AIdonnow at EXIST 2026: Multimodal Text Reconstruction and Sensor Fusion for TikTok Sexism Detection

Notebook for the EXIST Lab at CLEF 2026

Olaf Meneses Albalá[1,*], Javier Martínez Llinares[1]

[1]Universitat Politècnica de València, Valencia, Spain

Abstract

This paper describes our systems submitted to EXIST 2026 Task 3, which addresses sexism identification, source-intention detection, and sexism categorization in bilingual TikTok videos under a Learning with Disagreement evaluation protocol. Our approach converts each video into a rich text representation by combining cleaned automatic speech recognition, visual OCR, and frame-level visual descriptions generated by a Qwen3.5 vision-language model. We compare classical TF–IDF classifiers, XLM-RoBERTa large, and Qwen3-8B adapted with QLoRA; we also evaluate physiological sensor features from eye tracking, heart rate, and EEG through both classical feature concatenation and gated neural fusion. On the official test set, our submissions ranked first in Task 3.2 hard evaluation, first in Task 3.2 soft evaluation, and first in Task 3.3 hard evaluation. The results show that multimodal text reconstruction is consistently useful, hard ensembles transfer better than soft averaging, and physiological features help only in narrow task/model settings.

Keywords

sexism detection, multimodal classification, Learning with Disagreement, TikTok videos, sensor fusion, EXIST

1. Introduction

EXIST 2026 is the sixth edition of the sEXism Identification in Social neTworks shared task. The lab focuses on automatic sexism detection in social-media content and, in this edition, emphasizes Human-Centered AI by enriching multimedia instances with physiological and behavioral signals collected while subjects viewed the stimuli [1, 2]. The task is challenging for three reasons. First, sexist content in TikTok videos is often distributed across speech, captions, on-screen text, gestures, visual context, music, and platform conventions. Second, the data are bilingual, with Spanish and English videos and frequent social-media noise. Third, the official evaluation follows the Learning with Disagreement (LeWiDi) paradigm: systems are not only expected to predict a majority label but also to approximate the empirical distribution of annotator judgments.

We participated in the three video subtasks of EXIST 2026: Task 3.1, sexism identification; Task 3.2, source intention; and Task 3.3, sexism categorization. The subtasks are hierarchical. Task 3.1 asks whether a video is sexist. Task 3.2 distinguishes direct sexist content from judgemental or critical content, while retaining a non-sexist class. Task 3.3 is multilabel and assigns sexist videos to one or more categories such as stereotyping, objectification, sexual violence, and misogyny.

Our main design choice is to transform videos into structured text before classification. We use a three-stage audio-cleaning cascade followed by Whisper-large-v3 for speech transcription, and a frame-processing branch that extracts deduplicated frames and uses a Qwen3.5 vision-language model [3] to obtain OCR and visual descriptions. These fields are fused into two text variants: T+O (transcription plus OCR) and T+O+V (transcription plus OCR plus visual description). We then compare classical models, multilingual encoder fine-tuning, QLoRA adaptation of Qwen3-8B, and physiological sensor fusion.

The paper contributes a reproducible multimodal preprocessing pipeline, a comparison of classical, encoder-based, decoder-based, sensor-fusion, and ensemble systems under LeWiDi metrics, an official

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany

*Corresponding author.

omenalb@upv.es (O. M. Albalá); jmarlli@posgrado.upv.es (J. M. Llinares)

test-set analysis with three first-place results, and an empirical discussion of when physiological signals help or damage generalization.

2. Related Work

Sexism detection has evolved from binary text classification toward richer taxonomies that distinguish author intention and the type of sexist content. The EDOS shared task, for example, proposed a hierarchical explainable taxonomy for online sexism and emphasized that binary detection alone hides important differences in sexist expression [4]. EXIST follows a related direction, but places the task in a multilingual and increasingly multimodal setting. The 2025 overview describes the expansion of EXIST to tweets, memes, and TikTok videos, with three objectives across modalities: sexism identification, source intention, and sexism categorization [5].

The LeWiDi paradigm is central to recent EXIST editions. Rather than treating annotator disagreement as noise, LeWiDi preserves the distribution of human labels and evaluates whether systems can model uncertainty in subjective tasks. This is consistent with broader work on learning from disagreement, which argues that many NLP and vision tasks do not have a single unambiguous interpretation [6]. EXIST evaluates both hard predictions and soft distributions, using the Information Contrast Measure and its soft extension. ICM is grounded in information-theoretic similarity for hierarchical and multilabel classification [7].

Recent text-focused sexism systems commonly rely on multilingual transformer encoders, often XLM-RoBERTa, with task-specific heads, soft-label training, and post-processing for class imbalance. DUTH at EXIST 2025, for instance, used multilingual transformer models with explicit handling of soft labels and class imbalance [8]. This family is strong for bilingual social-media text, but it depends on the availability and quality of textual evidence extracted from the original post.

Multimodal sexism detection has increasingly used vision-language models and video-specific pipelines. CogniCIC and I2C-UHU-Altair explored combinations of text, visual media, VLMs, and LLMs for EXIST 2025 [9, 10]. ECA-SIMM-UVa treated TikTok videos through a segmentation-oriented "one is enough" strategy, reflecting the fact that a short visual or audio segment can determine the label [11]. UMUTeam at EXIST 2025 explored multimodal transformer architectures and soft-label learning across media types, including video models for TikTok subtasks [12]. Our approach is closer to a text-reconstruction pipeline: instead of directly training a video transformer, we convert audio and frames into semantically rich text and then use strong multilingual text models. This choice reduces training cost and makes the method robust when only a few thousand labeled videos are available.

