Jow at EXIST 2026: Lightweight Multimodal Sexism Detection Across Memes and Videos with Demographic-Aware Soft-Label Learning

Notebook for the EXIST Lab at CLEF 2026

Jonathan Santos da Silva[1,3,*], Rui Pedro Lopes[1,2], Ahmed Gamal Ali Ali Ibrahim[2] and Gleifer Vaz Alves[3]

[1]*Escola Superior de Tecnologia e Gestão, Instituto Politécnico de Bragança, Bragança, Portugal*

[2]*Research Center in Digitalization and Intelligent Robotics (CeDRI), Laboratório Associado para a Sustentabilidade e Tecnologia em Regiões de Montanha (SusTEC), Instituto Politécnico de Bragança, Campus de Santa Apolónia, 5300-253 Bragança, Portugal*

[3]*Universidade Tecnológica Federal do Paraná, Ponta Grossa - Paraná, Brasil*

Abstract

This paper describes the system submitted by the Jow team to EXIST 2026 (sEXism Identification in Social neTworks), covering Task 2 (memes) and Task 3 (TikTok videos). Both tasks share three subtasks: binary sexism detection (.1), source intention classification (.2), and multilabel fine-grained category classification (.3). The approach follows the Learning with Disagreement (LeWiDi) paradigm, training directly on soft label distributions rather than collapsed majority votes. For Task 2, the system combines XLM-RoBERTa-large with CLIP ViT-Large/14 and injects annotator demographic information via sum-normalised count vectors. For Task 3, it extends to video with VideoMAE-base (LoRA-adapted), Whisper-large-v3 audio features, and eye-tracking and heart-rate biometric signals. Both models operate well below one billion trainable parameters and were trained without large-scale compute, yet achieve competitive results: rank 4 out of 115 on Task 2.3 (soft) and rank 8 out of 114 on Task 2.2 (soft), among others.

Keywords

sexism detection, multimodal learning, learning with disagreement, soft labels, memes, TikTok videos, EXIST 2026, annotator demographics

1. Introduction

Sexism detection in online content has attracted considerable research attention, given the scale of gender-based harassment across social media platforms. The EXIST shared task series [1, 2] pushes this problem in two useful directions: multimodality, since memes and short-form videos carry meaning through the interaction of text and image in ways that defeat unimodal approaches; and subjectivity, since sexism is a culturally situated concept where annotators regularly disagree.

EXIST 2026 adds a third dimension: physiological data. A subset of the dataset was collected in a lab where participants wore eye trackers, heart rate monitors, and EEG headsets while viewing content. Connecting computational subjectivity models to these perceptual signals is largely unexplored territory.

The Jow team participated in Task 2 (English and Spanish memes with crowd-sourced soft labels) and Task 3 (TikTok videos with the same annotation scheme). Each task has three subtasks: binary sexism detection (2.1/3.1), source intention classification (2.2/3.2), and multilabel fine-grained category classification (2.3/3.3).

The central design choice is the *Learning with Disagreement* (LeWiDi) paradigm [3]: rather than treating disagreement between annotators as noise, the system trains on the raw distribution of votes

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

*Corresponding author.

jonathan224santos@gmail.com (J. S. d. Silva); rlopes@ipb.pt (R. P. Lopes); ahmed@ipb.pt (A. G. A. A. Ibrahim); gleifer@utfpr.edu.br (G. V. Alves)

and is evaluated with the Information-Contrast Model (ICM) [4], which rewards calibrated probabilistic predictions.

Both submitted systems are deliberately lightweight. Neither uses billion-parameter vision-language models or requires full fine-tuning of large backbones. Despite this, they produce competitive results, which is the main point this paper wants to make.

The paper is organised as follows. Section 2 describes the tasks and datasets. Section 3 reviews related work. Section 4 covers the Task 2 meme system. Section 5 covers the Task 3 video system. Section 6 presents and analyses the results. Section 7 concludes.

2. Task and Dataset Description

2.1. Task Structure

Task 2 and Task 3 share the same three-subtask structure applied to different media types.

Subtask .1 – Binary Sexism Detection. Given a meme or video, predict whether the content is sexist. The soft target is the normalised distribution of YES and NO votes across annotators; the hard target is the majority vote.

