Human-Centered Multimodal Fusion for Sexism Detection in TikTok Videos with Eye-Tracking, Heart Rate, and EEG Signals

Fusión multimodal centrada en el humano para la detección de sexismo en vídeos de TikTok mediante seguimiento ocular, ritmo cardíaco y señales EEG

Iván Arcos,[1] **Paolo Rosso**[1,2]
[1]PRHLT Research Center, Universitat Politècnica de València, Spain
[2]ValgrAI – Valencian Graduate School and Research Network of Artificial Intelligence, Spain
iarcgab@etsinf.upv.es, prosso@dsic.upv.es

Abstract: Sexism ranges from explicit misogyny to subtle forms of prejudice disguised as humor, irony, or seemingly harmless remarks. This challenge is particularly acute on short-form video platforms such as TikTok, where meaning emerges from the interaction of spoken audio, on-screen text, gestures, editing, and social context. To address this limitation, we propose a human-centered approach that augments content features with physiological data. We rely on a multimodal resource built on 2524 TikTok videos from the EXIST 2026 dataset, evaluated by 10 subjects across 15 sessions while Eye-Tracking (ET), Heart Rate (HR), and Electroencephalography (EEG) signals were recorded. Our analysis reveals significant physiological differences between sexist and non-sexist content, including longer reaction times, increased fixation patterns, and differential EEG signatures. Building on these findings, we propose a hierarchical multimodal fusion architecture that integrates physiological sequences with enriched textual-visual features derived from `Qwen3-VL-4B-Instruct` using four sampled frames per video. The complete model achieves an AUC of 0.790 in binary sexism detection and improves over the text+frames baseline by 3.9% in Task 1 (binary sexism detection), 6.5% in Task 2 (direct vs. judgmental sexism), and 6.8% in Task 3 (multiclass and multilabel sexism categories). These results show that human physiological responses provide a robust complementary signal for improving automated sexism moderation in social media videos.
Keywords: sexism detection, TikTok videos, eye-tracking, heart rate, EEG.

Resumen: El sexismo abarca desde la misoginia explícita hasta prejuicios sutiles disfrazados de humor, ironía o comentarios aparentemente inofensivos. Detectarlo resulta especialmente difícil en TikTok, donde el sentido surge del audio, el texto en pantalla, los gestos, la edición y el contexto social. Proponemos un enfoque centrado en el humano que complementa el contenido con datos fisiológicos. El recurso multimodal reúne 2524 vídeos de TikTok de EXIST 2026, evaluados por 10 sujetos en 15 sesiones con seguimiento ocular (ET), ritmo cardíaco (HR) y electroencefalografía (EEG). El análisis identifica diferencias fisiológicas significativas entre el contenido sexista y no sexista: mayores tiempos de reacción, más fijaciones y distintas firmas EEG. A partir de estos resultados, proponemos una arquitectura jerárquica de fusión multimodal que integra secuencias fisiológicas con características textuales-visuales enriquecidas mediante `Qwen3-VL-4B-Instruct` sobre cuatro fotogramas por vídeo. El modelo alcanza un AUC de 0.790 en la detección binaria y mejora respecto a texto+fotogramas un 3.9% en la Tarea 1 (detección binaria), un 6.5% en la Tarea 2 (sexismo directo frente a valorativo) y un 6.8% en la Tarea 3 (categorías multiclase y multietiqueta). Los resultados confirman que las respuestas fisiológicas humanas aportan una señal complementaria robusta para mejorar la moderación automatizada del sexismo en vídeos de redes sociales.
Palabras clave: detección de sexismo, vídeos de TikTok, seguimiento ocular, ritmo cardíaco, EEG.

1 Introduction

Sexism ranges from explicit misogyny to subtle prejudice expressed through stereotypes, objectification, irony, or humor. Detecting it is especially difficult on TikTok, where meaning emerges from spoken language, on-screen text, gestures, editing, and context. Although sexism detection has progressed from text-only to multimodal models (Arcos and Rosso, 2024;

De Grazia et al., 2026; Italiani et al., 2026), implicit and judgmental content remains challenging because a content-only classifier cannot observe the cognitive and affective response elicited in viewers.

Social platforms amplify gendered inequalities while exposing young audiences to rapidly changing audiovisual narratives. TikTok is especially influential because its recommendation mechanism and immersive format encourage repeated, context-light consumption. Harmfulness may depend on whether a speaker endorses a sexist statement, quotes it critically, or uses irony; the same surface cues can therefore support opposing interpretations. Robust moderation must capture both the stimulus and the interpretative effort it demands.

We therefore investigate physiological reactions as complementary evidence rather than as a replacement for human judgment. Eye-Tracking (ET), Heart Rate (HR), and Electroencephalography (EEG) can capture attention, arousal, and interpretative effort continuously. We use 2524 TikTok videos from EXIST 2026, evaluated by 10 subjects across 15 sessions, and combine their physiological responses with textual-visual representations generated from four frames by `Qwen3-VL-4B-Instruct`[1] through a hierarchical attention model. Audio integration is left for future work.

Unlike self-reports, these signals are continuous and largely involuntary. They may reveal hesitation, attention allocation, or emotional salience that annotators cannot easily verbalize. We use them only as an additional content-level training signal, never as a substitute for human judgment or as a proposed mechanism for monitoring users.

We ask whether physiological responses differ across sexism labels (**RQ1**), whether visual-semantic and physiological signals improve sexism classification (**RQ2**), and whether physiology significantly improves a strong text+frames baseline (**RQ3**). Our contributions are a synchronized human-centered resource, physiological evidence across three sexism tasks, and a hierarchical fusion model that yields consistent gains.

