UniKo at EXIST 2026: Integrating Physiological Sensor Signals Into a Multimodal Fusion Architecture for Sexism Detection in TikTok Videos*

Notebook for the EXIST Lab at CLEF 2026

Chinmay Marade[1,*], Marina Ernst[1] and Frank Hopfgartner[1]

[1]University of Koblenz, Universitätsstraße 1, 56070 Koblenz, Germany

Abstract

Sexist material circulates widely on social media platforms, where it has been linked to real harm for targeted individuals and to the wider entrenchment of gender inequality. Building detection tools that work across languages and content types is consequently a problem of concrete social urgency. This paper describes the UniKo system submitted to EXIST 2026 Subtasks 3.1 (Sexism Identification in Videos), 3.2 (Source Intention), and 3.3 (Sexism Categorisation) over TikTok videos, providing both hard and soft outputs for all three tasks. A distinguishing contribution of our work is the use of the EXIST 2026 physiological sensor recordings like eye-tracking, heart-rate, and EEG bandpower features captured while annotators watched each video as a dedicated fourth modality alongside text, speech, and visual frames. These recordings are encoded into a 324-dimensional vector representing per-annotator sensor responses together with their element-wise difference, which directly encodes inter-annotator disagreement as a trainable signal. All four modalities are projected to a shared 256-dimensional space and fused through a Gated Cross-Modal Attention module with per-task query vectors. Training combines hard-label cross-entropy, soft-label KL divergence weighted by annotator agreement and gender, focal loss for rare sexism categories, and offline data augmentation. On the official EXIST 2026 test set, our best hard-evaluation configuration achieves ICM-Hard = 0.2671 (F1 = 0.7586, **rank 19**) for Subtask 3.1 English, and Subtask 3.3 English reaches ICM-Hard = −0.2721 (F1 = 0.3306, **rank 12**), the strongest result across our 3.3 submissions.

Keywords

sexism detection, multimodal fusion, physiological signals, EEG, eye tracking, TikTok, learning with disagreement, EXIST 2026

1. Introduction

Short-form video has reshaped both how sexist content is produced and how quickly it reaches audiences. On TikTok, meaning is not carried by words alone: a backing track, a hand gesture, or text overlaid on screen can tip an otherwise neutral phrase into unambiguous sexism, or reframe an explicit statement as critique. No single channel records this fully i.e speech, visual framing, and the reaction of a viewer each contribute something the others miss.

Sexism on video-sharing platforms is sufficiently pervasive that automated approaches to its detection have become a research priority [1]. Documented effects of repeated exposure include the normalisation of gender-based discrimination and heightened impact on younger user groups; classifiers that function across languages and content formats would therefore be of practical value.

EXIST 2026 [2, 3] packages this problem as three interrelated classification tasks applied to a bilingual TikTok corpus of English and Spanish videos. What makes this edition unlike its predecessors is the collection of sensor data alongside human annotations: each training video was shown to two or three annotators, with two dedicated sensor observers whose eye movements, heart rate, and EEG activity were recorded throughout. Such recordings expose aspects of how viewers process content that explicit

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

*Corresponding author.

chinmaymarade@gmail.com (C. Marade); marinaernst@uni-koblenz.de (M. Ernst); hopfgartner@uni-koblenz.de (F. Hopfgartner)

0009-0005-2570-9432 (C. Marade); 0009-0001-7041-9419 (M. Ernst); 0000-0003-0380-6088 (F. Hopfgartner)

label choices do not capture the body's involuntary response to a stimulus and they ground evaluation in perceptual evidence that is, at least partly, measurable [4].

EXIST 2026 operates within the learning-with-disagreement paradigm [5], requiring models to predict the full distribution of annotator opinions. Evaluation uses the Information Contrast Measure (ICM) [6], a hierarchy-aware metric that penalises cross-level errors more severely than within-level errors.

This paper makes three contributions. First, we encode the physiological sensor data as a 324-dimensional vector capturing per-annotator readings and inter-annotator disagreement, treating it as a first-class modality. Second, we construct gender-weighted soft labels that account for documented sensitivity differences in annotation behaviour [7]. Third, we design a Gated Cross-Modal Attention (GCMA) module with per-task learnable query vectors, enabling each classification head to weight the modalities according to its own objective.

The paper is structured as follows. Section 2 reviews related work. Section 3 describes the tasks and data. Section 4 presents our approach. Section 5 defines the three submitted runs. Section 6 presents and analyses the official results. Section 7 concludes with directions for future work.

