UAB
Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Domene García, Luis. *MAVIS : Mapping the Limits of Semantic Embedding Alignment for Multimodal Video Misinformation Detection.* Treball de Final de Grau (Universitat Autònoma de Barcelona), 2026 (Grau en Intel·ligència Artificial)

This version is available at https://ddd.uab.cat/record/336301

under the terms of the  license.

MAVIS: Mapping the Limits of Semantic Embedding Alignment for Multimodal Video Misinformation Detection

Luis Domene García

June 28, 2026

Abstract

Short-form video misinformation is dominated not by deepfakes but by authentic footage *mis-accompanied*: a real clip paired with a false title or a statement taken out of context. We ask whether cross-channel semantic coherence, measured with off-the-shelf single-vector multimodal embeddings and cosine distance, can detect it without external evidence. Across three datasets (FakeVV, GroundLie360, M3A) and four encoders (Qwen3-VL 2B/8B, WAVE-7B, and the natively multimodal Gemini Embedding 2) we ask how far this signal can be pushed, from a direct within-pair comparison to a single, deployable detection threshold. The signal is real, but only as a *comparison*: the embeddings reliably tell which of two pairings is the authentic one. Turning that into a single threshold that must work across every video is far harder. On their own, then, these embeddings are not reliable misinformation detectors; but the cross-channel alignment they measure can still serve as a useful warning sign within a larger fact-checking system.

Keywords: multimodal misinformation, short-form video, cross-modal embedding alignment, out-of-context detection, bias audit, similarity over factuality

1 Introduction and Motivation

Short-form video platforms (TikTok, YouTube Shorts, Instagram Reels) have become a dominant channel for news consumption and, with it, a fertile ground for multimodal misinformation. Crucially, most deceptive short videos are not synthetic deepfakes but *cheap fakes*: authentic footage *mis-accompanied* — a real clip paired with a false title, a statement cropped out of context, or an image and a caption taken from unrelated news. In these cases each modality (video, audio, on-screen text) is genuine in isolation; the lie lives in how they are *combined* [1, 2].

Most fact-checking systems attack this problem *exogenously*, retrieving external evidence to judge veracity [3, 4]. This is effective for well-documented claims but scales poorly against the volume and virality of short video, and presupposes that a published ground truth already exists — which fails precisely for breaking news. An *endogenous* alternative asks instead whether the video is internally *coherent*: we model it as the channels that ought to agree — its visual content (V), the accompanying text such as a title or description (T), and the spoken audio through its automatic speech recognition (ASR) transcript — and look for semantic friction between them.

1.1 Scope of this work

MAVIS (*Multimodal Alignment Verification In Short-videos*) was initially conceived as a full endogenous *system* — a deployable tool that would segment a video, embed its channels, score their misalignment and emit a human-readable rationale. Two constraints reshaped this goal. The first is methodological and the more fundamental: the whole system rests on an untested premise — that cross-modal dissonance is a usable signal of misinformation at all. A tool that segments a video, embeds its channels and emits a rationale is only worth building once that premise has been validated; without such an analysis it would stand on unverified ground. The second is practical: engineering and validating a complete deployable system far exceeds the time and compute available for an undergraduate project. The project therefore deliberately narrowed its scope from *building a detector* to *rigorously characterising a method*: a systematic empirical study of **what off-the-shelf, single-vector multimodal embeddings can and cannot reveal about misinformation in video**. To our knowledge this analysis has not been carried out on video before; most prior similarity-based work targets static images. The deliverable

• Contact E-mail: Luis.Domene@autonoma.cat
• Supervised by: Ernest Valveny (Dept. of Computer Science, UAB)
• Academic Year 2025/26
• Code: `https://github.com/luisdomene4/MAVIS`

is therefore a *capability map* of how far off-the-shelf multimodal embeddings can characterise manipulations in short videos — which manipulation types semantic alignment detects, in which question format, at what cost, and where its reach ends — backed by a reproducible pipeline.

1.2 Problem statement and research questions

The central question is whether cross-channel semantic coherence, measured with pretrained multimodal embeddings and cosine distance, is sufficient to detect manipulation *without external evidence*. A positive answer would yield a cheap, scalable, source-free detector; the question is open because recent work on static images warns that similarity-based detectors may pass by exploiting surface-level shortcuts rather than factuality [5]. We test that suspicion systematically in video. Concretely:

- **RQ1.** Which quantifiable metrics capture cross-channel semantic alignment, and what do they *actually* measure?
- **RQ2.** Which manipulation types are separable by alignment, and under what conditions?
- **RQ3.** How much of the detection performance measured on these benchmarks is genuine signal rather than dataset-construction bias?

1.3 Objectives and contributions

The main objective is to design, implement and empirically evaluate a *zero-shot* multimodal-alignment analysis for short-video misinformation, and from it derive an honest map of its limits. The specific contributions are:

1. A systematic test of the "similarity-as-veracity" hypothesis in *video* (3 datasets $\times$ 4 encoders), where almost all prior evidence is on static images.
2. A **bias-audit methodology** exposing how much measured "detection" is construction artefact rather than signal.
3. An analysis of why an absolute cosine threshold fails where the within-pair comparison works.
4. A capability map of what semantic alignment can and cannot reveal about video misinformation.

1.4 Document structure

Section 2 reviews related work and positions MAVIS within the "similarity-over-factuality" debate. Section 3 presents the three datasets as an adversariality gradient. Section 4 details the pipeline, encoders, metrics and statistics. Section 5 reports the results (zero-shot ladder and probe/geometry audit). Section 6 synthesises the capability map and Section 7 concludes.

2 State of the Art and Background

Multimedia verification has evolved quickly to keep up with increasingly subtle deception. We review it along the axis most relevant to MAVIS: from exogenous, evidence-based checking to endogenous, internal-coherence approaches, and finally to the multimodal-embedding paradigm this work scrutinises.

2.1 From metadata to multimodal evidence

Early video verification relied on social context rather than content: corpora such as FVC-2018 [6] analysed propagation cascades and user comments. Content-based detection followed, split into two paradigms. The *exogenous* branch retrieves external articles and fact-checks to judge veracity — recent systems such as EvidSV [3] and TRUE/3MFact [4]. As those authors note, retrieval often lacks the reasoning needed for general video fact-checking, and depending on a pre-existing ground truth fails for breaking news and novel cheap fakes. In parallel, multimodal datasets such as FakeSV [1] framed the task as binary classification; later work (e.g. FakeVE [2]) argued binary labels are insufficient and pushed towards *explainable* analysis of *why* a video is deceptive.

2.2 Endogenous detection: granularity and grounding

Evaluating a video's internal semantic integrity is not new, but recent work has sharpened its granularity. GroundLie360 [7] argues that global-level checking is insufficient and introduces spatial/temporal *grounding* (bounding boxes and timestamps locating the contradiction). Official-NV [8] and FMNV [9] inject targeted misalignments into real news videos to force layer-level localisation, while OmniFake [10] and FakeVV/Fact-R1 [11] propose unified frameworks combining human-crafted and AI-synthesised content via cross-modal reasoning. MAVIS uses the *datasets* produced by this line (FakeVV, GroundLie360, and M3A [12]) but asks a different, methodological question about the embeddings themselves.

2.3 Multimodal embeddings: opportunities and limits

Multimodal embedding models map heterogeneous signals into a shared space, enabling cross-modal comparison through similarity. Contrastive foundations (CLIP [13], VideoCLIP [14], AudioCLIP [15], ImageBind [16], LanguageBind [17], InternVideo2 [18], VLM2Vec [19]) improve transfer and reduce reliance on task-specific training; video-LLM benchmarks such as Video-MME [20] chart their general capabilities. For misinformation this offers a unified interface and a cheap consistency test across channels — but as an *operational* convenience, not a guarantee of contradiction detection. Reported limitations are central to this thesis: global semantic gains do not transfer to fine-grained, temporal or compositional inconsistencies, and latent proximity is not factual correctness. The TRIANGLE area method [21] extends pairwise cosine to a tripartite measure, which we adopt and stress-test (Sec. 4).

A new opportunity: natively multimodal video embeddings. A key motivation for this study is the very recent arrival of *Gemini Embedding 2* (GE2) [22], released by Google in public preview on 10 March 2026 — its first *natively* multimodal embedding model, mapping text, images, video and audio into a single space (default dimensionality 3072, ingesting up to 120 s of video and 180 s of audio). Because it is so new, no published work characterises its behaviour on misinformation; MAVIS is, to our knowledge, the first to probe whether a natively multimodal video encoder opens detection doors that earlier text/image-centric encoders could not.

2.4 Similarity is not factuality — the gap MAVIS addresses

The closest critique to our work is "Similarity over Factuality" [5]: on static-image OOC benchmarks (NewsCLIPpings [23], COSMOS [24], VERITE [25]) a simple similarity baseline (MUSE) rivals complex architectures, revealing that detectors exploit *surface-level similarity shortcuts* rather than factuality. MAVIS extends this scrutiny to **video**, and inside the video, operating zero-shot and auditing the embedding geometry itself.

