồ sơ sẽ được đánh giá tại cuộc họp JVET ở Brisbane vào tháng 1/2027. JVET dự kiến có mô hình thử nghiệm đầu tiên vào nửa cuối năm 2027 và phiên bản đầu tiên của tiêu chuẩn trước cuối năm 2029, và gọi nỗ lực này là "có lẽ là nỗ lực lớn hơn bất kỳ lần nào trước đây để khởi động một tiêu chuẩn mới".

Hướng truyền thống đặt ra một ngưỡng cao. ECM, được xây dựng bằng xử lý tín hiệu truyền thống, tiết kiệm chừng 28% so với VVC trong cấu hình truy cập ngẫu nhiên. Nhóm thăm dò mạng nơ-ron của JVET báo cáo các

bộ lọc trong vòng lặp tiết kiệm khoảng 6% ở mức 10 nghìn phép nhân-cộng dồn mỗi pixel, 8% ở mức 22 nghìn, và 14% ở một điểm vận hành có độ phức tạp cao. Nhận định của tôi là H.267 nhiều khả năng sẽ là một lõi truyền thống cùng vài công cụ nơ-ron ở nơi lượng bit tiết kiệm được xứng với chi phí giải mã, và đợt đánh giá tháng 1/2027 sẽ cho thấy có bao nhiêu công cụ như vậy phân tích.

Đã có sẵn một cầu nối. Luồng VVC và HEVC có thể mang thông tin tăng cường bổ sung (SEI) mô tả một bộ lọc hậu xử lý nơ-ron để phía giải mã áp dụng, và các phiên bản mới của VVC cùng tiêu chuẩn SEI của nó (VSEI), hoàn tất vào tháng 10/2025, đã mở rộng các thông điệp này. Một luồng có thể cho biết nên chạy mạng nào sau khi giải mã mà không buộc mọi bộ giải mã phải chạy nó.

Ý nghĩa với doanh nghiệp trong mảng phát video

Hãy bán những gì hoạt động quanh codec ngay bây giờ: tiền xử lý phía mã hóa cho các dịch vụ, và khối hậu xử lý cùng dịch vụ tinh chỉnh cho chip TV và đầu thu. Đừng xây dựng một mảng kinh doanh phát video dựa trên luồng bit nơ-ron phải giải mã trên thiết bị của bên thứ ba trước thập niên 2030. Nếu bạn làm chip, doanh thu trước mắt là bộ giải mã AV2 và VVC; phần nơ-ron là một đường xử lý NPU để tăng cường chất lượng đặt bên cạnh chúng (Chương 8).

CHƯƠNG 03

NHIẾP ẢNH VÀ HÌNH ẢNH

03 Nhiếp ảnh và hình ảnh: nơ-ron trong camera, truyền thống trên web

Quy trình xử lý của camera đã là nơ-ron

Mọi bức ảnh từ một chiếc điện thoại cao cấp gần đây đều đã đi qua mạng nơ-ron trước khi được nén. Snapdragon 8 Gen 2 của Qualcomm, có trong điện thoại từ năm 2023, giới thiệu một "Cognitive ISP" chạy phân đoạn ngữ nghĩa thời gian thực khi chụp ảnh và quay video, điều chỉnh riêng khuôn mặt, tóc, bầu trời và hậu cảnh; Galaxy S23 Ultra của Samsung là điện thoại đầu tiên dùng nó cho camera selfie. Tháng 1/2026, tại CES, Chips&Media và Visionary.ai trình diễn thứ mà họ gọi là ISP hoàn toàn dựa trên AI đầu tiên: một quy trình xử lý ảnh đầu cuối chạy hoàn toàn trên NPU WAVE-N của Chips&Media, ở dấu phẩy động 16 bit, không cần ISP chức năng cố định, và có thể cập nhật qua mạng. "Các quy trình hiện có gắn thêm khối nơ-ron vào một ISP chức năng cố định. Chúng tôi thay thế hoàn toàn ISP truyền thống", CEO của Visionary.ai cho biết. Sau triển lãm, giải pháp được đưa đến các nhà sản xuất để đánh giá.

Sau đó, kết quả được lưu dưới dạng JPEG hoặc HEIF, những định dạng truyền thống mà thiết bị nào cũng mở được. Khuôn mẫu giống như trong phát video: nơ-ron bên trong quy trình của một công ty, chuẩn ở ranh giới.

JPEG AI: một tiêu chuẩn nơ-ron đã hoàn tất nhưng chưa có sản phẩm

JPEG AI là tiêu chuẩn mã hóa ảnh học máy đầu tiên của ISO/IEC và ITU-T (ISO/IEC 6048, ITU-T T.840). Phần 1, hệ thống mã hóa lõi, được ban hành năm 2025. Ủy ban JPEG báo cáo mức cải thiện nén từ 12,5% đến 27,9% so với mã hóa nội khung VVC, trên một dải độ phức tạp giải mã từ 7 đến 216 nghìn phép nhân-cộng dồn mỗi pixel. Thiết kế của nó gồm hai bộ mã hóa và ba bộ giải mã được huấn luyện chung, nên bên nhận chỉ cần triển khai một bộ giải mã mà vẫn đọc được mọi luồng. Đến tháng 5/2024, ủy ban đã báo cáo hai bản triển khai trên điện thoại thông

minh. Song song đó, vào tháng 12/2024, IEEE ban hành IEEE 1857.11-2024, một tiêu chuẩn mã hóa ảnh dựa trên mạng nơ-ron được thiết kế cho NPU.

Tính đến tháng 11/2026, tôi không tìm thấy camera điện thoại, ứng dụng ảnh hay trình duyệt nào ghi hoặc đọc JPEG AI theo mặc định.

Thứ web chọn thay vào đó: JPEG XL

Định dạng ảnh thực sự lan rộng trong năm 2026 là định dạng truyền thống. Ngày 6/10/2026, Google bật giải mã JPEG XL mặc định trong Chrome 155, dùng một bộ giải mã mới an toàn bộ nhớ viết bằng Rust, bốn năm sau khi gỡ bỏ hỗ trợ thử nghiệm. Safari đã giải mã JPEG XL từ năm 2023, và Mozilla cũng đang đi theo hướng đó. Lập luận của Google là khả năng nén tốt hơn JPEG, nén lại không mất dữ liệu các tệp JPEG hiện có, HDR và giải mã lũy tiến.

Bài học này mang tính khái quát. Một định dạng phải mở được ở mọi nơi sẽ thắng nhờ một bộ giải mã nhanh, tất định, chạy trên mọi CPU và có thể bảo đảm an toàn bộ nhớ. Bộ giải mã ảnh nơ-ron cần NPU hoặc GPU và số học chính xác giữa các thiết bị. Mã hóa nơ-ron thắng thế bên trong quy trình riêng của một thiết bị và trong những hệ thống mà cùng một công ty vừa mã hóa vừa giải mã, như một dịch vụ ảnh đám mây không bao giờ giao tệp nén cho bất kỳ ai khác.

Ý nghĩa với doanh nghiệp trong mảng xử lý ảnh

Tiền nằm ở quy trình ghi hình: IP ISP AI, NPU có kích cỡ phù hợp cho xử lý ảnh, và công việc tinh chỉnh mà các hãng camera mua. Với lưu trữ và phân phối, codec ảnh học máy ngày nay là công cụ cho hệ thống khép kín, không phải một định dạng để đặt cược.

CHƯƠNG 04



CUỘC GỌI VÀ HỌP TRỰC TUYẾN

04 Cuộc gọi và họp trực tuyến: nơi codec nơ-ron đầu cuối thực sự được triển khai

Âm thanh đi trước

Tháng 2/2021, Microsoft giới thiệu Satin, một codec giọng nói dùng AI, và cho biết nó đã được dùng cho mọi cuộc gọi hai người trên Teams và Skype. Satin cho giọng nói siêu băng rộng, băng thông âm thanh tới 16 kHz, ở 6 kbps, trong khi codec Silk trước đây của Skype là băng hẹp ở tốc độ đó. Phần bitrate tiết kiệm được dùng cho dữ liệu dự phòng, nên cuộc gọi chống chịu mất gói tốt hơn.

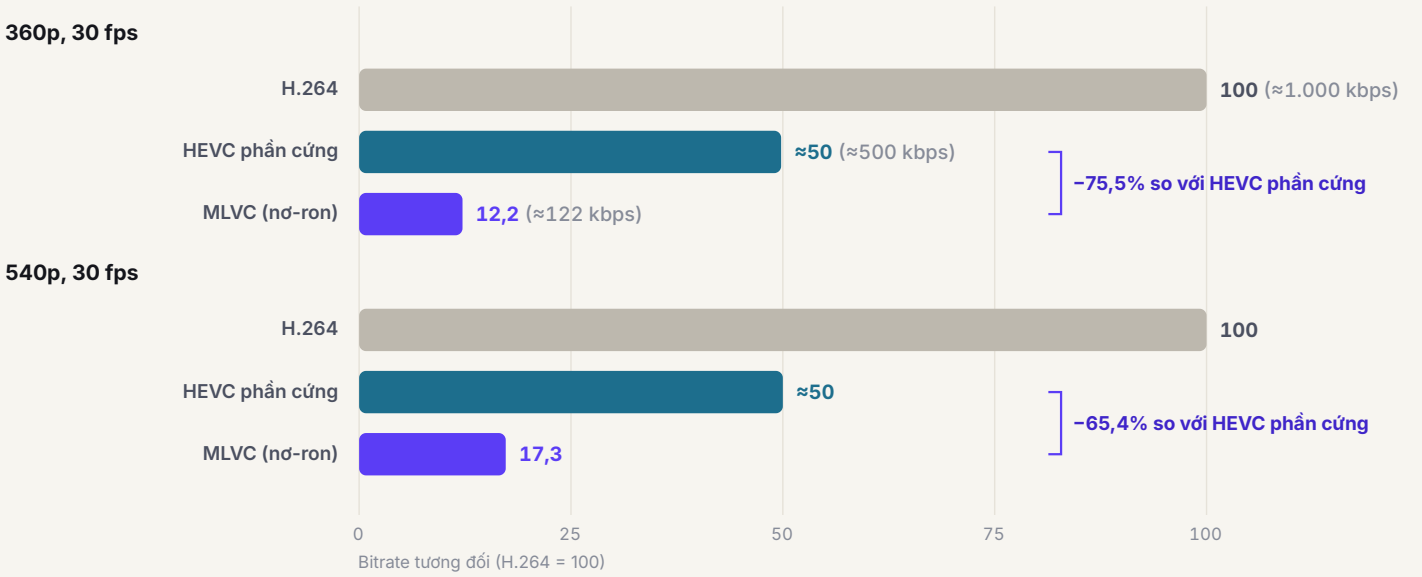
MLVC: codec video nơ-ron trong các cuộc gọi thực tế

Ngày 24/7/2026, Microsoft phát hành MLVC, codec video học máy của hãng, dưới dạng mã nguồn mở theo giấy phép MIT, kèm mã mô hình, trọng số đã huấn luyện, script huấn luyện và các công cụ chuyển đổi nhắm tới những NPU khác nhau. MLVC là phiên bản sản phẩm của họ DCVC mà Microsoft Research công bố từ năm 2021. Microsoft cho biết đang triển khai MLVC trên Teams cho cuộc gọi ngang hàng, với dữ liệu đo từ xa trực tiếp, thử nghiệm A/B và các codec truyền thống làm phương án dự phòng.