2. Related Work

2.1. Sexism and Hate-Speech Detection

Automated approaches to sexism detection evolved from early lexicon-based classifiers [1] to large pretrained transformers. XLM-RoBERTa [8] became a de facto backbone for cross-lingual tasks owing to its pretraining on 100 languages at scale. The EXIST shared-task series [9, 10, 11] established a systematic benchmark under the learning-with-disagreement framework, and demonstrated that models trained on full annotator distributions rather than majority-vote labels produce better-calibrated outputs on contested items [5].

2.2. Multimodal Video Analysis

The first work to examine sexism in TikTok videos as a multimodal problem was Arcos and Rosso [12], who built a system drawing on ASR transcripts, audio features, and sampled frames. Their results showed that fusing all three streams consistently beat classifiers that relied on text alone. Visual representations come from CLIP Radford et al. [13], a model trained by aligning images with natural-language descriptions at scale through a contrastive objective; the embeddings it produces transfer well to downstream tasks without any further fine-tuning. wav2vec 2.0 [14] captures prosodic and paralinguistic cues absent from transcripts, which is relevant for videos where tone of voice carries part of the sexist intent. Whisper [15] generates captions from raw audio through large-scale weak supervision, producing transcripts that are measurably cleaner than platform-side ASR output.

2.3. Physiological Signals in NLP Annotation

Recent work has begun to incorporate physiological sensor recordings into NLP annotation pipelines as an objective complement to subjective label assignment. Alacam et al. [16] showed that annotator gaze patterns during hate-speech labelling correlate with subjective annotation decisions, indicating that fixation and saccade statistics carry information about annotation uncertainty. Zhang et al. [17] found that EEG frequency-band features recorded during word-level reading predict cognitive engagement, establishing a precedent for using bandpower features in text-related classification. Arcos et al. [4] present the EXIST 2026 physiological experimental design and argue that autonomic and neural responses to sexist stimuli complement explicit label assignments.

2.4. Learning With Disagreement and ICM

Uma et al. [5] surveyed the learning-with-disagreement paradigm and showed that distributing training signal across the full set of annotator votes leads to better-calibrated models on contested items. The

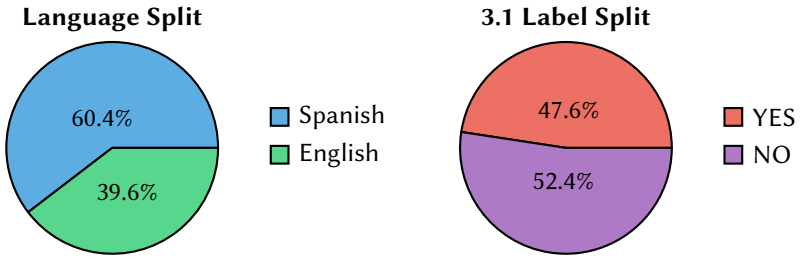

Figure 1: Training-set composition. Left: language distribution (Spanish 60.4 %, English 39.6 %). Right: Subtask 3.1 hard-label distribution (YES = 1,202, NO = 1,306). The near-balance in 3.1 contrasts sharply with the extreme minority-class problem in Subtask 3.3.

ICM metric [6] operationalises this by rewarding calibrated probability predictions and penalising misidentification of a non-sexist item as any sexist category more heavily than misidentification of the specific sexist type.

3. Task and Dataset

3.1. Task Definitions

Subtask 3.1 Binary sexism identification (YES/NO) for each video.

Subtask 3.2 Source-intention classification conditional on 3.1 = YES: DIRECT (the creator promotes sexism) or JUDGEMENTAL (the creator condemns it); non-sexist videos receive NO.

Subtask 3.3 Multi-label sexism categorisation into up to five non-exclusive types: Ideological-Inequality (IDEO), Stereotyping-Dominance (STEREO), Objectification (OBJ), Sexual-Violence (SV), and Misogyny-Non-Sexual-Violence (MIS); non-sexist videos receive NO.

The tasks form a strict hierarchy: a 3.1 = NO prediction suppresses the evaluation of 3.2 and 3.3. Evaluation is performed in two modes: *hard-hard* (majority-vote labels, ICM and F1) and *soft-soft* (annotator distribution, ICM-Soft and cross-entropy).

3.2. Dataset Overview

The EXIST 2026 Videos Dataset [2] contains 2,524 training and 674 test TikTok videos in English and Spanish, comprising 1,524 Spanish and 1,000 English training items, and 304 Spanish and 370 English test items. Each video is annotated by two to three raters drawn from a demographically diverse pool, with physiological data recorded from exactly two sensor observers per item. Figure 1 illustrates the language breakdown and the YES/NO ratio for Subtask 3.1 within the training split. Figure 2 plots per-category counts for Subtask 3.3, where the gap between the most and least frequent classes drives the augmentation approach described in Section 4.