This geometry has well-documented pathologies that any similarity-based detector must control. The *modality gap* [26] places each modality in a separate cone of the shared space, so even a perfectly matched triplet has cosine bounded away from 1; and *anisotropy* (failures of uniformity on the hypersphere [27]) inflates cosines even for unrelated items. Both differ from one encoder to another, so the absolute magnitude of a cosine is not comparable across models and is only meaningful relative to each encoder's own scale.

Positioning and novelty. Existing video misinformation detectors are *fine-tuned*: SV-FEND on FakeSV [1], the reasoning LLM Fact-R1 on FakeVV [11], and the closest, MAGIC3 [28], which models text–visual vs. text–audio consistency but via trained cross-attention. We cite MAGIC3's empirical asymmetry (real and fake videos differ in the *direction* of cross-modal consistency) as convergent motivation for our hypothesis. The triangle-area geometry we use is formalised by TRIANGLE [21], but there it is a *training objective*; MAVIS repurposes it as a *zero-shot diagnostic probe*. To our knowledge, no prior work studies the geometry of off-the-shelf single-vector embeddings as a misinformation signal in video, nor publishes a *capability map* of which manipulation types (out-of-context, entity substitution, false speech, temporal edit) embedding alignment can and cannot separate — the central contribution of this work.

3 DATASETS

We evaluate on three video misinformation datasets chosen as a deliberate *adversariality gradient* — from a trivial control to human-crafted lies designed to look coherent — and, as a differential of this work, we document their *construction biases*; for GroundLie360 in particular this is a bias we identify here, not one posed by its authors. All three are filtered to clips ≤ 120 s to share a common comparison with the API-limited GE2 encoder. Figure 1 shows, with real examples, how each dataset constructs its fakes.

3.1 FakeVV (Fact-R1)

FakeVV is the benchmark released with Fact-R1 [11] (news from YouTube Shorts). Its fakes are built by a non-random, CLIP-guided *entity substitution*: starting from a true title over a real video, one named entity (a person, location, event, or organization) is swapped for a plausible alternative while the rest of the title is left intact, so the false title stays grammatically perfect and topically credible — only its contrast with the video betrays it (location 373, person 310, event 220, organization 94). We use only its *test set*: 997 videos ($\approx$813 after the $\leq$ 120 s filter); each item provides the real video, its true and false title, and the manipulated-entity category, and we add a Whisper transcript per video.

3.2 GroundLie360

GroundLie360 [7] sources videos from the fact-checking site Snopes (8 platforms, 2016–2024). Here a fake is authentic footage made misleading either by the text that frames it or by an edit to the video itself, annotated with a rich, *grounded* 6-type taxonomy (Fig. 1): a *false_title* misrepresents the video's content; an *unsupported* title claims something the video neither shows nor supports; a *contradictory* title states the opposite of what the video shows; *false_speech* is a spoken lie in the audio; a *temporal_edit* re-orders or splices the footage to change its meaning; and *cgi* marks a synthetically altered region. The annotations are themselves grounded — temporal spans for the spoken lies, a scene segmentation, an edit mask, and CGI bounding boxes. The full set has 2,044 videos; our $\leq$ 120 s subset is 1,722 (899 fake / 823 real), with per-type counts false_title 689, unsupported 373, cgi 192, contradictory 159, false_speech 124, temporal_edit 77; videos often carry more than one type at once (a multi-label structure). We later show this dataset has a strong construction bias (Snopes-authored claim titles vs. original captions; Sec. 5.2).

3.3 M3A

M3A [12] is built from authentic content of 60 reputable outlets; it is the largest and multilingual dataset, and the only one with all four channels. It supplies two manipulation families that fit our framing. NEM (named-entity manipulation; subtypes person/location/organization/complete) swaps a named entity in the text — the textual analogue of FakeVV, but over the description rather than a headline. MM (multimodality mismatching) is *out-of-context re-pairing*: a real video is matched with the genuine text of a *different* story, so every channel is authentic on its own and only their pairing is wrong. Its video+audio population is 182,202; we materialise 20,000 clips ($\leq$ 120 s). The text modality is the `Summary` (a $\sim$28-word BART description of the post, *not* a headline). GE2 is limited to $\approx$2,340 clips by API quota, so the four-encoder comparison is run on the common intersection.

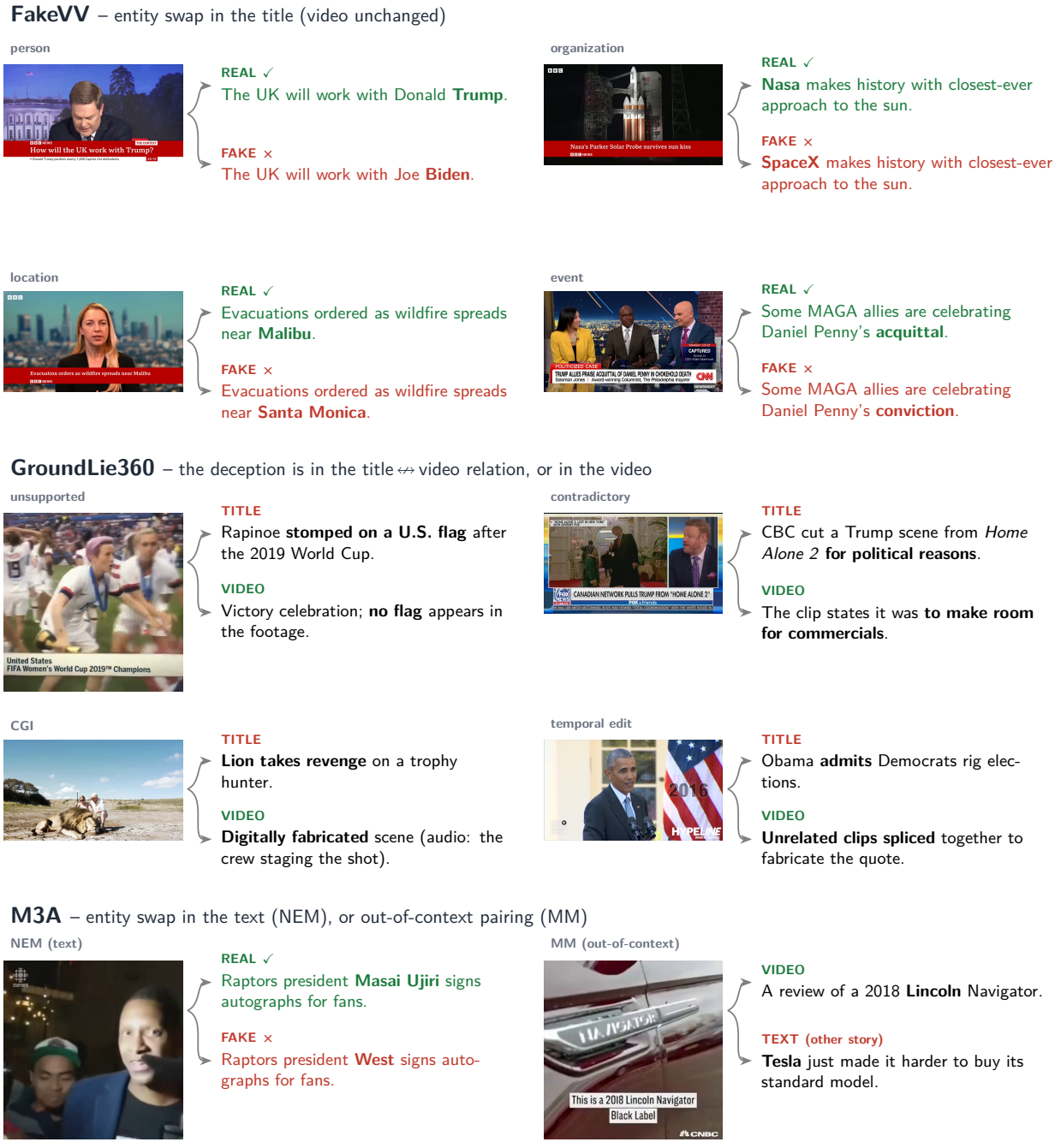

Fig. 1: **How each dataset builds its fakes. FakeVV**: one named entity in the title is swapped for a plausible alternative over an unchanged video. **GroundLie360**: authentic footage whose title is *unsupported* by or *contradicts* the video, or whose video itself is manipulated (*CGI*, *temporal* splice); four of the six grounded types are shown. **M3A**: a named entity is swapped in the text (NEM), or a real video is paired with real text from a different story (MM, out-of-context). In every case the manipulation lives in the *relation* between channels, not in any channel alone.

Encoder	Dim.	Channels used
Qwen3-VL 2B	2048	text, transcript, video
Qwen3-VL 8B	4096	text, transcript, video
WAVE-7B	3584	text, transcript, video
GE2 (Google)	3072	text, transcript, video

TABLE 1: Embedding encoders, output dimensionality and the channels used here. WAVE-7B also exposes audio and audio-visual channels, but they are out of scope (Sec. 6); GE2 is API-only and several experiments run it on a quota-limited subset.

