



AdShot

Benchmarking Multimodal LLMs for Video Advertisement Clipping

Can today's video-understanding AI cut a 30-second ad down to 15, the way a professional editor would?

Wen Xie · Om Rastogi · Sai Siddhartha Vivek Dhir Rangoju · Gijs Overgoor · Yakov Bart · Sarah Ostadabbas

Northeastern University · Southern Methodist University

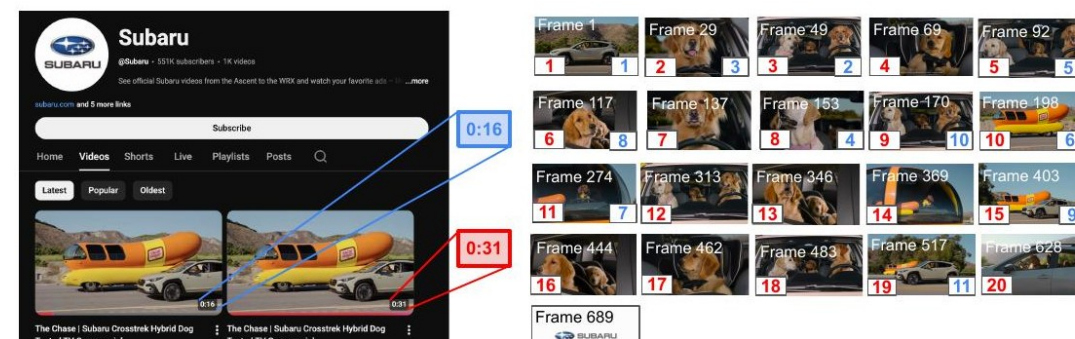
github.com/ostadabbas/AdShot · huggingface.co/datasets/omrastogi/adshot_benchmark

THE PROBLEM

Shrinking an ad is slow, skilled, manual work

- Advertisers ship the same campaign at several lengths, most commonly a 30s TV spot and a 15s cut.
- Making the short version means deciding, shot by shot, what to keep and what to drop.
- It's a persistent bottleneck, repeated across thousands of ads per brand.

A 30s Subaru ad (21 shots, red) cut to 15s by picking shots (blue)



The question

Can multimodal LLMs that excel at understanding video also make this editorial decision?

A cutdown reuses the brand's own footage.

It preserves exactly what a professional already judged effective, instead of generating new creative.

AdShot: 823 real long-short ad pairs

823

ad pairs

194

brands

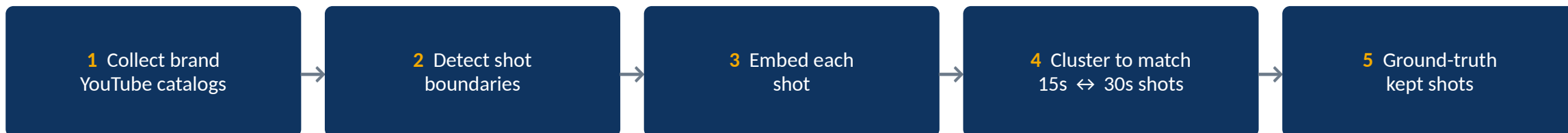
17

industries

1,646

videos

How each pair is built



98.2% automated-vs-human matching agreement · editors keep **~half** the shots · each shot labeled for **action, information, emotion, imagery**