3.3. Training-Set Statistics

Table 1 lists the hard-label counts alongside the class weights used in training and the augmentation targets for minority classes. The physiological annotation layer provides 24 eye-tracking summary statistics, 4 heart-rate statistics, and 80 EEG bandpower values (5 frequency bands × 16 channels) per subject, totalling 108 features per annotator.

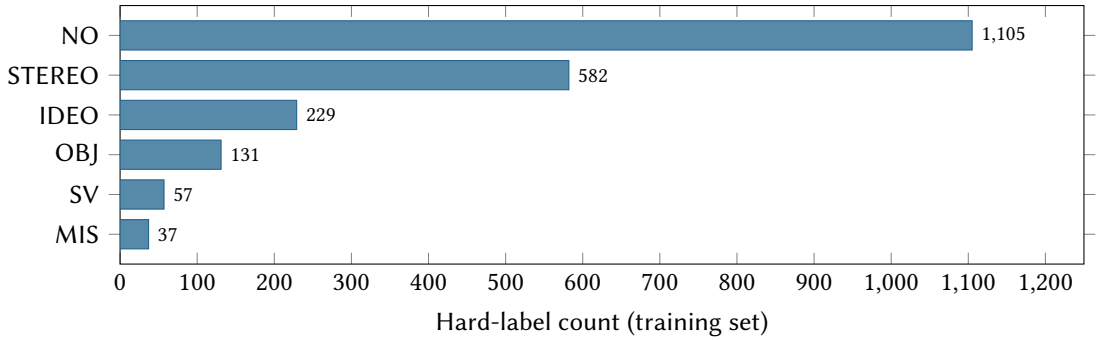

Figure 2: Subtask 3.3 hard-label counts in the training set (items can carry more than one label). STEREO = Stereotyping-Dominance, IDEO = Ideological-Inequality, OBJ = Objectification, SV = Sexual-Violence, MIS = Misogyny-Non-Sexual-Violence. The two rarest categories together account for fewer than 4 % of all training items.

Table 1
Training-set hard-label counts, class weights, and augmentation targets. Weights are inversely proportional to class frequency and normalised. Sixteen Task 3.1 items with tied votes and 74 UNKNOWN annotator votes are excluded from counts

Label / Class	Count	Weight	Aug. target
3.1 NO	1,306	0.63	—
3.1 YES	1,202	0.87	—
3.2 NO	1,306	0.63	—
3.2 DIRECT	736	1.12	—
3.2 JUDGEMENTAL	424	1.94	—
3.3 NO	1,105	0.32	—
3.3 STEREO	582	0.61	—
3.3 IDEO	229	1.56	200
3.3 OBJ	131	2.72	350
3.3 SV	57	6.26	500
3.3 MIS	37	9.64	500

4. System Description

4.1. Text Preprocessing

Raw captions from TikTok's built-in ASR contain considerable noise. Before tokenisation we run a bilingual cleaning pass that covers: NFC Unicode normalisation; removal of URLs and @mentions; hashtag splitting (*#word* → *word*); deletion of emoji text descriptions matched by a curated bilingual regular expression; collapsing of repeated characters; and stripping of bracketed noise tokens such as *[Music]*. Each item's language is then checked with *langdetect* and reconciled against the `lang` field in the JSON metadata; this step corrected 133 conflicts in the training data. We then ran Whisper large-v3 [15] over every video, generating fresh transcripts in the verified language to replace the platform captions as our primary text source.

4.2. Soft-Label Construction and Gender Weighting

Hard positives are determined by the EXIST 2026 video rule: a category is marked positive whenever more than one annotator selected it. Constructing soft labels required a decision about annotator weighting. We use gender-based weights on each annotator's vote before averaging, motivated by findings that demographic background shapes how people respond to sexist material [7]. Formally, for

an item with annotator set $\mathcal{A}$ and category c, the gender-weighted soft score is

$$\text{soft}(c) = \frac{\sum_{a \in \mathcal{A}} w_{g(a)} \cdot \mathbb{1}[a \text{ selected } c]}{\sum_{a \in \mathcal{A}} w_{g(a)}}, \tag{1}$$

where $g(a)$ is annotator a's reported gender, read from the dataset's `gender_annotators` field, and w_{female}=1.2, w_{male}=0.9. This field is fully populated for every training item (2,524/2,524) and takes only the values `F/M` in the released dataset, so no third case is observed in practice; for robustness against future data with additional categories, we define a neutral fallback w_{other}=1.0 for any annotator whose gender is unspecified, non-binary, or missing. When the 74 `UNKNOWN` Task 3.1 votes documented in Table 1 are excluded, the corresponding entries of `gender_annotators` are dropped at the same index to keep annotator–vote–gender alignment intact. Gender weighting as defined above is implemented against the verified `gender_annotators` field in the released dataset. An unweighted (w=1.0 for all annotators) baseline was not evaluated alongside this configuration. The isolated effect of gender weighting relative to that baseline has therefore not been established. The weighting also encodes a normative assumption worth stating directly: that female annotators' reactions should count more heavily toward the positive label, a defensible but not the only reading of the demographic-sensitivity literature cited above, and a separate question from whether the annotator pool itself is demographically representative.