MLVC chạy ở 540p và 360p với 30 khung hình/giây và được thiết kế để dùng không quá một nửa NPU trên Neural Engine của Apple cũng như NPU của Qualcomm và Intel, chia việc giữa CPU và NPU. Trong các thử nghiệm chủ quan theo ITU-T P.910 trên bộ Video Conferencing Dataset của chính Microsoft, nó cần ít bitrate hơn H.264 87,8% và ít hơn HEVC phần cứng 75,5% ở 360p, và lần lượt ít hơn 82,7% và 65,4% ở 540p theo nhà cung cấp. Ở 360p, đó là chừng 122 kbps trong khi H.264 cần khoảng 1 Mbps.

HÌNH 04 **Trong cuộc gọi, codec nơ-ron vượt HEVC phần cứng 65–76%**

Bitrate cần cho cùng mức chất lượng được chấm điểm trên bộ thử nghiệm hội nghị truyền hình của Microsoft, H.264 = 100. MLVC chạy trên NPU của laptop và điện thoại, tối đa 540p30, dùng không quá một nửa NPU.



Nguồn: Microsoft, công bố MLVC (24/7/2026): thử nghiệm chủ quan theo ITU-T P.910 trên bộ Video Conferencing Dataset của hãng. Cột HEVC được suy ra từ mức tiết kiệm đã công bố. Số liệu do nhà cung cấp công bố, không phải phép đo độc lập.

Có hai điều khiến con số trên tiêu đề bớt ấn tượng. Mức 88% là so với H.264; so với HEVC phần cứng vốn đã có trong chính những chiếc laptop đó, mức tiết kiệm là 65–76%, vẫn rất lớn. Và MLVC đã hy sinh khả năng nén để có thể triển khai: một phân tích độc lập chỉ ra rằng DCVC-RT được huấn luyện theo cảm nhận đạt 81,9% trên cùng thước đo mà MLVC đạt 75,5%. Cái giá của việc giải mã giống hệt nhau trên phần cứng không thuộc về Microsoft là khoảng sáu điểm.

Vì sao cuộc gọi là môi trường tự nhiên

Một cuộc gọi là một vòng khép kín. Hai đầu chạy cùng một ứng dụng, thỏa thuận khả năng của nhau khi bắt đầu cuộc gọi, và có thể chuyển về codec chuẩn giữa chừng. Độ phân giải khiêm tốn, mạng thường kém, và laptop ở cả hai đầu nay đều có NPU, như điện thoại đã có từ nhiều năm. Microsoft đặt NPU 40 TOPS làm ngưỡng cho PC AI của hãng, và Neural Engine của Apple M4 được công bố ở mức 38 TOPS. Mọi điều kiện cản trở codec nơ-ron trong phát video đều không có ở đây.

Các phản ví dụ: Opus và MLow

Không phải nền tảng gọi điện nào cũng đưa ra cùng lựa chọn, và lý do rất đáng học hỏi.

Opus, codec âm thanh mở trong WebRTC và phần lớn ứng dụng gọi điện, bổ sung học sâu ở phiên bản 1.5 (tháng 3/2024), lần đầu nó dùng mạng nơ-ron để xử lý hoặc tạo ra chính tín hiệu. Nhưng Opus thêm chúng dưới dạng tùy chọn: tăng cường phía giải mã (LACE và NoLACE) không cần thay đổi luồng bit, che giấu mất gói bằng mạng nơ-ron, và một cơ chế dư thừa sâu (DRED); tất cả đều tắt mặc định để thiết bị cũ không phải gánh thêm. Opus 1.6

thêm mở rộng băng thông bằng mạng nơ-ron và một mô hình DRED nhỏ hơn ba lần, và vẫn tương thích hoàn toàn với tiêu chuẩn năm 2012. Cách tiếp cận này là "trong tiêu chuẩn, tùy chọn và tương thích ngược".

Meta còn lùi xa hơn. Tháng 6/2024, hãng ra mắt MLow, một codec giọng nói bitrate thấp xây dựng trên xử lý tín hiệu truyền thống thay vì mạng nơ-ron, hiện có trong mọi cuộc gọi Instagram và Messenger và đang được triển khai trên WhatsApp. Lý do của Meta nằm ở thiết bị: hơn 20% cuộc gọi của hãng được thực hiện trên điện thoại ARMv7, và mỗi ngày có hàng chục triệu cuộc gọi WhatsApp chạy trên điện thoại hơn mười năm tuổi. Meta viết rằng codec nơ-ron, kể cả EnCodec của chính Meta, chỉ chạy ổn định trên các máy đắt tiền. MLow cho chất lượng gấp khoảng đôi Opus ở 6 kbps trong khi dùng ít hơn 10% tài nguyên tính toán.

Với một dịch vụ có người dùng trải từ laptop cao cấp đến điện thoại mười năm tuổi, ngân sách độ phức tạp là yếu tố quyết định. Codec nơ-ron được triển khai trước ở nơi lượng thiết bị đang dùng còn mới và nhà cung cấp kiểm soát cả hai đầu.

Ý nghĩa với doanh nghiệp trong mảng truyền thông liên lạc

Hội nghị truyền hình doanh nghiệp, hiện diện từ xa, hỗ trợ từ xa, đường truyền dẫn tín hiệu (contribution) và mạng riêng là những nơi mà codec nơ-ron đầu cuối ngày nay đã là một sản phẩm, với điều kiện cả hai đầu đều thuộc về bạn. Hãy lên kế hoạch cho ba việc ngay từ đầu: một codec dự phòng, kiểm thử giải mã trên mọi NPU bạn hỗ trợ, và đánh giá chất lượng trên nội dung của chính bạn thay vì bộ dữ liệu của nhà cung cấp.

CHƯƠNG 05

AI TẠO SINH

05 AI tạo sinh: nơi triển khai lớn nhất mà không ai gọi là codec

Mọi mô hình giọng nói và video đều chạy trên một codec học máy

Nơi triển khai nén nơ-ron đầu cuối lớn nhất không nằm trong bất kỳ tiêu chuẩn codec nào. Nó nằm bên trong AI tạo sinh. Một mô hình giọng nói không trực tiếp nghe hay tạo ra dạng sóng: nó đọc và ghi các token rời rạc của một codec âm thanh học máy. Mimi của Kyutai, phát hành năm 2024, chuyển giọng nói thành token ở 12,5 khung mỗi giây và 1,1 kbps, thấp hơn mọi codec điện thoại, kết hợp thông tin ngữ nghĩa và âm học trong một luồng. EnCodec của Meta, công bố năm 2022, là một ví dụ khác; sau này chính các kỹ sư của Meta mô tả những codec như vậy là xuất sắc ở bitrate thấp nhưng quá nặng cho điện thoại đời cũ (Chương 4).

Tạo video cũng hoạt động theo cách đó. Một mô hình khuếch tán hoặc transformer tạo nội dung trong không gian tiềm ẩn của một bộ tự mã hóa video học máy (VAE), và bộ giải mã của VAE biến kết quả thành pixel. VAE của LTX-Video nén video theo tỷ lệ 1:192: mỗi token tiềm ẩn đại diện cho một khối 32×32 pixel qua 8 khung hình, với 128 kênh. Bộ giải mã của nó còn thực hiện bước khử nhiễu cuối cùng, nên phần nào tự tạo chi tiết thay vì chỉ tái tạo.

Vì sao chúng không phải định dạng phân phối

Các codec này chạy trong mọi yêu cầu gửi tới một mô hình giọng nói hay video, nhưng không cái nào là định dạng mà bên khác giải mã. Mỗi mô hình có không gian tiềm ẩn riêng; không có kiểm soát bitrate, kiểm tra tuân thủ hay bảo đảm rằng một bộ giải mã khác sẽ cho ra cùng hình ảnh. Chúng được thiết kế cho chất lượng tạo sinh, không phải cho độ trung thực ở một bitrate cho trước, và chúng giải mã trên GPU của trung tâm dữ liệu. Thành

công của chúng cho thấy nén bằng học máy hoạt động ở quy mô lớn khi một công ty sở hữu bộ mã hóa, bộ giải mã và phần cứng.

Ý nghĩa với doanh nghiệp

Hai điểm thực tế. Thứ nhất, hiệu quả của bộ token hóa là một dòng chi phí trực tiếp với bất kỳ ai cung cấp video hoặc giọng nói tạo sinh: với một transformer áp dụng cơ chế chú ý trên toàn bộ token, khối lượng tính toán tăng nhanh hơn số token, nên một biểu diễn tiềm ẩn nén mạnh hơn sẽ giảm chi phí tạo sinh nhiều hơn mức tương ứng phân tích. Những kỹ sư codec hiểu thiết kế không gian tiềm ẩn, mô hình entropy và hiệu quả giải mã là nguồn lực quý với các đội AI tạo sinh. Thứ hai, ranh giới đang dịch chuyển: các tổ chức tiêu chuẩn đã bắt đầu làm việc về mã hóa cho máy thay vì cho con người. Video Coding for Machines (ISO/IEC 23888-2) của MPEG đạt mức dự thảo tiêu chuẩn quốc tế vào đầu năm 2026 và dự kiến hoàn tất vào cuối năm 2026, còn JPEG AI ngay từ đầu đã được thiết kế để phục vụ các tác vụ thị giác máy tính cũng như cho con người xem.