15 models, one editorial task

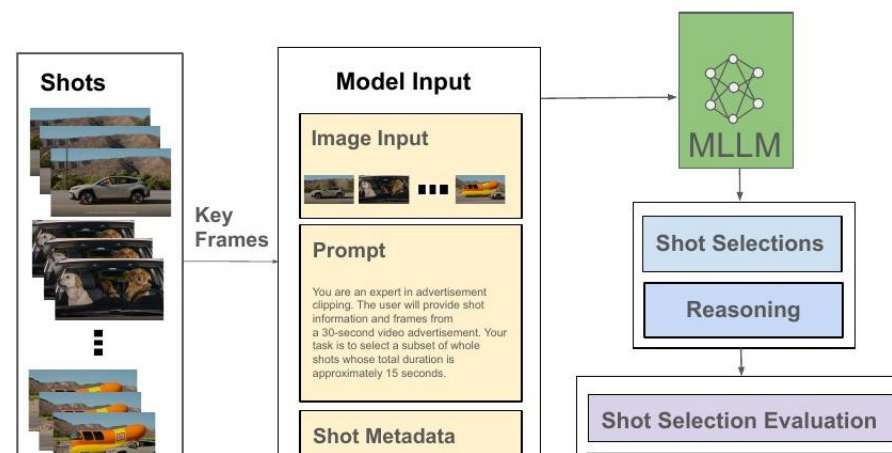
Input: keyframes + shot durations + a prompt to pick whole shots totaling ~15s.

Output: a JSON set of selected shots, scored against the professional edit.

Three axes of evaluation

- 1 Selection accuracy** Precision / Recall / F1 vs. the editor's shots
- 2 Duration adherence** Mean error against the 14–16s window
- 3 Content alignment** Does the mix of shot types match the pros?