[1] `https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct`

2 *Related Work*

Hate speech, misogyny, and sexism overlap but are not equivalent. Hate speech is broad (Nockleby, 2000); misogyny expresses hostility toward women; sexism additionally includes stereotyping, discrimination, and subtle or benevolent prejudice. This distinction matters computationally because implicit sexism may contain no overtly offensive lexical or visual marker.

Sexism detection has evolved from feature-based text classifiers using TF–IDF and n-grams (Jha and Mamidi, 2017; Pelle and Moreira, 2017) to neural representations (Gasparini et al., 2018; Grosz and Conde-Cespedes, 2020), transformers (Parikh et al., 2021), and multimodal systems (Rizzi et al., 2023; Arcos and Rosso, 2024). SemEval and EXIST have driven multilingual and fine-grained evaluation (Kirk et al., 2023; Plaza et al., 2025). Nevertheless, irony, in-group humor, and communicative stance remain difficult even for large models (Li et al., 2024; Abercrombie et al., 2024).

Recent work addresses this limitation through richer video and relational modeling. FineMuSe introduces hierarchical annotations for subtle communicative strategies (De Grazia et al., 2026); Qwen3-VL provides unified visual-textual representations (Bai et al., 2025); MultiHateLoc targets temporal localization (Sun et al., 2026); and MemeWeaver exploits inter-instance relations (Italiani et al., 2026). These developments motivate architectures that preserve temporal and hierarchical structure rather than treating a video as a static image-caption pair.

Physiological signals provide complementary evidence about perception. ET reflects attention and cognitive load (Khurana et al., 2023; Kar et al., 2025); HR reflects arousal (Azarbarzin et al., 2014); and EEG captures neural dynamics with high temporal resolution (Lin and Yang, 2024). Alpha activity is associated with atten-

tional gating, theta with cognitive control, and beta/gamma with top-down maintenance and feature integration. Prior human-centered NLP resources (Zermiani et al., 2024; Hollenstein et al., 2020) show that such signals can enrich language models without replacing human labels.

Cognitive feedback from EEG can encode preferences and interpretative effort (Harada and Oseki, 2025). In sexism detection, physiology-enhanced models for memes improve binary and difficult fine-grained cases (Arcos Gabaldón, Rosso, and Gomis Vicent, 2026). Our work extends this direction to short videos, where the stimulus is temporal, multilingual, and audiovisual, and where multiple subject responses must be fused for the same item.

3 A Multimodal Resource for Sexism Detection in Videos

To build and evaluate the proposed human-centered models, a multimodal setting is used in which physiological responses are recorded while subjects watch existing sexist and non-sexist videos annotated as part of the EXIST 2026 shared task[2].

3.1 EXIST 2026 TikTok Videos

We used the official training split of the EXIST 2026 shared task, a publicly available dataset containing 2524 short-form videos (**1524** in Spanish and **1000** in English). In EXIST, sexist videos are categorized by communicative intent: **Direct sexism** refers to content that itself expresses or promotes sexist ideas, whereas **Judgmental sexism** refers to content that portrays or condemns sexist situations or behaviors. This distinction is relevant because judgmental videos often require contextual interpretation of the speaker's stance. Videos are also annotated for fine-grained categories. Table 1 summarizes the dataset statistics.

The dataset is long-tailed: *Stereotyping and Dominance* is the dominant Task 3 category, whereas *Sexual Violence* and *Misogyny and Non-Sexual Violence* are rare. Fleiss' κ is 0.690 for T1, 0.799 for T2, and 0.430 for multilabel T3 (computed independently per category), confirming that fine-grained sexism is the most ambiguous setting. Tables 2–3 report the language and category breakdowns. Majority voting is used to derive the experimental labels.

Task	Label	Spanish	English
T1: Sexism ID	Sexist	747 (49.0%)	455 (45.5%)
	Non-Sexist	768 (50.4%)	538 (53.8%)
T2: Intention	Direct	495 (66.3%)	241 (53.0%)
	Judgmental	230 (30.8%)	194 (42.6%)
	Ties	22 (2.9%)	20 (4.4%)
T3: Category	Ideolog. Ineq.	141 (18.9%)	53 (11.6%)
	Stereotyping	320 (42.8%)	219 (48.1%)
	Objectification	50 (6.7%)	62 (13.6%)
	Sexual Violence	36 (4.8%)	12 (2.6%)
	Misogyny NSV	24 (3.2%)	7 (1.5%)
	Ties	176 (23.6%)	102 (22.4%)
Total Videos		**1524**	**1000**

Table 1: EXIST 2026 Video Dataset Statistics (TRAIN). Percentages for T2 and T3 are calculated over the sexist subset of each language.

The imbalance is strongest in T3: Stereotyping accounts for 42.8% of Spanish and 48.1% of English sexist videos, whereas Misogyny NSV accounts for only 3.2% and 1.5%. Ties are also frequent in T3, illustrating that fine-grained categories are not mutually obvious even to humans. Language differences must be interpreted carefully because subject nationalities and the number of valid judgments vary across sessions.

Agreement is substantial and almost identical across languages for T1. T2 reaches the highest agreement, indicating that annotators can distinguish direct from judgmental content reliably once sexism has been identified. T3 is markedly more subjective: Objectification has the highest overall category agreement, whereas rare Misogyny NSV has the lowest. These properties motivate both class-weighted learning and the search for complementary perceptual evidence.

3.2 Physiological Data Collection

Physiological data were collected from **10 gender-balanced subjects** of different nationalities across **15 sessions**. A resting baseline

[2] https://nlp.uned.es/exist2026/

Task	All	ES	EN
T1	0.6898	0.6899	0.6896
T2	0.7990	0.8370	0.7325
T3	0.4297	0.4982	0.3049

Table 2: Inter-annotator agreement (Fleiss' κ) for Tasks 1–3.