CHƯƠNG 06

TRÒ CHƠI VÀ ĐỒ HỌA

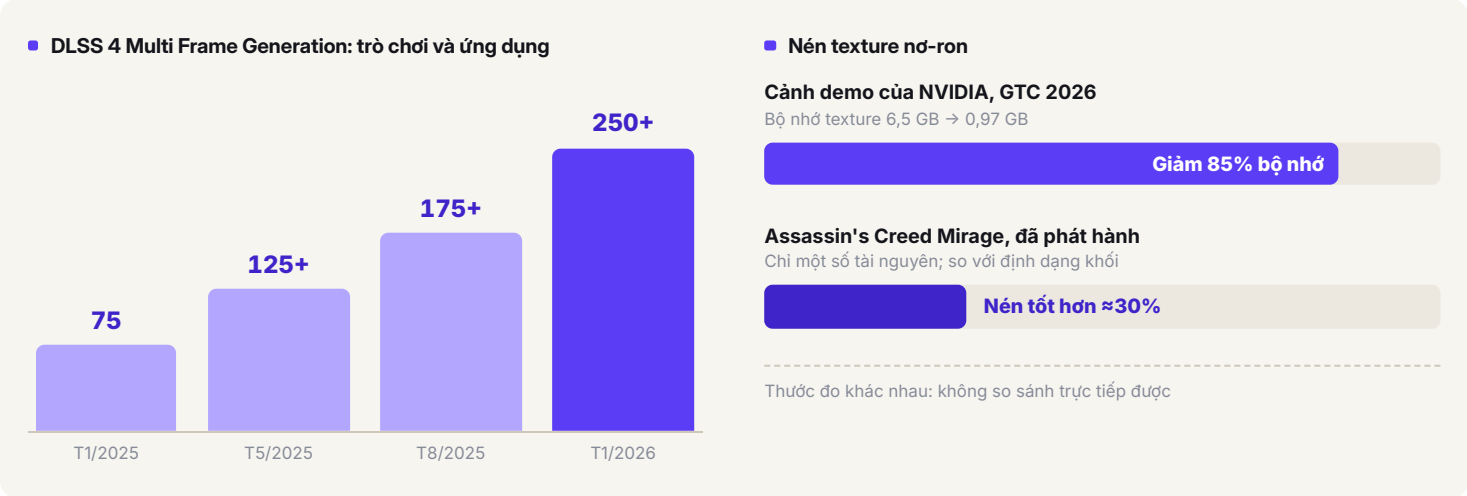
06 Trò chơi và đồ họa: tái tạo nơ-ron đã phổ biến, tài nguyên nơ-ron mới bắt đầu

Kết xuất ít pixel hơn và để mạng nơ-ron lấp phần còn lại

Trong đồ họa thời gian thực, mạng nơ-ron đã tái tạo phần lớn những gì người chơi nhìn thấy. DLSS của NVIDIA kết xuất ít pixel hơn và ít khung hình hơn, rồi dùng mạng nơ-ron để tạo phần còn lại. DLSS 4 với Multi Frame Generation ra mắt vào tháng 1/2025 với 75 trò chơi và ứng dụng hỗ trợ; NVIDIA đếm được hơn 125 vào tháng 5/2025, hơn 175 vào tháng 8 và hơn 250 vào tháng 1/2026, khi hãng công bố DLSS 4.5 với mô hình transformer thế hệ thứ hai cho siêu phân giải. Về bản chất đó là một bộ giải mã, chỉ khác tên gọi: "luồng bit" là một khung hình độ phân giải thấp cộng dữ liệu chuyển động, và mạng nơ-ron dựng lại toàn bộ hình ảnh.

HÌNH 05 **Khi một công ty kiểm soát cả hệ thống, tái tạo bằng mạng nơ-ron lan nhanh**

Trái: số trò chơi và ứng dụng hỗ trợ DLSS 4 Multi Frame Generation của NVIDIA. Phải: nén texture nơ-ron trong một cảnh demo và trong một trò chơi đã phát hành; trò chơi này chỉ dùng nó cho một số tài nguyên chọn lọc vì chi phí khi chạy.



Nguồn: NVIDIA (tháng 1/2025 đến tháng 1/2026; demo tại GTC 2026); Ubisoft La Forge (tháng 2/2026). Số liệu do nhà cung cấp công bố; hai số liệu texture đo những thứ khác nhau.

Texture nơ-ron đã được triển khai, một cách có chọn lọc

Texture được nén theo các định dạng khối cố định (BCn) mà GPU giải mã bằng phần cứng. Nén texture nơ-ron thay một bộ texture vật liệu bằng một biểu diễn tiềm ẩn gọn nhẹ cùng một mạng nhỏ tái tạo texel khi cần. Tháng 2/2026, phòng thí nghiệm La Forge của Ubisoft mô tả việc đưa nó vào Assassin's Creed Mirage, một trò chơi năm 2023, với tỷ lệ nén hiệu dụng cao hơn khoảng 30% so với các định dạng truyền thống ở cùng chất lượng hình ảnh theo nhà cung cấp. Nó không được dùng ở mọi nơi: "việc sử dụng bị giới hạn bởi hiệu năng khi chạy", nên Ubisoft áp dụng nó cho một nhóm tài nguyên có số lần xuất hiện cao và gây áp lực bộ nhớ lớn.

RTX Neural Texture Compression SDK của NVIDIA, đang ở bản beta trên GitHub, cho thấy mức trần. Trong một buổi trình diễn tại GTC 2026, một cảnh biệt thự dùng 6,5 GB bộ nhớ texture với BCn cần 970 MB với texture nơ-ron, ít hơn 85% theo nhà cung cấp. SDK có hai cách: giải mã sang BCn khi tải, giúp tiết kiệm dung lượng đĩa và kích thước tải về nhưng không tiết kiệm bộ nhớ video, hoặc giải mã khi lấy mẫu, giúp tiết kiệm bộ nhớ nhưng tốn thời gian shader ở mỗi khung hình.

API đang đến

Các API đồ họa đang bổ sung những phép toán ma trận mà giải mã nơ-ron cần bên trong shader. Microsoft phát hành Shader Model 6.9 vào tháng 2/2026; tính năng cooperative vector mà hãng cho xem trước năm 2025 đã được gộp vào một thiết kế đại số tuyến tính hợp nhất cho một shader model sau này, với DirectX Linear Algebra ra bản xem trước công khai vào tháng 4/2026. Khi tính năng đó có mặt trong mọi driver, một bộ giải mã texture nơ-ron có thể chạy trên GPU của bất kỳ hãng nào từ một đường mã duy nhất.

Ý nghĩa với doanh nghiệp

Trò chơi triển khai giải mã nơ-ron vì các điều kiện của một vòng khép kín được thỏa mãn: một engine, một quy trình tài nguyên, phần cứng tensor trong GPU và một định dạng dự phòng. Phần việc còn bỏ ngỏ nằm ở công cụ, quy trình tài nguyên mã hóa vật liệu nơ-ron, môi trường chạy liên hãng và, trên hết, thiết bị di động, nơi chi phí năng lượng cho mỗi texel giải mã là yếu tố quyết định.

CHƯƠNG 07

07

LIÊN KẾT CHUYÊN DỤNG

Liên kết chuyên dụng: vệ tinh, drone và robot

Đây là những trường hợp mà trên lý thuyết, mã hóa nơ-ron đầu cuối phù hợp nhất. Một chủ sở hữu kiểm soát cả hai đầu liên kết, băng thông khan hiếm và hay mất dữ liệu, nội dung đủ hẹp để huấn luyện riêng, và một hình ảnh trung thực về cảm nhận, thậm chí chỉ về ngữ nghĩa, thường là đủ. Codec nơ-ron chịu mất gói cho những liên kết như vậy là một hướng nghiên cứu sôi động, và việc khôi phục phía máy chủ cho nguồn cấp bitrate thấp đang được sử dụng.

Đó cũng là một thị trường nhỏ và ngày càng bị quản lý chặt so với video tiêu dùng phân tích. Tại Trung Quốc, Thông báo số 49 năm 2026 của Bộ Thương mại và hai cơ quan khác, ban hành ngày 19/10 và có hiệu lực từ ngày 18/11/2026, sửa đổi quy định kiểm soát xuất khẩu đối với 13 nhóm mặt hàng liên quan đến drone. Theo báo chí chuyên ngành, trong các mục công nghệ có một mục về nền tảng mặt đất không người lái được xác định một phần bằng các ngưỡng như nén ảnh ngữ nghĩa 75% và tái tạo ảnh trong vòng 150 mili giây. Bản thân nén ngữ nghĩa đã trở thành một thông số kiểm soát xuất khẩu. Tại Mỹ, ngày 22/12/2025, Ủy ban Truyền thông Liên bang đã đưa drone sản xuất ở nước ngoài và linh kiện quan trọng của chúng vào Covered List, nên các mẫu mới của nước ngoài không còn được cấp phép thiết bị.

Kết luận: một thị trường ngách vững chắc cho doanh nghiệp chuyên biệt có khách hàng công nghiệp hoặc chính phủ và có năng lực quản lý việc tuân thủ xuất khẩu ở cả hai phía, nhưng không phải thị trường đại chúng, và không phải nơi để biết mã hóa nơ-ron có hoạt động ở quy mô lớn hay không.

CHƯƠNG 08

08

CHIP

Góc nhìn từ chip: năng lượng, tính toán, bộ nhớ và độ chính xác

Đây là chương mà kinh nghiệm của chính tôi có ý nghĩa nhất. Tôi có nhiều năm nghiên cứu, phát triển và quản lý chip fabless, bao gồm thiết kế đường dữ liệu (datapath) ASIC full-custom và các dự án chip camera đa phương tiện như bộ tăng tốc codec video, và những câu hỏi một đội thiết kế chip đặt ra cho một codec mới không phải về BD-rate. Đó là: bao nhiêu joule mỗi pixel, bao nhiêu byte lưu lượng DRAM mỗi pixel, bao nhiêu cổng logic, và liệu mọi bộ giải mã trên thế giới có tính ra đúng cùng các bit hay không? Ở cả bốn câu hỏi, bộ giải mã nơ-ron năm 2026 đang ở vị trí của HEVC phần cứng trước khi có những con chip đầu tiên: hứa hẹn trên phần mềm, đắt đỏ trên silicon.