4 METHODOLOGY

4.1 Channels and core hypothesis

We treat each video as the confluence of information channels that ought to agree, and represent each channel by a *single* embedding vector per video, computed by a pretrained multimodal encoder. Vectors are L_2-normalised, so the cosine similarity $\cos(\mathbf{a}, \mathbf{b}) \in [-1, 1]$ measures their semantic alignment. The transversal hypothesis is simple: if two channels are coherent their embeddings point in similar directions (high cosine); a manipulation introduces semantic *mis-alignment* and should lower the cosine between the affected channels. The whole study tests how far this hypothesis holds, for which manipulation types, and in which reading format.

4.2 Encoders

We use four embedding models spanning open-source and proprietary, over three text and visual channels (Table 1). **Qwen3-VL-Embedding** (2B and 8B) [29] is an open-source bi-encoder built on the Qwen3-VL vision–language model: it maps text, images and video into a common space by reading out the model's last hidden state, and is trained in several stages (large-scale contrastive pre-training, multi-task fine-tuning, and distillation from a reranker), with Matryoshka representation learning for flexible dimensionality. **WAVE-7B** [30] is open-source and built on the Qwen2.5-Omni backbone; it is, to its authors' knowledge, the first LLM-based embedding to place text, audio, silent video and synchronised audio-visual streams in a single space, via a hierarchical feature-fusion strategy and joint multi-task training. **Gemini Embedding 2 (GE2)** [22] is Google's proprietary, API-only, natively multimodal model; it returns a Matryoshka-truncatable vector and is the only encoder we do not run locally. GE2's daily API quota restricts it to subsets, which is why several experiments report GE2 on a matched sample.

A practical note on WAVE. We rely on it only on FakeVV, where its text and video channels behave reliably. On M3A its video set could not be completed and was therefore withdrawn. Its raw *audio* and joint *audio-visual* channels, moreover, kept producing anomalous embeddings that repeated regeneration did not correct, so the acoustic signal is out of scope (Sec. 6) and speech is reached only through the Whisper transcript.

4.3 Reproducible pipeline and data architecture

The pipeline is designed so the analysis is fully regenerable from raw data. Deterministic, stratified indices select the video subset; embedding generation runs offline and writes *one SQLite database per (encoder × dataset)*, storing raw `float32` vectors as the single source of truth; the analysis notebooks read only from these databases and never recompute embeddings. The schema is shared across encoders and datasets (App. A.6).

4.4 Similarity measures

Every experiment is built on two similarity measures over L_2-normalised embeddings. The first is the *cosine similarity* $\cos(\mathbf{a}, \mathbf{b})$ between two channels — e.g. in FakeVV we directly compare the video embedding against the title embedding. The second generalises this to three channels at once: given video (V), text/title (T) and audio or its Whisper transcript (Tr), we place the three unit vectors on the hypersphere and measure the *area* of the triangle they span, following the TRIANGLE formulation [21]. Writing $\mathbf{u} = \mathbf{V} - \mathbf{T}$ and $\mathbf{w} = \mathbf{Tr} - \mathbf{T}$, the area is given by the Gram identity for $\|\mathbf{u} \times \mathbf{w}\|$ in high dimension (Eq. 1). The intuition: when the three modalities tell the same story their tips sit close together and the triangle is *small*; a manipulation pulls one vertex away and enlarges it.

$$A = \tfrac{1}{2}\sqrt{\|\mathbf{u}\|^2\|\mathbf{w}\|^2 - (\mathbf{u}\cdot\mathbf{w})^2}. \tag{1}$$

We initially also computed the triangle *perimeter* $\sum(1 - \cos)$, but it was near-equivalent to the area and dropped in favour of the area, consistent with [21].

4.5 Reading the signal: three levels of difficulty

We read these scores in three ways, ordered from the easiest to the hardest demand on a zero-shot embedding. The order is the backbone of the results.

(1) Paired comparison — "which of the two is the real one?" The simplest question. When a dataset gives, for the *same* video, a true and a false version of the text (FakeVV titles, M3A-NEM summaries), we check whether the channel embedding is closer to the true text than to the false one. The score is the fraction of videos answered correctly (Eq. 2); chance is 0.5. This measures the core thesis question — *given two options, can the embedding tell the real from the fake?* — and is the most favourable setting, so results here are usually positive.

$$\mathrm{acc} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left[\cos(c_i, t_i^+) > \cos(c_i, t_i^-)\right]. \tag{2}$$

Here N is the number of videos, c_i the anchor-channel embedding of video i (its video or transcript vector), t_i^+ and t_i^- the embeddings of its true and false text respectively, and $\mathbf{1}[\cdot]$ the indicator function (one when the inequality holds, zero otherwise).

(2) Population separation — "is there a similarity value that splits real from fake?" Harder, because now there is no paired alternative: each video carries only its own modalities, and within a dataset we know which are real and which are fake. We ask whether a single similarity value separates the two populations, and how well. We summarise this with the area under the ROC curve (AUC) — the curve traced by sweeping the decision threshold and plotting, at each setting, how many fakes are caught against how many real videos are wrongly flagged. The AUC has a simple reading: it is the probability that a random fake looks *less coherent* than a random real video (0.5 = no separation, 1 = perfect). Because GroundLie360 has no paired alternative, this is the standard reading there.

(3) Threshold classification — "could you actually deploy it?" The hardest test for a zero-shot score. Even if a separating value exists on the full population, a real detector must *learn* its threshold on some data and apply it to unseen data. We split each dataset $70\%/30\%$, fit the accuracy-maximising threshold on the train half (rule: *fake if* $\cos <$ *thr*), evaluate accuracy on the held-out half, and repeat 50 times, averaging. This is plain deployable classification — "if this similarity were all I had, how well would I classify?" — and, as the results show, it is where the zero-shot cosine collapses.

4.6 Intra-encoder comparison (key methodological decision)

All real-vs-fake comparisons are performed *within* a single encoder, never by comparing absolute cosines across encoders. This is an empirical necessity, not a convention: the modality gap [26] and anisotropy [27] differ so much between encoders (anisotropy ranging $\approx$0.27–0.81 across the four) that an absolute cosine of, say, 0.55 means different things in different spaces. We therefore characterise the geometry of each space and only compare *relative* shifts induced by manipulation inside the same encoder.

5 Results

Our experiments read these embeddings in two ways, and the results follow the same split. The first is *distance-based*: we take the raw cosine, with no training, and ask whether the geometry alone separates manipulated from authentic content (Sec. 5.1). The second is *classifier-based*: we train a probe on the same embeddings and ask not only whether learning recovers what the cosine misses, but whether any success is genuine signal or an artefact of how the benchmark was built (Sec. 5.2).

5.1 Zero-shot: the difficulty and format ladder

We present it in three parts. The first takes FakeVV — the first dataset we analysed — and follows it end to end: what the embedding does, where the signal breaks, and why, building the reading that the rest of the section reuses. The second runs that same reading on each of the two remaining datasets, M3A and GroundLie360, to see where the signal is strongest and where it vanishes. The third reads the overall picture together, drawing the general lesson its individual results share. Table 2 gathers all of it in one place: one row per condition (a manipulation type on a given dataset), ordered by how much signal the raw cosine carries, with its paired (2AFC), population (AUC) and deployable-threshold scores.

FakeVV: the full causal chain (reference case)

FakeVV is our *reference case* — the first dataset we analysed and the one we use to walk the three reading regimes of Sec. 4 in full. It is not the easiest condition: a single-entity swap is a small, surgical change, harder to read than a whole-modality re-pairing. We start here for methodological clarity, not because the manipulation is trivial: each clip is a real news short whose true title has one entity — person, location, organisation or event — swapped for a plausible alternative, so the only inconsistency is the contrast with the video. The pattern it reveals recurs on every other dataset.

Read as a within-pair 2AFC, the video embedding prefers the true title in 74–91% of clips, depending on the encoder (GE2 0.909, both Qwen $\approx$0.82, WAVE 0.74; per-encoder numbers in Table 4). Every encoder's real–fake gap is statistically significant — $p \ll 0.001$, i.e. a gap this large is virtually impossible if the title carried no signal. The signal also survives through the Whisper transcript instead of the video (0.75–0.82, for the three encoders that carry that channel): a single swapped entity already leaves a measurable cross-channel trace. The trace is uneven by type — *location* swaps are easiest (largest visual-context shift), organisation and event intermediate, and *person* hardest, since a swapped name barely changes the scene.

The picture changes the moment we leave the comparative regime. Asked instead as a deployable threshold — one cut learned on a split and applied to the rest — accuracy falls to near chance (0.55–0.63; GE2 0.633), and the population AUC of the same cosine is only 0.59–0.67. The strong comparative signal does not translate into an absolute one; this *high-paired / low-absolute* gap is the format effect that recurs across the whole ladder.