Subtask .2 – Source Intention Classification. Predict whether sexism is *direct* (the author endorses it), *judgemental* (the author critiques it), or *none* (non-sexist). Both soft and hard targets are provided.

Subtask .3 – Fine-grained Category Classification. Assign one or more of six sexism categories: *ideological-inequality*, *stereotyping-dominance*, *objectification*, *sexual-violence*, *misogyny-non-sexual-violence*, and a combined category. The soft target is the annotator vote fraction per label.

2.2. Datasets

Task 2 provides memes in English and Spanish, each annotated by six crowd-sourced workers. Every meme comes with annotator demographic metadata (age group, gender, educational background, ethnicity, country of origin). A subset was also shown to lab participants wearing biometric sensors, yielding physiological signals per stimulus. Importantly, the biometric participants are not the same people as the crowd-sourced annotators, so the biosignals reflect perceptual responses rather than annotation-level disagreement.

Task 3 provides TikTok videos in English and Spanish with an equivalent annotation structure and the same biometric sub-corpus design.

2.3. Evaluation Metrics

Both tasks are evaluated with the Information-Contrast Model (ICM) [4]:

ICM-Soft compares predicted probability distributions against gold soft label distributions. This is the primary metric.

ICM-Hard compares hard predictions (argmax or thresholded) against majority-vote gold labels.

ICM-Norm normalises either variant to $[0, 1]$ where 1.0 is the gold system and 0.0 is the worst observed baseline. All reported scores use ICM-Norm. Macro F1 (subtasks .2 and .3) and F1-YES (subtask .1) are reported as supplementary metrics for the hard evaluation.

3. Related Work

3.1. Sexism Detection

Transformer-based models pretrained on large multilingual corpora have become the standard approach to sexism detection [5]. Memes complicate this because the sexist signal is often split across image and text, or carried through irony that only reads correctly when both modalities are considered together [6]. Short-form video adds temporal dynamics, speech, and audio on top of that.

3.2. Learning with Disagreement

Uma et al. [3] survey the case for treating annotator disagreement as a signal rather than noise. Plank [7] shows empirically that soft labels carry information that majority votes discard, and that models trained with them generalise better to contested instances. EXIST has operationalised this at competition scale since 2023.

3.3. Multimodal Hate Speech Detection

The Hateful Memes challenge [6] established that meme classification requires genuine cross-modal reasoning. One consistent finding is that simple CLS concatenation often matches or beats cross-modal attention on small and medium datasets, because attention tends to collapse on limited data. CLIP [8] works well as a visual backbone here because its contrastive pretraining already aligns vision and language representations.

For video, VideoMAE [9] provides strong temporal representations from a compact architecture. LoRA [10] makes fine-tuning feasible under memory constraints by updating only low-rank residuals. Whisper [11] contributes acoustic and prosodic information beyond what the transcript text alone captures.

3.4. Annotator Modelling

Encoding annotator demographics rather than treating all annotators as interchangeable has been shown to improve label aggregation quality [12]. One practical constraint is that systems built on learned annotator identity embeddings cannot generalise to unseen annotators at test time. Demographic embeddings avoid this problem by capturing group-level variation rather than individual identity [13, 12].

4. Task 2 System: Meme Sexism Detection

4.1. Architecture Overview

The Task 2 system is a joint multi-task model with a shared encoder trunk and three separate output heads. The text branch is XLM-RoBERTa-large [14], chosen for its cross-lingual coverage of English and Spanish. The visual branch is CLIP ViT-Large/14 [8], which receives the full meme image; the text is provided as a separate raw string from the dataset. Both branches produce 1024-dimensional representations, which are independently projected to 512 dimensions and then concatenated into a 1024-dimensional fused vector. Annotator demographics are then injected additively. Three task-specific linear heads produce predictions for subtasks 2.1, 2.2, and 2.3. Figure 1 illustrates the architecture.

4.2. Visual Encoder and Freezing Strategy

The EXIST 2026 meme dataset is relatively small, so fine-tuning all of CLIP ViT-Large/14 would almost certainly overfit. In Jow_1 (base model), only the last two transformer layers of CLIP are unfrozen. In Jow_2, this extends to four layers, combined with upweighting the Task 2.1 loss. Keeping most CLIP parameters frozen acts as an effective regulariser while still allowing some domain adaptation.