3. Task and Data

The EXIST 2026 video dataset contains 2,524 training videos and 674 official test videos. Both Spanish and English are represented in the training and test splits, with 1,524 Spanish and 1,000 English training videos, and 304 Spanish and 370 English test videos. Table 1 summarizes the split.

Table 1
Dataset split for the EXIST 2026 video subtasks.

Split	All	ES	EN
Train	2,524	1,524	1,000
Test	674	304	370

Task 3.1 is a binary classification problem with labels YES and NO. Task 3.2 is a hierarchical single-label problem with labels NO, DIRECT, and JUDGEMENTAL. Task 3.3 is a hierarchical multilabel problem: each video can be assigned no sexism or one or more sexist categories. The hard training label distribution is shown in Table 2. Task 3.1 is close to balanced, while Task 3.3 is highly imbalanced.

The dominant sexist category is STEREOTYPING-DOMINANCE, whereas SEXUAL-VIOLENCE and MISOGYNY-NON-SEXUAL-VIOLENCE are rare.

Table 2
Training-set hard-label distribution after applying the same majority-vote rules used in our pipeline. Task 3.3 is multilabel, so category counts can overlap. Abbreviations: IDEO. = IDEOLOGICAL-INEQUALITY; STEREO. = STEREOTYPING-DOMINANCE; OBJ. = OBJECTIFICATION; SEX.-V. = SEXUAL-VIOLENCE; MISOG. = MISOGYNY-NON-SEXUAL-VIOLENCE.

Task 3.1: Identification

Label	All	ES	EN
NO	1,313	774	539
YES	1,211	750	461

Task 3.2: Intention

Label	All	ES	EN
NO	1,328	775	553
DIRECT	760	512	248
JUDG.	436	237	199

Task 3.3: Categorization

Label	All	ES	EN
NO	1,549	940	609
IDEO.	230	146	84
STEREO.	582	330	252
OBJ.	131	58	73
SEX.-V.	57	39	18
MISOG.	38	29	9

3.1. Soft Labels

For Task 3.1 and Task 3.2, soft labels are probability distributions over mutually exclusive classes. For Task 3.3, labels are not mutually exclusive; we therefore treat each category independently and compute the fraction of annotators who selected it. As a result, the probabilities for Task 3.3 do not need to sum to one.

Hard labels are used for conventional hard evaluation and for training some models. For Task 3.1 and Task 3.2, we use majority vote after mapping the placeholder "-" to NO and ignoring UNKNOWN. For Task 3.3, a category is included in the hard label set when it appears in more than half of the valid annotator label sets; otherwise the instance is assigned NO.

3.2. Physiological Features

EXIST 2026 provides physiological features collected while subjects viewed the stimuli. The available modalities are eye tracking (ET), heart rate (HR), and EEG. The released variables are aggregate descriptors of the stimulus-viewing interval rather than raw time series. We aggregate features across subjects by mean and standard deviation. This produces up to 216 numerical features per video: 24 ET variables, 4 HR variables, and 80 EEG variables, each summarized across viewers.

The motivation is that physiological responses may capture cognitive effort, arousal, or attention patterns that are not explicit in the video text. The risk is that these signals may reflect viewer reaction to emotionally salient content rather than the target label. This distinction is particularly important for Task 3.2, where the model must infer the creator's intention rather than whether viewers reacted strongly.

Figure 1 provides a descriptive view of the most class-separating physiological features for Task 3.1. Values are standardized within each feature, so the plot compares relative class profiles rather than raw units. In this profile, the largest YES/NO separations appear in eye-tracking features, heart-rate separations are smaller, and the selected EEG bandpower features show weaker class differences. This pattern motivated the sensor-fusion experiments, but it should be interpreted cautiously because the features are aggregated across viewers and do not include temporal dynamics.

3.3. Evaluation

EXIST reports hard-hard and soft-soft evaluations. Hard-hard compares hard system predictions with hard ground truth and reports ICM-Hard, normalized ICM-Hard, and F1. Soft-soft compares predicted probability distributions with soft ground truth and reports ICM-Soft, normalized ICM-Soft, and cross

