NeverChorizoInMyPaella at EXIST 2026: Multimodal Sexism Identification and Categorisation in TikTok Videos via Sequential Cascade Training and Visual Cross-Attention

Notebook for the EXIST Lab at CLEF 2026

Giuseppe Di-Palma[1,†], Raúl Hernández-López[2,*,†]

[1]University of Pisa, Largo Bruno Pontecorvo 3, 56127 Pisa, Italy
[2]Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

Abstract

This paper describes the participation of the *NeverChorizoInMyPaella* team in the EXIST 2026 competitive challenge, specifically focusing on Tasks 3.1, 3.2, and 3.3 for sexism identification and categorisation in TikTok videos under the Learning with Disagreement paradigm. Building upon insights gathered from preliminary social media benchmarks on tweets and memes, we propose a multimodal framework that integrates textual OCR features, filtered automatic speech recognition (ASR) transcripts, key visual frames, and physiological signals (Eye-Tracking, Heart Rate, and EEG). To mitigate computational inefficiencies and leverage the hierarchical dependencies between the tasks, we implement a Sequential Cascade Training pipeline in combination with Visual Cross-Attention. Our experiments show that applying a strict confidence filter to ASR transcription outputs improves classification performance by suppressing background music hallucinations, achieving a top macro-F_1 score of 0.821 on hard targets and 0.791 on soft targets for primary sexism detection. Moreover, our approach ranked 1st overall in Task 3.3 (Soft-Soft), indicating that our fusion strategy is effective at modeling annotator subjectivity and disagreement in multi-label natural language scenarios.

Keywords

Sexism Detection, TikTok Videos, Multimodal, Learning with Disagreement, Natural Language Processing

1. Introduction

The rapid proliferation of short-video sharing platforms, most notably TikTok, has reshaped online content consumption in recent years. At the same time, it has amplified the spread of hostile behaviour, including online sexism and misogyny [1]. Detecting sexism within this medium is challenging because of its multimodal nature, where text, audio, visual streams, and implicit contextual cues interact. Furthermore, traditional evaluation paradigms that enforce a single ground truth often fail to capture the inherent subjectivity of human annotation. To address these limitations, the EXIST 2026 shared task [7, 8] adopts the *Learning with Disagreement* framework, requiring models to predict both hard labels (majority vote) and soft labels (complete annotator probability distributions).

Our team, *NeverChorizoInMyPaella*, participated in the TikTok video track challenge, covering Task 3.1 (Binary Sexism Identification), Task 3.2 (Source Intention), and Task 3.3 (Sexism Type Categorisation) (see Section 3.1). The approach presented in this work is the result of an incremental development process. In preliminary phases, classic machine learning baselines (TF-IDF with linear classifiers such as Logistic Regression or LinearSVC) and compact transformer architectures (xlm-roberta-base, bert-base-multilingual-cased, cardiffnlp/twitter-xlm-roberta-base) were evaluated on the text-only tweets

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
g.dipalma6@studenti.unipi.it (G. Di-Palma); raherlpe@upv.edu.es (R. Hernández-López)
0009-0009-0504-5172 (G. Di-Palma); 0009-0009-4894-1057 (R. Hernández-López)


dataset and static multimedia image-based memes dataset.[1]

During this exploratory work, we investigated multilingual approaches to sexism detection, finding that Cross-Modal Language Encoding [10] via CLIP [11] and other visual fusion strategies [3] gave competitive results. These early stages confirmed that while engineered sparse representations remain robust benchmarks, large-scale pre-trained transformers [2] optimised for social media text are essential for capturing implicit bias and sarcasm.

For the final TikTok challenge, we introduce a unified multimodal system that leverages a `twitter-xlm-roberta-large-2022` [14] backbone. The architecture uses cross-attention fusion to align visual representations (extracted via SigLIP) with speech transcripts (extracted via Whisper large-v3) and OCR text. Additionally, human physiological responses (Eye-Tracking, Heart Rate, and EEG data) are integrated via a Multi-Layer Perceptron (MLP) to explore human-centered signals. To exploit the natural subset relationship among the three tasks, we structure our training workflow as a Sequential Cascade Pipeline, minimizing the computational overhead of training separate models for nested labels.