Each CLS token is first projected to 512 dimensions, and the two projections are then concatenated to form a 1024-dimensional fused representation:

$$h_{\text{fused}} = [\, W_{\text{text}} \cdot \text{CLS}_{\text{text}};\ W_{\text{vis}} \cdot \text{CLS}_{\text{vision}} \,] \tag{1}$$

where $W_{\text{text}}, W_{\text{vis}} \in \mathbb{R}^{512\times1024}$ and $h_{\text{fused}} \in \mathbb{R}^{1024}$.

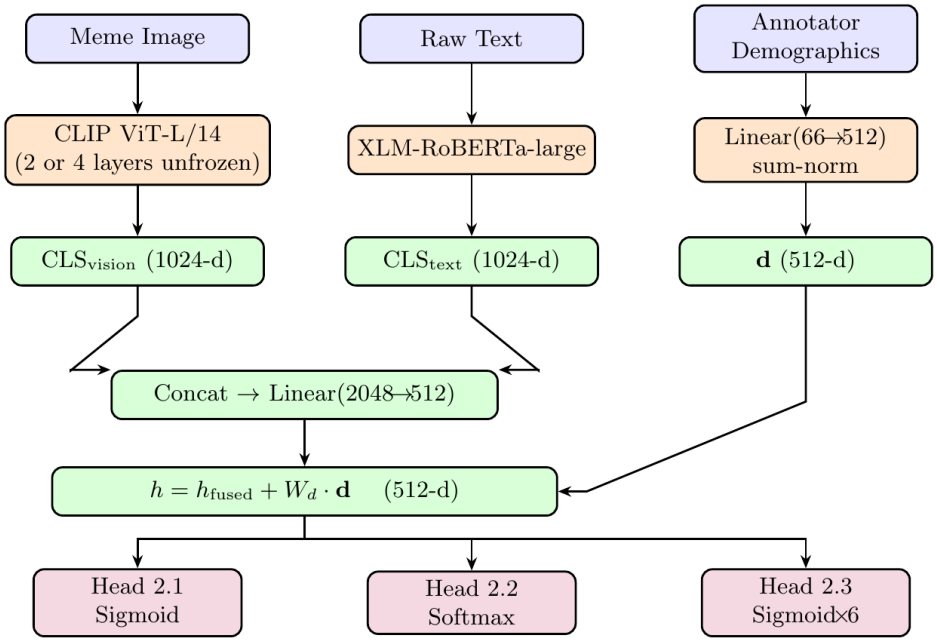

Figure 1: Task 2 architecture. The meme image is fed to CLIP; the text is fed to XLM-RoBERTa-large as a raw string. CLS tokens are projected to 512-d each, concatenated to 1024-d, and then annotator demographics are injected additively. In Jow_2, the last 4 CLIP layers are unfrozen (vs. 2 in Jow_1); in Jow_3, the additive injection is replaced by a learned sigmoid gate.

4.3. Fusion Strategy

An earlier version of the system used cross-modal attention between the text and vision CLS tokens. On the development set, attention weights collapsed to near-uniform distributions, which is a known failure mode on small datasets. CLS concatenation with a linear projection was therefore adopted for all three runs. It proved more stable and easier to tune.

4.4. Annotator Demographic Injection

The dataset provides demographic metadata for each annotator: age group, gender, educational background, ethnicity, and country of origin. These are encoded as a count vector (approximately 66-dimensional): each entry records how many annotators for a given instance belong to that demographic category.

Sum normalisation (dividing each count by the total number of annotators) was found to be critical. L2 normalisation destroys the compositional signal – an instance annotated by three women and one annotated by one woman would map to the same representation. Sum normalisation preserves that difference.

Following Nowakowski et al. [15], the demographic vector is projected to 1024 dimensions and added to the fused representation:

$$h = h_{\text{fused}} + W_d \cdot \text{demo} + b_d \tag{2}$$

Injecting at full fused dimensionality consistently outperformed lower-dimensional bottleneck designs in development experiments.

In Jow_3, the plain additive injection is replaced by a learned sigmoid gate: a small linear layer maps the demographic vector to a 1024-dimensional weight vector via sigmoid, which then scales the demographic projection before addition. This allows the model to vary the influence of each demographic dimension per instance.