Category	All	Spanish	English
Ideological Inequality	0.4336	0.5628	0.2351
Stereotyping Dominance	0.4874	0.5641	0.3184
Objectification	0.5049	0.5355	0.4464
Sexual Violence	0.4736	0.5152	0.3970
Misogyny NSV	0.2487	0.3136	0.1276

Table 3: Inter-annotator agreement for Task 3 (Fleiss' κ, multilabel). Agreement computed independently per category as a binary decision among annotators who marked the item as sexist in Task 1.

preceded the trials, and a 3-second neutral interval reduced carry-over effects. Trials were aligned to stimulus onset; because several subjects could watch the same video, the model treats their responses as a variable-length sequence. Data were collected with informed consent and anonymized.

Subjects watched each complete video before responding. The resting baseline and fixed inter-stimulus interval support trial segmentation and band-power correction. The many-to-one relation between subjects and videos is important: rather than averaging away individual variation, the fusion model receives the available per-subject observations as a sequence and learns their relative relevance.

The following devices were used:

- **Eye-Tracking:** Pupil Labs Neon glasses recorded binocular gaze data at 200 Hz.
- **Heart Rate:** A Garmin Venu 3 watch continuously recorded heart-rate signals.
- **Electroencephalography:** A 16-channel OpenBCI Cyton headset recorded neural activity at 250 Hz using the 10–20 system.

3.3 Physiological Feature Extraction

For **ET**, summary statistics over fixations, saccades, blinks, and pupil dynamics captured visual exploration (Dalveren and Cagiltay, 2018; Takemoto, Nakazawa, and Kumada, 2022). **EEG** was converted to μV, band-pass filtered at 0.5–40 Hz, and baseline corrected. Welch PSD features were integrated over Delta, Theta, Alpha, Beta, and Gamma bands for 16 channels, then harmonized using Box–Cox transformation, ComBat (Fortin et al., 2018), Winsorization, and robust z-scoring. **HR** summary statistics and **Reaction Time** provided arousal and interpretative-effort measures.

ET aggregates included mean, dispersion, extrema, and counts for each event type, representing both viewing intensity and scan variability. EEG time-domain statistics were combined with band power from Delta (0.5–4 Hz), Theta (4–8 Hz), Alpha (8–13 Hz), Beta (13–30 Hz), and Gamma (30–40 Hz). Harmonization reduces session- and subject-dependent scale shifts while retaining condition-related variation. RT was measured from video onset until the response to proceed, and therefore combines viewing and decision time as a coarse measure of interpretative effort.

This process yielded a multimodal resource linking video content to synchronized physiological responses. To support reproducibility, the preprocessing code for ET/HR/EEG, the exact prompts used for the vision–language model, and the random seeds used in all experiments are publicly available.[3]

4 Physiological Correlates of Sexism Perception

Our analysis (**RQ1**) revealed significant physiological markers, showing that the human brain and body respond differently to sexist content.

[3] https://github.com/ivanarcos02/exist2026-human-centered-sexism-tiktok

4.1 Cognitive Load and Autonomic Arousal

Table 4 shows a monotonic non-sexist → direct → judgmental gradient in reaction time, fixations, saccades, and blinks. This pattern is consistent with increased cognitive load and visual search (De Rivecourt et al., 2008; Marquart, Cabrall, and de Winter, 2015; Dalveren and Cagiltay, 2018; Takemoto, Nakazawa, and Kumada, 2022): judgmental content additionally requires inference of the author's critical stance. Longer responses are consistent with detection-response workload (van Winsum, 2018), while increased fixation and saccade activity indicates higher workload and more demanding visual search (Dalveren and Cagiltay, 2018; Takemoto, Nakazawa, and Kumada, 2022; Young and Hulleman, 2013; Marquart, Cabrall, and de Winter, 2015; De Rivecourt et al., 2008). HR also differed across conditions, indicating autonomic modulation (Berntson et al., 1997; Kim et al., 2018; Delliaux et al., 2019), while blink duration was marginally shorter for sexist stimuli (Benedetto et al., 2011; Marquart, Cabrall, and de Winter, 2015).

4.2 Category-Specific and Neural Signatures

Significant Task 1 differences were negative in Delta, Theta, and Alpha over central/parietal sites, consistent with event-related desynchronization and increased engagement (Pfurtscheller and Lopes da Silva, 1999; Foxe and Snyder, 2011; Cavanagh and Frank, 2014; Harmony, 2013). Task 2 differences were positive in frontal/right-lateral Beta and Gamma, consistent with stronger top-down contextual processing (Engel and Fries, 2010; Spitzer and Haegens, 2017; Fries, 2015; Bastos et al., 2015). Sexual Violence showed right-lateral negative differences, whereas Joy showed positive frontal differences, in line with affective-processing evidence (Codispoti, De Cesarei, and Ferrari, 2023; Luther et al., 2023; Aftanas et al., 2001; Ahern and Schwartz, 1985). These are scalp-level distributed patterns, not source-localization results.

For **Task 1**, the negative low-frequency differences indicate lower band power for sexist videos. Under the ERD/ERS framework this should not be read as reduced processing: desynchronization can reflect stronger recruitment of task-relevant neural populations. The central/-parietal distribution and Alpha modulation are compatible with increased attentional allocation and suppression of irrelevant information.

For **Task 2**, every significant difference was positive and concentrated in frontal or right-lateral Beta/Gamma sites. Judgmental content requires recognition of the depicted sexism and inference that the speaker condemns it. The observed higher-frequency synchronization is therefore consistent with the additional top-down maintenance and contextual disambiguation required by this second-order judgment.

The **Sexual Violence** contrast showed negative differences at significant sites, especially on the right, consistent with enhanced engagement for aversive stimuli. In contrast, **Joy** produced positive frontal/fronto-lateral differences. Together, these contrasts suggest that the EEG features capture emotional salience more generally rather than merely a binary response to negative material.