Subtask 3.3 is multi-label, so each annotator's selections are counted separately for each category; the resulting soft fractions can therefore sum beyond 1.0, which is permitted under the task specification.

4.3. Modality Feature Extraction

Text. We encode each Whisper transcript with XLM-RoBERTa-large [8], a 560M-parameter multilingual model, taking the `[CLS]` representation at a maximum of 256 sub-word tokens (1,024-dim, matching the model's standard hidden size). The text encoder is fine-tuned end-to-end during training, using a lower learning rate than the classification heads (Section 4, Training and Calibration).
Audio. Audio tracks are downsampled to 16 kHz and passed through wav2vec2-large-xlsr-53 [14]; we average the hidden states across the time axis to obtain a 1,024-dim embedding. Where audio is absent, we substitute a zero vector. To reduce the risk of any one modality dominating training, we apply a 7.5 % dropout to each non-text modality stream.
Video. Eight frames are sampled uniformly across each video. CLIP ViT-L/14 [13] encodes each frame (768-dim); the eight embeddings are averaged into a single video token.
Fine-tuning status. Only the text encoder is fine-tuned jointly with the rest of the network. Audio (wav2vec2) and video (CLIP) embeddings are extracted once as an offline pre-processing step and cached; the corresponding encoders are kept frozen and are not updated during training. All four modality features (text `[CLS]`, audio, video, and physiological) are precomputed or produced fresh per batch as described above before entering the shared 256-dim projection MLPs.
Physiological. Per-subject 108-dim physiological vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ are concatenated with their element-wise difference: $\mathbf{s} = [\mathbf{u}_1 \parallel \mathbf{u}_2 \parallel (\mathbf{u}_1 - \mathbf{u}_2)]$, giving 324 dimensions. The difference term directly encodes inter-annotator disagreement. Median imputation handles missing values; RobustScaler normalises each feature.

4.4. Architecture

Figure 3 illustrates the full system pipeline. Each modality vector is projected to 256-dim via a two-layer MLP (LayerNorm, GELU, Dropout p=0.15). The four 256-dim tokens enter the **Gated Cross-Modal Attention** (GCMA) module, which performs four operations in sequence:

1. **Modality gating.** Each token is weighted by a sigmoid scalar derived from the full concatenated representation, allowing the network to down-weight modalities that are missing or uninformative rather than zeroing them out entirely.

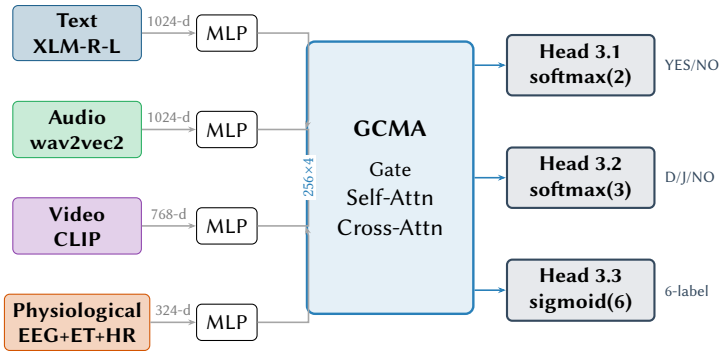

Figure 3: UniKo system architecture. Four modality encoders project to a shared 256-dim space via MLPs. The Gated Cross-Modal Attention (GCMA) module produces one global and three task-specific fused tokens; each head receives a 512-dim concatenation of both. D = DIRECT, J = JUDGEMENTAL.

