EDGE RESEARCH · TECHNICAL REPORT 01

Fingerprinting live TV in 72 bytes

How a television can say exactly which moment it is showing, on any input, while the picture itself never leaves the set.

Edge Research, Edge Video AI Version 1.0 · 15 Nov 2026

Status: Edge Research technical report. Not peer reviewed. It describes Edge's own systems and internal measurements; every number carries its scope. Third-party work is cited with links.

ABSTRACT

We describe the visual fingerprint Edge Video AI uses to recognise the exact moment playing on a connected TV. The TV turns a captured frame into a 576-bit perceptual fingerprint, 72 bytes in all, and sends only those bytes, about every five seconds while a listed programme is on. The picture never leaves the TV. On our side, the fingerprint is compared with clips the vision language model (VLM) has already fingerprinted, and the answer names the programme, the clip and the second. Because the fingerprint is computed from the rendered picture, the same method works whether the programme arrives over HDMI, free-to-air or through the TV's own apps. We set out the design goals, the fingerprint and matching method at a level suitable for publication, the system around it, the payload arithmetic at scale, and the limits we know about. Match-rate results from partner testing are held under partner confidentiality and are not reported here.

1. Introduction

A television is the one device in the home that always knows what is on the screen, and the one that usually cannot say it. Channel guides describe schedules, not moments. Streaming apps know what they are playing, but a set-top box on an HDMI cable or a broadcast arriving through an antenna is opaque to the TV's software. If you want something on the screen to respond to the moment, a recipe card when the dish is plated, a shopping prompt when the product appears, you first need to know which moment it is, to the second.

Automatic content recognition (ACR) solves the first half of that problem by comparing a compact signature of what is playing with signatures of known content. This report describes Edge Video AI's visual fingerprint: a 72-byte signature of one frame that a TV can compute in integer arithmetic and send in a single small request. It is the part of our system that runs on

the television. Everything expensive, including the vision language model, runs once per channel on our side.

We make three claims, each scoped carefully in the sections that follow:

- A 72-byte fingerprint of a frame is enough to locate the moment within registered content, while the image itself stays on the TV.
- Computing the fingerprint from the rendered picture makes recognition independent of the input: HDMI, free-to-air and built-in apps look the same to it.
- Separating perception (once per channel) from recognition (per TV) means model cost grows with the number of channels, not the number of screens.

2. Background and related work

Fingerprinting media is an old and well-studied problem. In audio, Haitsma and Kalker described a robust fingerprint built from the signs of energy differences across spectral bands and time, matched by bit error rate [1]. Wang's description of the Shazam system used constellations of spectrogram peaks hashed into pairs, which made search over very large catalogues fast and tolerant of noise [2]. Both papers set the pattern we follow: reduce the signal to a small, robust code; compare codes with a cheap distance; and push the cost of indexing to the server.

For video, Oostveen, Kalker and Haitsma proposed fingerprints built from the differences of mean luminance between blocks of a frame and between consecutive frames, with a database strategy for fast lookup [3]. Perceptual image hashing, such as the robust image hashing of Venkatesan et al., aims for codes that change little under compression, scaling and small edits, and change a lot between different images [4]. More recent work on image copy detection trains neural descriptors for the same purpose, as in the 2021 Image Similarity Challenge [5] and the self-supervised SSCD descriptor [6]. These learned descriptors are more robust, but they are larger and need more compute on the device than a television can always spare.

Search over fingerprints is a nearest-neighbour problem. Locality-sensitive hashing gives sub-linear approximate search for Hamming-like distances [7], and GPU libraries such as FAISS make exhaustive and approximate search practical at billion scale [8]. Our matching sits in this family.

What is different in our setting is the deployment constraint rather than the maths. The fingerprint has to run inside TV firmware on many different chips, produce identical bits on all of them, add almost nothing to network traffic, and never send an image. Those constraints shape every choice below.

3. Design goals

Goal	Why it matters	How the design meets it
Private by construction	Viewers should not have to trust a policy to know their picture stays at home.	Only a 72-byte code is sent. The image is discarded on the TV.
Tiny on the wire	Millions of TVs sending regularly must not become a traffic problem.	One 72-byte body per request, about every five seconds, only while a listed programme is on.
Identical on every chip	A fingerprint that differs by processor cannot be matched reliably.	Integer arithmetic only. No floating point, no hardware-specific maths.
Input-agnostic	Much live TV arrives over HDMI or an antenna, where the TV's software has no metadata.	Computed from the rendered picture, after the input has been decoded.
Cheap to match	Recognition must not need a model call per TV.	Bit-level distance against a prepared clip index; no AI runs at view time.