Năng lượng mỗi pixel: cách nhau một bậc độ lớn

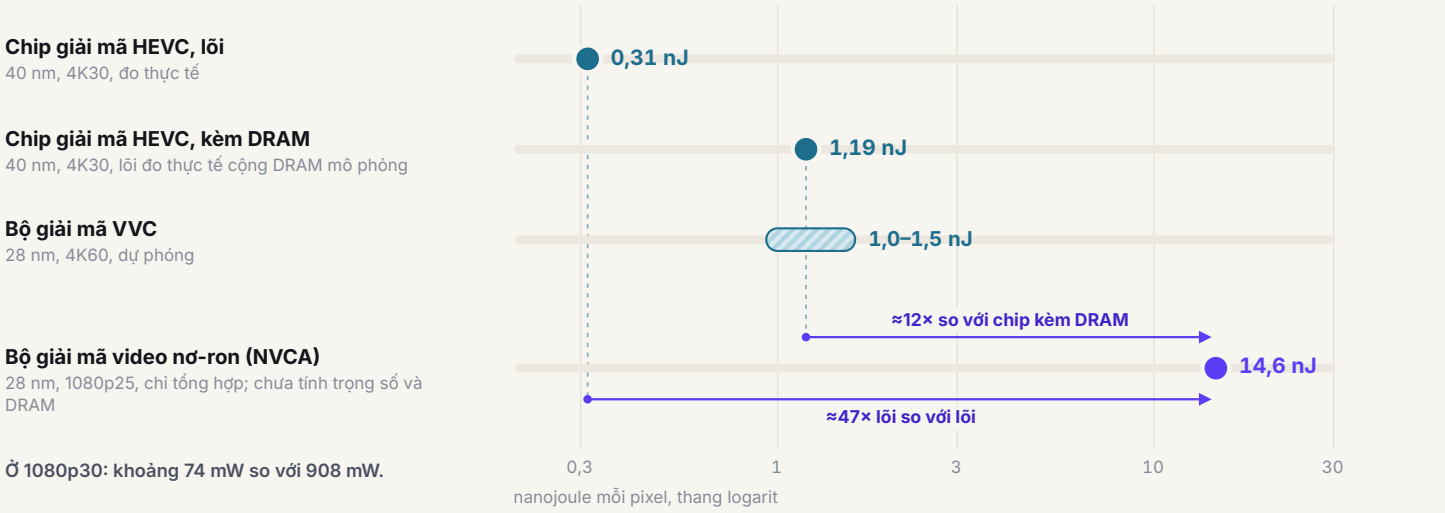
Rất ít bộ giải mã nơ-ron từng được thiết kế thành chip, nên phép so sánh dựa trên một số ít thiết kế đã công bố. Mốc so sánh phía truyền thống là chip giải mã HEVC của MIT (Tikekar và cộng sự, ISSCC 2013 và IEEE JSSC

2014): 40 nm, 1,77 mm², giải mã 4K ở 30 khung hình/giây với xung nhịp 200 MHz. Lõi của nó tiêu thụ 78 mW, tức 0,31 nJ mỗi pixel; khi tính cả công suất DRAM mô phỏng, hệ thống cần 1,19 nJ mỗi pixel. Với VVC, số liệu của chip thương mại là thông tin độc quyền; một phân tích trên Electronic Design năm 2026, khi quy đổi chip của MIT sang 28 nm và tính đến độ phức tạp cao hơn của VVC, dự phóng 1–1,5 nJ mỗi pixel cho một bộ giải mã 4K60 khoảng một triệu cổng.

Ở phía nơ-ron, theo cách đếm của Electronic Design, thiết kế ASIC giải mã video nơ-ron đầu cuối duy nhất đã công bố là NVCA (Zhang và cộng sự, DATE 2024), được tổng hợp trên tiến trình TSMC 28 nm với 5,01 triệu cổng, chưa tính trọng số vốn nằm ngoài chip. Nó giải mã 1080p ở 25 khung hình/giây với 760 mW, tức 14,6 nJ mỗi pixel.

HÌNH 06 **Năng lượng mỗi pixel giải mã: bộ giải mã chức năng cố định dẫn trước một bậc độ lớn**

Thiết kế chip giải mã video nơ-ron duy nhất đã công bố cần năng lượng mỗi pixel gấp khoảng 12 lần một chip giải mã HEVC đã đo khi tính cả DRAM, và khoảng 47 lần riêng phần lõi. Lưu lượng trọng số và DRAM của chính thiết kế nơ-ron chưa được tính.



Nguồn: Tikekar và cộng sự, IEEE JSSC 2014 (MIT); NVCA, Zhang và cộng sự, DATE 2024 (arXiv 2312.10716); khoảng giá trị của VVC là dự phóng của Electronic Design (2026). Chip HEVC dùng tiến trình cũ hơn (40 nm so với 28 nm), điều này có lợi cho thiết kế nơ-ron.

Mức đó gấp khoảng 12 lần chip HEVC khi tính cả DRAM, và khoảng 47 lần riêng phần lõi của nó. Phép so sánh có lợi cho NVCA theo ba cách: chip HEVC dùng tiến trình cũ hơn, số liệu của NVCA bỏ qua lưu lượng trọng số và DRAM, và NVCA là kết quả tổng hợp chứ không phải silicon đã đo thực tế. Ở 1080p30, chênh lệch là khoảng 74 mW so với chừng 900 mW: một bộ giải mã không ai để ý, so với một bộ giải mã lộ rõ trong ngân sách pin và nhiệt của điện thoại.

Tính toán mỗi pixel: bộ giải mã nơ-ron đòi hỏi gì ở NPU

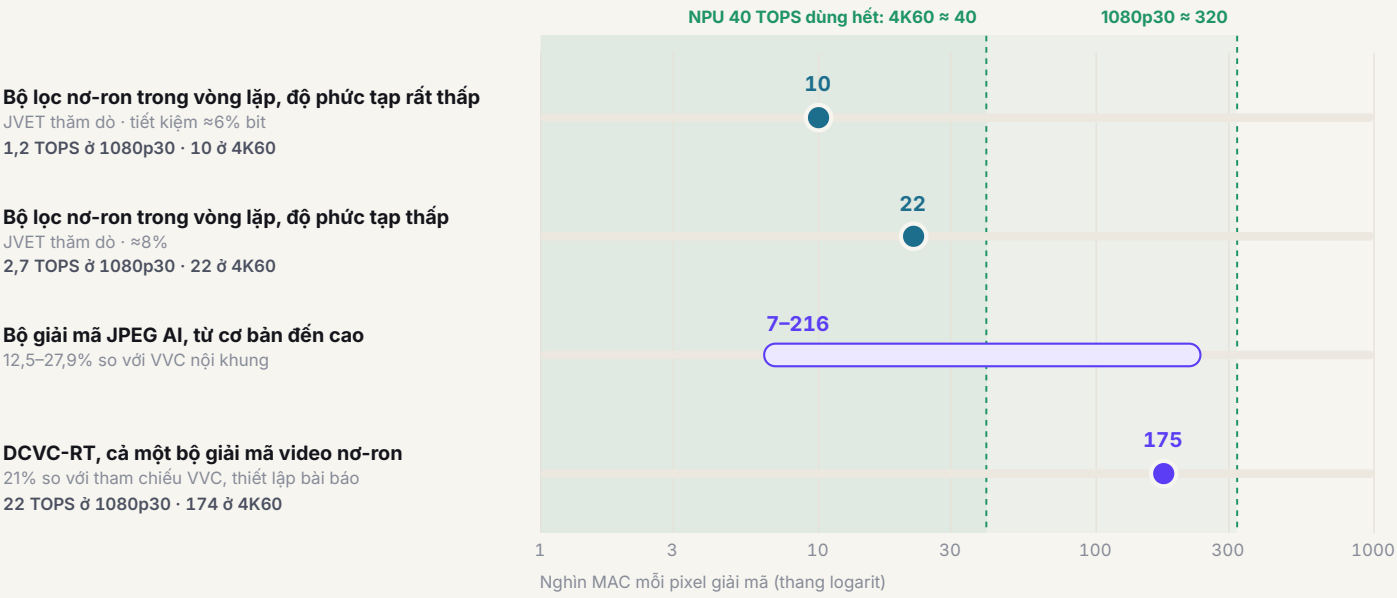
Codec video nơ-ron đã công bố nhanh nhất mà tôi tìm thấy là DCVC-RT của Microsoft Research (Jia và cộng sự, CVPR 2025). Nó mã hóa và giải mã 1080p ở 125 và 113 khung hình/giây trên GPU NVIDIA A100, đồng thời tiết kiệm 21% bitrate so với bộ mã hóa tham chiếu của VVC trong thiết lập thử nghiệm của bài báo, và nó dùng số học số nguyên để các thiết bị khác nhau giải mã nhất quán. Tài liệu bổ sung của nó liệt kê độ phức tạp của từng mô-đun. Cộng bốn mô-đun mà bộ giải mã phải chạy (bộ giải mã, bộ trích xuất đặc trưng, mô hình entropy và khối tái tạo), tôi có khoảng 175 nghìn phép nhân-cộng dồn cho mỗi pixel giải mã ước tính.

Ở 1080p30, đó là khoảng 11 nghìn tỷ phép nhân-cộng dồn mỗi giây, tức khoảng 22 TOPS với hai phép toán mỗi MAC, chưa tính bất kỳ tổn hao hiệu suất sử dụng nào; 4K60 cần gấp tám lần ước tính. Để so sánh, Microsoft định

nghĩa PC AI của hãng bằng một NPU tối thiểu 40 TOPS, còn Apple công bố Neural Engine của M4 ở mức 38 TOPS. Đó là các con số đỉnh, được chia sẻ với mọi việc khác mà thiết bị làm.

HÌNH 07 **Một công cụ nơ-ron tốn hàng chục nghìn MAC mỗi pixel; cả một bộ giải mã nơ-ron tốn hàng trăm nghìn**

Số phép nhân-cộng dồn của bộ giải mã cho mỗi pixel, thang logarit, so với mức mà một NPU 40 TOPS dùng hết công suất có thể đáp ứng ở 4K60 và 1080p30. Các mức cải thiện được tính so với những mốc khác nhau nên không so sánh trực tiếp được.



Nguồn: JVET qua chủ tịch SC 29 (tháng 4/2026); ủy ban JPEG (2024); DCVC-RT, tài liệu bổ sung CVPR 2025, tác giả cộng các mô-đun giải mã; định nghĩa Copilot+ PC của Microsoft. Tác giả tính: hai phép toán mỗi MAC, dùng hết công suất.