Across tasks, low-frequency desynchronization tracks engagement with sexist or aversive content, whereas higher-frequency synchronization distinguishes contextual and evaluative processing. This complementary structure motivates the band-resolved features used by the fusion architecture.

4.3 Agreement Between Physiological Subjects and External Annotators

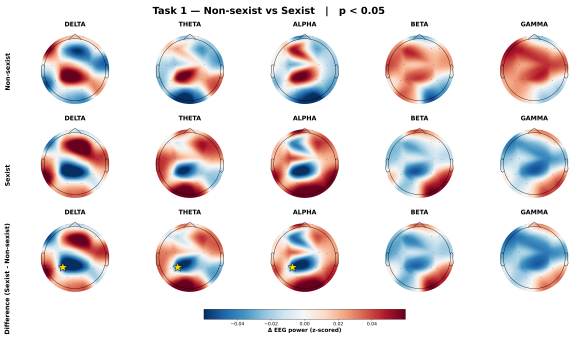
Cohen's κ between physiological-subject majority labels and external EXIST labels was 0.619 for T1 and 0.817 for T2 (Table 5). Thus, the signal-producing subjects were broadly aligned with the official annotations rather than idiosyncratic outliers.

5 Multimodal Fusion Model

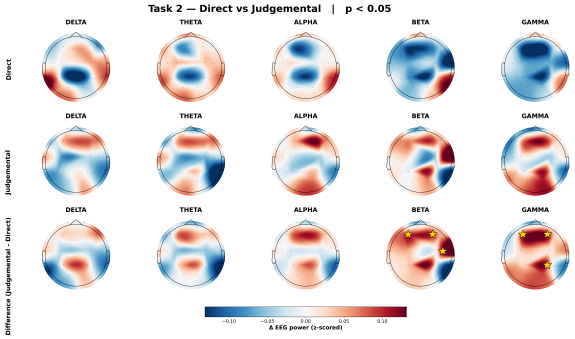
Figure 2 shows the hierarchical fusion model. Text remains the backbone, while physiology

Feature	Non-Sex.	Direct	Judg.	p
React. Time (s)	23.64±20.30	25.81±16.18	34.19±21.74	0.00000***
Fixation Count	58.13±37.09	62.53±34.10	74.86±49.61	0.00000***
Saccade Count	57.26±37.27	61.34±34.59	73.88±49.81	0.00000***
Blink Count	7.31±7.04	7.79±7.31	9.58±9.56	0.00000***
HR Std (bpm)	1.94±1.36	2.12±1.52	2.09±1.36	0.00011***
HR Mean (bpm)	66.92±10.76	66.36±9.99	67.47±11.37	0.04601*
Blink Dur. (ms)	252.59±34.37	249.94±33.83	250.59±33.14	0.05001

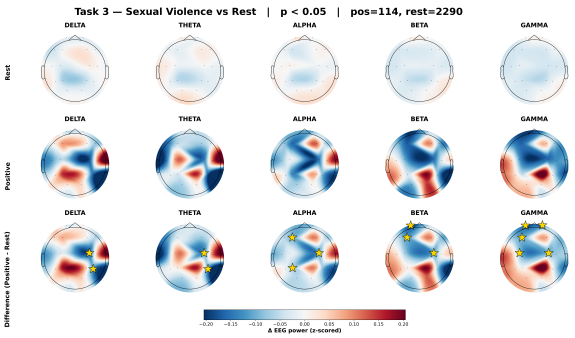
Table 4: Key physiological features across sexism levels (Mean ± SD). p-values from one-way ANOVA.



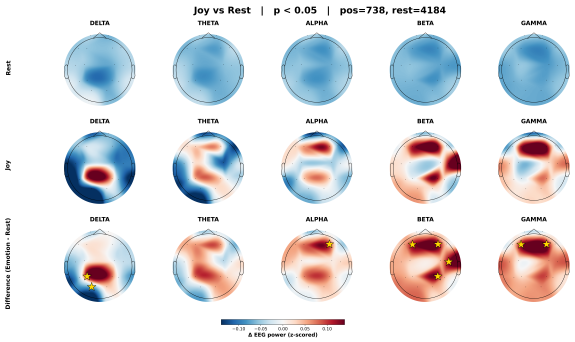
(a) T1: Sexist vs. Non-Sexist.



(b) T2: Direct vs. Judgmental.



(c) T3: Sexual Violence vs. Rest.



(d) Emotion: Joy vs. Rest.

Figure 1: EEG topographic maps of band-power differences across four representative contrasts. In each subfigure, rows correspond to the two conditions and their difference, and columns correspond to Delta, Theta, Alpha, Beta, and Gamma bands. Red indicates positive differences and blue indicates negative differences; yellow stars mark statistically significant electrodes ($p < 0.05$).

reweights and enriches its representation. The model accepts a variable-length, order-invariant sequence of subject responses.

This asymmetric design prevents noisy sensor observations from replacing the semantic representation. Instead, physiology selects the textual or visually grounded evidence most compatible with each subject's response. Because the number of available observations differs by video, all aggregation operations are masked and invariant to subject order.

Enriched Content Representation. Four sampled frames are processed jointly by `Qwen3-VL-4B-Instruct`, prompted to return a 1–2 sentence visual summary and a brief sexism analysis using only visible evidence. This interpretable description is concatenated with transcript/OCR text.

