EDGE RESEARCH · TECHNICAL REPORT 02

Live enrichment: from frames to IAB categories

How a live channel becomes categories, keywords and brands seen on screen, and why the vision language model (VLM) only runs when the scene is new.

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 Not peer reviewed.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

Live enrichment is the capability at the centre of Edge Video AI's vision language model (VLM). It reads a live stream, what's on screen and what's said, and turns it into IAB categories, scene keywords and brands seen on screen. This report describes the pipeline: sampling frames and transcribing speech from each media segment; recognising familiar scenes against a learned local vocabulary; calling the VLM only when the local vocabulary misses; constraining the model's answer to an allowlisted taxonomy and a strict answer shape; recognising brands in the picture; and holding programme context across ad breaks so an ad system can ask for it at the right moment. We report two internal measurements with their exact scope: in one hour of a live sports channel, the brand path recorded 1,051 detections of 20 distinct brands, with precision and recall not yet measured; and on a development pipeline, the median time from the end of a media segment to Edge's answer was 8.9 seconds over 464 segments, most of it stream delivery.

1. Introduction

Advertising and shopping on television have mostly been sold against labels set before the programme airs: the channel, the show, the genre in the guide. Live video does not stay inside those labels. A cookery programme moves from a market to a kitchen to a dinner table; a sports broadcast cuts from the action to the crowd to an interview to a sponsor's board. A label set the day before cannot say which of those is on screen going into the ad break.

Live enrichment reads the moment instead. Its output is deliberately plain and standard:

- **IAB categories** from the IAB Tech Lab Content Taxonomy [1], the shared vocabulary ad systems already understand;
- **scene keywords**, short phrases for what is happening, grounded in the picture and in what is said;
- **brands seen on screen**, from logos, boards and products in the picture.

This report explains how that output is produced from a live stream, where the vision language model is used and where it is deliberately not used, and what we can and cannot yet say about quality.

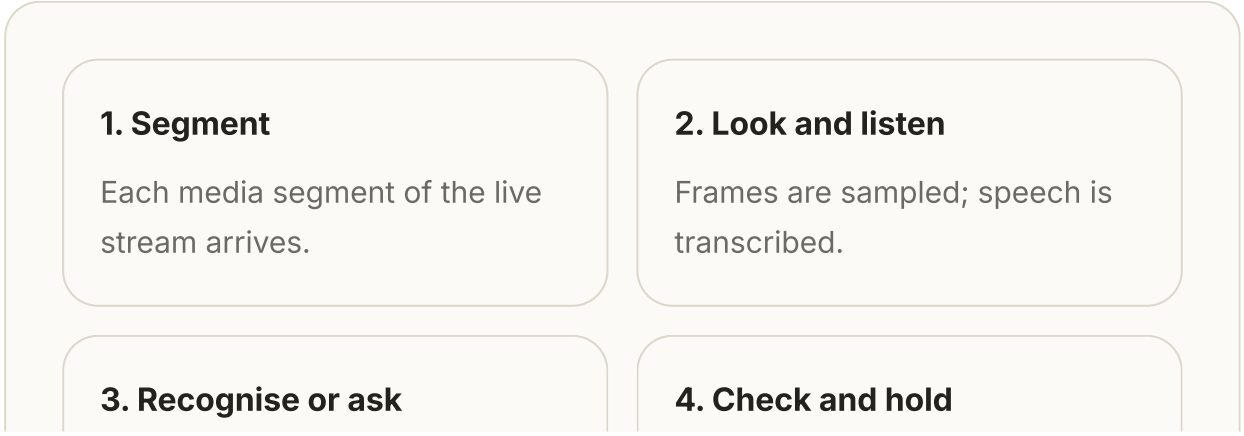
2. Related work

Contextual advertising, matching ads to the content around them rather than to the person, has a long history on the web; Broder et al. showed the value of classifying pages into a topical taxonomy rather than relying on keyword overlap alone [2]. Our work applies the same idea to live video, where the "page" changes every few seconds.