Nấc thang trong Hình 7 là cách hữu ích để nhìn nhận lĩnh vực này. Những công cụ nơ-ron tốn hàng chục nghìn MAC mỗi pixel, như các bộ lọc trong vòng lặp của JVET (tiết kiệm khoảng 6% ở mức 10 nghìn, 8% ở mức 22 nghìn), là khả thi bên trong một bộ giải mã chuẩn tương lai. JPEG AI trải trên cả dải này, từ 7 đến 216 nghìn MAC mỗi pixel cho mức cải thiện 12,5–27,9% so với VVC nội khung. Cả một bộ giải mã video nơ-ron với chừng hai trăm nghìn MAC mỗi pixel thuộc về NPU, ở độ phân giải khiêm tốn, trong một vòng khép kín. Đó chính là vị trí của MLVC: 540p30, dùng không quá một nửa NPU.

Lưu lượng bộ nhớ, chứ không phải MAC, thường mới là giới hạn thực sự

Nhóm tác giả DCVC-RT nhận thấy thứ giới hạn tốc độ không phải là tính toán mà là "chi phí vận hành": nhập xuất bộ nhớ và số lần gọi hàm. Họ loại bỏ các mô-đun chuyển động tường minh và xử lý đa tỷ lệ phần lớn vì lý do đó. Nhóm thiết kế NVCA xây dựng một luồng dữ liệu đặc biệt để cắt giảm lưu lượng bộ nhớ ngoài chip từ các đặc trưng chuyển động và phần dư. Người thiết kế chip sẽ nhận ra khuôn mẫu này. Trong chip HEVC của MIT, DRAM chiếm khoảng ba phần tư năng lượng hệ thống (0,88 trên 1,19 nJ mỗi pixel), ngay cả sau khi một bộ nhớ đệm đã cắt giảm 67% băng thông DRAM phân tích. Một bộ giải mã nơ-ron giữ các bản đồ đặc trưng nhiều kênh có kích thước lớn giữa các khung hình sẽ khiến vấn đề bộ nhớ tệ hơn, chứ không tốt hơn.

Độ chính xác: mọi bộ giải mã phải tính ra cùng các bit

Mã hóa entropy chỉ hoạt động nếu mô hình xác suất của bộ giải mã khớp chính xác với của bộ mã hóa. Trong codec truyền thống, mô hình đó là một bảng hoặc một tập nhỏ quy tắc số nguyên. Trong codec nơ-ron, nó là một

mạng nơ-ron, và một khác biệt dấu phẩy động giữa hai thiết bị, do kernel khác, thứ tự cộng dồn khác hay NPU khác, có thể khiến bộ giải mã đọc sai ký hiệu và mất luồng từ điểm đó trở đi.

Bất kỳ ai đã triển khai hay chuẩn hóa một codec nơ-ron đều phải dùng kỹ thuật để loại bỏ vấn đề này. DCVC-RT chuyển mô hình của nó sang số nguyên để mã hóa nhất quán giữa các thiết bị. MLVC dùng thiết kế hyperprior dùng chung thang đo để chịu được sai lệch giữa các nền tảng, điều có lẽ là một phần lý do khiến nó mất khả năng nén so với DCVC-RT phân tích. Nhóm thăm dò mạng nơ-ron của JVET chạy mọi mạng qua SADL, một thư viện nhỏ suy luận bằng số nguyên 16 bit. JPEG AI định nghĩa các điểm tuân thủ trong miền tiềm ẩn, miền phần dư và miền pixel. Với một đội thiết kế chip, đây chính là kỷ luật của các phép biến đổi số nguyên trong HEVC: tiêu chuẩn phải định nghĩa chính xác phép tính, điều này ràng buộc thiết kế mạng và loại trừ các lối tắt riêng của từng nhà cung cấp.

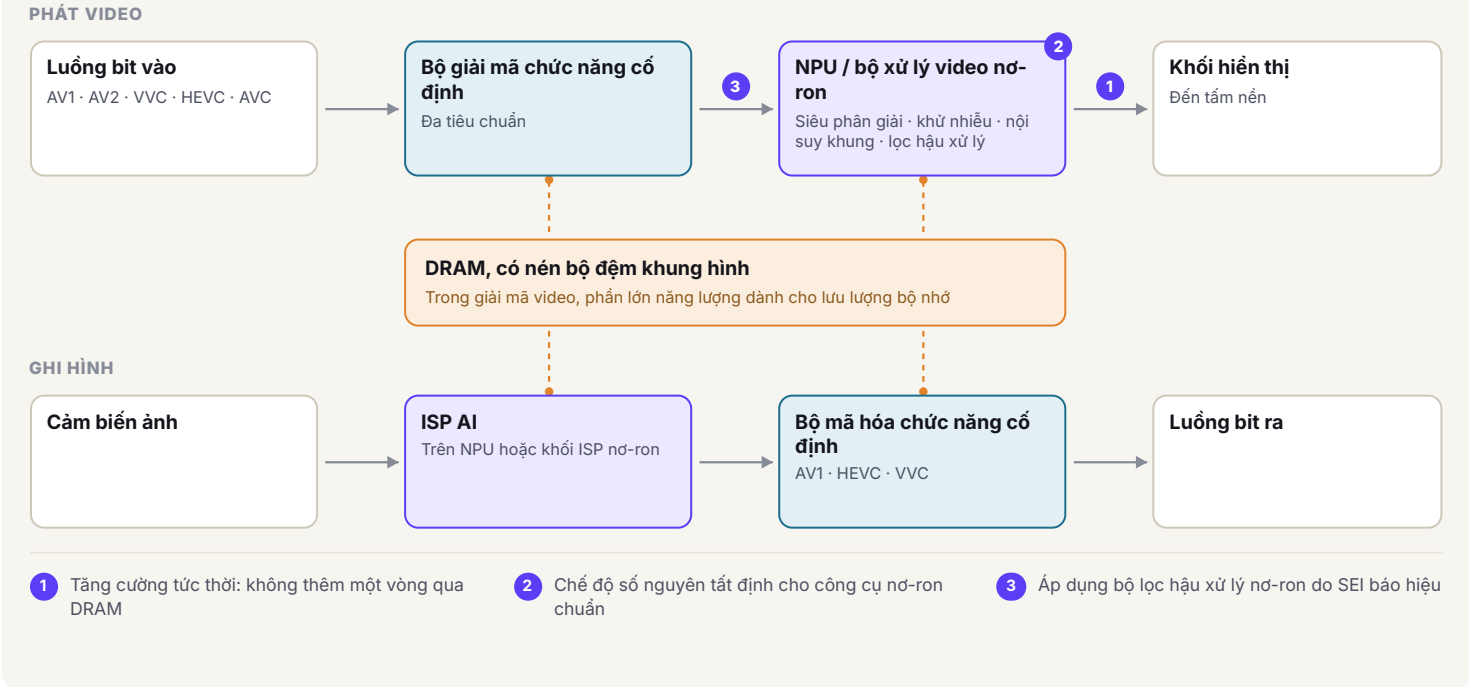
Diện tích không phải trở ngại; kiểm chứng và vòng đời mới là trở ngại

5 triệu cổng của NVCA so với khoảng một triệu của một bộ giải mã VVC theo dự phóng trông có vẻ lớn, nhưng trên một SoC điện thoại hay TV hiện đại vốn đã mang một NPU nhiều TOPS, vài triệu cổng không phải là điều quyết định thiết kế. Thứ thực sự tốn tiền là kiểm chứng với một bộ kiểm tra tuân thủ, giữ cho một bộ giải mã lập trình được luôn đúng qua các bản cập nhật phần mềm trong suốt vòng đời thiết bị, và công suất ở 4K60 HDR. Những chi phí đó chỉ có ý nghĩa khi một tiêu chuẩn đã cố định những gì phải xây dựng.

Kiến trúc đang thắng thế: codec chức năng cố định cộng NPU

HÌNH 08 **Phân chia đã vào sản phẩm: phần cứng chức năng cố định cho luồng bit, khối xử lý nơ-ron cho pixel**

Một SoC đa phương tiện năm 2026. Khối codec giải mã mọi tiêu chuẩn với chưa đến một nanojoule mỗi pixel; NPU làm những việc mạng nơ-ron giỏi và tiêu chuẩn không ràng buộc. Lưu lượng DRAM mới là chi phí quan trọng.



Nguồn: Góc nhìn kiến trúc của tác giả, dựa trên thông tin sản phẩm do Tencent (Canghai V2), Chips&Media, Allegro DVT và Samsung công bố (2026).

Những con chip bán được trong năm 2026 kết hợp một codec đa tiêu chuẩn chức năng cố định với một khối xử lý nơ-ron đặt bên cạnh. Khối codec xử lý AV1, AV2, VVC, HEVC và AVC với chưa đến một nanojoule mỗi pixel. Khối xử lý nơ-ron đảm nhận những việc mạng nơ-ron giỏi và tiêu chuẩn không ràng buộc: khử nhiễu và ISP ở phía ghi

hình, siêu phân giải, nội suy khung hình và tăng cường chất lượng ở phía hiển thị. Hai khối dùng chung cơ chế nén bộ đệm khung hình và lịch truy cập DRAM, vì lưu lượng bộ nhớ là chi phí quan trọng.

Thị trường cho thấy rõ sự phân chia này. ASIC codec video Canghai (沧海) của Tencent, tự phát triển từ năm 2019, đã có hơn 100.000 chip thế hệ đầu được triển khai trên Tencent Cloud và các dịch vụ phát trực tiếp, theo yêu cầu và trò chơi trên đám mây của chính hãng; thế hệ thứ hai, được công bố vào tháng 5/2026 là đã cấp nguồn chạy thử và bước vào sản xuất hàng loạt để đưa vào phục vụ trong nửa cuối năm, hỗ trợ H.265 và H.266 trong một kiến trúc, với khả năng nén tốt hơn thế hệ trước lần lượt hơn 10% và 30%, thông lượng gấp khoảng đôi, và tăng cường hình ảnh trên chip như siêu phân giải. Chips&Media bán song song IP codec và một NPU hướng video, WAVE-N. Allegro DVT bán bộ giải mã AV2 và đã bổ sung IP xử lý video nơ-ron đầu tiên của hãng. Các bộ xử lý TV của Samsung kết hợp bộ giải mã chuẩn với hàng trăm mạng nâng độ phân giải.