The cause is geometric: the real- and fake-cosine distributions overlap almost completely because the absolute cosine *level* varies far more between videos than the true–fake gap within a video. Across videos the two cosines are near-perfectly correlated (Pearson r of 0.87–0.95) and rise together, so the discriminative gap ($\sim$0.03) is dwarfed by the between-video spread. A within-pair comparison cancels this per-video level as a *common mode*; an absolute threshold cannot, and conditioning the threshold on each video (per-video mean or z-score) fails to recover it.

The same pattern on more adversarial data

On the two harder datasets, the same paired-vs-absolute pattern recurs, but its severity now tracks how *cross-modal* the manipulation is.

M3A — where the absolute signal survives. Out-of-context re-pairing (MM, Sec. 3) is the strongest signal in the study. $\cos(\text{video}, \text{summary})$ separates real from re-paired clips with a 2AFC of 0.81–0.92 and, unlike FakeVV, the absolute reading holds up: AUC 0.70–0.85 and a deployable

TABLE 2: Difficulty/adversariality ladder, ordered by *zero-shot signal*: the signal degrades from a trivial control to the human-crafted floor, and the strong comparative (2AFC) signal does not survive as an absolute one (AUC, deployable threshold). Ranges span the four encoders; GroundLie carries no paired alternative.

Step	Condition	Dataset (exp.)	Paired 2AFC	AUC	Threshold
1	Random out-of-context	M3A control	≈0.99	—	—
2	Similarity out-of-context	M3A MM (A1)	0.81–0.92	0.70–0.85	0.66–0.78
3	Entity swap, title	FakeVV (E1)	0.74–0.91	0.59–0.67	0.55–0.63
4	Entity swap, summary	M3A NEM (B)	0.76–0.93	0.59–0.73	0.57–0.68
5	Cross-modal mismatch	GL360 contra./unsupp.	—	0.56–0.67	∼0.57
6	Plausible lie, all types	GL360 (E1, global)	—	0.44–0.55	0.48–0.54

threshold of 0.66–0.78 (GE2 0.777, i.e. $\sim 78\%$ of clips classified correctly from a single cosine). A random-video control inverts the paired score, confirming the signal is genuinely video-grounded. Entity substitution in the summary (NEM, the textual analogue of FakeVV) reproduces the familiar pattern instead — a strong 2AFC of 0.76–0.93 but a modest threshold (~ 0.61) — because, as in FakeVV, the swap is a small within-pair perturbation.

GroundLie360 — the floor. On the human-crafted Snopes lies, zero-shot $\cos(\text{title}, \text{video})$ is at or below chance (AUC 0.44–0.55; the open-source encoders invert, only GE2 reaches 0.55). This is consistent with the taxonomy (Sec. 3): *false_title* and *false_speech* are unimodal factual lies and *temporal_edit* a localised cut, none of which a global cosine has a causal path to. The signal that does exist follows the *modality* of the manipulation: only the genuinely cross-modal types — *contradictory* (a text–video semantic mismatch) and *unsupported* (a headline the video does not support) — carry a weak coherence signal.

What the ladder reveals

The ladder is ordered by *zero-shot signal*, and that order is itself a result: it follows neither the dataset nor the human severity of the lie, but the *magnitude and cross-modal nature of the semantic displacement* a manipulation produces in embedding space.

At the top, a random out-of-context re-pairing (step 1) swaps whole modalities from unrelated events — an enormous, incoherent displacement — and is trivially separable; it serves only as a control. A similarity-guided swap (step 2, M3A MM) still replaces an entire modality, but with a topically close one, so the displacement is large yet coherent: this is the strongest *usable* signal and the only condition that survives as an absolute threshold. Entity substitution (steps 3–4) is the opposite extreme — a single token changes, displacing the vector by only ~ 0.03, which a within-pair comparison detects but an absolute threshold cannot separate from the between-video spread (the mechanism diagnosed for FakeVV in Sec. 5.1). At the bottom, the human-crafted GroundLie lies separate only where the manipulation is genuinely cross-modal: *contradictory* and *unsupported* content carry a weak but real signal (step 5; GE2 AUC up to 0.73 on single-type cases), whereas *false_title* and *false_speech* — unimodal factual lies whose modalities still cohere — and *temporal_edit* — a localised cut — have no cross-modal path and sit at or below chance (step 6).

Two readings follow. First, detectability tracks the *size of the cross-modal displacement*, not the deceptiveness of the lie to a human: the cruder the modality break the easier it is, while the most surgical entity swaps are the hardest for an absolute score. Second, even where a displacement exists it is reliably read only in *comparative* format; a deployable threshold needs that displacement to exceed the between-video variance, which only the out-of-context swaps achieve.

5.2 Probe audit: signal versus construction bias

Is the space too poor, or the cosine too blunt?

The zero-shot collapse leaves two possibilities: the embeddings do not encode the manipulation, or they do but the cosine does not order it. We test the second by training linear probes on the GroundLie embeddings (50 grouped $70/30$ repetitions; groups are Snopes *events*, so no title leaks across folds) and reading the same two regimes off the learned score: separability (AUC) and operating accuracy. We compare several *feature sets*, each isolating where a signal could live: the single-channel embeddings (*title*, *video*, *transcript*) as controls; their full *concatenation* (a representational ceiling); and a learned per-dimension reweighting of the title–video product ($V \odot T$, which strictly generalises the cosine). As no-embedding controls we add a plain TF-IDF over the title words and thirteen style-only features (length, punctuation, digits, capitalisation).

At first the result looks like a strong supervised detector: a probe on the title embedding reaches AUC 0.82, on the full concatenation 0.86, and a learned reweighted cosine ($V \odot T$) a comparable 0.83, far above the zero-shot 0.44–0.55.

The controls dismantle that reading. Plain TF-IDF over the raw title words (no embeddings at all) already reaches 0.77, and the thirteen style-only features reach 0.64. Decisively, the fakes whose title is *actually true* (the lie lives in the speech or the edit) still separate at 0.79 (down from 0.85 on the full set): the probe reads style, not falsity. **It is not detection; it is dataset-construction bias.** The cause is structural: GroundLie fakes carry the Snopes *claim* (declarative, "A video authentically showed…") while the reals keep the original caption (TikTok *fyp/foryou*), so text style correlates with the label by construction. The same control fails in every modality: GroundLie is multi-label, so each probe can be re-run keeping only the fakes that do *not* lie in that modality. The transcript still separates fakes whose speech is true at 0.68 (and a plain TF-IDF of the ASR words, 0.69, beats all three transcript embeddings), and the

video still separates fakes whose footage is never CGI- or edit-manipulated at 0.68–0.71, as high as on the full set: it reads viral-versus-news content, not manipulation.

Only the relational coherence survives, and it survives two ways. Of all the feature sets, only the low-dimensional cross-channel coherence (the cosines and distances among the channels, a shallow classifier over cross-channel similarities in the spirit of MUSE [5]) behaves like a genuine measure. *Within* GroundLie it separates the fakes that are actually cross-modal (contradictory, unsupported, where a title–video incoherence exists) better than the rest, GE2 0.668 vs. 0.573, the signature of a coherence measure, not a lexical shortcut. On those same rows the title still scores higher (0.86–0.90): the lexical shortcut wins *inside* the dataset even where the lie is genuinely cross-modal. *Across* benchmarks the contrast is decisive: training each probe on one dataset and testing it unchanged on the other, every content and learned probe collapses to near chance (the title to 0.57, the concatenation to 0.57, the reweighted cosine to 0.56; on Qwen, to 0.49–0.53), losing essentially all of its in-dataset advantage, while the bare coherence transfers *intact* for GE2 (0.67, matching its native value). The large supervised number is benchmark memory; only the small zero-shot coherence is portable. The honest reading: no unimodal probe AUC measures veracity; it measures how separable the dataset's fake and real classes are, which here is set by construction style. The fair baseline is TF-IDF, and the one genuine cross-modal signal is real but modest and encoder-dependent.

Why per-type AUCs mislead, and the control that fixes it

A single AUC per fake type is misleading. Because the types are topically segregated, every cell of the type×modality matrix (Table 3) is high even with no causal path — the *title* "predicts" a CGI edit it cannot see almost as well as it predicts a false title. The right control asks, per type, whether the modality the lie *lives in* beats that bias floor. Only two do (**bold** in Table 3): **video** detects *CGI* (the one strong genuinely visual signal) and the **ASR transcript** detects *false_speech*. Everywhere else the causal modality stays at or below its row's floor: *temporal_edit* is missed even by video, and the cross-modal types (*contradictory*, *unsupported*) lift only the relational coherence — weakly, above chance but below the lexical floor of their rows, exactly where a small genuine signal, never a shortcut, belongs.

TABLE 3: Type×modality matrix (GE2; bias audit). AUC, reals vs. all fakes of each type. The embedding *title* column is the bias floor; **bold** = channels that rise clearly above it (video→cgi, ASR→false_speech).