4.5. Training Objectives

Subtask 2.1 uses binary cross-entropy with soft float targets (the fraction of annotators voting YES). Subtask 2.2 uses soft cross-entropy via log-softmax over the three intention classes. Subtask 2.3 uses per-label binary cross-entropy (sigmoid), treating each of the six categories independently with the annotator vote fraction as the soft target. All three heads are trained jointly with a learning rate of 10^{-5} and $\texttt{warmup\_steps} = 0.1$. In Jow_2 and Jow_3, the subtask 2.1 loss is multiplied by 2.0 to prioritise the primary binary detection signal.

4.6. Per-class Threshold Tuning

For the hard prediction in subtask 2.3, a threshold of 0.5 per label performs poorly because category prevalences are uneven. Greedy coordinate search over the development set finds per-class thresholds $\tau_c \in \{0.05, 0.10, \ldots, 0.60\}$ (12 candidates per class, 72 PyEvALL evaluations total) maximising ICM-Hard-Norm. This produced consistent improvements over the uniform 0.5 default.

On the Task 2 training logs, per-class threshold tuning consistently outperformed the uniform 0.5 default at every epoch, with the best tuned ICM-Hard-Norm reaching 0.3948 compared to 0.3761 for the default threshold, a gain of 0.019 points from post-processing alone.

The thresholds from the best validation epoch were then used directly during inference to produce the final hard predictions for submission.

4.7. Summary of Task 2 Runs

- **Jow_1**: Base model. XLM-RoBERTa-large + CLIP ViT-L/14 (2 unfrozen layers) + sum-normalised demographic injection + uniform loss weighting.
- **Jow_2**: 4 unfrozen CLIP layers + Task 2.1 upweighted loss.
- **Jow_3**: Same as Jow_2 but with a learned sigmoid gate on the demographic injection instead of plain addition.

5. Task 3 System: TikTok Video Sexism Detection

5.1. Architecture Overview

The Task 3 system handles four modalities: video frames (VideoMAE-base with LoRA), audio (Whisper-large-v3 encoder, pre-extracted), transcript text (XLM-RoBERTa-large with LoRA), and physiological signals (Bio-MLP). Annotator demographics follow the same mechanism as Task 2. The video transcript is provided directly by the dataset as a separate text string, not derived at runtime from Whisper. Figure 2 illustrates the architecture.

VideoMAE-base [9] processes 16 uniformly sampled frames per video, each resized to 224×224 pixels. To keep memory usage feasible, LoRA adapters [10] are applied to the query and value projections of VideoMAE's self-attention layers ($r = 16$, $\alpha = 32$). All other VideoMAE parameters are frozen. This substantially reduces the number of trainable video parameters while preserving the temporal representations from pretraining.

Whisper-large-v3 [11] is used as an audio feature extractor only. Mean-pooled final encoder hidden states (1280-dimensional) are extracted offline and stored to disk before training begins, so Whisper is never loaded during training. A linear projection maps these 1280-d features to 768 dimensions before fusion. This removes Whisper's memory footprint entirely while keeping acoustic and intonation information that the transcript text does not carry.

The text input is the video transcript provided directly by the dataset as a raw string. It is processed by XLM-RoBERTa-large with the same LoRA configuration as the video encoder. No language-specific preprocessing is needed since XLM-RoBERTa handles English and Spanish natively.