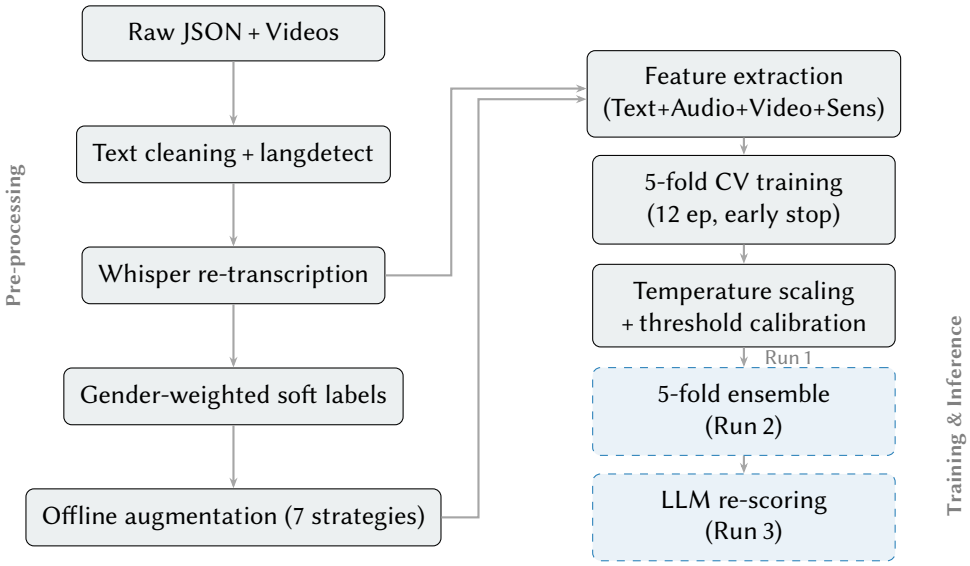

Figure 4: Training and inference pipeline. Pre-processing (left) feeds cleaned transcripts and soft labels into feature extraction. After 5-fold CV training, temperature scaling yields Run 1 (best checkpoint); ensemble averaging produces Run 2; LLM re-scoring of uncertain items produces Run 3.

2. **Self-attention.** A four-head attention layer operates over all four modality tokens simultaneously, letting each token attend to the others and pick up dependencies that span more than one input channel.
3. **Task-specific cross-attention.** Three learnable 256-dim query vectors—one per subtask—each attend to the fused token set, so each head learns its own modality weighting.
4. **Head input construction.** The global pooled token (256-dim) is concatenated with the task-specific attended token (256-dim), forming a 512-dim input for each classification head.

Three heads are applied: Linear(512→2) + softmax for 3.1; Linear(512→3) + softmax (NO / DIRECT / JUDGEMENTAL) for 3.2; Linear(512→6) + sigmoid for 3.3. At inference, hard predictions enforce hierarchical consistency: if 3.1 = NO then 3.2 is forced to NO and 3.3 to [NO], preventing hierarchy violations that ICM penalises heavily.

4.5. Training Loss

The training objective combines four terms:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + w_{\text{ann}}\,\mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{focal}} + \mathcal{L}_{\text{mixup}}, \tag{2}$$

Table 2
Illustrative augmentation outputs. Source (English): *"Every woman on this app just wants attention from men and never talks about anything else. She should just stay quiet and look pretty instead."* Changes from the source are underlined.

Strategy	Output
Token shuffle	Every woman on this app just wants attention men from and never talks about anything else. She should just stay quiet and look pretty instead.
Token deletion	woman on this app just attention from men and never talks anything else. She should just stay quiet pretty instead.
Token repeat	Every woman on on this app just wants attention from men and never talks about anything else. She should just stay quiet and look pretty instead.
Span crop	Every woman on this app just wants attention from men and never talks about anything else. She should just stay quiet
Sentence swap	She should just stay quiet and look pretty instead. Every woman on this app just wants attention from men and never talks about anything else.
Emphasis insertion	Every woman on this app just wants attention from absolutely men and never talks about anything else. She should just stay quiet and look pretty instead.
Synonym substitution	Every girl on this app just wants attention from males and never talks about anything else. She should just stay quiet and look pretty instead.

where $\mathscr{L}_{\text{CE}}$ is class-weighted cross-entropy with label smoothing ε=0.05; $\mathscr{L}_{\text{KL}}$ is soft-label KL divergence weighted by w_{ann}=n_{ann}/3; $\mathscr{L}_{\text{focal}}$ is focal loss (γ=2) on the 3.3 sigmoid head; and $\mathscr{L}_{\text{mixup}}$ is soft-label mixup (α=0.2). Task weights are 50 % for 3.1 and 25 % each for 3.2 and 3.3.

4.6. Data Augmentation

Seven augmentation strategies cycle across minority-class items without network access: (1) token shuffle (swap two adjacent tokens); (2) token deletion (10 % rate); (3) token repeat (duplicate one token); (4) span crop (85 % span from a random start point); (5) sentence swap (exchange two adjacent sentences); (6) emphasis-word insertion (a language-detection heuristic inserts one of a small set of English or Spanish emphasis adverbs, e.g. *clearly*/*claramente*, *really*/*realmente*); (7) dictionary-based synonym substitution, replacing a fixed set of sexism-domain terms (e.g. *woman*/*mujer*, *body*/*cuerpo*) with a hand-curated bilingual synonym list, at a per-occurrence probability of 0.4. Strategies are applied in round-robin fashion across synthetic examples for a given source item; only the transcript text is perturbed; audio, video, and physiological features are copied unchanged from the source item. A WeightedRandomSampler additionally applies oversampling (4× for the two rarest categories, SV and MIS; 2× for OBJ) per epoch.