	TF-IDF		Embedding			
Fake type	title	tr.	title	tr.	video	rel.
false_title	0.795	0.672	0.892	0.644	0.714	0.617
false_speech	0.747	**0.846**	0.775	0.785	0.744	0.796
temporal_edit	0.743	0.742	0.848	0.713	0.825	0.587
cgi	0.859	0.752	0.875	0.778	**0.884**	0.730
contradictory	0.744	0.711	0.858	0.665	0.705	0.676
unsupported	0.814	0.695	0.896	0.692	0.739	0.675

6 Discussion — The capability map

The findings reorganise into a capability map: what cross-channel embedding alignment can and cannot reveal about video misinformation. One boundary frames the whole map. No embedding, of any modality, can judge whether a claim is *true*; factual verification is not a perception problem and lies outside what a similarity between channels can measure. What the alignment can measure is semantic *(mis)alignment*: whether a video and the text around it belong together. A mismatch often co-occurs with misinformation, and that is what makes the signal useful, but it is not the same as detecting falsity. With that boundary fixed, the capability sorts into four bands.

- **Measurable zero-shot, as a comparative signal.** The clearest capability is flagging out-of-context content: a video paired with text that does not belong to it. Read as a within-pair choice, off-the-shelf cosines separate re-pairings and entity substitutions well above chance (M3A out-of-context 2AFC 0.81–0.92; FakeVV 0.74–0.91; M3A entity swaps 0.76–0.93). The signal is genuine, but it is a *ranking* between two alternatives, not a verdict on a single clip.
- **The same signal does not survive as an absolute threshold.** This is one finding about one signal, not a second capability that contradicts the first. The very cosine that ranks a fake pairing below its authentic counterpart cannot fix a single global cut-off: the per-pair similarity level, set by topic, dwarfs the manipulation, so the substitutions that scored 0.74–0.93 in comparison collapse to 0.55–0.63 once one threshold must serve every pair, and no probe, non-linearity or calibration recovers the gap. The one exception is M3A out-of-context, strong enough to also work as a threshold (up to 0.78).
- **Recoverable with supervision, but only after subtracting construction bias.** A trained classifier separates GroundLie fakes well, yet most of that separability is the dataset's writing style, not the lie (Sec. 5.2). Once that bias floor is subtracted, three signals survive, each because it has a real causal path: CGI from the video pixels (0.89–0.95 on pure cases, the one strong genuinely visual detection); false speech from the ASR transcript, where the lie is literally in the words (TF-IDF 0.85); and a weak but genuine cross-modal coherence (GE2 ∼0.67), the only supervised feature that passes both a within-dataset validity test and cross-dataset transfer, and only on the aligned encoder. Everything else the probe appears to "detect" is the lexical shortcut.
- **Out of reach for single-vector embeddings.** A single global vector per channel cannot localise anything in time or space, so a whole class of tasks is simply unavailable: detecting or localising temporal edits, pointing to the exact frame of a malicious cut or the exact moment a spoken lie is uttered, and fine visual forensics beyond coarse CGI. And, underneath all of it, none of these is factual verification: whether a true-looking, well-aligned video actually misinforms is not something perception alone can settle.

6.1 Limitations

The strongest caveat is the benchmarks themselves. GroundLie360 is real-world and qualitatively rich, but modest in size, multi-label, and, as Section 5.2 shows, shaped by a construction bias that any supervised number must be read against; FakeVV and M3A are far larger but algorithmically generated, so they need not mirror how real misinformation is produced. The conclusions hold for what these datasets represent, and whether the cross-modal dissonance signal survives on larger, more realistic corpora is open.

A second, structural limitation is the single-vector bottleneck: one global embedding per channel cannot localise a lie in time or space, which is why temporal edits and the exact moment of a spoken claim fall outside its reach. We did probe temporality by segmenting each video into scenes and scoring abrupt drops in adjacent-scene similarity, but the adaptation is awkward and weak (AUC 0.43), and over half of the temporal edits are boundary truncations that leave no visual discontinuity to find at all.

Finally, the audio channel is out of scope: the WAVE audio/AV embeddings carry a token-handling bug and, even after regeneration, remain anomalous enough that we do not rely on them in any experiment, so the spoken and acoustic signal is left for future work. A few sub-analyses also rest on small samples capped by API quota, and the geographic robustness check carries a residual language confound from Whisper auto-detection; these are flagged in place rather than aggregated.

7 Conclusions and Future Work

This work set out to build an endogenous video-misinformation detector. Under the constraints encountered the objective shifted, and rather than force a detector the evidence did not support, the project turned to rigorously characterising what off-the-shelf single-vector multimodal embeddings can and cannot reveal about video misinformation. That adaptation is itself part of the contribution: an exploratory study of a still largely open question, on a problem far less charted than its image-text predecessor.

The central result is a *negative one with a measured cause*, and the nuance matters. Cross-channel semantic similarity measures topical affinity, not veracity, so it does not work as a directly deployable detector. But it is negative about *deployment*, not about the underlying relation: a genuine link between cross-modal dissonance and manipulated video does exist, at least on the benchmarks tested. The signal holds as a within-pair comparison and dissolves only when forced into one absolute threshold, because each manipulation shifts the embedding by far less than ordinary variation between videos, and no probe, non-linearity or calibration closes that gap.

The second contribution is methodological: a bias-audit protocol (TF-IDF and style floors, a type$\times$modality control matrix, and within- and cross-dataset validity tests) that quantifies how much of a supervised detector's apparent success on these benchmarks is construction artefact rather than signal, and isolates the small, encoder-dependent cross-modal coherence that genuinely survives. The resulting capability map, separating what is measurable as a comparison from what bias inflates and what is structurally out of reach, is the practical takeaway.

Future work follows from that map. The portable kernel, a cross-modal dissonance that survives the cross-dataset transfer where every lexical shortcut collapses, suggests the cosine, not the idea, is the bottleneck, and that specialised similarity measures could extract more. Two roles follow naturally: as a feature, the dissonance score is one more cue a richer learned detector can weigh; and in the agentic direction automated fact-checking is now taking, where a language model reasons over a claim and checks it against external evidence, a cheap single-vector dissonance score can serve as a flag alongside other signals and as a triage filter before an expensive verification is launched. Validating the signal on larger, non-synthetic data, and extending it to the audio channel left out here, are the natural next steps.

Acknowledgments

I thank my supervisor, Ernest Valveny, for his guidance throughout this project, and the Escola d'Enginyeria (UAB) for the GPU computing cluster without which the large-scale embedding generation behind this study would not have been possible.

To my family in Villena (Alicante): thank you for your patience and your unconditional support, for holding home steady across the distance and through the long silences while I was in Barcelona. None of this would mean anything without you. And to every one of the extraordinary people I have been lucky enough to meet here in Catalonia, who walked this stretch of the road beside me and quietly made me who I am today: thank you.

This is not an ending, only the first of many beginnings.

References

[1] P. Qi, Y. Bu, J. Cao, W. Ji, R. Shui, J. Xiao, D. Wang, and T.-S. Chua, "Fakesv: A multimodal benchmark with rich social context for fake news detection on short video platforms," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, pp. 14 444–14 452, June 2023. [Online]. Available: https://doi.org/10.1609/aaai.v37i12.26689

[2] L. Chen, Z. Qian, P. Li, and Q. Zhu, "Multimodal fake news video explanation: Dataset, analysis and evaluation," *arXiv preprint arXiv:2501.08514*, 2025. [Online]. Available: https://arxiv.org/abs/2501.08514

[3] Z. Zeng, J. Wu, M. Luo, X. Kong, Z. Ma, G. Dai, and Q. Zheng, "Understand, refine and summarize: Multi-view knowledge progressive enhancement learning for fake news video detection," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25. ACM, Oct. 2025, pp. 9216–9225. [Online]. Available: https://doi.org/10.1145/3746027.3754551

[4] K. Niu, D. Xu, B. Yang, W. Liu, and Z. Wang, "Pioneering explainable video fact-checking with a new dataset and multi-role multimodal model approach," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 27, pp.