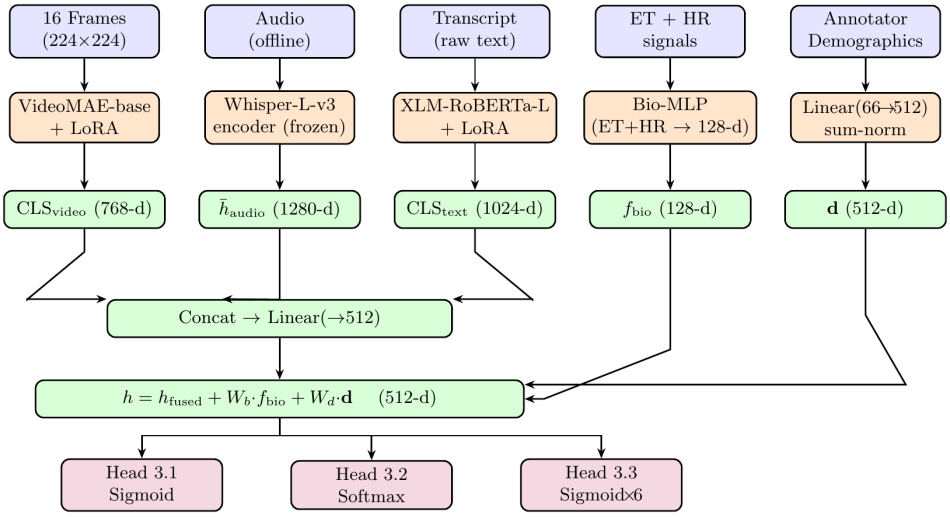

Figure 2: Task 3 architecture. Video frames go to VideoMAE-base (LoRA); audio features are pre-extracted offline with Whisper-large-v3 (frozen); the transcript text is provided separately by the dataset. Bio features (ET + HR) and annotator demographics are injected additively. In Jow_2, SWA averages checkpoints from epochs 8–14. In Jow_3, cross-modal attention replaces CLS concatenation between text and video.

5.2. Biometric Feature Integration

The physiological sub-corpus provides eye-tracking (ET) and heart rate (HR) signals from 2–4 lab participants per stimulus. EEG was considered but dropped: the data had a 70.6% dead channel rate (values in the millions rather than microvolts), making it unusable.

For each stimulus with biometric data, two participants' signals are used. Each participant contributes a 16-dimensional vector covering 12 eye-tracking features (pupil diameter mean/std for both eyes, blink count and duration, fixation count, duration mean/std, saccade count and duration, and reaction time) and 4 heart rate features (mean, std, max, min). Reaction times are clipped at 120 seconds and imputed per language when missing.

An inter-viewer agreement score (CVPC) is computed as the cosine similarity between the two participants' robust-scaled feature vectors. This scalar is concatenated to the mean of the two scaled vectors, giving a 17-dimensional input to the Bio-MLP. Features are normalised using a `RobustScaler` fitted on the training set only, which uses median and interquartile range instead of mean and standard deviation, making it less sensitive to the extreme outliers present in raw eye-tracking data.

Note that chi-squared was not used for feature selection here: it is designed for discrete data and gives incorrect results on continuous signals like ET and HR. Mutual information is the appropriate substitute.

The Bio-MLP is a two-layer network (`Linear`(17→64), GELU, `Dropout`(0.2), `Linear`(64→32)), producing a 32-dimensional biometric representation. For instances without biometric data, this component is set to zero.

5.3. Fusion Strategy and Training Configuration

The four modality representations are concatenated: video (768-d), audio (768-d), text (1024-d), and bio (32-d), giving a 2592-dimensional vector that is projected to 1024 dimensions by a two-layer fusion MLP with GELU activation. In Jow_1 and Jow_2, this concatenation is the sole fusion mechanism. In Jow_3, cross-modal attention between text and video replaces the raw CLS concatenation for those two modalities: text attends to video and video attends to text, and the attended representations enter the concatenation in place of the original CLS tokens.

During training, one randomly selected modality is dropped per step with probability 0.15, acting as

a regulariser that prevents the model from relying too heavily on any single input.

All Task 3 runs use AdamW with LoRA parameters at 10^{-4} and all other parameters (fusion, heads, bio MLP, audio projection) at 10^{-3}, with weight decay 10^{-2} and `warmup_ratio` $= 0.1$ on a cosine schedule. The combined loss is $\mathcal{L} = \mathcal{L}_A + 0.7\,\mathcal{L}_B + 0.7\,\mathcal{L}_C$, which gives Task A (binary sexism) more weight than intention and category classification. PyEvALL [4] is integrated into `compute_metrics` via `tempfile.TemporaryDirectory()` for JSON I/O.

5.4. Summary of Task 3 Runs

- **Jow_1**: Base model. VideoMAE-base (LoRA) + Whisper audio (frozen, offline, 1280→768 projected) + XLM-RoBERTa-large (LoRA) + Bio-MLP (ET + HR, 32-d) + annotator demographics + concatenation fusion.
- **Jow_2**: Same as Jow_1 with stochastic weight averaging (SWA), which averages model checkpoints from epochs 8 through 14 (7 averaged snapshots over 15 total training epochs).
- **Jow_3**: Cross-modal attention between text and video, replacing CLS concatenation as the fusion mechanism.