Example outputs. Table 2 shows each strategy applied to an illustrative English sentence (constructed for this example, not drawn from the dataset, to avoid reproducing annotated items). Deletion, cropping, and repetition perturb surface form while leaving content words intact; sentence swap reorders clauses without changing their content; emphasis insertion and synonym substitution make small, targeted lexical substitutions.

Label preservation and cross-modal consistency. Because these transforms operate only on the transcript while audio, video, and physiological vectors are inherited unchanged from the source item, we assume that label preservation holds: the perturbations are local (single-word swaps, deletions, insertions) and are not expected to change the sexism category of the underlying item.

Augmentation and cross-validation folds. Synthetic items are generated once from the full training set, before the 5-fold split described in Section 4 (Training and Calibration), and the same fixed pool

of synthetic items is then added to the training portion of every fold. Because each synthetic item is a paraphrase of a single source video, and that source video may fall in a different fold's validation split, a fold's training set can contain a synthetic derivative of a video that is held out for validation in that same fold. Augmentation was not constrained to respect fold boundaries in this submission. A corrected version would generate fold-specific augmentation pools after the split, or track source-item provenance and exclude any synthetic item whose source falls in the current validation fold.

4.7. Training and Calibration

We train with AdamW (encoder LR = 8×10^{-6}, head LR = 4×10^{-5}, weight decay 0.01), cosine-decay schedule with 15 % warm-up, effective batch size 32 (8×4 gradient accumulation), mixed precision (fp16), on an NVIDIA A100 40 GB. Five-fold stratified cross-validation (language × 3.1 label) runs for up to 12 epochs with early stopping (patience 3). Post-training, per-task temperatures $T_{3.1}, T_{3.2}, T_{3.3}$ are fitted with LBFGS on the validation fold; the 3.1 decision threshold is swept over $[0.30, 0.70]$ to maximise binary F1.

5. Submitted Runs

We submitted three runs for all subtasks in both hard and soft modes, yielding 18 files (3 runs × 3 subtasks × hard+soft), all validated against the official format-validation script. The pipeline is illustrated in Figure 4.

Run 1. Best single checkpoint, threshold = 0.5, temperature scaling.

Run 2. Five-fold ensemble: outputs averaged across five checkpoints; calibrated threshold applied.

Run 3. Run 2 with LLM re-scoring for items where the ensemble's Subtask 3.1 P(YES) falls in an uncertain band (nominally $[0.35, 0.65]$, reported as $[0.38, 0.62]$ after threshold calibration shifts the operating point). Each uncertain item is re-queried with a single fixed system prompt that requests strict JSON output covering all three subtasks in one call (label set and required keys given verbatim; sampling temperature 0.05, max. 250 output tokens), sent through a fixed provider chain in order of first success: Gemini 2.0 Flash, then Gemini 1.5 Flash, then Groq `llama-3.3-70b-versatile`, `llama-3.1-8b-instant`, and `gemma2-9b-it`. If every provider in the chain fails a network error, malformed JSON, or missing required keys the item's score is left unchanged at its Run 2 ensemble prediction rather than falling back to any offline keyword-based method. When a provider does succeed, it returns three normalised probability distributions, one per subtask, which are clipped to $[0, 1]$ and re-normalised to sum to 1 before being blended with the model's own scores at a ratio of 55 %: 45 % (model:LLM) for Subtask 3.1 and 50 %: 50 % for Subtasks 3.2 and 3.3, with no additional calibration step applied. The blended soft scores are then re-thresholded to produce hard labels: the calibrated decision threshold for 3.1, arg-max with a DIRECT/JUDGEMENTAL tie-break for 3.2, and a per-label threshold of 0.5 with an arg-max fallback when no label clears it for 3.3.

6. Results and Analysis

6.1. Official Test-Set Results

Tables 3 and 4 report the official EXIST 2026 scores as notified by the organisers. Bold marks the best value per column within each subtask block. Positive ICM values are prefixed with + to distinguish them from negative values.

Table 3
Hard-hard evaluation on the EXIST 2026 test set. ICM-H = ICM-Hard; Norm = ICM-Hard Normalised. Bold = best value per metric within each subtask block.