2. Related Work

Prior work on sexism detection in social media has moved towards multimodal and relational reasoning. MemeWeaver [9] proposes a graph-based architecture that models inter-meme relationships for sexism and misogyny detection in image-text pairs, showing that relational context across items provides complementary signals per-item classification. Our system extends this direction to short-form video, fusing confidence-filtered ASR transcripts, speech-aligned visual frames, and physiological signals within a cascade architecture.

ECA-SIMM-UVa [5] at EXIST 2025 addressed TikTok video sexism detection, by decomposing each video into three independent channels: text (OCR and WhisperX transcriptions), audio (Wav2Vec-Large), and visual frames (ViViT). Despite three-modal fusion, text remains the dominant signal, with audio and video contributing minimally; their system ranked 1st in hard evaluation but mid-table in soft evaluation, exposing the difference between hard classification and calibrated distribution prediction in video content.

On the meme track, Arcos [3] presents a systematic ablation progressing from text-only XLM-RoBERTa through BLIP-generated caption augmentation and GPT-4o description injection, up to direct ViT+RoBERTa embedding fusion (3rd place overall in hard evaluation). A key finding is that soft-label training slightly reduces discriminative power in the best-performing model, and that intent classification is substantially harder than binary detection, which informed our cascade design. UMUTeam [6] at EXIST 2025 presents a fully multimodal system across all nine EXIST 2025 subtasks using XLM-RoBERTa, ViT, and VideoMAE with LeWiDi-style MSE/BCE soft-label training, reaching top-10 in the tweet hard track and competitive rankings across the meme and video subtasks.

3. Dataset and Pre-processing

3.1. Tasks

The EXIST 2026 shared task [7, 8] defines three nested sexism identification subtasks, each operating over the same TikTok video pool:

1. **Task 3.1 — Sexism Detection.** A binary classification task that determines whether a given TikTok video contains any form of sexist content, assigning one of two labels: *Sexist* or *Non-Sexist*.

[1] These baselines were not carried forward to the TikTok video track, because preliminary tests showed a substantial performance drop on video content, motivating the shift to the large-scale multimodal transformer pipeline presented in this paper. As the two problems operate over different modalities and datasets, we do not include a dedicated appendix for the tweets/memes results here.

2. **Task 3.2 — Source Intention Categorisation.** Conditioned on a sexism-positive prediction, this task categorises content according to the creator's underlying communicative intent:
    - *Judgemental Sexism*: the video critiques, denounces, or reports sexist behaviour without endorsing it.
    - *Direct Sexism*: the video explicitly embodies, promotes, or perpetuates sexist attitudes.
    - *No*: No sexism detected.
3. **Task 3.3 — Sexism Type Categorisation.** Also conditioned on a sexism-positive prediction, this task assigns each video to one or more of five sexism categories (plus "no") reflecting the nature of the sexist expression:
    - *Ideological and Inequality*: content that undermines women's rights or advocates for gender inequality.
    - *Stereotyping and Dominance*: content that reinforces gender-role stereotypes or asserts male dominance.
    - *Objectification*: content that reduces women to sexual objects for male gratification.
    - *Sexual Violence*: content that trivializes, promotes, or glorifies sexual harassment or assault.
    - *Misogyny and Non-sexual Violence*: content that expresses hostility, hatred, or threats of physical violence toward women.
    - *No*: No sexism detected.

3.2. Language Imbalance and Data Augmentation

The provided EXIST 2026 TikTok dataset consists of 2,524 multi-modal video samples. An initial exploratory analysis revealed a sharp language imbalance: 1,524 videos are Spanish-dominant (60.4%), while only 1,000 are English-dominant (39.6%). This imbalance risks cross-lingual overfitting during hyperparameter optimization. Table 1 summarises the hard-label distribution of the original training split.

Table 1
Label distribution in the original training set (2,524 videos: 1,524 ES, 1,000 EN; majority-vote hard labels). The **All** percentages, and Tasks 3.2 and 3.3 generally, are relative to the 1,202 sexist entries; the **ES** and **EN** percentages are relative to that language's own subset (1,524 ES / 1,000 EN videos for Task 3.1; 747 ES / 455 EN sexist videos for Tasks 3.2 and 3.3).