6. Results and Analysis

6.1. Official Results

Tables 1 to 4 report the official EXIST 2026 leaderboard results for all submitted Jow runs. ICM-Norm scores are used throughout; higher is better. The gold system and the majority-class baseline are included for reference.

Table 1
Task 2 – Soft Evaluation (ICM-Soft-Norm). Total systems: 141 (2.1), 114 (2.2), 115 (2.3).

Subtask	System	Rank	ICM-Soft-Norm
2.1	Gold	0	1.0000
	Jow_1	**10**	**0.5017**
	Jow_2	15	0.4829
	Jow_3	16	0.4820
	Majority baseline	140	0.1212
2.2	Gold	0	1.0000
	Jow_1	**8**	**0.3971**
	Jow_3	11	0.3788
	Jow_2	18	0.3660
	Majority baseline	112	0.0000
2.3	Gold	0	1.0000
	Jow_2	**4**	**0.2768**
	Jow_3	5	0.2765
	Jow_1	9	0.2545
	Majority baseline	74	0.0000

6.2. Task 2 Analysis

Subtask 2.1. Jow_1 placed rank 10 out of 141 under soft evaluation (0.5017) and rank 20 under hard evaluation (0.6090, F1-YES $= 0.7403$). Both are well above the majority-class baseline. Somewhat unexpectedly, the base model outperforms both variants on this subtask. Unfreezing more CLIP layers (Jow_2) and gating the demographics (Jow_3) do not help for binary detection, even though they help for subtask 2.3. This probably reflects the fact that binary classification needs less visual granularity than fine-grained category assignment.

Table 2
Task 2 – Hard Evaluation (ICM-Hard-Norm). Total systems: 214 (2.1), 183 (2.2), 184 (2.3).

Subtask	System	Rank	ICM-Hard-Norm
2.1	Gold	0	1.0000
	Jow_1	**20**	**0.6090**
	Jow_3	51	0.5524
	Jow_2	53	0.5515
	Majority baseline	202	0.2947
2.2	Gold	0	1.0000
	Jow_1	**20**	**0.4515**
	Jow_3	32	0.4086
	Jow_2	40	0.3914
	Majority baseline	167	0.1369
2.3	Gold	0	1.0000
	Jow_2	**28**	**0.3220**
	Jow_3	32	0.3093
	Jow_1	35	0.3053
	Majority baseline	140	0.0703

Table 3
Task 3 – Soft Evaluation (ICM-Soft-Norm). Total systems: 60 (3.1), 50 (3.2), 52 (3.3).

Subtask	System	Rank	ICM-Soft-Norm
3.1	Gold	0	1.0000
	Jow_3	**23**	**0.5002**
	Jow_1	29	0.4779
	Jow_2	31	0.4671
	Majority baseline	57	0.2740
3.2	Gold	0	1.0000
	Jow_3	**23**	**0.3122**
	Jow_1	26	0.3075
	Jow_2	30	0.2769
	Majority baseline	46	0.1663
3.3	Gold	0	1.0000
	Jow_3	**31**	**0.0746**
	Jow_1	32	0.0736
	Jow_2	36	0.0547
	Majority baseline	25	0.0931

Jow_1's cross-entropy (0.9005) is lower than several higher-ranked soft systems. ICM-Soft rewards confident correct predictions, so models that produce cautious near-uniform distributions can score lower on ICM even when their cross-entropy is fine. This is a good reminder that cross-entropy and ICM are not measuring exactly the same thing.

Subtask 2.2. Jow_1 placed rank 8 out of 114 under soft evaluation (0.3971) and rank 20 under hard evaluation (0.4515, Macro F1 $= 0.4900$). Rank 8 is the strongest relative result across all Jow runs. Source intention is a hard three-class problem – the *judgemental* class in particular requires detecting irony and pragmatic cues that are not usually explicit in a meme. Getting to rank 8 with a model under 700M trainable parameters is a reasonable outcome.

The soft and hard rankings are close for this subtask, which suggests the model's predicted distributions are reasonably calibrated for the intention classes.