Task	Lang	Run	Rank	ICM-H	Norm	F1
	ALL	3	23	+0.1932	0.5975	0.7185
3.1	ES	3	29	+0.0798	0.5414	0.6617
	EN	3	**19**	**+0.2671**	**0.6336**	**0.7586**
	ALL	3	**28**	**+0.0239**	**0.5090**	**0.6035**
3.2	ES	3	34	−0.0982	0.4599	0.5953
	EN	3	27	+0.1008	0.5362	0.6091
	ALL	3	38	−0.5169	0.3328	0.2926
3.3	ES	1	37	−0.7301	0.2326	0.2993
	EN	3	**12**	**−0.2721**	**0.4158**	**0.3306**

Table 4
Soft-soft evaluation on the EXIST 2026 test set. CE = Cross-Entropy (lower is better). Bold = best value per metric within each subtask block. "—" denotes a degenerate distribution (Section 6.3).

Task	Lang	Run	Rank	ICM-S	Norm	CE
	ALL	1	39	−0.4559	0.4200	0.8343
3.1	ES	1	40	−0.8550	0.3482	0.8745
	EN	1	**40**	**−0.2246**	**0.4610**	**0.8012**
	ALL	1	35	−2.5466	0.2288	1.2049
3.2	ES	1	38	−3.8478	0.1370	1.1639
	EN	1	**34**	**−2.0224**	**0.2760**	**1.2385**
	ALL	3	46	−18.103	0.000	—
3.3	ES	3	46	−16.807	0.000	—
	EN	3	47	−27.352	0.000	—

6.2. Cross-Validation Results

Five-fold cross-validation on the training set yielded: $F1_{3.1}$=0.7294±0.017, $F1_{3.2}$=0.6310±0.046, $F1_{3.3}$=0.3228±0.074. Test-set values (0.7185, 0.6035, 0.2926) are modestly lower, as expected when moving from held-out training folds to the fully unseen test partition. The small gap for Subtasks 3.1 and 3.2 indicates stable generalisation; the larger drop in 3.3 reflects the difficulty of transferring to very rare categories on a new set of videos.

6.3. Analysis

Hard evaluation. The single best result in our submission comes from Subtask 3.1 evaluated on English data: ICM-Hard = 0.2671, Norm = 0.6336, F1 = 0.7586, **rank 19**. To contextualise this, the majority-class baseline in the corresponding EXIST 2025 video task sat at ICM = −0.4244 [11], and more than half of all entered systems failed to beat it; our score sits 0.69 ICM points above that threshold. Looking at all languages together, Subtask 3.1 returned ICM = +0.1932 (**rank 23**) and Subtask 3.2 returned ICM = +0.0239 (**rank 28**)—both positive, which is relatively uncommon on this track given how many 2025 submissions finished below zero. On Subtask 3.3, the English partition gave us **rank 12** (Norm = 0.4158, F1 = 0.3306), our highest-placed result across all three categorisation submissions.

Run comparison. Among the three submitted configurations, Run 3 (ensemble with LLM re-scoring) produced the best hard-evaluation numbers for Subtasks 3.1 and 3.2 in every language partition, which indicates that routing uncertain predictions through an LLM is beneficial. We note this benefit is not

attributable to any single model: the provider chain tries Gemini before Groq's Llama and Gemma variants, so uncertain items were resolved by whichever provider first returned a valid response, not uniformly by one model. The picture reverses for Spanish Subtask 3.3: Run 1 (single checkpoint) beat both ensemble variants there, pointing to a loss of precision when the LLM assigns probability mass to rare multi-label categories it has little prior exposure to. For soft evaluation Run 1 again came out on top across the board; the logit-averaging step in Run 2 and Run 3 appears to flatten the predicted distributions enough to hurt ICM-Soft.

Soft evaluation. Soft scores for Subtasks 3.1 and 3.2 are negative, though this pattern is not unexpected: ICM-Soft figures across the board were very low in the analogous 2025 video evaluation [11]. The Subtask 3.3 soft results (ICM-Soft Norm = 0, CE absent) reflect a submission-format error: the JSON output used thresholded binary values rather than raw sigmoid probabilities. The PyEvALL evaluator treated this as a degenerate distribution and assigned the minimum normalised score. This is a correctable post-processing issue rather than a model deficiency; passing raw sigmoid outputs would have produced valid soft scores.

Language gap. English consistently outperforms Spanish across all subtasks. The hard Subtask 3.1 gap is 0.19 ICM points (EN 0.2671 vs. ES 0.0798). A likely contributor is transcript quality: Whisper large-v3 produces cleaner transcripts for informal English video speech than for informal spoken Spanish, which plausibly propagates into weaker text representations for the Spanish partition.