Tôi sẽ xây dựng gì, theo từng vai trò

Đội SoC và chip TV. Duy trì bộ giải mã chức năng cố định cho mọi tiêu chuẩn phải phát, gồm AV2 ngay từ bây giờ và VVC ở nơi phát sóng yêu cầu. Định cỡ NPU cho hậu xử lý ở độ phân giải và tốc độ khung hình hiển thị, và thiết kế đường đi từ bộ giải mã qua NPU đến màn hình sao cho bước tăng cường không thêm một vòng qua DRAM. Hỗ trợ SEI cho bộ lọc hậu xử lý nơ-ron. Cung cấp các chế độ số nguyên tất định trong NPU. Coi H.267 là điểm khởi đầu thiết kế từ năm 2029 trở đi, và giữ lại một phần khả năng lập trình cho các công cụ nơ-ron của nó.

Nhà cung cấp IP. Bán bộ giải mã AV2 và VVC ngay bây giờ, cùng những thành phần mà xử lý nơ-ron cần quanh chúng: bộ tăng tốc lọc hậu xử lý, nén bộ đệm khung hình và công cụ kiểm tra tuân thủ.

Nền tảng video. Tiền xử lý nơ-ron ở bộ mã hóa hoàn vốn trước tiên, vì nó chạy một lần cho mỗi nội dung và không cần gì từ thiết bị. Dùng siêu phân giải phía máy khách ở nơi thiết bị cho phép. Dành codec nơ-ron đầu cuối cho các vòng khép kín mà bạn kiểm soát.

CHƯƠNG 09

TIÊU CHUẨN VÀ BẰNG SÁNG CHẾ

09 Tiêu chuẩn và bằng sáng chế: nơi công cụ nơ-ron trở thành đối tượng cấp phép

Lộ trình các tiêu chuẩn

Tiêu chuẩn	Tổ chức	Thành phần nơ-ron	Tình trạng, tháng 11/2026
H.267 (tên tạm; thế hệ sau VVC)	JVET (ITU-T và ISO/IEC)	Còn mở. Công cụ truyền thống (ECM) và công cụ nơ-ron (bộ lọc trong vòng lặp, dự đoán nội khung, siêu phân giải) đều là ứng viên	Kêu gọi đề xuất tháng 7/2026; đánh giá hồ sơ tháng 1/2027; mô hình thử nghiệm đầu tiên nửa cuối năm 2027; phiên bản 1 cuối năm 2029
VVC và HEVC, với VSEI	JVET	Bộ lọc hậu xử lý nơ-ron có thể được mô tả trong siêu dữ liệu SEI; việc áp dụng là tùy chọn với bộ giải mã	Đang có hiệu lực; các thông điệp được mở rộng trong phiên bản tháng 10/2025
AV2	AOMedia	Công cụ truyền thống, miễn phí bản quyền	Phiên bản 1.0.0 ngày 28/5/2026; IP giải mã được cấp phép từ tháng 6/2026; YouTube dự kiến phát trên web đầu năm 2027

Tiêu chuẩn	Tổ chức	Thành phần nơ-ron	Tình trạng, tháng 11/2026
JPEG AI	JPEG (ISO/IEC và ITU-T)	Codec ảnh học máy đầu cuối, cho thị giác của người và máy	Phần 1 ban hành năm 2025; phiên bản thứ hai đang được thảo luận; chưa thấy sản phẩm hỗ trợ mặc định
IEEE 1857.11	IEEE	Mã hóa ảnh dựa trên mạng nơ-ron, thiết kế cho NPU	Ban hành tháng 12/2024
Video Coding for Machines	MPEG	Mã hóa cho máy phân tích thay vì để xem	Dự thảo tiêu chuẩn quốc tế đầu năm 2026; dự kiến hoàn tất cuối năm 2026
Phần mở rộng ML của Opus	IETF	Dư thừa sâu, tăng cường phía giải mã, mở rộng băng thông, đều là tùy chọn	Có trong libopus 1.5 và 1.6; phần dư thừa sâu đang được chuẩn hóa

Có hai điều nổi bật. Thứ nhất, thành phần nơ-ron đi vào tiêu chuẩn ở phần rìa trước khi vào phần lõi: dưới dạng bộ lọc hậu xử lý tùy chọn, tăng cường tùy chọn và mã hóa ảnh, tất cả đều giữ nguyên đường giải mã chính. Thứ hai, tiêu chuẩn duy nhất có thể đưa công cụ nơ-ron vào đường giải mã của mọi màn hình, H.267, còn cách phiên bản đầu tiên ba năm.

Ai đang có vị thế về bằng sáng chế

Công cụ trong tiêu chuẩn trở thành bằng sáng chế có thể cấp phép, nên quá trình H.267 cũng là một cuộc đua bằng sáng chế, và các công cụ nơ-ron là một phần của cuộc đua đó. Bức tranh toàn cảnh năm 2025 của GreyB về 3.916 bằng sáng chế thế hệ sau VVC đặt ByteDance ở vị trí dẫn đầu, với 645 họ bằng sáng chế và 251 đóng góp cho JVET, đồng thời ghi nhận các quan hệ hợp tác giữa đại học và doanh nghiệp như Xiaomi với Đại học Xidian, với hơn 20 họ bằng sáng chế. Tháng 3/2026, MLex đưa tin Tencent, Alibaba và ByteDance đang nổi lên như những thành viên tích cực trong các nhóm bằng sáng chế codec video. Tháng 5/2026, Sauvegarder Investment Management mua 126 tài sản bằng sáng chế codec video từ Alibaba, bao trùm AV1, AV2 và Video Coding for Machines. Thương vụ InterDigital mua Deep Render vào tháng 10/2025 đã bổ sung các bằng sáng chế mã hóa video AI của Deep Render vào danh mục của hãng.

Số tiền theo từng thiết bị hiện rõ trong bảng phí của các nhóm bằng sáng chế. Mức phí của Access Advance cho VVC (tính đến ngày 1/7/2026, cho bên được cấp phép tuân thủ quy định và sử dụng nhãn hiệu của tổ chức này) là 0,50 USD mỗi điện thoại di động, máy tính bảng hoặc laptop tại các thị trường Khu vực 1 như Mỹ, EU và Nhật Bản, và 0,25 USD tại Khu vực 2; với TV 4K, từ 1,00 đến 1,50 USD tùy giá bán tại Khu vực 1; kèm mức trần hằng năm cho mỗi doanh nghiệp là 75 triệu USD. AV1 và AV2 được các thành viên AOMedia cung cấp miễn phí bản quyền. Dù H.267 chứa gì, các công cụ của nó cũng sẽ được định giá theo cách đó, và đó là lý do các công ty có bằng sáng chế mã hóa nơ-ron đang đóng góp cho JVET ngay từ bây giờ phân tích.

CHƯƠNG 10

AI KIẾM ĐƯỢC TIỀN

10 Ai kiếm được tiền, và doanh nghiệp có thể tham gia ở đâu

Dòng tiền năm 2026

HÌNH 09 **Ai kiếm được tiền từ mã hóa nơ-ron năm 2026**

Tiền đi theo các tiêu chuẩn và nền tảng. Tôi không tìm thấy công ty độc lập nào sống được nhờ cấp phép một codec video nơ-ron đầu cuối.

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)

Nguồn: Thông tin công bố của các công ty và nhóm bằng sáng chế được trích dẫn trong báo cáo; tình trạng tính đến ngày 12/11/2026. Phí của nhóm: Access Advance, VVC, Khu vực 1, bên tuân thủ và dùng nhãn hiệu, theo biểu phí ngày 1/7/2026.

Nền tảng giữ lại phần tiết kiệm. Netflix, YouTube, Meta, Microsoft và Tencent tự xây dựng xử lý nơ-ron và giữ lại lợi ích dưới dạng băng thông thấp hơn, dung lượng lưu trữ ít hơn và chất lượng trên mỗi bit tốt hơn. Khi việc đó giúp thúc đẩy áp dụng, họ cho không mã nguồn: Microsoft phát hành MLVC theo giấy phép MIT.

Hãng thiết bị và chip bán sự khác biệt. ISP camera và bộ xử lý hình ảnh của TV là nơi mạng nơ-ron chạm tới mọi người dùng, và chúng được bán như các tính năng: Cognitive ISP của Qualcomm, các bộ xử lý nâng độ phân giải của Samsung.

Nhà cung cấp IP bán bộ giải mã cho mỗi tiêu chuẩn mới, kèm NPU. Trong vòng ba tuần sau khi có đặc tả AV2, Chips&Media, Allegro DVT và VeriSilicon đều đã công bố khách hàng cho bộ giải mã AV2. Chính những nhà cung cấp này nay bán các khối xử lý nơ-ron bên cạnh codec.

Hãng GPU xây dựng nền tảng. DLSS và nén texture nơ-ron biến phần cứng tensor của NVIDIA thành một phần trong cách làm ra trò chơi.

Chủ sở hữu bằng sáng chế thu phí theo thiết bị, thông qua các nhóm bằng sáng chế cho VVC và HEVC và, sau năm 2029, cho H.267.

Các startup codec nơ-ron đã được mua lại. WaveOne về tay Apple năm 2023, iSIZE về Sony Interactive Entertainment năm 2023, và Deep Render về InterDigital năm 2025, đều không công bố điều khoản. Trong mỗi trường hợp, một nền tảng hoặc một bên cấp phép đã mua lại đội ngũ và bằng sáng chế. Tôi không tìm thấy công ty độc lập nào trong năm 2026 sống được nhờ cấp phép một codec video nơ-ron đầu cuối.

Doanh nghiệp có thể tham gia ở đâu

Hướng đi	Ai mua	Vì sao là lúc này	Rủi ro chính
Tiền xử lý nơ-ron ở bộ mã hóa (giảm độ phân giải, khử nhiễu, tinh chỉnh theo từng nội dung)	Nền tảng phát trực tuyến và video người dùng, đài truyền hình	Không cần thay đổi thiết bị; đã được chứng minh ở Netflix	Các nền tảng lớn tự xây dựng
IP và tinh chỉnh tăng cường phía hiển thị (siêu phân giải, nội suy khung hình, chất lượng hình ảnh)	Hãng chip cho TV, đầu thu và điện thoại	NPU đã có sẵn trong thiết bị; khác biệt dễ thấy	Bị hàng hóa hóa; hãng chip tự làm
Codec nơ-ron vòng khép kín (hội nghị truyền hình, vận hành từ xa, liên kết riêng)	Hãng giải pháp liên lạc doanh nghiệp, công nghiệp và hiện diện từ xa	MLVC chứng minh tính khả thi, và mã nguồn đã mở	Phải kiểm soát cả hai đầu; NPU phân mảnh
Chip hoặc IP codec cộng NPU (AV2, VVC và hậu xử lý)	Hãng SoC	Chu kỳ AV2 đang bắt đầu	Các hãng đương nhiệm mạnh
Công cụ 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

Với phần lớn doanh nghiệp, hai hàng đầu tiên là nơi có doanh thu trong năm 2027. Hàng thứ ba thú vị nhất với một công ty sản phẩm tập trung. Hàng cuối cùng đòi hỏi kiên nhẫn và một chiến lược cấp phép.