Task	Label	All		ES		EN	
		N	%	N	%	N	%
3.1	YES (sexist)	1,202	47.6	747	49.0	455	45.5
	NO	1,306	51.7	768	50.4	538	53.8
3.2	DIRECT	736	61.2	495	66.3	241	53.0
	JUDGEMENTAL	424	35.3	230	30.8	194	42.6
3.3	Stereotyping-Dominance	582	48.4	330	44.2	252	55.4
	Ideological-Inequality	229	19.1	146	19.5	83	18.2
	Objectification	131	10.9	58	7.8	73	16.0
	Sexual-Violence	57	4.7	39	5.2	18	4.0
	Misogyny-Non-Sexual-Violence	37	3.1	28	3.7	9	2.0

The Task 3.1 percentages do not sum to 100% because instances with an exact annotator tie were discarded (Section 3.4); Task 3.2 likewise omits a 3.5% tie remainder, and Task 3.3, being multi-label, does not need to sum to 100%. Sexism prevalence itself is comparable across languages, with Spanish videos (49%) leading English ones (45.5%) by 3.5 percentage points. Moreover, the Task 3.3 categories exhibit marked cross-lingual differences: Misogyny-Non-Sexual-Violence accounts for only 3.1% of sexist entries (with 3.7% Spanish against 2.0% English); Objectification (16% English against 7.8% Spanish);

and STEREOTYPING-DOMINANCE (55.4% English against 44.2% Spanish) is more prevalent in English than in Spanish; by contrast, IDEOLOGICAL-INEQUALITY (18.2% English against 19.5% Spanish) and SEXUAL-VIOLENCE (4.0% English against 5.2% Spanish) remain comparatively balanced across languages. Taken together, these cross-lingual disparities and scarcity of the smallest category motivate the class-targeted augmentation discussed in Section 6.

To balance the language distribution, we applied data augmentation via back-translation [20, 21] prior to optimization. We executed a stratified selection of Spanish samples, using the Task 3.1 majority vote label as the stratification key to preserve the underlying class distribution. These text blocks were translated into synthetic English using the Google Translate API. Each generated sample was assigned a unique identifier, inheriting the visual frames and physiological signals of its native Spanish source while substituting the text field with the English translation. This augmentation produced approximately 524 new English-dominant instances, establishing a fully balanced expanded dataset of 3,048 samples ($\approx 50\%$ Spanish, $\approx 50\%$ English). The augmented pool was subsequently divided into structured splits: 80% for training (2,438 samples), 10% for validation (305 samples), and 10% for internal testing (305 samples).

3.3. Text Normalization

Textual data from OCR and metadata underwent uniform normalization. Uniform resource locators (URLs), user handles, and social hashtags were mapped to explicit placeholder tokens, namely `[URL]`, `[USER]`, and `[HASHTAG]`. Whitespaces were collapsed, and characters were lowercased to ensure stable token boundaries.

3.4. Disagreement Mapping

Regarding target formulation, the hard target for Task 3.1 was derived via majority voting from the annotators, computed over 744 and 1,780 instances annotated by three and two annotators, respectively. Instances resulting in an exact tie were discarded to maintain binary separability. Ties occurred in 16 instances (0.6% of the training set, 9 Spanish and 7 English), in each case, one of the three annotators recorded an UNKNOWN label, reducing the vote to a 1–1 split between the remaining two valid annotations. For Tasks 3.2 and 3.3, soft targets were preserved as continuous vector distributions reflecting the raw fraction of annotators who assigned each specific category, fully adhering to the learning with disagreement paradigm.

4. Methodology

Our methodology focuses on a multimodal fusion architecture and a structured multi-stage training pipeline, as illustrated in Figure 1 and Figure 2 respectively.

4.1. Multimodal Feature Extraction

Textual & ASR Stream: Audio streams were extracted from the TikTok MP4 containers into mono WAV files at a 16 kHz sampling rate. Audio transcriptions were generated using the `Whisper large-v3` architecture [13, 15]. Preliminary experiments indicated that raw ASR outputs degraded downstream classification performance due to phonetic hallucinations triggered by TikTok background music and sound effects. To counteract this, we implemented a strict confidence filter: individual words were retained if and only if their Whisper token probability satisfied $p \geq 0.6$. Entire audio segments were completely discarded if they contained fewer than two valid words. The resulting filtered ASR text was directly concatenated with the normalized OCR text block.