Table 4
Task 3 – Hard Evaluation (ICM-Hard-Norm). Total systems: 75 (3.1), 70 (3.2), 69 (3.3).

Subtask	System	Rank	ICM-Hard-Norm
	Gold	0	1.0000
	Jow_2	**31**	**0.5788**
3.1	Jow_3	40	0.5637
	Jow_1	60	0.5248
	Majority baseline	74	0.2858
	Gold	0	1.0000
	Jow_3	**42**	**0.4702**
3.2	Jow_2	45	0.4633
	Jow_1	46	0.4484
	Majority baseline	69	0.2155
	Gold	0	1.0000
	Jow_1	**24**	**0.3784**
3.3	Jow_3	29	0.3661
	Jow_2	40	0.3316
	Majority baseline	52	0.1916

Subtask 2.3. Jow_2 and Jow_3 placed ranks 4 and 5 out of 115 under soft evaluation (0.2768 and 0.2765). This is the best collective result across the whole submission. It is worth noting that on this leaderboard, most systems score below the majority-class baseline under soft evaluation, including higher-capacity approaches. The Jow runs are consistently above it.

Under hard evaluation, the picture is less good: Jow_2 drops to rank 28 (0.3220). This is the clearest example of the soft/hard tension. Per-class threshold tuning commits probability mass to specific label assignments, which improves hard precision but flattens the distribution, hurting ICM-Soft. The two objectives pull in opposite directions, and there is no obvious way to optimise both simultaneously.

Architectural variants. Jow_1 leads on subtasks 2.1 and 2.2; Jow_2 and Jow_3 lead on 2.3. More visual capacity (4 unfrozen layers) and gated demographics are most useful for the granular multilabel task, where the model needs finer visual discrimination across six interdependent categories.

6.3. Task 3 Analysis

Subtask 3.1. Jow_3 placed rank 23 out of 60 under soft evaluation (0.5002); Jow_2 placed rank 31 under hard evaluation (0.5788). Sitting near the median on a task this complex – four modalities, multiple languages, video content – is a reasonable result for a sub-billion-parameter system.

The same pattern from Task 2 repeats: cross-modal attention (Jow_3) helps soft scores; SWA (Jow_2) helps hard scores. Cross-attention produces better-calibrated distributions; SWA's weight averaging reduces variance and stabilises hard decisions.

Subtask 3.2. Jow_3 placed rank 23 out of 50 under soft evaluation (0.3122) and rank 42 under hard evaluation (0.4702). The relative performance mirrors subtask 2.2 closely, which suggests that intention classification difficulty is roughly consistent across memes and short videos.

Subtask 3.3. This is the hardest subtask. Under soft evaluation, all three Jow runs fall below the majority-class baseline (0.0931), which is the global pattern on this leaderboard. Predicting a soft multilabel distribution across six co-occurring categories for a short video is genuinely difficult, and most participating systems cannot beat the baseline. Under hard evaluation, though, Jow_1 places rank 24 out of 69 (0.3784, Macro F1 $= 0.3787$), well above the majority-class baseline of 0.1916.

Biometric features. Eye-tracking and heart rate signals are included in all three Task 3 runs, but no formal ablation was run to isolate their contribution. This is a limitation. The competitive results relative to systems without biosignals are circumstantial evidence of utility, but a proper ablation is needed to confirm it.

Parameter efficiency. The key numbers: VideoMAE-base is 87M parameters; XLM-RoBERTa-large is 560M; CLIP ViT-L/14 is 307M; Whisper-large-v3 encoder is 637M but is inference-only with no gradient computation during training. No ensemble methods, no instruction-tuned VLMs, no full fine-tuning. The rank 4 result on Task 2.3 soft comes from a model with well under 700M trainable parameters. That is the main practical takeaway.

7. Conclusion and Future Work

The Jow system participated in EXIST 2026 Tasks 2 and 3, covering six subtasks across memes and TikTok videos. Both systems are built on the Learning with Disagreement paradigm and deliberately kept lightweight.

For Task 2, XLM-RoBERTa-large combined with CLIP ViT-L/14 and sum-normalised demographic injection reached rank 8 out of 114 on subtask 2.2 soft and ranks 4 and 5 out of 115 on subtask 2.3 soft. For Task 3, adding VideoMAE (LoRA), Whisper audio features, and biometric signals reached rank 23 out of 60 on subtask 3.1 soft, among other competitive placements.