4. The fingerprint



Figure 1. One frame becomes 72 bytes (576 bits). Frame from Edge's explainer film; each tile stands for one byte.

The TV captures a frame of the picture it is showing and reduces it to a small, fixed-size brightness image. It then summarises the coarse layout of brightness over a fixed grid of regions. Each region contributes one bit: whether it is brighter or darker than a local reference computed from its neighbours. The result is 576 bits, which is 72 bytes, or 144 hexadecimal characters when written out.

Three properties follow from this construction. First, it is *robust to what does not matter*: changes in compression, a slightly different scaler, mild colour shifts and small overlays move few bits, because each bit depends on relative brightness over a whole region rather than on individual pixels. Second, it is *sensitive to what does*: a change of shot or of scene changes the layout of light and dark, and so changes many bits. Third, it is *deterministic*: because every step uses integer arithmetic, two different TV processors given the same picture produce the same 72 bytes.

Throughout this report we describe the fingerprint at the level above. The exact region layout, reference computation and weighting are part of Edge's own implementation and are not needed to follow the system design.

A useful way to picture the size: one hour of a listed programme, with one fingerprint about every five seconds, is 720 fingerprints, or about 52 KB of payload (720 × 72 bytes = 51,840 bytes). That is smaller than a single typical photo.

5. System design

5.1 Before air: the VLM holds the fingerprinted clips

Recognition needs something to recognise. Ahead of air, a clip is registered with the information that should travel with it: the programme it belongs to, the second at which a moment happens, and what should appear on screen at that moment. The clip is fingerprinted frame by frame with the same method the TV uses, and the fingerprints are held by the VLM as its index of fingerprinted clips. The same step lets a brand register its own ad (Section 5.4).

5.2 On air: the TV sends 72 bytes

While a listed programme is on, Edge's ACR on the TV captures the picture and sends the 72-byte fingerprint over HTTPS, about every five seconds in current deployments. The request body is the raw 72 bytes. Nothing else about the picture is sent.

5.3 Matching: the clip and the second

Edge's ACR talks to the VLM, which holds the fingerprinted clips. A matcher compares the incoming code with the clip index using a weighted bit-level distance and a threshold. If one clip and one position are clearly closest, the answer is that programme, that clip and the number of seconds until the registered moment. If nothing is close enough, or two candidates are equally close, the answer is an explicit no-match rather than a guess. The TV can then count down to the moment and show what was registered for it.

No model runs at this point. The VLM did its work when the clip was registered and when the channel was read at the source; recognition on a TV is a lookup.

5.4 The same path, for an advertiser's own ad

Because registration is just "fingerprint every second of a video and store what goes with it", an ad can be registered the same way as a programme clip. A brand uploads its ad or pastes a link; every second of the ad is fingerprinted; the brand adds where viewers should go, such as a product page, a coupon or a store finder. When the ad plays on a TV, the fingerprint matches the exact second and the screen can offer the brand's next step. We describe this flow in more detail in Technical report 04.

6. Why the input does not matter

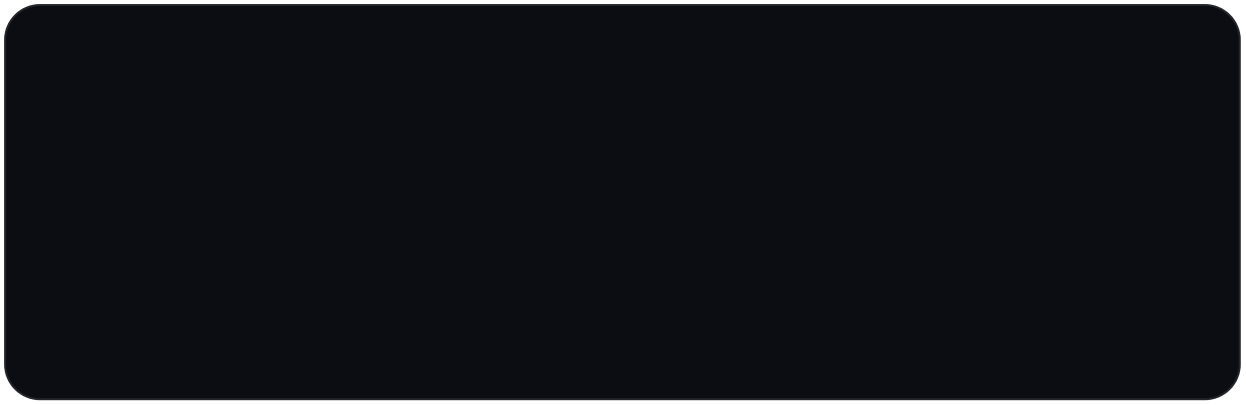


Figure 2. The fingerprint is computed from the rendered picture, so the source of the signal does not change the method.