The evaluation pipeline

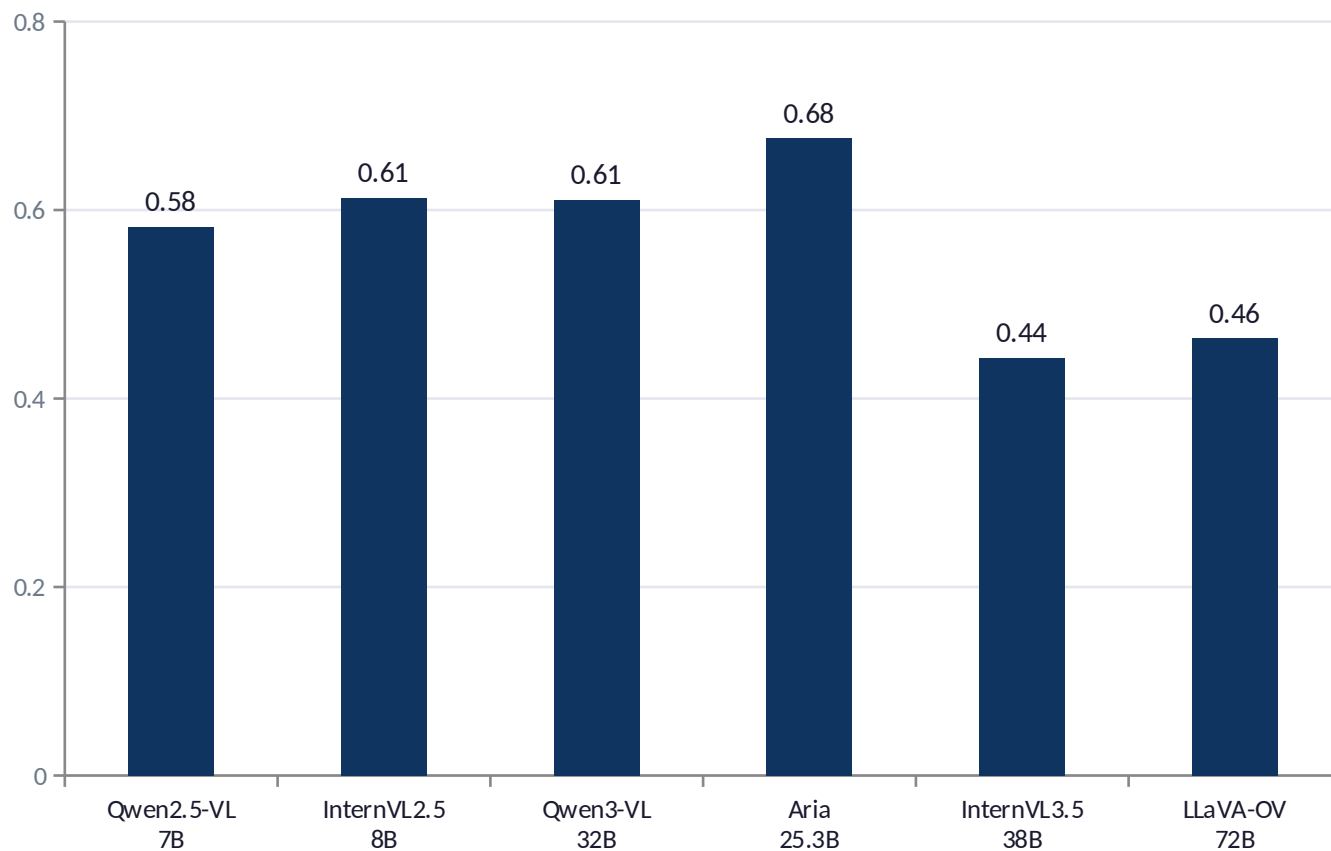


13 video-only + 2 audio-visual models

Qwen2.5/3-VL · InternVL2.5/3.5 · LLaVA-OneVision · Aria · MiniCPM-V · Molmo2 · Qwen2.5-Omni · MiniCPM-o
7B to 72B · zero-shot · greedy decoding

FINDING 1

Architecture matters more than scale



Bigger ≠ better

- A 7B model (0.582) beats a 72B one (0.464).
- Best F1 comes from Aria, a 25B mixture-of-experts, not the largest model.
- Within a family, scaling up doesn't reliably help.

Fit to fast-cut ad video matters more than parameter count.

The skill is real, but not universal

9 of 15

models beat a content-free random baseline. 6 fall below it.

The 6 sub-baseline models share one signature: **decent precision but far too little recall**: they select too few shots. The failure spans 8B to 72B, so it's about fit, not size.

Duration is hard to hit

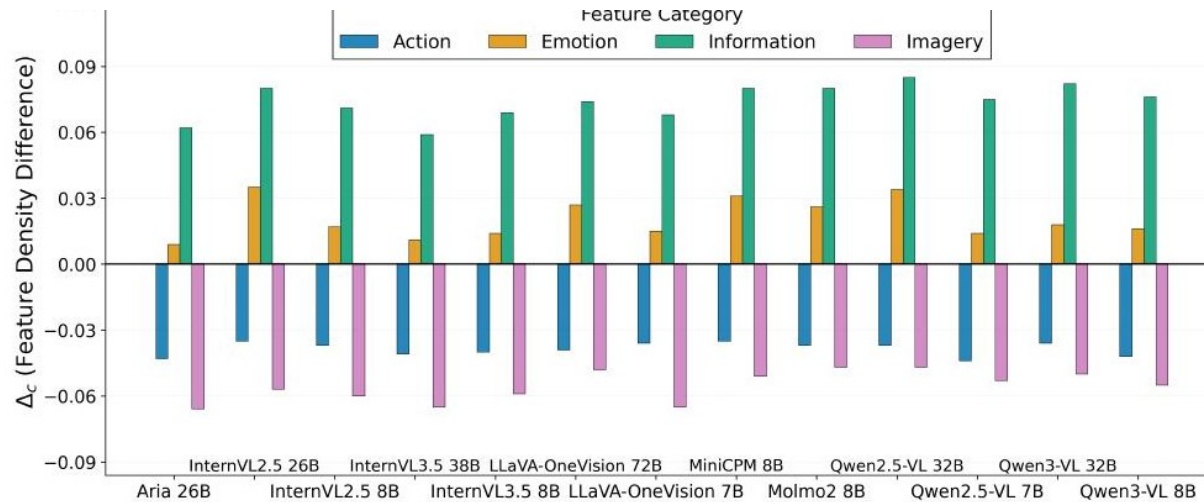
5.39s mean duration error

against a tight 14–16s target window.