Visual Stream via Cross-Attention: Rather than uniformly sampling video frames, which introduces heavy semantic redundancy, frame extraction was dynamically aligned with speech activity.

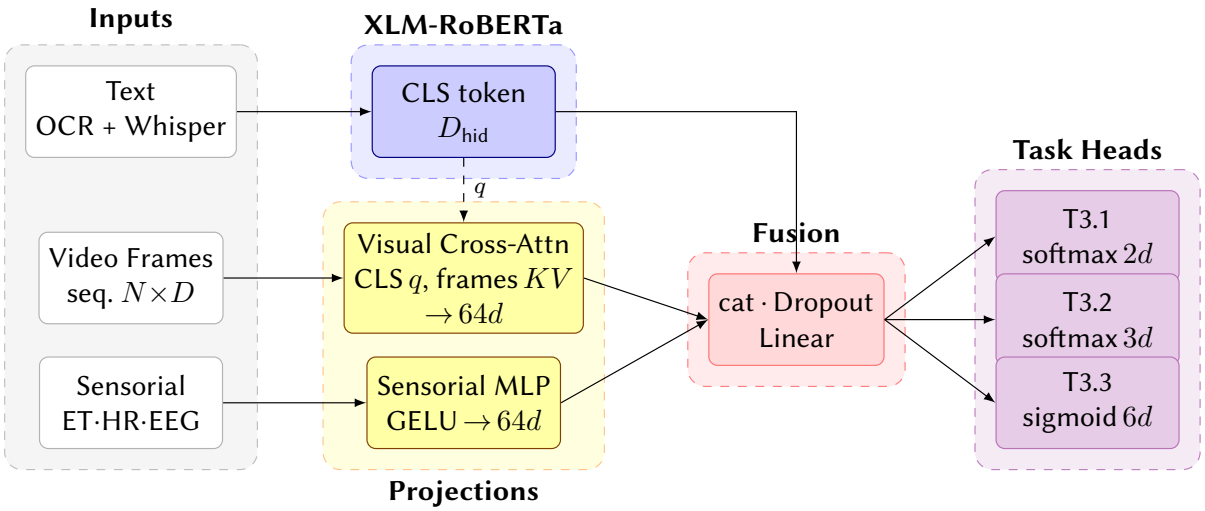

Figure 1: Multimodal architecture. Text and ASR transcripts are encoded by the XLM-RoBERTa backbone; the CLS token serves both as a direct fusion input and as the (visual) cross-attention query (q) over sequential visual frames. Physiological signals are projected by a shallow GELU-MLP. All streams are concatenated and routed to three task-specific classification heads.

Timestamps corresponding to valid words from the Whisper output were grouped into speech segments by merging consecutive words whose inter-word gap did not exceed 0.5 seconds. For each speech segment, a single representative frame was sampled at its midpoint. A maximum cap of 20 frames per video was enforced. These frames were encoded using a pre-trained SigLIP model (`google/siglip-base-patch16-224`) [12], which applies sigmoid pairwise loss [16], rather than the softmax contrastive loss used by CLIP [11]. SigLIP worked better on preliminary tests for categorisation.

For semantic fusion, we bypassed simple mean-pooling in favor of a Multi-Head Cross-Attention mechanism [18] (`attn` mode). Here, the contextualized `[CLS]` token embedding from the textual transformer serves as the *Query* (Q), while the sequence of visual frame embeddings ($N \times 768$) is projected to match the hidden dimension and serves as the *Keys* (K) and *Values* (V). For videos entirely lacking spoken content, a fallback mechanism uniformly extracted 5 frames across the total video duration.

Physiological Stream: The dataset includes human-centered physiological recordings collected during video consumption: Eye-Tracking (ET, 24 dimensions), Heart Rate (HR, 4 dimensions), and Electroencephalography (EEG, 80 dimensions across 16 channels and 5 frequency bands) [7, 4]. Missing biometric segments were mean-imputed and standardized via L_2 normalization. The composite 108-dimensional vector was processed by a shallow Multi-Layer Perceptron (MLP) mapping to a 64-dimensional dense space, using a Layer Normalization layer and the Gaussian Error Linear Unit (GELU) [22] activation function. GELU was selected over standard ReLU to prevent neuron death on negative coefficients arising from standardization.