CHƯƠNG 11

11 TUYÊN BỐ VÀ BẰNG CHỨNG
Những tuyên bố bạn sẽ nghe, và bằng chứng nói gì

Tuyên bố	Bằng chứng cho thấy gì
"Codec nơ-ron đã vượt VVC."	Trong phòng thí nghiệm thì đúng. DCVC-RT tiết kiệm 21% so với bộ mã hóa tham chiếu của VVC ở 1080p trong thiết lập của bài báo; JPEG AI tiết kiệm 12,5–27,9% so với VVC nội khung. Bộ giải mã tốn khoảng 175 nghìn MAC mỗi pixel, tức năng lượng của một NPU, so với một phần nhỏ của một nanojoule trên phần cứng chức năng cố định.
"Codec AI sẽ thay thế H.26x và AV1 trong vài năm tới."	Không phải trong phát video. Tiêu chuẩn đầu tiên có thể mang công cụ nơ-ron đặt mục tiêu cuối năm 2029, chip theo sau vào khoảng năm 2031 và phổ cập vào giữa thập niên 2030 ước tính. Codec đang đến lúc này, AV2, là codec truyền thống.
"MLVC giảm 88% băng thông."	So với H.264 ở 360p, trên bộ thử nghiệm của chính Microsoft. So với HEVC phần cứng, mức đó là 75,5% ở 360p và 65,4% ở 540p, trong các cuộc gọi mà Microsoft kiểm soát cả hai đầu.
"Texture nơ-ron giảm 85–96% bộ nhớ video."	Trong các cảnh trình diễn. Trò chơi đã phát hành đạt mức nén hiệu dụng tốt hơn khoảng 30% và chỉ dùng texture nơ-ron cho một số tài nguyên chọn lọc vì chi phí khi chạy.
"JPEG AI là tiêu chuẩn ISO, nên camera và trình duyệt sẽ áp dụng."	Tôi không tìm thấy sản phẩm nào hỗ trợ mặc định. Định dạng web chọn trong năm 2026, JPEG XL trong Chrome 155, là định dạng truyền thống.
"Giải mã nơ-ron quá đắt cho điện thoại."	Đúng một phần. MLVC giải mã 540p30 trong phạm vi một nửa NPU; 4K60 với một mô hình cỡ DCVC-RT sẽ cần khoảng 175 TOPS năng lực tính toán liên tục ước tính.
"Liên kết vệ tinh và drone là thị trường lớn cho video nơ-ron."	Phù hợp nhất về kỹ thuật, nhưng nhỏ và ngày càng bị kiểm soát xuất khẩu ở cả hai bờ Thái Bình Dương.
"Các nền tảng lớn sẽ mua giấy phép codec nơ-ron từ một startup."	Họ tự xây dựng (Netflix, Microsoft, Tencent) hoặc mua luôn startup (Apple, Sony, InterDigital).

Một phép thử đơn giản giúp nhìn thấu phần lớn các lời chào hàng: 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 một kết quả nghiên cứu, không phải sản phẩm.

CHƯƠNG 12

12

CÁC MỐC CẦN THEO DÕI

Những gì cần theo dõi, 2027 đến 2030

Thời điểm	Mốc	Vì sao quan trọng
Tháng 1/2027	JVET đánh giá các hồ sơ H.267 tại Brisbane	Bao nhiêu hồ sơ hàng đầu dựa vào công cụ nơ-ron, và với chi phí mỗi pixel bao nhiêu
Đầu năm 2027	Những luồng AV2 trên web đầu tiên của YouTube; bộ giải mã AV2 trong các SoC đã công bố	Chu kỳ truyền thống tiếp theo định ra ngân sách thiết bị cho đến khoảng năm 2030
Nửa cuối năm 2027	Mô hình thử nghiệm đầu tiên của H.267	Công cụ nơ-ron nằm trong thiết kế lõi hay chỉ ở mức thăm dò
2027	MLVC ra ngoài cuộc gọi ngang hàng trên Teams; các nhà cung cấp khác áp dụng mã nguồn mở này	Phép thử đầu tiên của một codec nơ-ron mở bên ngoài chủ sở hữu của nó
2027–2028	Sản phẩm đầu tiên ghi hoặc đọc JPEG AI theo mặc định	Định dạng nơ-ron đầu tiên được dùng xuyên ranh giới giữa các công ty
2027–2028	DirectX Linear Algebra trong các driver chính thức; texture nơ-ron trên nhiều trò chơi hơn và trên thiết bị di động	Giải mã nơ-ron liên hãng bên trong đồ họa
Liên tục	Ngân sách NPU trong điện thoại và PC	Ở mức khoảng 100 TOPS liên tục, một bộ giải mã 1080p30 cỡ DCVC-RT sẽ chiếm khoảng một phần năm NPU thay vì một nửa phân tích
Cuối năm 2029	H.267 phiên bản 1	Khởi đầu chu kỳ chip cho công cụ nơ-ron trong phát video đại chúng

CHƯƠNG 13

13 Câu hỏi mở

Một số câu hỏi trong báo cáo này chưa có câu trả lời công khai, và đó là những câu tôi thấy thú vị nhất về mặt kỹ thuật.

Độ chính xác ở quy mô lớn. Một bộ giải mã nơ-ron có thể tiến gần đến mức nào tới số học chỉ dùng số nguyên, chính xác đến từng bit, trên NPU của các hãng khác nhau trước khi đánh mất phần lớn lợi thế nén? Mức đánh đổi sáu điểm của MLVC là một điểm dữ liệu; một tiêu chuẩn cần câu trả lời tổng quát.

Bộ nhớ trước MAC. Hậu xử lý nơ-ron tốn bao nhiêu lưu lượng DRAM mỗi pixel khi chạy tức thời giữa bộ giải mã phần cứng và màn hình, và tập dữ liệu làm việc của nó có thể thu nhỏ đến mức nào?

Giá của một công cụ nơ-ron trong H.267. Ở bao nhiêu MAC và nanojoule mỗi pixel thì một bộ lọc nơ-ron trong vòng lặp không còn bù đắp được bằng lượng bit nó tiết kiệm trong một bộ giải mã 4K60?

Vòng khép kín ngoài cuộc gọi. Những hệ thống khép kín nào khác, từ lưu trữ ảnh đám mây và phát trực tuyến trò chơi đến robot và liên kết truyền dẫn riêng, có thể áp dụng codec nơ-ron đầu cuối trước khi H.267 xuất hiện?

Bộ giải mã biết tạo sinh. Khi bộ giải mã tự tạo chi tiết thay vì tái tạo, như VAE video đã làm, nên đo lường và đặc tả độ trung thực như thế nào?

Đây là những lĩnh vực chúng ta có thể cùng khám phá trong những năm tới, qua trao đổi kỹ thuật và hợp tác kỹ thuật: i@tanshuai.com.

Giới hạn của báo cáo này

Đây là nghiên cứu tài liệu dựa trên các văn bản công khai đến ngày 12/11/2026. Tôi không thực hiện phòng vấn trực tiếp hay phép đo độc lập nào cho ấn bản này. Một số con số quan trọng là tuyên bố của nhà cung cấp hoặc của tác giả trên bộ thử nghiệm riêng của họ, như mức tiết kiệm bitrate của MLVC và tỷ lệ nén texture nơ-ron; chúng được gắn nhãn tương ứng. Các ước tính tính toán của tôi cho DCVC-RT cộng các số liệu theo từng mô-đun mà nhóm tác giả công bố và bỏ qua tổn hao hiệu suất sử dụng, nên tải NPU thực tế sẽ cao hơn. Phép so sánh năng lượng dựa trên một vài thiết kế đã công bố ở các tiến trình khác nhau; số liệu của bộ giải mã VVC và AV2 thương mại là thông tin độc quyền. Lộ trình tiêu chuẩn là mục tiêu do chính các ủy ban đặt ra. Các mốc H.267 sau năm 2029 là ước tính của tôi dựa trên lịch sử của AV1 và VVC.

Về tác giả

Tan Shuai là chuyên gia công nghệ và nhà sáng lập, với hơn hai mươi năm làm việc trong các lĩnh vực chip, phần cứng và phần mềm hệ thống, và 20 bằng sáng chế đã được cấp. Ông có nhiều năm nghiên cứu, phát triển và quản lý chip fabless, bao gồm thiết kế đường dữ liệu (datapath) ASIC full-custom và các dự án chip camera đa phương tiện như bộ tăng tốc codec video, và đó cũng là lăng kính của báo cáo này: năng lượng mỗi pixel, lưu lượng bộ nhớ, và những gì một tiêu chuẩn cho phép bộ giải mã cố định trên silicon. Công việc đã đưa ông qua Thung lũng Silicon, New York, Chicago, Đài Loan, Vùng Vịnh Lớn (Thâm Quyến và Hồng Kông), Việt Nam và Malaysia. Hiện ông làm việc với vai trò nhà nghiên cứu và phát triển AI.

Trao đổi kỹ thuật: i@tanshuai.com · tanshuai.com

Ghi chú và tuyên bố miễn trừ

Báo cáo này là nghiên cứu tài liệu độc lập dựa trên các văn bản công khai có đến ngày 12/11/2026. Báo cáo chỉ nhằm mục đích cung cấp thông tin và không phải là tư vấn đầu tư, pháp lý hay kỹ thuật. Sản phẩm, lộ trình tiêu chuẩn, mức phí của các nhóm bằng sáng chế và quy định xuất khẩu đều có thể thay đổi; hãy đối chiếu mọi thông tin bạn dựa vào với nguồn gốc. Các số liệu gắn nhãn theo nhà cung cấp, ước tính hoặc phân tích cần được đọc đúng như vậy. Tên công ty và sản phẩm thuộc về chủ sở hữu của chúng và chỉ được dùng để nhận diện; không công ty nào thẩm định hay tài trợ cho báo cáo này.

© 2026 Tan Shuai. Bạn có thể chia sẻ báo cáo này nguyên vẹn, kèm ghi nguồn.

Nguồn tham khảo

Tất cả được truy cập trong khoảng từ ngày 3/11 đến ngày 12/11/2026. Sắp xếp theo chương.