Generating a constrained description rather than directly concatenating raw frame vectors

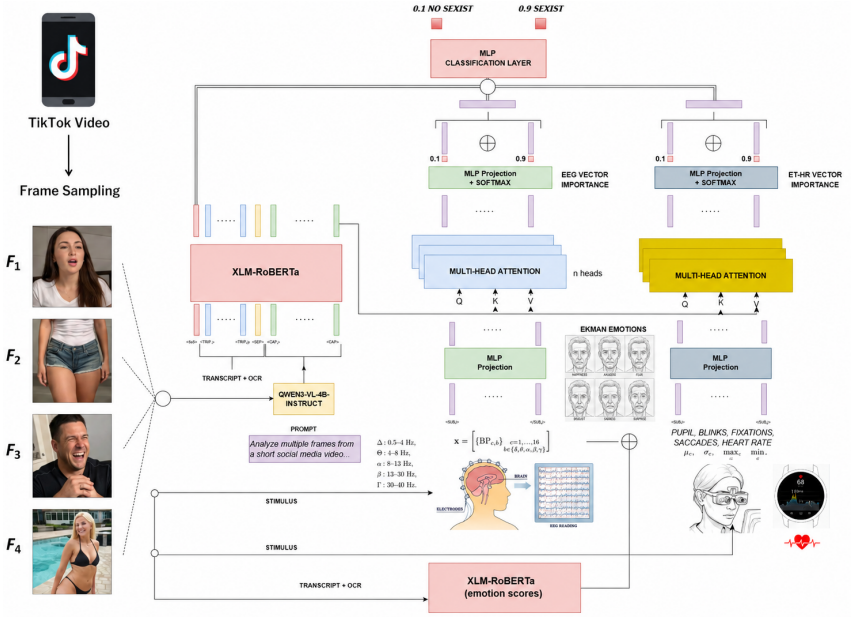


Figure 2: Architecture of the final hierarchical attention-based fusion model. It integrates enriched video text (transcript/OCR + VLM-generated frame description) with sequences of physiological reactions from several subjects (EEG and ET/HR). Cross-attention mechanisms allow the model to learn correlations between specific textual tokens and physiological responses.

Task	Lang.	N	Accuracy	κ
T1	ALL	1927	0.815	0.619
	EN	711	0.802	0.543
	ES	1216	0.822	0.644
T2	ALL	988	0.917	0.817
	EN	364	0.852	0.703
	ES	624	0.955	0.886

Table 5: Agreement between GOLD_EXIST (external annotators) and GOLD_SUBJECT (physiological subjects) for the EXIST 2026 TikTok video training set. Cohen's κ is computed on videos with valid, non-tied labels from both annotation groups.

provides a common linguistic space for both languages and makes the visual contribution auditable. The four frames capture the temporal progression at modest computational cost, while the prompt discourages unsupported inference and enforces a predictable output format.

Text Encoder. `XLM-RoBERTa` encodes the original and generated bilingual text jointly.

Physiological Fusion Core. EEG and ET/HR vectors form subject sequences. In independent multi-head cross-attention layers, physiology is the **Query** and XLM-R tokens are the **Key/-Value**, allowing each response to identify the content that best explains it.

Using physiology as Query lets the perceptual trace "ask" which content tokens account for it. EEG and ET/HR use separate projections because their dimensionalities and statistical properties differ. Their attended representations are aggregated only after modality-specific cross-attention.

Prompt used for `Qwen3-VL-4B-Instruct`

```
Analyze the multiple frames using only visible evidence. (1) Summarize the video in 1-2 sentences, including on-screen text. (2) State briefly whether sexist elements are present and why. Return exactly: Description: <1-2 sentences> Sexism Analysis: <sexism or not + brief reason>
```

Classification Head and Training Strategy. Text-aware physiological representations are

combined by a learned weighted sum, concatenated with XLM-R's `[CLS]`, and classified by an MLP. Training first freezes XLM-R while learning fusion layers, then fine-tunes the full network with discriminative learning rates and a class-weighted loss.

The frozen phase stabilizes the randomly initialized attention and classification components. End-to-end fine-tuning then adapts the multilingual encoder without overwhelming it with initially noisy gradients. Class weights are particularly important for the long-tailed T3 labels shown in Table 1.

6 Results and Analysis

We use **10-fold cross-validation**. Table 6 is an incremental ablation over frames, EEG, and ET/HR. The full model reaches AUCs of **0.790**, **0.788**, and **0.703** for T1–T3, improving over Text+Frames by **3.9%**, **6.5%**, and **6.8%**, respectively. The larger T2/T3 gains indicate that physiology helps both intent inference and low-agreement fine-grained categorization.

Task 1 has the strongest content-only baseline and consequently the smallest physiological gain. Task 3 has the lowest inter-annotator agreement ($\kappa = 0.430$); here physiology provides evidence where consensus labels are uncertain. Task 2 differs: agreement is high ($\kappa = 0.799$), yet physiology yields a 6.5% gain because direct versus judgmental sexism requires communicative-intent inference. The frontal Beta/Gamma signatures in Section 4 are consistent with this increased contextual demand. Thus, physiology helps both where labels are ambiguous and where content alone underspecifies the intended meaning.

Welch tests comparing Full against Text+Frames are significant for every task ($p < 0.001$). Incremental additions of EEG and ET/HR are also significant; only Text→Text+Frames for T2 is not ($p = 0.403$), supporting the interpretation that intention relies more on contextual processing than on explicit visual cues. Category-level F1 gains are largest for Ideological Inequality, Objectification, and Misogyny NSV, while rare Sexual Violence results are less stable.

The incremental tests clarify where each modality contributes. Frames produce large T1 and T3 gains but virtually no T2 gain, whereas EEG significantly improves all three tasks. Adding ET/HR on top of EEG remains significant, showing that ocular/autonomic features are not redundant with neural activity. The Full versus Text+Frames comparison is the central test because it isolates the complete physiological contribution from the enriched content baseline.

At category level, Full raises Ideological Inequality F1 from 0.645 to 0.705, Objectification from 0.441 to 0.506, and Misogyny NSV from 0.294 to 0.349. Stereotyping changes less because it is frequent and already well represented by content. Sexual Violence is the exception: Text+Frames remains stronger, and the small class size makes fold-level estimates unstable. These results caution against interpreting the aggregate T3 improvement as uniform across categories.