28 276–28 283, Apr. 2025. [Online]. Available: https://doi.org/10.1609/aaai.v39i27.35048

[5] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, "Similarity over factuality: Are we making progress on multimodal out-of-context misinformation detection?" in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2025. [Online]. Available: https://arxiv.org/abs/2407.13488

[6] O. Papadopoulou, M. Zampoglou, S. Papadopoulos, and I. Kompatsiaris, "A corpus of debunked and verified user-generated videos," *Online Information Review*, vol. 43, no. 1, pp. 72–88, Feb. 2019. [Online]. Available: https://doi.org/10.1108/OIR-03-2018-0101

[7] B. Yang, D. Xu, K. Niu, W. Liu, Z. Wang, and M. Kankanhalli, "A new dataset and benchmark for grounding multimodal misinformation," in *Proceedings of the 33rd ACM International Conference on Multimedia*, ser. MM '25. ACM, Oct. 2025, pp. 12 571–12 577. [Online]. Available: https://doi.org/10.1145/3746027.3758191

[8] Y. Wang, L. Chen, Z. Qian, and P. Li, "Official-nv: An llm-generated news video dataset for multimodal fake news detection," *arXiv preprint arXiv:2407.19493*, 2024. [Online]. Available: https://arxiv.org/abs/2407.19493

[9] Y. Wang, Z. Qian, and P. Li, *FMNV: A Dataset of Media-Published News Videos for Fake News Detection*. Springer Nature Singapore, 2025, pp. 321–332. [Online]. Available: https://doi.org/10.1007/978-981-96-9794-6_27

[10] H. Li, Y. Wang, S. Tang, L. Wu, L. Cheng, and Z. Zhong, "Towards unified multimodal misinformation detection in social media: A benchmark dataset and baseline," *arXiv preprint arXiv:2509.25991*, 2025. [Online]. Available: https://arxiv.org/abs/2509.25991

[11] F. Zhang, D. Li, Q. Zhang, J. Chen, G. Liu, J. Lin, J. Yan, J. Liu, and Z.-J. Zha, "Fact-r1: Towards explainable video misinformation detection with deep reasoning," *arXiv preprint arXiv:2505.16836*, 2025. [Online]. Available: https://arxiv.org/abs/2505.16836

[12] Q. Xu, H. Chen, H. Du, H. Zhang, S. Łukasik, T. Zhu, and X. Yu, "M3A: A multimodal misinformation dataset for media authenticity analysis," *Computer Vision and Image Understanding*, vol. 249, p. 104205, 2024. [Online]. Available: https://doi.org/10.1016/j.cviu.2024.104205

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[14] H. Xu, G. Ghosh, P.-Y. Huang, D. Okhonko, A. Aghajanyan, F. Metze, L. Zettlemoyer, and C. Feichtenhofer, "Videoclip: Contrastive pre-training for zero-shot video-text understanding," *arXiv preprint arXiv:2109.14084*, 2021. [Online]. Available: https://arxiv.org/abs/2109.14084

[15] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," *arXiv preprint arXiv:2106.13043*, 2021. [Online]. Available: https://arxiv.org/abs/2106.13043

[16] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "Imagebind: One embedding space to bind them all," *arXiv preprint arXiv:2305.05665*, 2023. [Online]. Available: https://arxiv.org/abs/2305.05665

[17] B. Zhu, B. Lin, M. Ning, Y. Yan, J. Cui, H. Wang, Y. Pang, W. Jiang, J. Zhang, Z. Li, W. Zhang, Z. Li, W. Liu, and L. Yuan, "Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment," *arXiv preprint arXiv:2310.01852*, 2023. [Online]. Available: https://arxiv.org/abs/2310.01852

[18] Y. Wang, K. Li, X. Li, J. Yu, Y. He, C. Wang, G. Chen, B. Pei, Z. Yan, R. Zheng, J. Xu, Z. Wang, Y. Shi, T. Jiang, S. Li, H. Zhang, Y. Huang, Y. Qiao, Y. Wang, and L. Wang, "Internvideo2: Scaling foundation models for multimodal video understanding," *arXiv preprint arXiv:2403.15377*, 2024. [Online]. Available: https://arxiv.org/abs/2403.15377

[19] Z. Jiang, R. Meng, X. Yang, S. Yavuz, Y. Zhou, and W. Chen, "Vlm2vec: Training vision-language models for massive multimodal embedding tasks," *arXiv preprint arXiv:2410.05160*, 2024. [Online]. Available: https://arxiv.org/abs/2410.05160

[20] C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang, P. Chen, Y. Li, S. Lin, S. Zhao, K. Li, T. Xu, X. Zheng, E. Chen, C. Shan, R. He, and X. Sun, "Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis," *arXiv preprint arXiv:2405.21075*, 2024. [Online]. Available: https://arxiv.org/abs/2405.21075

[21] G. Cicchetti, E. Grassucci, and D. Comminiello, "A triangle enables multimodal alignment beyond cosine similarity," 2025. [Online]. Available: https://arxiv.org/abs/2509.24734

[22] Google, "Gemini embedding 2: Our first natively multimodal embedding model," Google Blog, Mar. 2026, public preview released 10 March 2026. [Online]. Available: https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-embedding-2/

[23] G. Luo, T. Darrell, and A. Rohrbach, "Newsclippings: Automatic generation of out-of-context multimodal media," in *Proceedings of the 2021 Conference on*

Empirical Methods in Natural Language Processing (EMNLP), 2021. [Online]. Available: https://arxiv.org/abs/2104.05893

[24] S. Aneja, C. Bregler, and M. Nießner, "Cosmos: Catching out-of-context misinformation with self-supervised learning," in *arXiv preprint arXiv:2101.06278*, 2021. [Online]. Available: https://arxiv.org/abs/2101.06278

[25] S.-I. Papadopoulos, C. Koutlis, S. Papadopoulos, and P. C. Petrantonakis, "Verite: A robust benchmark for multimodal misinformation detection accounting for unimodal bias," *International Journal of Multimedia Information Retrieval*, vol. 13, 2024. [Online]. Available: https://doi.org/10.1007/s13735-023-00312-6

[26] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou, "Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [Online]. Available: https://arxiv.org/abs/2203.02053

[27] T. Wang and P. Isola, "Understanding contrastive representation learning through alignment and uniformity on the hypersphere," in *International Conference on Machine Learning (ICML)*, 2020. [Online]. Available: https://arxiv.org/abs/2005.10242

[28] C. Tian, Y. Wang, C. Yang, J. Guan, Z. Lin, Y. Liu, X. Chen, and Q. Ho, "Exposing cross-modal consistency for fake news detection in short-form videos," *arXiv preprint arXiv:2603.14992*, 2026. [Online]. Available: https://arxiv.org/abs/2603.14992

[29] M. Li, Y. Zhang, D. Long, K. Chen, S. Song, S. Bai, Z. Yang, P. Xie, A. Yang, D. Liu, J. Zhou, and J. Lin, "Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking," *arXiv preprint*, 2026.

[30] C. Tang, Q. Xiao, K. Mei, T. Wang, F. Rao, and C. Zhang, "WAVE: Learning unified & versatile audio-visual embeddings with multimodal LLM," *arXiv preprint arXiv:2509.21990*, 2025.

APPENDIX

A.1 Experiment catalogue — FakeVV

Setup. FakeVV test set; each clip has a true title, a false title (one entity swapped: person/location/organisation/event) and a Whisper transcript. Four encoders (Qwen3-VL 2B/8B, WAVE-7B, GE2); all cosines on L_2-normalised global vectors. Each experiment builds its own id set as the intersection of the channels it needs, which is why N differs across edges.

E1 — video↔title (2AFC). For ids = {video, text_real, text_fake} present in the model's DB, compute $\cos(V, T^+)$, $\cos(V, T^-)$ and report $\text{acc} = \text{mean}[\cos(V, T^+) > \cos(V, T^-)]$ with the mean real and fake cosines, the gap, a paired t test, p, Cohen's d and a bootstrap CI (`summarize_paired`). $N = 813$ for all four encoders: the video embedding was materialised only on the $\leq$120 s subset common to every encoder (GE2's 120 s video limit), so the video edge is fixed to this matched set.

E2 — transcript↔title (2AFC). Same metric over ids = {transcript, text_real, text_fake}; this edge needs no video, so N rises to 990–993 (near-full test set with a valid transcript). WAVE is skipped (transcript/title channels absent in its DB). The N difference vs. E1 is purely a channel-availability artefact, not a change of sample.

Per-category 2AFC. E1 accuracy partitioned by swapped-entity type, averaged across models (ordering location > organisation ≈ event > person).

E1b / E2b — threshold classification. 50 repetitions; per repetition a stratified 70/30 split; `fit_threshold` picks the train-accuracy-maximising cut under the rule *fake if* $\cos < \text{thr}$; test accuracy is averaged (mean±std). $N_{\text{vid}} = 813$ (E1b) / ∼990 (E2b). Reports the collapse to 0.55–0.63.

E1c — variance diagnosis. Per video: within-pair gap $\cos_{\text{real}} - \cos_{\text{fake}}$; level $(\cos_{\text{real}} + \cos_{\text{fake}})/2$; between-video σ of the level; ratio gap$/\sigma$; the across-video correlation $\text{corr}(\cos_{\text{real}}, \cos_{\text{fake}})$; and raw vs. per-video-mean-calibrated threshold accuracy. Shows a correlation of 0.87–0.95, a ratio 0.18–0.63, and that calibration does not help (GE2 0.630→0.615). Both readings rest on the same numerator (the within-pair gap): the paired test cancels each video's shared level, while a global threshold must absorb the between-video spread of that level as noise — same signal, different denominator.

E1d — per-video-conditioned threshold. Ladder raw → calibrated ($-$per-video μ) → z-score ($\div$per-video σ), with μ, σ from a reference corpus of true titles in train (leave-one-out, $C = V\,\text{Tr}^\top$). None recovers the paired signal.