The recurring pattern across both tasks is that soft and hard evaluation rankings diverge: what helps calibrated distribution prediction (cross-modal attention, soft training) does not always help hard decision accuracy (which benefits more from SWA and threshold tuning), and vice versa. This tension is worth tracking in future work.

The main point is straightforward: sub-billion-parameter models trained with soft-label objectives on a single GPU are competitive in this evaluation setting. That gap to billion-parameter systems is narrower than might be expected.

7.1. Future Work

Annotator reliability modelling. The Dawid-Skene EM algorithm [16] can estimate per-annotator reliability and produce re-weighted soft labels that down-weight unreliable annotators. Integrating this into the training pipeline, using demographic embeddings instead of identity embeddings so that it transfers to unseen annotators at test time, is a natural next step.

Biometric features for memes. Task 2 also has physiological signals, but they were not used in the submitted system. Applying the same biometric integration from Task 3 to Task 2 would give a controlled cross-media comparison and may improve the model's representation of perceptual ambiguity.

Better video modelling. VideoMAE-base is modest in capacity. Larger video backbones with more efficient adaptation, or architectures that explicitly model the sequential structure of short-form content, could give substantial gains on Task 3.

Formal ablations. The contribution of biometrics, demographics, LoRA, SWA, and cross-modal attention was not formally isolated. Systematic ablation experiments are needed before drawing strong conclusions about any individual component.

Calibration objectives. Temperature scaling or Dirichlet calibration applied post-hoc could help narrow the soft/hard divergence without requiring retraining.

Acknowledgments

The authors thank the EXIST 2026 organisers for maintaining the task infrastructure and the physiological annotation sub-corpus. This work was carried out in the context of a Master's programme in Artificial Intelligence and Data Science at the Instituto Politécnico de Bragança.

This work was supported by FCT - Fundação para a Ciência e Tecnologia, I.P. by projects: CeDRI, UID/05757/2025 (DOI: 10.54499/UID/05757/2025) and UID/PRR/05757/2025 (DOI: 10.54499/UID/PRR/05757/2025); SusTEC, LA/P/0007/2020 (DOI: 10.54499/LA/P/0007/2020).

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] 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.

[4] E. Amigó, 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), Dublin, Ireland, 2022, pp. 5809–5819. doi:10.18653/v1/2022.acl-long.399.

[5] Z. Waseem, 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.

[6] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020, pp. 2611–2624.

[7] B. Plank, The "problem" of human label variation: On ground truth in data, modeling and evaluation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 10671–10682. doi:10.18653/v1/2022.emnlp-main.731.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), volume 139 of *PMLR*, 2021, pp. 8748–8763.

[9] Z. Tong, Y. Song, J. Wang, L. Wang, VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training, in: Advances in Neural Information Processing Systems (NeurIPS), 2022.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: Proceedings of the International Conference on Learning Representations (ICLR), 2022.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: Proceedings of the 40th International Conference on Machine Learning (ICML), volume 202 of *PMLR*, 2023, pp. 28492–28518.

[12] A. Mostafazadeh Davani, M. Díaz, V. Prabhakaran, Dealing with disagreements: Looking beyond the majority vote in subjective annotations, Transactions of the Association for Computational Linguistics 10 (2022) 92–110. doi:10.1162/tacl_a_00449.

[13] M. van der Meer, N. Falk, P. K. Murukannaiah, E. Liscio, Annotator-centric active learning for subjective NLP tasks, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP), Miami, Florida, USA, 2024, pp. 18537–18555. doi:10.18653/v1/2024.emnlp-main.1031.

[14] 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.

[15] N. Nowakowski, L. Calogiuri, E. Egyed-Zsigmond, D. Nurbakova, J. Erbani, S. Calabretto, GrootWatch at EXIST 2025: Automatic sexism detection on social networks – classification of tweets and memes, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-4038/paper_160.pdf.

[16] A. P. Dawid, A. M. Skene, Maximum likelihood estimation of observer error-rates using the EM algorithm, Journal of the Royal Statistical Society: Series C (Applied Statistics) 28 (1979) 20–28. doi:10.2307/2346806.