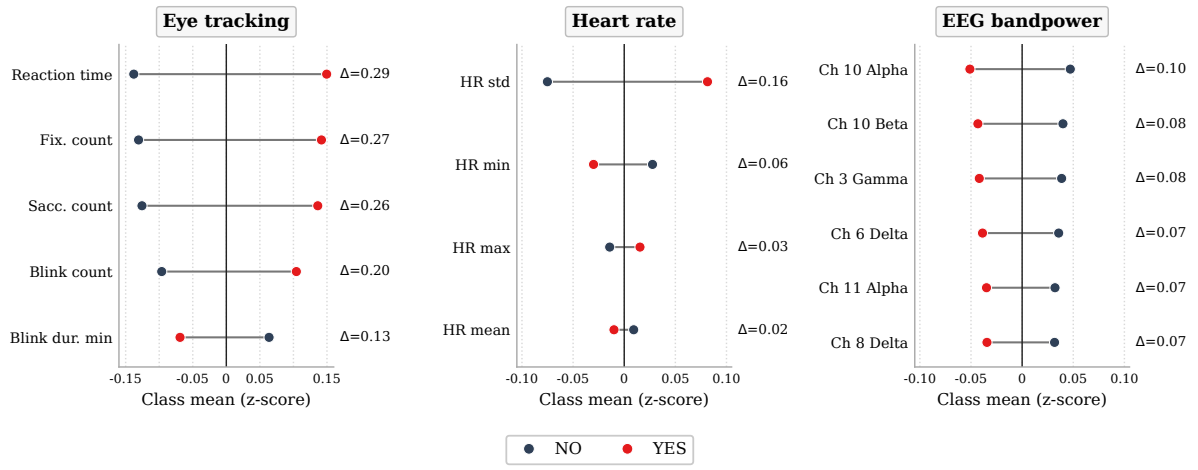

Figure 1: Standardized Task 3.1 class profiles for selected physiological features. Points show mean z-scores for NO and YES videos; connectors and Δ values show the absolute YES–NO separation for each feature. Each panel displays the features with the largest separation within that modality.

entropy when applicable. In this paper we focus on ICM-Norm for hard submissions and ICM-Soft-Norm for soft submissions, because these are the official normalized ICM metrics used for ranking.

4. System Overview

Figure 2 summarizes the system. The preprocessing pipeline runs once and stores text artifacts for all subsequent experiments. The model-selection pipeline then compares classical and neural classifiers, performs sensor ablation, retrains selected models on the full training set, and builds hard and soft submission runs.

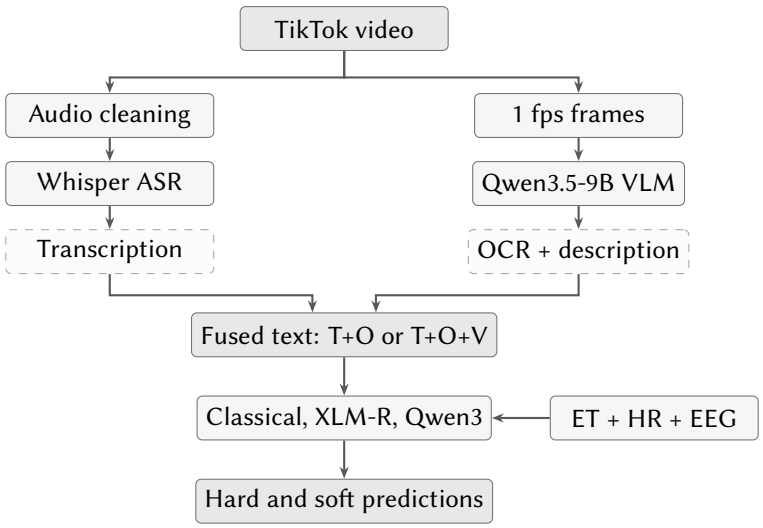

Figure 2: Overview of our multimodal pipeline. Audio and video are converted into text artifacts, then fused with optional physiological features before classification.

The experimental design follows four phases aligned with the blocks in Figure 2:

1. Use the *Fused text* block to sweep T+O and T+O+V with the *Classical, XLM-R, Qwen3* block, keeping *ET + HR + EEG* disabled.
2. Add the *ET + HR + EEG* block through sensor ablations (none, EEG-only, and all sensors) on the best text/model configuration for each pipeline and task.

3. Retrain the top two classical and top two deep-learning configurations from the *Classical, XLM-R, Qwen3* block on the full training set.
4. Build final hard and soft predictions by grid-searching ensemble weights over the retrained systems.

The GPU-heavy stages used one NVIDIA L40S GPU with 48 GB of memory. Intermediate ASR, OCR, visual-description, validation-prediction, and submission artifacts were cached for reproducibility and ensemble selection.

5. Video-to-Text Processing

5.1. Audio Extraction and ASR

Videos often contain music, echo, background speech, and platform audio artifacts. We therefore use a three-stage audio cascade before transcription. First, we isolate vocals from background audio with a BS-RoFormer model [13]. Second, we denoise the vocal channel with a Mel-Band RoFormer denoising model [14]. Third, we apply a Mel-Band RoFormer de-reverberation model to reduce echo. Vocal stems with RMS energy below 0.01, approximately −40 dBFS, are discarded to avoid hallucinated ASR output.

The cleaned audio is transcribed with Whisper-large-v3 [15]. We run Whisper in language-aware mode using the dataset language field. The transcription post-processing stage collapses excessive word repetitions, normalizes elongated artifacts, and replaces common hallucination phrases with a no-speech marker. These steps are important because TikTok audio frequently includes music-only segments or repeated short phrases.

5.2. Frame Description and OCR

The visual branch extracts frames at one frame per second; in the released videos, the 99th percentile is 62 candidate frames before filtering and 60 retained frames after filtering. Near-duplicate frames are removed with brightness, color-histogram, and dHash filters, and at most 150 frames per video are passed to the VLM. Qwen3.5-9B-AWQ [16] then processes the deduplicated frames as an INT8-quantized offline vLLM batch job and extracts two complementary JSON fields: visible text and a scene/people description. The OCR field captures captions, overlays, and embedded screenshots; the description field captures visual information that may indicate role stereotypes, objectification, or contextual irony.

The prompt is deliberately task-neutral and asks for observable evidence rather than label predictions. The exact prompt used for frame-level extraction is shown below.