Lưu lượng, thiết bị và phạm vi

- Ericsson, *Mobility Report*, tháng 6/2026 (video chiếm khoảng 75% lưu lượng dữ liệu di động vào cuối năm 2025): https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/june-2026 ; bản tóm tắt: 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, lượng xuất xưởng điện thoại thông minh toàn cầu, quý 4 và cả năm 2025 (1,26 tỷ), ngày 13/1/2026: https://www.idc.com/resource-center/press-releases/q425mobilephonetop5/
- TrendForce, dự báo lượng xuất xưởng TV năm 2025 (195,71 triệu): 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, dự báo TV năm 2025 (208,7 triệu): https://iconnect007.com/article/145954/omdia-forecasts-flat-tv-shipments-in-2025-but-mini-led-poised-for-explosive-growth-at-829/145951/ein

Phát video và phát trực tuyến

- Netflix Technology Blog, "AV1 — Now Powering 30% of Netflix Streaming", tháng 12/2025: 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", tháng 11/2022: 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, ngày 28/5/2026: 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" (tỷ lệ AV1 trong kho nội dung của 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" (DTV+ đi vào vận hành thương mại, tháng 6/2026): https://mpegh.com/brazil/ ; CartaCapital về đầu thu tích hợp trong TV: https://www.cartacapital.com.br/?p=628337 ; Olhar Digital về VVC trong 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", ngày 29/10/2025: 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, ngày 16/10/2025: https://atscaleconference.com/how-meta-deployed-super-resolution-at-scale-to-transform-video-quality/
- Samsung, 8K TV highlights (NQ8 AI Gen3, 768 mạng nơ-ron): https://www.samsung.com/us/tvs/8k-tv/highlights/

Nhiếp ảnh và hình ảnh

- 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 và Visionary.ai, ISP hoàn toàn dựa trên AI, ngày 5/1/2026: 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/ ; phỏng vấn trên PetaPixel: https://petapixel.com/2026/01/18/a-look-at-the-worlds-first-full-ai-based-image-signal-processor
- Ủy ban JPEG, thông cáo báo chí cuộc họp lần thứ 103 (mức cải thiện, độ phức tạp và các bản triển khai trên điện thoại thông minh của JPEG AI): https://jpeg.org/items/20240507_press.html
- G. J. Sullivan, giới thiệu về ISO/IEC JTC 1/SC 29, tháng 4/2026 (tình trạng JPEG AI; số liệu ECM và 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/
- Nhóm công tác IEEE 1857.11 (mã hóa ảnh dựa trên mạng nơ-ron): https://sagroups.ieee.org/fvc
- Chrome 155 và JPEG XL, ngày 6/10/2026: https://xenospectrum.com/en/chrome-155-jpeg-xl-rust/

Cuộc gọi và họp trực tuyến

- Microsoft, "Satin: Microsoft's latest AI-powered audio codec for real-time communications", ngày 17/2/2021: 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)", ngày 24/7/2026: https://techcommunity.microsoft.com/blog/linuxandopensourceblog/announcing-the-open-source-release-of-ml-video-codec-mlvc/4539875 ; mã nguồn: 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 ; phân tích của iReplay: https://ireplay.tv/blog/mlvc-microsoft-neural-video-codec-open-source-streaming/

- Ghi chú phát hành 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", ngày 13/6/2024: https://engineering.fb.com/2024/06/13/web/mlow-metas-low-bitrate-audio-codec/
- Microsoft, Copilot+ PC (NPU từ 40 TOPS trở lên): https://www.microsoft.com/en-us/windows/copilot-plus-pcs ; Apple, thông báo ra mắt M4 (Neural Engine 38 TOPS): https://www.apple.com/newsroom/2024/05/apple-introduces-m4-chip/

AI tạo sinh

- Kyutai, thẻ mô hình Mimi (12,5 Hz, 1,1 kbps): https://huggingface.co/kyutai/mimi
- Lightricks, "LTX-Video: Realtime Video Latent Diffusion" (VAE video 1:192): https://arxiv.org/abs/2501.00103
- Thông cáo báo chí cuộc họp MPEG lần thứ 153 (Video Coding for Machines ở giai đoạn DIS): https://www.mpeg.org/wp-content/uploads/2026/03/press_release_153.pdf

Trò chơi và đồ họa

- NVIDIA, mức độ áp dụng DLSS 4: khi ra mắt và tại Computex 2025 (75 và hơn 125): https://www.nvidia.com/en-us/geforce/news/125-dlss-4-multi-frame-gen-games-more-announced-computex-2025/ ; Gamescom 2025 (hơn 175): https://www.nvidia.com/en-us/geforce/news/gamescom-2025-dlss-4-full-ray-tracing-game-announcements/ ; CES 2026 (hơn 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", tháng 2/2026: https://www.ubisoft.com/en-us/studio/laforge/news/415slmB3ZGzv8d2LhO9QT/shipping-neural-texture-compression-in-assassins-creed-mirage
- TweakTown về bản demo texture nơ-ron của NVIDIA tại GTC 2026 (từ 6,5 GB xuống 970 MB): https://www.tweaktown.com/news/110886/nvidias-neural-texture-compression-cuts-vram-usage-from-6-5gb-down-to-970mb/index.html
- Blog Microsoft DirectX, Shader Model 6.9 bản chính thức, ngày 26/2/2026: https://devblogs.microsoft.com/directx/shader-model-6-9-retail-and-more/ ; về cooperative vector: https://devblogs.microsoft.com/directx/shader-model-6-9-and-the-future-of-cooperative-vector/ ; kế hoạch tại GDC 2026: https://windowsreport.com/microsoft-showcases-neural-rendering-and-dxr-2-0-plans-for-directx-at-gdc-2026/

Liên kết chuyên dụng

- Global Trade Alert về Thông báo số 49 năm 2026 của Trung Quốc: 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 (ngưỡng cho nền tảng mặt đất không người lái): 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 đưa drone và linh kiện sản xuất ở nước ngoài vào Covered List, ngày 22/12/2025: https://dronelife.com/2025/12/22/fcc-adds-foreign-made-drones-and-components-to-covered-list-citing-national-security-risks/

Chip

- M. Tikekar và cộng sự, "A 249-Mpixel/s HEVC Video-Decoder Chip for 4K Ultra-HD Applications", IEEE JSSC 2014: https://dspace.mit.edu/handle/1721.1/93876 ; trang chip của MIT (78 mW ở 4K30): https://eecsg.mit.edu/chips/hevc-video-decoder-for-4k-ultra-hd-applications/
- S. Zhang và cộng sự, 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 và cộng sự, "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 ; tài liệu bổ sung (độ phức tạp từng mô-đun): https://openaccess.thecvf.com/content/CVPR2025/supplemental/Jia_Towards_Practical_Real-Time_CVPR_2025_supplemental.pdf
- NNVC và thư viện suy luận số nguyên SADL: https://arxiv.org/abs/2309.05846
- Tencent Canghai V2, ngày 27/5/2026 (Shanghai Securities News qua 10jqka): https://stock.10jqka.com.cn/20260527/c677016020.shtml ; tóm tắt tính năng: https://www.besthub.dev/articles/tencent-s-canghai-v2-chip-enters-mass-production-what-the-new-capabilities-mean-for-video-encoding-bcfc18dbfa29
- Chips&Media, IP giải mã AV2, ngày 12/5/2026: https://www.design-reuse.com/news/202530501-chips-media-completes-development-of-next-gen-av2-hw-decoder-ip/ ; giấy phép AV2, ngày 17/6/2026: https://www.edge-ai-vision.com/?p=58270
- Allegro DVT, IP giải mã AV2, ngày 16/6/2026: https://www.edge-ai-vision.com/2026/06/allegro-dvts-pulsar-decoder-ip-adds-support-for-av2-video-codec/
- VeriSilicon, VC9800D hỗ trợ AV2, ngày 10/6/2026: https://www.businesswirechina.com/en/news/63176.html

Tiêu chuẩn và bằng sáng chế

- Thông cáo báo chí cuộc họp MPEG lần thứ 155 (lời kêu gọi đề xuất chung cho thế hệ sau VVC), ngày 5/8/2026: https://www.mpeg.org/wp-content/uploads/2026/08/ag3w254SN26474Press-Release-MPEG155.pdf
- Thông cáo báo chí cuộc họp MPEG lần thứ 152 (kết quả kêu gọi bằng chứng; các phiên bản VVC và VSEI), ngày 30/10/2025: https://www.mpeg.org/wp-content/uploads/2025/11/Press-Release-MPEG152.pdf ; chuyên mục MPEG của 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, bức tranh bằng sáng chế thế hệ sau VVC (H.267): https://greyb.com/blog/exclusive-video-codec-landscape/ ; về cấp phép AV2 và thương vụ bán danh mục bằng sáng chế của Alibaba: https://greyb.com/blog/av2-licensing/
- MLex, "Fragmented codec licensing exposes need for clearer rules for Chinese tech majors", ngày 27/3/2026: https://www.mlex.com/mlex/articles/2458402/fragmented-codec-licensing-exposes-need-for-clearer-rules-for-chinese-tech-majors
- Access Advance, mức phí bản quyền MCBA và VVC Advance tính đến ngày 1/7/2026: https://accessadvance.com/wp-content/uploads/2025/07/MCBA-Program-Overview-September-2026.pdf

Tiền và các công ty

- InterDigital mua lại Deep Render, ngày 30/10/2025: https://www.advanced-television.com/2025/10/30/interdigital-acquires-deep-render
- TechCrunch, "Apple acquired a startup using AI to compress videos", ngày 27/3/2023: https://techcrunch.com/2023/03/27/apple-acquired-a-startup-using-ai-to-compress-videos
- Neowin, "Sony acquires AI video startup iSIZE", tháng 11/2023: https://www.neowin.net/news/sony-acquires-ai-video-startup-isize-to-enhance-playstation-streaming/

NÉN BẰNG HỌC MÁY · PHÁT VIDEO × CHIP

Mã hóa nơ-ron năm 2026

Triển khai quanh codec, chứ không thay cho codec

Mọi con số quan trọng trong báo cáo này đều kèm nguồn và thời điểm. Con số do tác giả tự suy ra được gắn nhãn “ước tính”; con số chỉ do bên bán công bố được gắn nhãn “theo nhà cung cấp”.

Tan Shuai

Chuyên gia công nghệ và nhà sáng lập · Nhà nghiên cứu và phát triển AI

tanshuai.com · i@tanshuai.com

ẤN BẢN CÔNG KHAI · Số liệu tính đến ngày 12/11/2026