Three-regime self-contained block. On the common ids with all channels ($N \approx$ 806–809), recomputes per edge (E1 video–title, E2 transcript–title, and the title·video·transcript **triangle** $\text{area}_{\text{true}}$) the three regimes: 2AFC, AUC (Mann–Whitney) and grouped threshold. The triangle does not beat the single video edge for GE2 (2AFC 0.911 vs. 0.909); it lifts Qwen's 2AFC (0.82→0.87) only because transcript–title is itself a separate edge.

A.2 Experiment catalogue — GroundLie360

Setup. 2,044 target videos from Snopes (995 real / 1,049 fake), filtered to the $\leq$120 s subset (1,722); six fake types annotated as a multi-label vector (a video may carry several). Every title↔video pair is a genuine real-world post (no synthetic re-pairing), but real and fake are *disjoint* video sets here (unlike FakeVV, where both share one clip); the resulting content confound is audited in App. A.5. Fakeness convention: for a single pairwise edge (e.g. title↔video) high $= -\cos$ (less aligned = more suspect); for the triangle dissonance measure, high $= +$area. All comparisons intra-encoder.

E1 — $\cos(\text{title}, \text{video})$**, AUC by type.** The negative cosine is the per-video fakeness score; AUC (Mann–Whitney) is computed against the real population, overall and per fake type. $N_{\text{fake}} \approx 899$. Zero-shot; reproduces the FakeVV edge on a harder dataset. Only GE2 clears chance (overall 0.551; best on contradictory 0.610), and CGI inverts

(< 0.5) for every encoder.

E2 — false-speech textual window. For videos with false_speech$=1$ and false_title$=0$ (true headline, false spoken claim), the transcript is split into the annotated lying span (`window_fake`) and the surrounding clean text (`window_clean`); we test whether the clean text aligns better with the (true) title: $\Delta = \cos(T_{\text{clean}}, \text{title}) - \cos(T_{\text{fake}}, \text{title})$, $\Delta > 0$ expected. $N \approx 17$ (three videos have empty clean windows). The effect is tiny and near chance ($\Delta = +0.042$ for Qwen2B, $+0.022$ for GE2), so it is reported for completeness only.

E3 — adjacent-scene coherence (temporal-edit detection). Per video, the mean cosine between consecutive scene embeddings; edited vs. non-edited videos are compared at the video level. *Scope:* only *internal* cuts (between two scenes) are visible; end-trims have no neighbouring scene and are structurally invisible.

E3a — cut localisation. Given a video known to be edited ($N = 24$ with an internal annotated cut), rank its transitions by coherence and check whether the annotated cut is the least coherent: hit@1 (fraction where it is the single least coherent, strict top-1) and mean normalised rank ($0 =$ always lowest, $0.5 =$ chance). A weak, noisy localisation signal exists (Qwen2B hit@1 0.58 vs. chance 0.39, mean rank 0.21; GE2 mean rank 0.27 but hit@1 only 0.42). Localisation, not detection.

E3b — anomalous-cut detection. Per video, z-score its transition coherences (removing each video's own level) and take the deepest dip ($-\min z$); AUC of this dip vs. the mean-coherence baseline. Both stay at chance and the dip is no better than the baseline (AUC 0.43–0.48 vs. 0.49–0.54): edited videos are not detectable, since end-trims — the dominant edit — leave no neighbouring scene.

E6b — title·video·transcript triangle. $\text{area}_{\text{true}}$ as fakeness; AUC overall and by type, the gain over E1, and pure (single-type) vs. marginal (any co-occurrence) AUC. Finding: the triangle recovers CGI (gain $+0.20$ to $+0.27$) and helps the cross-modal types (contradictory, unsupported), but inverts false_speech (-0.20 to -0.30, since the spoken lie is self-consistent); temporal_edit stays at chance. Only GE2 shows a consistent signal across types.

Type×modality matrix (negative control). Each modality (title, video, transcript) and metadata field is scored against each of the six fake types in isolation; a cell that scores high without a causal path to that type marks the bias floor. Full matrix in Table 3.

A.3 Experiment catalogue — M3A

Setup. 20,000 clips ($\leq$120 s) from the $\sim$182k video+audio population. The text channel is the *summary* (a $\sim$28-word BART description of the post, not a headline). Two manipulation families from the M3A taxonomy: NEM (named-entity swap in the summary: person / location / organisation / complete) and MM (multimodality mismatching: re-pairing the text with another event's summary by cosine similarity, $\geq$30-day gap). All scores use $-\cos$ as the fakeness direction. GE2 runs on a quota-limited subset, and WAVE is omitted: its M3A video set is incomplete ($15{,}294/20{,}000$) and its NEM/MM scores sit at chance (≤ 0.52).

A1 — $\cos(\text{video}, \text{summary})$**.** Real vs. MM (text-swap), in the three regimes (paired 2AFC $\rightarrow$ AUC $\rightarrow$ threshold). Isolates the single edge the mismatch breaks. N=4,389 (open-source) / 2,177 (GE2). This is the strongest absolute signal in the study (2AFC 0.805–0.920, AUC 0.699–0.846), yet even here the deployable threshold reaches only 0.655–0.777 (Table 4).

A3 — summary·video·transcript triangle. $\text{area}_{\text{true}}$ (consistent with the rest of the study), same three regimes. Adding the transcript edge does not help: AUC falls to 0.636–0.685 vs. A1's 0.699–0.846, so the single broken edge already carries the signal.

B — NEM by subtype. 2AFC and threshold per entity type over the summary; open-source only (GE2 has no NEM embeddings). The signal is monotone in how much text changes: *complete* (whole-entity replacement) is easiest (2AFC 0.90–0.93), *organisation* hardest (0.76–0.80); a single entity swap barely moves the embedding, so the threshold stays near chance (0.57–0.61, Table 5).

Control N2 (random video). Re-running B's paired test with a *random* video in place of the matched one collapses the 2AFC below chance (0.36–0.42 across subtypes and encoders): if the NEM text merely "sounded wrong" the accuracy would survive the swap, so the paired signal is genuinely grounded in the matched video, not a text artefact.

Lexical caveat (TF-IDF). A bag-of-words classifier on the summary text alone (real vs. NEM, GroupKFold by video) reaches AUC 0.73–0.85: the NEM generator leaves a lexical fingerprint, so part of the apparent NEM detection is available without any cross-modal embedding.

P4 — OOD robustness. AUC of the A1 signal stratified by geography and topic (Qwen3-VL 2B reference). It holds across strata: geography AUC 0.63–0.74, topic AUC 0.66–0.73, with no stratum collapsing to chance (Table 9).

A.4 Extended results tables

This section gives the full per-(dataset, encoder, condition) numbers behind the ladder (Table 2) and the probe audit (Sec. 5.2). All cosines are on L_2-normalised global vectors; "2AFC" is the within-pair accuracy, "AUC" the population Mann–Whitney AUC of the (fakeness) score, and "Thr" the deployable-threshold accuracy (mean over 50 stratified $70/30$ splits). WAVE is reported only where its embeddings are reliable: its FakeVV channels were regenerated after the token fix, but its M3A video set is still incomplete ($15{,}294/20{,}000$) and its NEM/MM scores collapse to chance, so WAVE is omitted from the M3A rows.

TABLE 4: Zero-shot, paired datasets (FakeVV and M3A), per encoder and edge. FakeVV rows use the common-channel subset ($N \approx 806$–809); the video↔title edge on the full set gives WAVE 2AFC 0.738 (N=813). M3A MM is reported both as the single $\cos(\text{video}, \text{summary})$ edge (A1) and the summary·video·transcript triangle (A3); M3A NEM is the summary entity swap (TOTAL over subtypes).

Condition	Encoder	2AFC	AUC	Thr
FakeVV — video↔title (E1)				
	Qwen3-VL 2B	0.821	0.595	0.565
	Qwen3-VL 8B	0.818	0.590	0.560
	GE2	0.909	0.673	0.635
FakeVV — transcript↔title (E2)				
	Qwen3-VL 2B	0.767	0.591	0.567
	Qwen3-VL 8B	0.748	0.585	0.564
	GE2	0.819	0.601	0.570
FakeVV — title·video·transcript triangle				
	Qwen3-VL 2B	0.868	0.593	0.565
	Qwen3-VL 8B	0.869	0.600	0.569
	GE2	0.911	0.619	0.591
M3A MM — out-of-context, single edge cos*(V,summary) (A1)*				
	Qwen3-VL 2B	0.805	0.699	0.655
	Qwen3-VL 8B	0.832	0.718	0.665
	GE2	0.920	0.846	0.777
M3A MM — out-of-context, summary·video·transcript triangle (A3)				
	Qwen3-VL 2B	0.821	0.636	0.591
	Qwen3-VL 8B	0.861	0.645	0.603
	GE2	0.906	0.685	0.629
M3A NEM — summary entity swap, TOTAL (B)				
	Qwen3-VL 2B	0.828	0.650	0.612
	Qwen3-VL 8B	0.859	0.643	0.607

TABLE 5: M3A NEM 2AFC by entity subtype (open-source encoders; WAVE collapses to chance, ≤ 0.52). *complete* (whole-entity replacement) is easiest; *organisation* hardest.