Frame-description prompt

These are frames extracted at 1 frame per second from a TikTok video, shown in chronological order. They represent the full video content.

Describe this TikTok video briefly. Focus on: people present, their actions, any text overlays, and the general scene.

Return a JSON object with:

- `"text"`: all text visible throughout the video, output raw without changing content or language. Use one breakline `\n` as separator between different text elements. If no text, use null.
- `"description"`: description of the video focusing on people present, their gender, actions across the video, and overall scene.

5.3. Text Fusion

We compare two fused text configurations. T+O concatenates transcription and OCR. T+O+V concatenates transcription, OCR, and visual description. Each field is prefixed with an explicit role marker, such

as `Transcription:`, `OCR:`, and `Description:`. Fields are separated with the model separator token when available.

OCR can be redundant when the speaker reads captions aloud. If OCR has at least 75% word overlap with the transcription and contains at least 15 words, we replace it with `[MATCHES_TRANSCRIPTION]`. This reduces duplicated text without removing the information that the text appeared visually.

For example, one English training video produced the following shortened and de-identified fields after preprocessing:

> **T**: "You all right? I'm going to miss this. Leave me alone."
> **O**: "How my gf wants me to react when a girl comes up to me"
> **V**: "A woman in an orange dress stands on a city street at night; a man nearby acts out an exaggerated, flustered reaction."

T+O serializes only T and O, while T+O+V serializes T, O, and V with role prefixes (`Transcription:`, `OCR:`, `Description:`) and model separators. If O substantially duplicates T, O is replaced by `[MATCHES_TRANSCRIPTION]` before serialization.

6. Models

6.1. Classical Models

Classical systems use TF–IDF features with a maximum vocabulary of 20,000 unigrams and bigrams and sublinear term frequency. We evaluate logistic regression, linear SVM, and XGBoost [17]. Logistic regression and SVM use balanced class weights. For Task 3.3, each model is wrapped in a one-vs-rest multilabel strategy. During model selection, classical configurations are evaluated with stratified 5-fold cross-validation, and soft outputs are obtained from class-probability estimates.

Classical preprocessing is aggressive: emojis are converted to textual tokens, URLs are removed, mentions are normalized to `@USER`, text is lowercased, punctuation is stripped, Spanish and English stopwords are removed, and one-character tokens are filtered. This representation is intentionally sparse and robust for small-data baselines.

6.2. Deep Learning Models

The deep-learning systems use softer preprocessing. URLs and mentions are normalized, TikTok branding artifacts and social-media UI strings are removed, date/time expressions are replaced with placeholders, and casing and emojis are otherwise preserved.

XLM-R large. XLM-RoBERTa large [18] is fully fine-tuned with a linear head over the CLS representation and its 512-token architectural limit. We sweep learning rates $\{5, 10, 20, 30\} \times 10^{-6}$ for up to 10 epochs.

Qwen3-8B. Qwen3-8B-Base [19] is adapted with 4-bit NF4 QLoRA [20]: fp16 compute, rank 16, $\alpha = 32$, dropout 0.1, and attention/MLP projection targets. We pool the final non-padding token hidden state. Sequence lengths are capped at 650 tokens for T+O and 800 for T+O+V, and the learning-rate sweep $\{1, 2, 3, 5\} \times 10^{-4}$ is run on Task 3.1 with T+O+V and reused later. Qwen3 is trained for up to 5 epochs.

All neural configurations use an 80/20 train/validation split with a fixed random seed, batch size 16, weight decay 0.01, cosine scheduling, 6% warmup, and validation-ICM-Soft-Norm checkpoint selection. Early stopping uses patience 3 for XLM-R and patience 2 for Qwen3. For Task 3.1 and Task 3.2, both models optimize KL-divergence against annotator soft distributions; for Task 3.3, they use binary cross-entropy with logits against per-category agreement fractions, followed by per-class threshold search on validation in $[0.2, 0.8]$.

6.3. Sensor Fusion

For classical models, sensor features are standardized and concatenated with TF–IDF features. For neural models, sensor features are normalized with a robust scaler, clipped to $[-5, 5]$, and compressed to 32 dimensions by a small MLP. EEG-only features use a hidden size of 48; all-sensor features use a hidden size of 64. The sensor vector is then projected to the encoder hidden dimension.

The neural fusion module computes a sigmoid gate from the concatenation of text and sensor representations:

$$\mathbf{h} = g\mathbf{h}_{text} + (1 - g)\mathbf{h}_{sensor}.$$

This allows the classifier to downweight sensors when they are noisy or irrelevant. Figure 3 shows the module. The design is intentionally simple because the sensor features are aggregated descriptors, not synchronized time series.

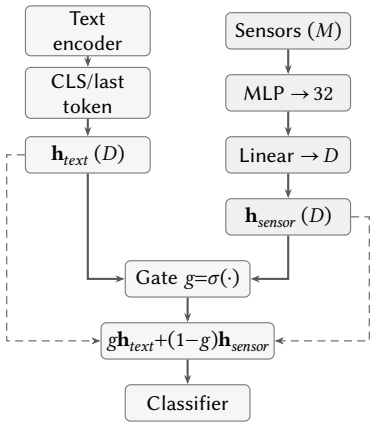

Figure 3: Gated neural sensor-fusion module. Text and sensor representations are projected to the same dimension; dashed lines show that both representations enter the weighted sum after the learned gate.