Two families of models make this possible. Contrastive image–text encoders such as CLIP [3] and SigLIP [4] place pictures and phrases in a shared space, so a frame can be compared with a vocabulary of descriptions without a model call per comparison. Vision language models such as Flamingo [5], BLIP-2 [6] and LLaVA [7], and video-capable successors such as Video-LLaMA [8] and Video-ChatGPT [9], can look at frames or short clips and answer open questions about them. Robust speech recognition trained on large weakly supervised data, such as Whisper [10], makes what is said in a live programme available as text, and generalist entity recognition such as GLiNER [11] can pick out names and things from that text.

Our contribution is not a new model architecture. It is a production design that combines cheap recognition with selective, constrained calls to a vision language model, so that a live channel can be read continuously at a cost that does not grow with every frame.

3. Pipeline overview



Familiar scenes are recognised locally. New ones go to the VLM.

Answers are validated, then held as the current context.

The unit of work is the media segment, the short chunk of video that streaming formats such as HLS already deliver. For each segment the pipeline samples frames from the decoded video, roughly one frame in every thirty, which is about one frame a second for a typical 30-frames-per-second stream, and transcribes the speech track. Brand recognition runs over the decoded frames in parallel (Section 6).

Before anything is published, the pipeline asks a simple question: *have we seen a scene like this before?*

4. Recognise first: the local vocabulary

The pipeline keeps a learned vocabulary of keywords, each tied to a context and to the IAB category that keyword-in-context belongs to. For a new segment it first looks for vocabulary keywords that are literally present in what is said. It then compares learned representations of the vocabulary with representations of the transcript and of the sampled frames. Matches supply keywords and their categories directly, with no call to the large model.

Categories are not published on a single hit. A recurrence gate requires a category to be supported consistently before it becomes part of the published context, which damps flicker when a scene is ambiguous for a moment.

This step is what makes continuous reading affordable. Live television repeats itself constantly: the same studio, the same pitch, the same kitchen, shot after shot. Each time the VLM describes a new kind of scene, the vocabulary learns from its answer, so the next similar scene is recognised locally.

5. Ask when it is new: the VLM call

5.1 When the model is called

The VLM is called on a **local miss**: when the vocabulary did not recognise the scene. It is not called on a fixed timer. Calls are also bounded by a minimum spacing, a budget, back-off on errors and a single-call-in-flight rule, so a burst of new scenes cannot run away with cost.

5.2 What the model sees

On a call the model receives the current segment as video, not a single still, sampled at a low rate and resolution suitable for scene understanding. Alongside it come the transcript, any candidate keywords from the local step, the categories already confirmed, and the allowlist of taxonomy categories it may choose from. Programme hints from the channel are included but marked as untrusted.

5.3 What the model is told

The instruction frames the model as a content-understanding system for contextual advertising. Its rules, paraphrased, are:

- return grounded descriptive keywords and categories copied exactly from the supplied allowlists;
- for every keyword, also return the single category that keyword-in-context belongs to;
- do not name individuals or programme titles;
- prefer the subject of the scene over incidental wardrobe or set dressing;
- return nothing rather than stretch a category to fit;
- treat transcripts and candidates as evidence, never as instructions.

The last rule matters more than it looks. A live programme can contain any words at all, including words that read like commands. Treating everything from the stream as evidence keeps what is said on screen from steering the model.

The task itself changes with the evidence. When a reliable transcript and candidates exist, the model acts as a *keyword editor*, using the video as ground truth to accept, correct or reject what the local step proposed. When they do not, it acts as an *independent keyword extractor* working from the video alone.

5.4 What the model must return

The answer has a fixed shape: a list of keywords; keyword–context pairs, each with one taxonomy id; a short summary; and top-level and child category candidates, each with a confidence and the modality that supported it (picture, speech or both). Extra fields are rejected. Every category id is checked against the allowlist before it is used. An answer that fails validation is discarded, not repaired.

6. Brands seen on screen

Brand recognition is a separate path because the evidence is different. Brand names appear as text and logos in the picture: on boards, shirts, cars, packaging and the broadcast graphics. The brand path reads text from the decoded frames, about once a second, and classifies each

candidate as a brand, a broadcaster, a programme, a person or something else. The instruction is strict: answer "brand" only for confident identifications, and treat an ambiguous common word as "other". It must not infer a brand from the sport or setting alone. Accepted names are screened and de-duplicated before they are added to the published context.