Physiological modality. The positive ICM results for Subtasks 3.1 and 3.2 are consistent with the physiological modality contributing discriminative signal beyond the transcript alone, though this remains an open question rather than a demonstrated result: no ablation compares the full four-modality system against an equivalent text + audio + video variant with the physiological stream removed, and no experiment isolates the individual contribution of EEG, eye-tracking, or heart-rate features. The GCMA gating mechanism is designed to suppress physiological tokens when EEG or eye-tracking data are weak or absent, which may partially explain why performance does not degrade even though physiological coverage is incomplete for some items. A same-architecture ablation (full system vs. physiological-ablated, and per-sub-modality leave-one-out) is the most direct way to resolve this and is our first priority for follow-up work.

7. Conclusion

We have presented UniKo, a four-modality system for EXIST 2026 that incorporates the physiological sensor data introduced in this edition as a dedicated fusion modality alongside text, speech, and visual frames. A Gated Cross-Modal Attention module with per-task query vectors allows the three classification heads to weight the modalities independently, reflecting the different information needs of binary identification, intention detection, and multi-label categorisation. Gender-weighted soft labels, annotator-agreement-scaled KL divergence, focal loss, and seven-strategy offline augmentation address the learning-with-disagreement setting and severe class imbalance.

On the official test set, UniKo achieves positive ICM for Subtasks 3.1 and 3.2 overall—both above the 2025 majority-class baseline [11]—and attains **rank 19** on English Subtask 3.1 (F1 = 0.7586) and **rank 12** on English Subtask 3.3 for hard evaluation. For soft evaluation the best result is ICM-Soft-Norm = 0.4610 (**rank 40**) for Subtask 3.1 English. The Subtask 3.3 soft submission collapsed to the floor due to a correctable output-formatting error.

Three concrete directions emerge from this analysis for future work. First, and highest priority, an ablation study should compare the full four-modality system against a text + audio + video variant with the physiological stream removed, and separately isolate the contribution of each physiological sub-modality—EEG α-band activity, pupil dilation, heart-rate variability—to each subtask and language, to determine whether the full 324-dimensional vector is necessary or whether a subset suffices. This

directly addresses the current gap between the positive ICM results reported here and a demonstrated causal contribution from the physiological modality (Section 6.3). Second, ICM-optimised threshold calibration [6] could replace the F1-optimised calibration used here, better aligning the training and evaluation objectives. Third, retrieval-augmented synthesis from external sexism corpora [9] could supply more diverse training instances for SV and MIS, complementing the surface-level augmentation strategies employed in this work.

Acknowledgments

The authors thank the EXIST 2026 organizing committee for providing the dataset, the format-validation tooling, and the timely notification of results. The authors also gratefully acknowledge the support and resources provided by the University of Koblenz, which contributed to the successful completion of this work.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly in order to: grammar and spelling check. The authors also used Claude (Anthropic) in order to: grammar and spelling assistance. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content. See also https://ceur-ws.org/GenAI/Policy.html.

References

[1] Z. Talat, D. Hovy, Hateful symbols or hateful people? predictive features for hate speech detection on twitter, in: Proceedings of the NAACL student research workshop, 2016, pp. 88–93.

[2] L. Plaza, J. C. 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), Springer, 2026.

[3] L. Plaza, J. C. 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. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[4] I. Arcos, P. Rosso, E. Gomis-Vicent, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and eeg signals, arXiv preprint arXiv:2602.23862 (2026).

[5] 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.

[6] E. Amigo, A. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers), 2022, pp. 5809–5819.

[7] Z. Talat, Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter, in: Proceedings of the first workshop on NLP and computational social science, 2016, pp. 138–142.

[8] 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.

[9] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of exist 2023-learning with disagreement for sexism identification and characterization (extended overview)., CLEF (Working Notes) (2023) 813–854.

[10] L. Plaza, J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of exist 2024—learning with disagreement for sexism identification and characterization in tweets and memes, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 93–117.

[11] L. Plaza, J. C. 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: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025, pp. 1690–1735.

[12] I. Arcos, P. Rosso, Sexism identification on tiktok: a multimodal ai approach with text, audio, and video, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2024, pp. 61–73.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[14] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems 33 (2020) 12449–12460.

[15] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518.

[16] Ö. Alacam, S. Hoeken, S. Zarrieß, Eyes don't lie: Subjective hate annotation and detection with gaze, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 187–205.

[17] Y. Zhang, Q. Li, S. Nahata, T. Jamal, S.-K. Cheng, G. Cauwenberghs, T.-P. Jung, Integrating large language model, eeg, and eye-tracking for word-level neural state classification in reading comprehension, IEEE Transactions on Neural Systems and Rehabilitation Engineering 32 (2024) 3465–3475.