- Most models overshoot, keeping too much.
- Overshoots are easy to trim; undershoots drop key content.
- Selection quality and timing are separable; a useful model needs both.

FINDING 3

Systematic content bias, and an audio signal



Δ feature density vs. professional editors, across all 13 video-only models

Green/amber above zero = over-selected · blue/pink below = under-selected

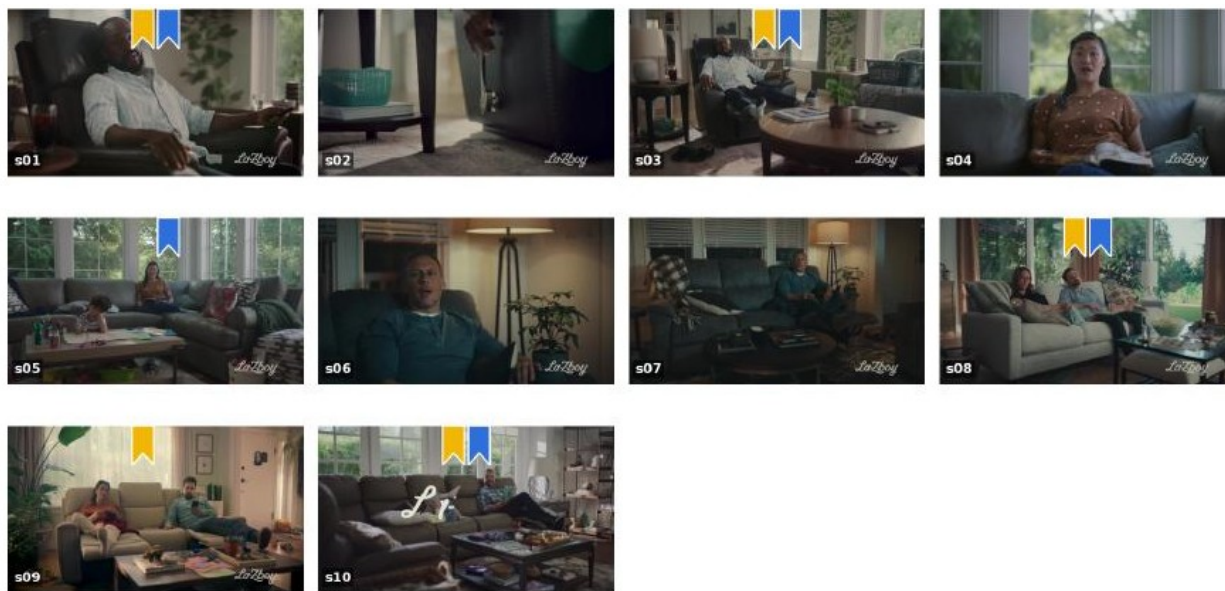
Every model favors **information and emotion**, under-selecting **action and imagery**.
Size doesn't fix it.

Audio helps

Two audio-visual models nearly match the best video-only model at **a third of the parameters**.

Controlled ablation: removing sound drops F1 by up to **0.083**, driven by recall.

Where model and editor agree, and don't



■ Qwen3VL-32B selected

■ Golden shot / GT

La-Z-Boy · Retail

La-Z-Boy ad (Retail, 10 shots): Qwen3-VL 32B selection vs. the professional edit

How to read it

■ **Both markers** model and editor agree (correct)

■ **Blue only** model kept a shot the editor dropped (over-selection)

■ **Yellow only** model missed a shot the editor kept

Errors skew toward over-selection, matching the overshoot in duration error.

Takeaways

Assist, don't replace

Current MLLMs can speed up ad cutdowns but aren't ready to run them unsupervised.

A new kind of eval

AdShot measures editorial decision-making, not comprehension, a gap existing benchmarks miss.

Diagnosis, not just a score

Reveals who beats chance, who overshoots duration, and a shared content bias across all models.

Publicly released: benchmark, code, and Croissant metadata

github.com/ostadabbas/AdShot • huggingface.co/datasets/omrastogi/adshot_benchmark