7 Discussion

Ocular, autonomic, and neural responses differ across sexism labels (**RQ1**), and their integration significantly improves all tasks (**RQ2–RQ3**). The behavioral measures reveal a monotonic cognitive-load gradient, while EEG separates engagement-related low-frequency desynchronization from the higher-frequency contextual processing of judgmental sexism.

The pattern of model gains is informative. Physiology helps T3 where labels are uncertain and T2 where intent inference exceeds surface content, while T1 gains are smaller because its content baseline is stronger. Agreement with external annotators also indicates that the physiological subjects broadly share the dataset's interpretation of sexism. Physiological evidence should therefore be treated as complementary, content-aligned supervision during development, not as a replacement for annotation.

The alignment between physiological con-

Model	T1	T2	T3
Text	0.705±0.013	0.735±0.012	0.596±0.016
Text + Frames	0.760±0.012	0.740±0.014	0.658±0.015
Text + Frames + EEG	0.776±0.011	0.772±0.011	0.679±0.014
Full (+ ET + HR)	**0.790±0.011**	**0.788±0.011**	**0.703±0.013**

Table 6: Multimodal model performance (AUC) across all tasks. Rows form an incremental ablation: each adds one group of signals.

Comparison	T1	T2	T3
Text → Text+Frames	$\Delta = +0.055$, t=9.83 p<0.001***	$\Delta = +0.005$, t=0.86 p=0.403$^{\text{n.s.}}$	$\Delta = +0.062$, t=8.94 p<0.001***
Text+Frames → +EEG	$\Delta = +0.016$, t=3.11 p=0.006**	$\Delta = +0.032$, t=5.68 p<0.001***	$\Delta = +0.021$, t=3.24 p=0.005**
+EEG → Full (+ET+HR)	$\Delta = +0.014$, t=2.85 p=0.011*	$\Delta = +0.016$, t=3.25 p=0.004**	$\Delta = +0.024$, t=3.97 p<0.001***
Text+Frames → Full	$\Delta = +0.030$, t=5.83 p<0.001***	$\Delta = +0.048$, t=8.53 p<0.001***	$\Delta = +0.045$, t=7.17 p<0.001***
Text → Full	$\Delta = +0.085$, t=15.78 p<0.001***	$\Delta = +0.053$, t=10.30 p<0.001***	$\Delta = +0.107$, t=16.41 p<0.001***

Table 7: Statistical significance of AUC differences between model pairs (Welch's two-sample t-test, $n = 10$ folds). Δ is the difference in mean AUC. $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$, n.s. $p\geq0.05$.

Category	Text	+Frames	+EEG	Full
Ideolog. Ineq.	0.603±0.017	0.645±0.017	0.664±0.015	**0.705±0.014**
Stereotyping	0.830±0.012	0.830±0.012	**0.858±0.011**	0.855±0.011
Objectif.	0.412±0.020	0.441±0.019	0.484±0.018	**0.506±0.017**
Sexual Viol.	0.299±0.022	**0.345±0.020**	0.296±0.021	0.321±0.020
Misogyny NSV	0.273±0.021	0.294±0.020	0.342±0.019	**0.349±0.019**

Table 8: F1-score for Task 3 categories (10-fold cross-validation).

trasts and downstream gains strengthens this interpretation. Judgmental clips elicit longer viewing and stronger frontal Beta/Gamma activity, and T2 is precisely the task for which frames alone do not help but physiology does. Likewise, the ambiguity of T3 corresponds to larger improvements for categories whose content cues are contextual rather than explicit. Although these correspondences are not causal proof, they provide convergent behavioral, neural, and predictive evidence.

The most plausible deployment trains with physiological supervision but performs inference on content alone. This allows the model to internalize perceptual cues without requiring sensors or monitoring end users, and preserves a clear separation between research signals and operational moderation.

8 Limitations

The study involves only **10 subjects across 15 sessions**; many viewing instances do not eliminate subject-specific effects. Laboratory viewing may also differ from natural mobile consumption. Majority voting discards informative disagreement and produces a consensus label that is not identical to each subject's judgment. Although subject/external agreement is high, physiology is evidence about perception rather than direct ground truth. EEG results are scalp-level patterns, not source localization. Finally, the current model omits audio. Larger, naturalistic, multilingual studies are therefore needed.

The fixed presentation protocol may also introduce order or fatigue effects despite resting baselines and neutral intervals. Harmonization mitigates scale shifts but cannot remove every session-specific confound. Moreover, a system trained with sensitive physiological data requires

careful governance even if deployment uses content alone.

The majority-vote gold labels were produced by external annotators who are different from the subjects in the physiological recordings. Consequently, the signals should not be interpreted as direct ground truth for each subject's personal judgment, but as perceptual evidence generalized to an external consensus. This mismatch between subject-specific perception and dataset-level labels remains an important limitation despite the high agreement observed between both groups.

9 Conclusion and Future Work

This work validates a human-centered paradigm for sexism detection in short social media videos. The results indicate that physiological responses to TikTok videos provide a rich and objective signal of perception that complements content analysis. By integrating EEG, ET, and HR with enriched textual-visual representations generated from video frames, our multimodal fusion model achieves consistent gains over strong baselines across binary detection, intention classification, and fine-grained categorization.

Physiological supervision complements textual-visual content for TikTok sexism detection. EEG, ET, and HR improve Text+Frames by **3.9%**, **6.5%**, and **6.8%** for T1–T3. The physiological contrasts also explain why judgmental and emotionally salient content demands additional interpretation, providing an interpretable link between human perception and model gains.