6.4. Ensembles

Submitted hard ensembles blend soft probability outputs and then derive hard labels by argmax or per-class thresholding. Candidate ensembles include two or three models, with family diversity constraints to avoid averaging only near-identical systems. Two-way ensembles require different families; three-way ensembles allow at most two members from the same family if a third family is present. Blend weights are grid-searched on validation, optimizing ICM-Norm for hard submissions and ICM-Soft-Norm for soft submissions.

7. Validation Results

Table 3 summarizes the strongest validation candidates for each concrete model and task. The LR, SVM, and XGBoost rows come from stratified 5-fold cross-validation, whereas the XLM-R and Qwen3 rows come from the 80/20 validation split described above, so the table should be used to understand trends and candidate selection rather than to claim statistically precise cross-model differences.

These validation results should be read as model-selection diagnostics rather than as a strict controlled comparison between model families. Classical models are deterministic TF–IDF pipelines evaluated after feature and model sweeps, whereas neural models include stochastic fine-tuning, checkpoint selection, and task-specific threshold search for Task 3.3. Consequently, small differences between classical and deep-learning rows are interpreted cautiously; the validation scores are most reliable for comparing candidates within the same pipeline, while the official test set is the basis for stronger cross-family conclusions.

Table 3
Best validation candidates by task and model. F1 is macro-F1, and ICM-Soft-Norm is the PyEvALL validation ICM-Soft-Norm.

Model	Best configuration	F1	ICM-Soft-Norm
Task 3.1: Identification			
LR	LR, T+O/none	0.680	0.321
SVM	SVM, T+O+V/none	0.666	0.376
XGBoost	XGBoost, T+O+V/ALL	0.663	0.385
XLM-R	XLM-R, T+O+V/none	0.742	0.565
Qwen3	Qwen3, T+O+V/ALL	0.745	0.581
Task 3.2: Intention			
LR	LR, T+O+V/none	0.577	0.192
SVM	SVM, T+O+V/none	0.581	0.278
XGBoost	XGBoost, T+O+V/none	0.509	0.268
XLM-R	XLM-R, T+O+V/none	0.665	0.480
Qwen3	Qwen3, T+O+V/none	0.706	0.499
Task 3.3: Categorization			
LR	LR, T+O+V/none	0.293	0.000
SVM	SVM, T+O+V/none	0.255	0.071
XGBoost	XGBoost, T+O+V/none	0.204	0.070
XLM-R	XLM-R, T+O+V/ALL	0.416	0.196
Qwen3	Qwen3, T+O+V/none	0.409	0.122

Three patterns are clear. First, T+O+V is selected for most best rows, suggesting that visual descriptions add information beyond speech and OCR. Second, model families behave differently by task: Qwen3 is strongest for Task 3.1 and Task 3.2, whereas XLM-R gives the best soft score for multilabel Task 3.3. Third, F1 and ICM-Soft-Norm are not interchangeable. In Task 3.1, the logistic-regression row has the highest classical F1 in Table 3 (0.680), but its ICM-Soft-Norm is lower (0.321) than SVM and XGBoost because its probabilities are less aligned with the annotator distributions. This is a useful reminder that LeWiDi systems should optimize calibration, not only hard-label accuracy. Qwen3 also illustrates this point in Task 3.3: despite competitive macro-F1, its ICM-Soft-Norm is much lower than XLM-R's, likely because the multilabel BCE objective and dominant categories lead to probability mass that is poorly calibrated for the soft LeWiDi target.

7.1. Sensor Ablation

Figure 4 shows the sensor ablation as a change relative to the text-only baseline. Sensors do not provide a uniformly useful signal: all sensors slightly improve Qwen3 on Task 3.1 from 0.566 to 0.581 ICM-Soft-Norm and XLM-R on Task 3.3 from 0.188 to 0.196, but the strongest effects are negative. For Task 3.2, EEG-only features reduce Qwen3 by 0.369 ICM-Soft-Norm and SVM by 0.132. We interpret this as a mismatch between signal and label: physiological arousal may correlate with whether content feels offensive, but source intention requires distinguishing sexist intent from critical or judgemental discussion. Viewer reaction alone does not reliably encode that distinction.