Encoder	person	location	org.	complete	TOTAL
Qwen3-VL 2B	0.807	0.815	0.759	0.904	0.828
Qwen3-VL 8B	0.829	0.858	0.798	0.928	0.859

TABLE 6: GroundLie360 zero-shot $\cos(\text{title}, \text{video})$ AUC by fake type (high $= -\cos$). At or below chance for every type and encoder except GE2, and only the cross-modal types (*contradictory*, *unsupported*) lift GE2 clearly. *cgi* inverts (fakes *more* aligned, AUC < 0.5): a CGI clip is a single internally coherent fabricated scene, so the cross-modal cosine sees alignment, not a mismatch — coherence cannot flag it (visual forensics can, Table 3).

Fake type	N	Qw2B	Qw8B	WAVE	GE2
ALL fake	899	0.442	0.462	0.489	0.551
false_title	689	0.443	0.457	0.497	0.554
false_speech	124	0.475	0.505	0.524	0.553
temporal_edit	77	0.444	0.445	0.476	0.555
cgi	192	0.348	0.395	0.381	0.496
contradictory	159	0.478	0.486	0.517	0.610
unsupported	373	0.465	0.479	0.524	0.568

A.5 Dataset construction bias — full audit

These tables are the full evidence that the high GroundLie360 unimodal AUCs measure construction style, not veracity (Sec. 5.2). All values are AUC against the real population; "no-manip. subset" keeps only the fakes that do *not* lie

TABLE 7: Supervised linear probe, full results ($50 \times 70/30$ grouped splits; group = video in FakeVV/M3A, Snopes event in GroundLie360; full dimension, C=0.1, class-balanced). AUC and operating accuracy (ACC); chance = 0.500. `video`/`transcript` are no-leak controls (identical in the paired sets $\Rightarrow$ 0.5). `had. edge` = $V \odot T$ and `had. tri` = $[V \odot T, V \odot \mathrm{Tr}, T \odot \mathrm{Tr}]$ are the *full-dimension* learned Hadamard (reweighted cosine), *not* the low-dimensional $[\cos, \mathrm{dist}]$ coherence of Tables 8/3. The high unimodal AUCs are construction bias (Table 11); these learned $V \odot T$ sets score high here but collapse on transfer like the content probes (Table 8, `had.`). Only the bare coherence (Table 8, `rel.`) is portable. Bold = best feature set per encoder within each dataset.

		Qwen3-VL 2B		Qwen3-VL 8B		GE2	
Dataset	Feature set	AUC	ACC	AUC	ACC	AUC	ACC
GroundLie360	video	0.669	0.636	0.681	0.645	0.670	0.636
	title	0.736	0.681	0.757	0.693	0.818	0.744
	transcript	0.600	0.578	0.590	0.573	0.612	0.584
	concat	**0.787**	**0.718**	0.791	0.710	**0.864**	**0.781**
	had. edge	0.756	0.687	0.769	0.702	0.832	0.757
	had. tri	0.786	0.712	**0.805**	**0.723**	0.858	0.779
M3A	video	0.500	0.500	0.500	0.500	0.500	0.500
	title	0.523	0.516	0.499	0.488	0.610	0.579
	transcript	0.500	0.500	0.500	0.500	0.500	0.500
	concat	0.539	0.528	0.540	0.526	0.601	0.572
	had. edge	0.672	0.623	0.665	0.617	0.835	0.758
	had. tri	**0.706**	**0.645**	**0.749**	**0.680**	**0.870**	**0.790**
FakeVV	video	0.500	0.500	0.500	0.500	0.500	0.500
	title	**0.663**	**0.619**	**0.687**	**0.636**	0.717	0.658
	concat	0.662	0.618	0.683	0.633	0.727	0.664
	had. edge	0.638	0.601	0.684	0.631	**0.771**	**0.702**

TABLE 8: Cross-dataset transfer (probe AUC): train on one benchmark, test unchanged on the other. Every content/learned probe loses its in-dataset advantage and collapses to $\sim$0.5; only the bare coherence (`rel.`) transfers, and only for GE2, where the transferred value equals the native zero-shot ($0.672 \approx 0.671$). `had.` = full-dimension $V \odot T$.

Encoder	Feat.	intra-GL	intra-FVV	GL$\rightarrow$FVV	FVV$\rightarrow$GL
GE2	title	0.856	0.752	0.569	0.567
	concat	0.872	0.739	0.569	0.597
	rel.	0.546	0.671	0.672	0.551
	had.	0.869	0.771	0.557	0.581
Qw2B	title	0.767	0.661	0.506	0.529
	concat	0.780	0.646	0.522	0.527
	rel.	0.554	0.594	0.405	0.442
	had.	0.801	0.637	0.487	0.519
Qw8B	title	0.769	0.675	0.501	0.525
	concat	0.801	0.656	0.523	0.519
	rel.	0.529	0.589	0.410	0.462
	had.	0.829	0.687	0.524	0.509

in that modality, so a genuine detector should drop there and a style shortcut should not. These audit probes use a 5-fold StratifiedKFold split (event-ungrouped), so their absolute AUCs run $\sim$0.02–0.03 above the event-grouped probe of Sec. 5.2 (e.g. title 0.85 here vs. 0.82 there); the evidence is the relative drop within each column, not the absolute level.

A.6 Reproducibility

The analysis is regenerable from the raw videos through three deterministic stages. *(i) Subset selection.* A seeded, stratified index selects the $\leq$120 s clips of each dataset (stratifying by manipulation type and, for M3A, by geography/topic), so the exact id sets used by every experiment are fixed and reproducible. *(ii) Embedding generation.* Each encoder embeds every channel offline and writes *one SQLite database per (encoder $\times$ dataset)*. Each database holds the raw `float32` global vectors keyed by raw video id and channel (`video`, `text`/`summary`/`title`, `transcript`), plus the per-scene segment embeddings used by the temporal experiments; these vectors are the single source of truth and are never recomputed downstream. *(iii) Analysis.* A single master notebook reads only from these databases and produces every number, table and figure in this report, so re-running it on the published databases reproduces the results exactly. The Python environment is pinned in `environment.yml`. The complete source code is publicly available at `https://github.com/luisdomene4/MAVIS`.

TABLE 9: P4 — out-of-distribution robustness of the M3A out-of-context signal ($\cos(\text{video}, \text{summary})$, Qwen3-VL 2B) by geography stratum ($N \geq 100$); AUC with bootstrap 95% CI (2000 resamples). No catastrophic multilingual collapse; the signal only dips where the cosine is weakest (International).

Geography	N	AUC	95% CI
North America	2331	0.711	[0.696, 0.726]
UK	816	0.685	[0.659, 0.711]
Asia	459	0.687	[0.654, 0.723]
Australia	232	0.703	[0.655, 0.750]
Europe	224	0.714	[0.667, 0.760]
International	190	0.630	[0.574, 0.683]
Middle East	137	0.743	[0.685, 0.798]

TABLE 10: Title-channel audit (GE2; N=1606, 829 fake / 777 real). A plain TF-IDF over the raw title words rivals the embedding, 13 style-only features (no word identity) already reach 0.64, and the fakes whose title is *actually true* still separate at 0.79 — the probe reads declarative Snopes-claim style, not the lie.

Feature	AUC
Title embedding	0.853
after de-duplication (-66)	0.838
TF-IDF words (word 1–2-grams)	0.770
TF-IDF characters	0.787
Style only (13 format features)	0.637

TABLE 11: Per-modality bias audit (GE2 / Qwen3-VL 2B / Qwen3-VL 8B). Every unimodal channel separates the no-manipulation fakes almost as well as the full set; metadata is a non-confound; only the relational coherence behaves like a measure (drops on the non-cross-modal fakes). Here `Relational` = the low-dimensional $[\cos, \text{dist}]$ coherence scalars (not the $V \odot T$ Hadamard). The ASR TF-IDF (0.692, no embedding) beats all three transcript embeddings.

Probe	GE2	Qw2B	Qw8B
Video, full	0.716	0.696	0.698
Video, no-manip. subset	0.709	0.683	0.676
Transcript, full	0.674	0.632	0.634
Transcript, true-speech subset	0.683	0.649	0.650
Title, full	0.853	0.767	0.766
Title, true-title subset	0.791	0.706	0.726
Concat	0.873	0.789	0.787
Concat, no title	0.739	0.700	0.688
Relational, full	0.625	0.595	0.601
Relational, cross-modal fakes	0.668	0.610	0.612
Relational, non-cross-modal	0.573	0.588	0.598
Metadata / lexical (encoder-independent)			
Duration		0.564	
Platform		0.528	
Year		0.486	
ASR transcript TF-IDF		0.692	