These gains show that physiological signals are not merely auxiliary information, but a meaningful source of complementary evidence in cases where content alone remains ambiguous. The physiological analysis also offers a more interpretable perspective on why the task is hard. Judgmental sexism requires more contextual disambiguation than direct sexism; sexual violence triggers strong right-lateralized desynchronization patterns; and emotionally salient content such as joy yields robust positive oscillatory differences. These findings reinforce the value of incorporating human responses into computational models of harmful content.

Future work will model disagreement rather than collapsing it through majority voting, align reactions to specific video segments, incorporate audio, and evaluate larger multilingual collections in more realistic moderation settings.

Modeling subjective disagreement is especially important in subtle and context-dependent cases of sexism. Finer temporal alignment would allow the model to identify not only whether a video is difficult to interpret, but also which moments trigger greater cognitive or emotional responses. Audio is another central source of meaning in TikTok videos and may provide cues not captured by text or sampled frames.

10 Ethical Considerations

Subjects gave informed consent for anonymous research use. The UPV Research Ethics Committee approved the study (P13_26-12-2025); stimuli came from EXIST, and demographic/physiological data were anonymized. Physiological measurements are sensitive data requiring consent and strict access control. They are intended for model development, not end-user profiling; operational systems should rely on content-level inference.

The physiological data were collected from 10 subjects of different nationalities. The resulting models and resources are intended only for systems that prioritize transparency, user agency, privacy protection, and robust appeal processes. Physiological signals should be understood as a research and model-development resource, not as a mechanism for profiling end users in deployment settings. Their collection, storage, and reuse must remain governed by explicit consent and strict access control, and any operational moderation system derived from this work should avoid continuous monitoring of individuals.

Acknowledgments

This work was done in the framework of the research project *Multimodal and Multisensor data-centric AI models (ANNOTATE-MULTI2)* funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU (Grant PID2024-156022OB-C32).

References

Abercrombie, G., N. Vitsakis, A. Jiang, and I. Konstas. 2024. Revisiting annotation of online gender-based violence. In G. Abercrombie, V. Basile, D. Bernadi, S. Dudy, S. Frenda, L. Havens, and S. Tonelli, editors, *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, pages 31–41, Torino, Italia, May. ELRA and ICCL.

Aftanas, L. I., A. A. Varlamov, S. V. Pavlov, V. P. Makhnev, and N. V. Reva. 2001. Event-related synchronization and desynchronization during affective processing: emergence of valence-related time-dependent hemispheric asymmetries in theta and upper alpha band. *International Journal of Neuroscience*, 110(3-4):197–219.

Ahern, G. L. and G. E. Schwartz. 1985. Differential lateralization for positive and negative emotion in the human brain: Eeg spectral analysis. *Neuropsychologia*, 23(6):745–755.

Arcos, I. and P. Rosso. 2024. Sexism identification on TikTok: A multimodal AI approach with text, audio, and video. In L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 61–73, Cham. Springer Nature Switzerland.

Arcos Gabaldón, I., P. Rosso, and E. Gomis Vicent. 2026. Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and eeg signals. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 9830–9840. European Language Resources Association (ELRA).

Azarbarzin, A., M. Ostrowski, P. Hanly, and M. Younes. 2014. Relationship between arousal intensity and heart rate response to arousal. *Sleep*, 37:645–653, 06.

Bai, S., Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu. 2025. Qwen3-vl technical report.

Bastos, A. M., J. Vezoli, C. A. Bosman, J.-M. Schoffelen, R. Oostenveld, J. R. Dowdall, P. De Weerd, H. Kennedy, and P. Fries. 2015. Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, 85(2):390–401.

Benedetto, S., M. Pedrotti, L. Minin, T. Baccino, A. Re, and R. Montanari. 2011. Driver workload and eye blink duration. *Transportation Research Part F: Traffic Psychology and Behaviour*, 14(3):199–208.

Berntson, G. G., J. T. Bigger, D. L. Eckberg, P. Grossman, P. G. Kaufmann, M. Malik, H. N. Nagaraja, S. W. Porges, J. P. Saul, P. H. Stone, and M. W. van der Molen. 1997. Heart rate variability: Origins, methods, and interpretive caveats. *Psychophysiology*, 34(6):623–648.

Cavanagh, J. F. and M. J. Frank. 2014. Frontal theta as a mechanism for cognitive control. *Trends in Cognitive Sciences*, 18(8):414–421.

Codispoti, M., A. De Cesarei, and V. Ferrari. 2023. Alpha-band oscillations and emotion: A review of studies on picture perception. *Psychophysiology*, 60(12):e14438.

Dalveren, G. G. M. and N. E. Cagiltay. 2018. Using eye-movement events to determine the mental workload of surgical residents. *Journal of Eye Movement Research*, 11(4).

De Grazia, L., D. Sánchez Villegas, D. Elliott, M. Farrús, and M. Taulé. 2026. Beyond binary classification: Detecting fine-grained sexism in social media videos.

De Rivecourt, M., M. N. Kuperus, W. J. Post, and L. J. M. Mulder. 2008. Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics*, 51(9):1295–1319.

Delliaux, S., A. Delaforge, J.-C. Deharo, and G. Chaumet. 2019. Mental workload alters heart rate variability, lowering non-linear dynamics. *Frontiers in Physiology*, 10:565.

Engel, A. K. and P. Fries. 2010. Beta-band oscillations–signalling the status quo? *Current Opinion in Neurobiology*, 20(2):156–165.

Fortin, J.-P., N. Cullen, Y. I. Sheline, W. D. Taylor, I. Aselcioglu, P. A. Cook, P. Adams, C. Cooper, M. Fava, P. J. McGrath, M. McInnis, M. L. Phillips, M. H. Trivedi, M. M. Weissman, and R. T. Shinohara. 2018. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120.

Foxe, J. J. and A. C. Snyder. 2011. The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Frontiers in Psychology*, 2:154.

Fries, P. 2015. Rhythms for cognition: Communication through coherence. *Neuron*, 88(1):220–235.