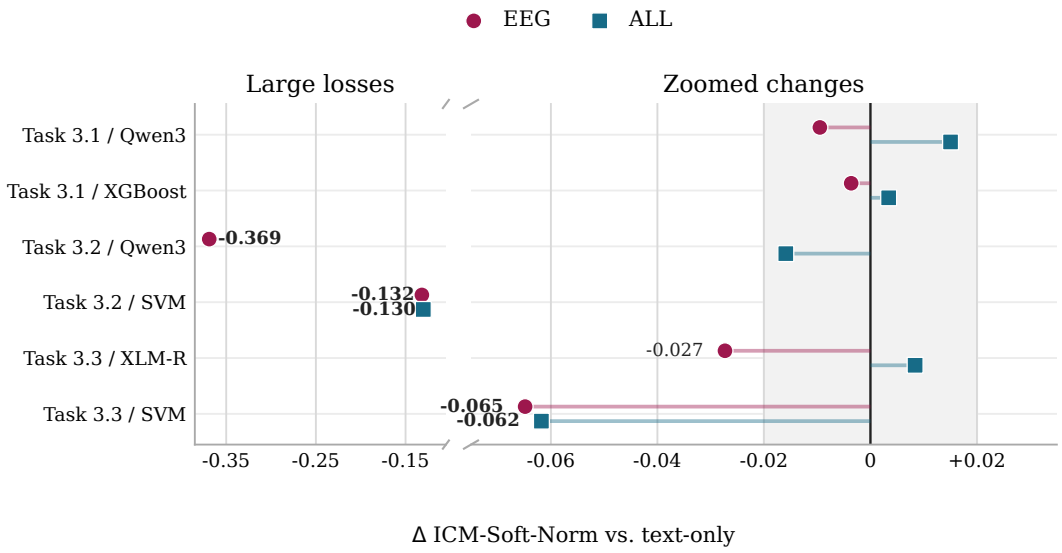

Figure 4: Sensor contribution relative to the same text-only configuration. Markers show the change in validation ICM-Soft-Norm after adding EEG-only or all physiological features. The left panel isolates large negative effects, while the right panel zooms in on smaller changes; the shaded band marks changes within ±0.02 normalized ICM.

These deltas should be treated as weak evidence rather than as a statistically supported sensor effect. The positive changes are marginal, and we did not test significance over multiple validation splits or random seeds. Because system selection was automatic and metric-driven, a sensor-enhanced configuration could be retained even when it improved validation by only a few thousandths of normalized ICM. If selecting the systems again, we would treat those cases skeptically and require either a larger validation margin or more consistent gains across tasks. The result may also reflect the feature representation: our pipeline aggregates physiological signals into fixed vectors, which may discard temporal alignment, viewer-specific variation, or event-level reactions. Under this processing, sensors more often degraded performance than improved it.

7.2. Submitted Runs

After validation, selected systems are retrained on the full training set and used to generate the final submissions. Table 4 lists the 18 submitted runs: three hard and three soft runs for each task. Run labels H1–H3 and S1–S3 are assigned independently within each task by validation rank. The table also defines each submitted system, so validation rank, run label, and ensemble composition can be read together.

Hard ensembles consistently outperform the individual validation candidates in ICM-Norm and therefore become the hard submissions for all three tasks. For soft submissions, the best individual model remains S1 in every task. Ensemble blending is retained only for S2 and S3 of Task 3.1 and Task 3.2, where it gives validation scores close to the best individual model; for Task 3.3, the candidate soft ensembles flatten the output distributions and do not improve over individual XLM-R models.

Table 4
Submitted systems, ensemble definitions, and validation scores used to assign run labels. Hard scores are shown under validation ICM-Norm and soft scores under validation ICM-Soft-Norm. Abbreviations: Q=Qwen3-8B, X=XLM-R large, G=XGBoost, S=SVM; O=T+O/none, V=T+O+V/none, A=T+O+V/ALL, E=T+O+V/EEG.

Task	Run	System definition	Val. ICM-Norm	Val. ICM-Soft-Norm
Task 3.1	H1	Hard ensemble: $.40Q_A + .20Q_O + .40X_V$	0.685	–
	H2	Hard ensemble: $.10G_V + .50Q_A + .40X_O$	0.682	–
	H3	Hard ensemble: $.10G_E + .50Q_A + .40X_O$	0.682	–
	S1	Qwen3, T+O+V/ALL	–	0.581
	S2	Soft ensemble: $.90Q_A + .10X_V$	–	0.580
	S3	Soft ensemble: $.80Q_A + .10Q_V + .10X_V$	–	0.578
Task 3.2	H1	Hard ensemble: $.40G_O + .30Q_V + .30Q_A$	0.631	–
	H2	Hard ensemble: $.10G_O + .70Q_A + .20X_V$	0.629	–
	H3	Hard ensemble: $.10S_O + .70Q_A + .20X_V$	0.628	–
	S1	Qwen3, T+O+V/none	–	0.499
	S2	Soft ensemble: $.90Q_V + .10X_V$	–	0.497
	S3	Soft ensemble: $.80Q_V + .10Q_A + .10X_V$	–	0.496
Task 3.3	H1	Hard ensemble: $.30Q_V + .60X_V + .10X_E$	0.465	–
	H2	Hard ensemble: $.50Q_V + .50X_V$	0.463	–
	H3	Hard ensemble: $.10G_O + .20Q_V + .70X_V$	0.462	–
	S1	XLM-R, T+O+V/ALL	–	0.196
	S2	XLM-R, T+O+V/none	–	0.188
	S3	XLM-R, T+O/none	–	0.178

8. Official Test Results

The organizers provided official leaderboards for each task and evaluation setting. Table 5 reports the best submitted run per task and evaluation. The ranking column is the official run-level ranking, and Table 4 defines the submitted systems behind each H/S run label.

Table 5
Best official result for our submitted runs per task and evaluation. Hard rows report ICM-Norm and soft rows report ICM-Soft-Norm. Official first-place ranks and normalized ICM scores are shown in bold. In the auxiliary column, bold is used only for the best reported F1. The last column is F1 for hard evaluations and cross entropy for soft evaluations when the organizer sheet reports it. The Task 3.3 soft sheet did not include cross entropy.