A TV knows a great deal about content played by its own apps and almost nothing about content arriving through an HDMI port from a set-top box, a streaming stick or a console, or through an antenna. Recognition from the rendered picture closes that gap. The fingerprint is computed after decoding, from what the viewer actually sees, so it works the same on HDMI, on free-to-air and in the TV's built-in apps.

7. Arithmetic at scale

We separate two costs that are often confused. **Perception**, reading a channel with the vision language model, happens once per channel at the source. **Recognition**, matching a TV's fingerprint, happens per TV, but is a small lookup. The arithmetic below is payload only; transport overhead such as TLS and HTTP headers is not included and has not been measured here.

Scenario	Arithmetic	Result
One TV, one hour of a listed programme	3,600 s ÷ 5 s = 720 fingerprints × 72 B	about 52 KB
Per 1,000,000 TVs watching listed programmes at once	1,000,000 ÷ 5 s = 200,000 per second × 72 B	about 14.4 MB/s, about 115 Mbit/s
Model reads for a channel	Once at the source, shared by every TV watching it	Independent of audience size

The consequence is the main architectural point of this report: adding viewers adds lookups, not model calls. Matcher throughput per server has not yet been benchmarked publicly by us, and we do not quote a figure for it.

8. Evaluation and what we report

We have run the fingerprint in laboratory testing and on air with partners. Those results include match rates and timing, and they belong partly to the partners concerned; we will publish them when they are cleared for publication. In this report we limit ourselves to properties that follow from the design and can be checked by anyone who implements it: the payload size, the integer-only computation, the explicit no-match behaviour, and the input independence.

Time resolution deserves a specific note. With one fingerprint about every five seconds, the delay before the system knows the moment is bounded by the interval between fingerprints plus the time for one lookup. The interval is a deployment setting, not a property of the fingerprint; a shorter interval improves time resolution at the cost of more requests.

9. Limitations

- **Registration comes first.** The method recognises content that has been fingerprinted ahead of time. It does not, by itself, describe content it has never seen. That is the vision language model's job (see Technical report 02).
- **Low-information frames.** Black frames, slates, and long static graphics carry little structure and can be ambiguous. The matcher returns no-match rather than guessing.
- **Heavy overlays.** Picture-in-picture, large menus and split screens change the brightness layout and can prevent a match until they clear.
- **Cadence.** Time resolution is limited by the interval between fingerprints, as above.
- **Published measurements.** Partner match-rate results are not included in this version.

10. Privacy

The design keeps the image on the television. What leaves the set is a 72-byte code that summarises coarse brightness and cannot be turned back into the picture. It is sent only while a listed programme is on. We discuss our privacy approach more broadly on our privacy by design page.

References

1. J. Haitsma and T. Kalker. A Highly Robust Audio Fingerprinting System. ISMIR 2002. ismir2002.ismir.net/proceedings/02-FP04-2.pdf
2. A. L.-C. Wang. An Industrial-Strength Audio Search Algorithm. ISMIR 2003. ee.columbia.edu/~dpwe/papers/Wang03-shazam.pdf
3. J. Oostveen, T. Kalker and J. Haitsma. Feature Extraction and a Database Strategy for Video Fingerprinting. Recent Advances in Visual Information Systems (VISUAL 2002), LNCS 2314. doi:10.1007/3-540-45925-1_11
4. R. Venkatesan, S.-M. Koon, M. H. Jakubowski and P. Moulin. Robust Image Hashing. Proceedings of ICIP 2000. doi:10.1109/ICIP.2000.899541
5. M. Douze et al. The 2021 Image Similarity Dataset and Challenge. 2021. arXiv:2106.09672
6. E. Pizzi et al. A Self-Supervised Descriptor for Image Copy Detection. CVPR 2022. arXiv:2202.10261
7. P. Indyk and R. Motwani. Approximate Nearest Neighbors: Towards Removing the Curse of Dimensionality. STOC 1998. doi:10.1145/276698.276876
8. J. Johnson, M. Douze and H. Jégou. Billion-scale Similarity Search with GPUs. 2017. arXiv:1702.08734