Gasparini, F., E. Fersini, Z. Perjési, M. Anzovino, E. Messina, and M. Pievani. 2018. Multimodal classification of sexist advertisements. In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE)*, volume 1, pages 399–406. SciTePress.

Grosz, D. and P. Conde-Cespedes. 2020. Automatic detection of sexist statements commonly used at the workplace. In *Trends and Applications in Knowledge Discovery and Data Mining: PAKDD 2020 Workshops, DSFN, GII, BDM, LDRC and LBD, Singapore, May 11–14, 2020, Revised Selected Papers*, page 104–115, Berlin, Heidelberg. Springer-Verlag.

Harada, Y. and Y. Oseki. 2025. Cognitive feedback: Decoding human feedback from cognitive signals. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 209–219, Suzhou, China. Association for Computational Linguistics.

Harmony, T. 2013. The functional significance of delta oscillations in cognitive processing. *Frontiers in Integrative Neuroscience*, 7:83.

Hollenstein, N., M. Troendle, C. Zhang, and N. Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146, Marseille, France, May. European Language Resources Association.

Italiani, P., D. Gimeno-Gómez, L. Ragazzi, G. Moro, and P. Rosso. 2026. Memeweaver: Inter-meme graph reasoning for sexism and misogyny detection. In *Findings of the Association for Computational Linguistics: EACL*. Also available as arXiv:2601.08684.

Jha, A. and R. Mamidi. 2017. When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data. In D. Hovy, S. Volkova, D. Bamman, D. Jurgens, B. O'Connor, O. Tsur, and A. S. Doğruöz, editors, *Proceedings of the Second Workshop on NLP and Computational Social Science*, pages 7–16, Vancouver, Canada, August. Association for Computational Linguistics.

Kar, A., I. Pathak, S. Mukherjee, and S. Chatterjee. 2025. Decoding complexity and emotion: A computational linguistic approach to sentence comprehensibility and writer affect. *International Journal of Innovative Science and Research Technology*, pages 423–428, 07.

Khurana, V., Y. Kumar, N. Hollenstein, R. Kumar, and B. Krishnamurthy. 2023. Synthesizing human gaze feedback for improved NLP performance. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1895–1908, Dubrovnik, Croatia, May. Association for Computational Linguistics.

Kim, H. W., E. J. Cheon, D. S. Bai, Y. H. Lee, and B. M. Koo. 2018. Stress and heart rate variability: A meta-analysis and review of the literature. *Psychiatry Investigation*, 15(3):235–245.

Kirk, H., W. Yin, B. Vidgen, and P. Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In A. K. Ojha, A. S. Doğruöz, G. Da San Martino, H. Tayyar Madabushi, R. Kumar, and E. Sartori, editors, *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada, July. Association for Computational Linguistics.

Li, L., L. Fan, S. Atreja, and L. Hemphill. 2024. "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Trans. Web*, 18(2).

Lin, P. and L. Yang. 2024. Building a virtual psychological counselor by integrating EEG emotion detection with large-scale NLP models. In A. Wang, editor, *Third International Conference on Biological Engineering and Medical Science (ICBioMed2023)*, volume 12924, page 129241X. International Society for Optics and Photonics, SPIE.

Luther, L., J. M. Horschig, J. M. van Peer, K. Roelofs, O. Jensen, and M. A. Hagenaars. 2023. Oscillatory brain responses to emotional stimuli are effects related to events rather than states. *Frontiers in Human Neuroscience*, 16:868549.

Marquart, G., C. Cabrall, and J. de Winter. 2015. Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, 3:2854–2861.

Nockleby, J. T. 2000. *Hate Speech*. Macmillan, New York, 2 edition.

Parikh, P., H. Abburi, N. Chhaya, M. Gupta, and V. Varma. 2021. Categorizing sexism and misogyny through neural approaches. *ACM Trans. Web*, 15(4).

Pelle, R. P. and V. P. Moreira. 2017. Offensive comments in the brazilian web: a dataset and baseline results. In *Proceedings of the 6th Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, pages 510–519, São Paulo, Brazil. Sociedade Brasileira de Computação.

Pfurtscheller, G. and F. H. Lopes da Silva. 1999. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, 110(11):1842–1857.

Plaza, L., J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, and R. Morante. 2025. EXIST 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and TikTok videos. In C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, and N. Tonellotto, editors, *Advances in Information Retrieval*, pages 442–449, Cham. Springer Nature Switzerland.

Rizzi, G., F. Gasparini, A. Saibene, P. Rosso, and E. Fersini. 2023. Recognizing misogynous memes: Biased models and tricky archetypes. *Information Processing & Management*, 60(5):103474.

Spitzer, B. and S. Haegens. 2017. Beyond the status quo: A role for beta oscillations in endogenous content (re)activation. *eNeuro*, 4(4):ENEURO.0170–17.2017.

Sun, Q., T. Chen, Y. Zhang, Y. Zhang, J. Yue, J. Jiao, and Z. Fu. 2026. Multihateloc: Towards temporal localisation of multimodal hate content in online videos. In *Proceedings of the ACM Web Conference (WWW)*. Also available as arXiv:2512.10408.

Takemoto, A., A. Nakazawa, and T. Kumada. 2022. Non-goal-driven eye movement after visual search task. *Journal of Eye Movement Research*, 15(2).

van Winsum, W. 2018. The effects of cognitive and visual workload on peripheral detection in the detection response task. *Human Factors*, 60(6):855–869.

Young, A. H. and J. Hulleman. 2013. Eye movements reveal how task difficulty moulds visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 39(1):168–190.

Zermiani, F., P. Dhar, E. Sood, F. Kögel, A. Bulling, and M. Wirzberger. 2024. InteRead: An eye tracking dataset of interrupted reading. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9154–9169, Torino, Italia, May. ELRA and ICCL.