Run	Eval.	Rank	Norm. ICM	F1 / CE
Task 3.1: Identification				
1	Hard	14	0.6239	0.7150
1	Soft	4	0.5791	0.8436
Task 3.2: Intention				
2	Hard	**1**	**0.6262**	**0.7065**
1	Soft	**1**	**0.4590**	1.0930
Task 3.3: Categorization				
1	Hard	**1**	**0.4624**	0.3750
2	Soft	7	0.1522	–

Our strongest official results were in Task 3.2 and Task 3.3. In Task 3.2, our submissions ranked first in both hard and soft evaluation. The best hard run was AIdonnow_2 with 0.6262 ICM-Norm and 0.7065 F1. The best soft run was AIdonnow_1 with 0.4590 ICM-Soft-Norm and 1.0930 cross entropy. In Task 3.3, our submission ranked first in hard evaluation with 0.4624 ICM-Norm and 0.3750 F1. The soft Task 3.3 ranking was lower: our best run was AIdonnow_2, ranked seventh with 0.1522 ICM-Soft-Norm. In Task 3.1, our best soft run ranked fourth with 0.5791 ICM-Soft-Norm, while our best hard run ranked fourteenth with 0.6239 ICM-Norm.

The official test results reveal that validation ranking transferred unevenly. Figure 5 compares

validation and official scores for all submitted runs, separating hard and soft submissions because they use different normalized ICM metrics. The best validation hard run for Task 3.2 was run 1, but official test performance was highest for runs 2 and 3. Conversely, Task 3.3 hard validation was predictive: runs 1 and 2 placed first and second officially. The soft ensembles for Task 3.1 and Task 3.2 did not generalize as expected: although their validation ICM-Soft-Norm was very close to the best individual run, their official scores dropped sharply. This supports our decision to treat small validation differences cautiously.

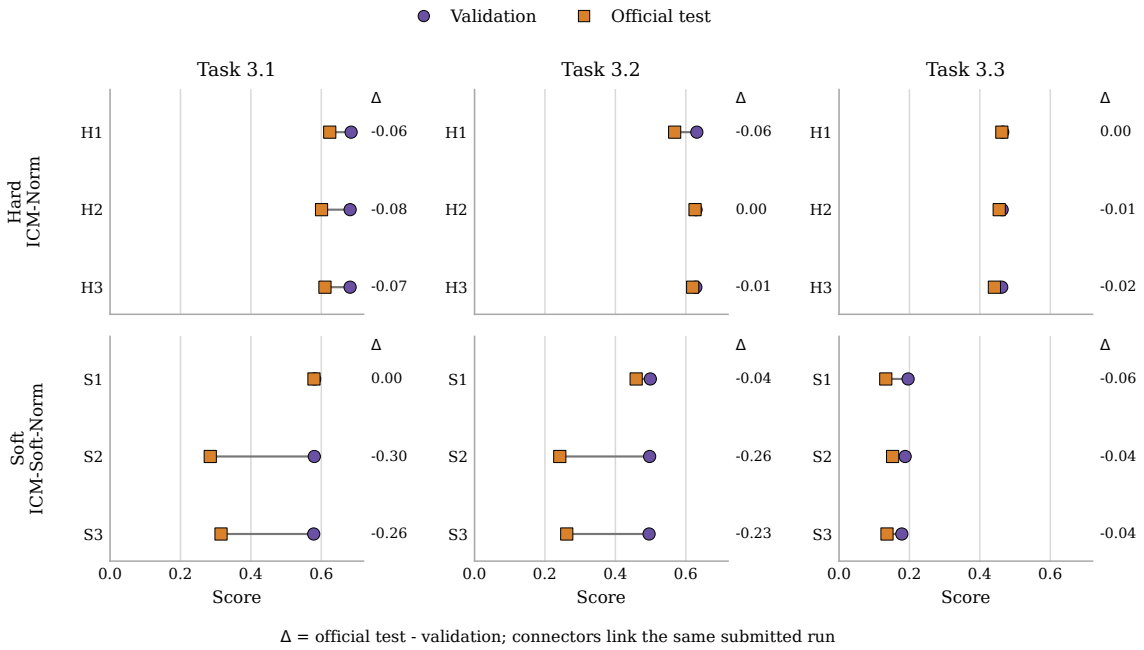

Figure 5: Validation-transfer diagnostic for submitted runs. Hard panels report ICM-Norm, while soft panels report ICM-Soft-Norm; each connector runs from validation to official test for the same submitted run, and Δ reports official test minus validation. H1–H3 and S1–S3 are run numbers within each evaluation setting.

9. Discussion

The results support the central video-to-text design. T+O+V outperformed T+O across most classical and neural configurations, indicating that frame descriptions add information beyond ASR and OCR. This is plausible for TikTok videos, where sexism may be expressed through visual framing, screenshots, gestures, or the relation between a speaker and an on-screen caption. The trade-off is that one-frame-per-second descriptions are lossy and can miss fast visual changes, sarcasm, music-driven meaning, or audiovisual contrast.

The largest discrepancy between validation and official testing was in soft ensembling. Hard ensembles combined complementary decision boundaries and led to first-place official hard rankings in Task 3.2 and Task 3.3. Soft averaging was less stable: the validation gains were small, and the best official soft submissions were individual models. This suggests that probability averaging can flatten distributions too much for contested examples, even when it looks competitive on a small validation split.

9.1. Limitations and Negative Findings

The main methodological limitation is that the video model is not trained end-to-end. Each video is first compressed into ASR, OCR, and a visual description, so errors in any stage become fixed classifier inputs. Direct video models remain a natural direction when more data and computation are available.

The validation protocol was useful for candidate selection, but not for definitive cross-family comparisons. Classical systems use deterministic cross-validation, whereas neural systems use stochastic fine-tuning, checkpoint selection, and task-specific threshold tuning. Small validation margins should therefore be interpreted cautiously.