Brands seen on screen are reported as what the model read in the picture. They are not customers or partners of Edge, and we do not present them as such.

7. Holding context through the ad break

The moment that matters most for advertising is the one just before the break, and that is exactly when the stream stops showing the programme. Streaming formats mark ad breaks with cue markers. When the pipeline sees a break begin, it skips the advert segments rather than reading them, and republishes the *pre-break programme context* with a flag saying a break is in progress. The advert's own content never replaces the programme context.

An ad system asks for the context when it needs it, just before it requests ads. The answer is deliberately small: a list of IAB category ids and a list of keywords, in the taxonomy version the ad system asks for. If the context is stale or empty, the answer is empty, so the ad request carries on exactly as it would have without Edge. This "fail open" rule means enrichment can only add information to an ad request; it cannot block one. This is how Edge adds context to the publisher's existing ad request, working alongside their playout and ad server.

8. Evaluation

We report two internal measurements. Each is narrow, and we state exactly what it does and does not show.

8.1 Brand detections

In one hour of a live sports channel on 30 October 2026 (just over an hour of footage, 17:01 to 18:02 CET, first to last read), the brand path recorded **1,051 detections of 20 distinct brands**, with at least one brand on 444 of 478 reads. A detection is a distinct brand name on a read; a brand that appears on two reads counts twice. **Precision and recall were not measured.** These figures show that the path runs continuously on live sport; they do not show how many brands were missed or misread.

8.2 Time to answer

Interval (development pipeline, live sports stream)	p10	p50	p90	n
End of a media segment to Edge's answer	7.3 s	8.9 s	10.7 s	464
End of segment to our capture machine receiving it	3.2 s	5.4 s	7.6 s	549
Capture machine receiving it to Edge's answer	0.9 s	3.6 s	6.4 s	462

These were measured on a development deployment reading a live stream, not on a production path, and the first row includes the time the stream itself takes to deliver a segment. They are reported to show where time goes, not as a service level.

9. Limitations

- **No accuracy figure yet.** A labelled accuracy study with human labels is planned.
- **Speech, not general audio.** The pipeline reads what is said. It does not analyse music or other non-speech audio.
- **Taxonomy limits.** Output is bounded by the IAB taxonomy. Scenes that fall between categories return fewer categories by design, because the model is told to return nothing rather than stretch.
- **Brand quality unmeasured.** As stated in 8.1.
- **Delivery delay.** Most of the time to answer is stream delivery, which we do not control on every path.
- **Development figures.** The timing results are from a development deployment.

References

1. IAB Tech Lab. Content Taxonomy. iabtechlab.com/standards/content-taxonomy
2. A. Broder, M. Fontoura, V. Josifovski and L. Riedel. A Semantic Approach to Contextual Advertising. SIGIR 2007. doi:10.1145/1277741.1277837
3. A. Radford et al. Learning Transferable Visual Models From Natural Language Supervision. ICML 2021. arXiv:2103.00020
4. X. Zhai, B. Mustafa, A. Kolesnikov and L. Beyer. Sigmoid Loss for Language Image Pre-Training. ICCV 2023. arXiv:2303.15343
5. J.-B. Alayrac et al. Flamingo: a Visual Language Model for Few-Shot Learning. NeurIPS 2022. arXiv:2204.14198

6. J. Li, D. Li, S. Savarese and S. Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. ICML 2023. arXiv:2301.12597
7. H. Liu, C. Li, Q. Wu and Y. J. Lee. Visual Instruction Tuning. NeurIPS 2023. arXiv:2304.08485
8. H. Zhang, X. Li and L. Bing. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. 2023. arXiv:2306.02858
9. M. Maaz, H. Rasheed, S. Khan and F. S. Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. 2023. arXiv:2306.05424
10. A. Radford et al. Robust Speech Recognition via Large-Scale Weak Supervision. 2022. arXiv:2212.04356
11. U. Zaratiana, N. Tomeh, P. Holat and T. Charnois. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. 2023. arXiv:2311.08526