Physiological features were useful only in restricted cases. Aggregated viewer responses may help estimate uncertainty or emotional salience, but they are not a direct substitute for understanding the video creator's intention. This explains why sensor fusion was harmful for Task 3.2, where distinguishing sexist intent from critical or judgemental content is central.

Task behavior also differed. Task 3.1 produced a strong official soft score but weaker hard ranking, suggesting that the model captured disagreement better than it chose the final threshold. Task 3.2 was the strongest task overall, while Task 3.3 remained difficult for soft prediction because the minority categories contain very few training examples.

10. Conclusion

We presented our systems for EXIST 2026 Task 3. Our best systems combine offline multimodal text reconstruction, multilingual text models, selective sensor fusion, and validation-optimized ensembles. On the official test set, we ranked first in Task 3.2 hard evaluation, first in Task 3.2 soft evaluation, and first in Task 3.3 hard evaluation. The main lesson is that reconstructing TikTok videos into rich text is a strong practical strategy for this setting, while soft calibration and physiological fusion require more careful modeling. Future work should focus on probability calibration for Task 3.3, targeted data augmentation or specialized losses for rare categories, temporal video modeling, and sensor architectures that represent individual viewer responses instead of only aggregate descriptors.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI Codex in order to assist with LaTeX drafting, table and figure generation, grammar revision, and consistency checking. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), 2026.

[2] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media (extended overview), in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[3] Qwen Team, Qwen3.5-9B: Hugging face model card, https://huggingface.co/Qwen/Qwen3.5-9B, 2026. Accessed: 2026-05-31.

[4] H. Kirk, W. Yin, B. Vidgen, P. Röttger, SemEval-2023 task 10: Explainable detection of online sexism, in: Proceedings of the 17th International Workshop on Semantic Evaluation, 2023, pp. 2193–2210. doi:10.18653/v1/2023.semeval-1.305.

[5] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and tiktok videos (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 1690–1735. URL: https://ceur-ws.org/Vol-4038/paper_135.pdf.

[6] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470. doi:10.1613/jair.1.12752.

[7] E. Amigó, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022, pp. 5809–5819. doi:10.18653/v1/2022.acl-long.399.

[8] G. Arampatzis, V. Perifanis, S. Symeonidis, A. Arampatzis, DUTH at EXIST 2025: Multilingual sexism detection with soft labels and transformers, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 1793–1800. URL: https://ceur-ws.org/Vol-4038/paper_140.pdf.

[9] T. Alcantara, O. García-Vázquez, H. Calvo, J. E. Valdez-Rodríguez, CogniCIC at EXIST 2025: Identifying sexist content in text and visual media using transformers and generative AI models, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 1761–1776. URL: https://ceur-ws.org/Vol-4038/paper_138.pdf.

[10] M. Guerrero-García, F. Carrillo-García, J. Mata-Vázquez, V. Pachón-Álvarez, I2C-UHU-Altair at EXIST 2025: Multimodal sexism detection and classification using advanced vision-language models BLIP2 and Qwen, large language models, and learning with disagreement frameworks, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 1978–1993. URL: https://ceur-ws.org/Vol-4038/paper_153.pdf.

[11] D. Fernández García, E. Amigó Cabrera, V. Cardeñoso Payo, ECA-SIMM-UVa at EXIST 2025: A segmentation oriented approach to sexism detection in TikTok videos based on a "one is enough" paradigm, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 1947–1958. URL: https://ceur-ws.org/Vol-4038/paper_151.pdf.

[12] R. Pan, T. Bernal-Beltrán, J. A. García-Díaz, R. Valencia-García, UMUTeam at EXIST 2025: Multimodal transformer architectures and soft-label learning for sexism detection, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, 2025, pp. 2084–2099. URL: https://ceur-ws.org/Vol-4038/paper_161.pdf.

[13] W.-T. Lu, J.-C. Wang, Q. Kong, Y.-N. Hung, Music source separation with band-split RoPE transformer, arXiv preprint arXiv:2309.02612 (2023). URL: https://arxiv.org/abs/2309.02612. doi:10.48550/arXiv.2309.02612.

[14] J.-C. Wang, W.-T. Lu, M. Won, Mel-band RoFormer for music source separation, arXiv preprint arXiv:2310.01809 (2023). URL: https://arxiv.org/abs/2310.01809. doi:10.48550/arXiv.2310.01809.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, Proceedings of Machine Learning Research 202 (2023) 28492–28518. URL: https://proceedings.mlr.press/v202/radford23a.html.

[16] cyankiwi, Qwen3.5-9B-AWQ-BF16-INT8: Hugging face model card, https://huggingface.co/cyankiwi/Qwen3.5-9B-AWQ-BF16-INT8, 2026. Accessed: 2026-05-31.

[17] T. Chen, C. Guestrin, XGBoost: A scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi:10.1145/2939672.2939785.

[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. doi:10.18653/v1/2020.acl-main.747.

[19] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al., Qwen3 technical report, arXiv preprint arXiv:2505.09388 (2025). URL: https://arxiv.org/abs/2505.09388. doi:10.48550/arXiv.2505.09388.

[20] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, QLoRA: Efficient finetuning of quantized LLMs, in: Advances in Neural Information Processing Systems, vol-

ume 36, 2023, pp. 10088–10115. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract.html.