4.2. Sequential Cascade Training Pipeline

To optimise prediction across all three tasks simultaneously without training three distinct, uncoordinated heavy models, we designed a Sequential Cascade Training pipeline. This design exploits the logical hierarchy of the labels: a video can only have a source intention (Task 3.2) or a sexism type (Task 3.3) if it is fundamentally classified as sexist (Task 3.1).

Particularly, for the multi-label Task 3.3 head, the BCE objective is weighted with a per-class `pos_weight` $= n_{\text{neg}}/n_{\text{pos}}$ to counteract the severe category imbalance, preventing rare types such as Misogyny-Non-Sexual-Violence from collapsing to near-zero recall.

The pipeline operates as follows:

1. **Gate Phase (Task 3.1):** The core multimodal transformer is fully trained on Task 3.1 to distinguish sexist from non-sexist content using standard Cross-Entropy loss for hard labels and Soft Cross-Entropy (Kullback-Leibler divergence) for soft label configurations.
2. **Downstream Warm-Start (Tasks 3.2 & 3.3):** The optimal model checkpoint from Task 3.1 is frozen in its lower layers and used to initialize the weights for the Task 3.2 and Task 3.3 heads. Training for these subsequent heads is conditioned strictly on samples that possess active positive sexism assignments.
3. **Runtime Gating Interpolation:** During soft-label inference at runtime, the downstream classification vectors are dynamically scaled by the primary gating confidence of Task 3.1 to ensure mathematical consistency. The final adjusted distribution $\tilde{q}_i^{(t)}$ for a given downstream category t is computed as:

$$\tilde{q}_i^{(t)} = P(\text{YES} \mid \text{T3.1})_i \cdot \hat{p}_i^{(t)} + P(\text{NO} \mid \text{T3.1})_i \cdot e_{\text{NO}}^{(t)} \tag{1}$$

where $P(\text{YES} \mid \text{T3.1})_i$ is the probability of sexism predicted by the first stage, $\hat{p}_i^{(t)}$ is the raw downstream network output, and $e_{\text{NO}}^{(t)}$ represents a non-sexist default assignment vector.

4.3. Implementation Details

All experiments were carried out using two hardware configurations. Optuna hyperparameters search and final training were performed on a server equipped with an NVIDIA L40S GPU (48 GB VRAM). Local training of draft/base models was run on a GeForce RTX 5080 (16 GB VRAM).

5. Experimental Results and Discussion

The experimental evaluation of our framework shows that our multimodal sequential pipeline is effective. We structured our analysis into optimization results, internal test set validation, and final competition test set performance.

5.1. Hyperparameter Optimization and Configuration Ranking

Hyperparameter optimization was executed in a two-phase routine via the *Optuna* [19] framework across 15 trials with a `MedianPruner` scheduler, comparing `cardiffnlp/twitter-xlm-roberta-large-2022` [2] against standard `xlm-roberta-large` [17]. The top three configurations were subsequently refit on the unified training and validation split before final inference.

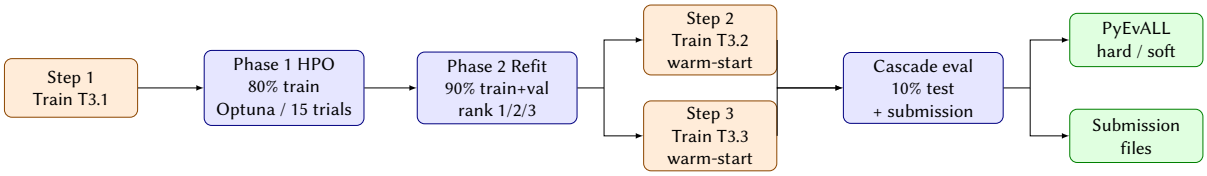

Figure 2: Sequential Cascade Training pipeline. Task 3.1 is trained first (Step 1) and optimised via Optuna HPO on 80% of the data (Phase 1). The top-3 checkpoints are refit on 90% of the data (Phase 2) and used to warm-start the Task 3.2 and Task 3.3 heads (Steps 2–3). Both downstream heads are evaluated on a 10% internal hold-out and submitted for final scoring.

Table 2 summarizes the top 3 configurations selected during the Phase-1 Optuna hyperparameter optimization for each of the three subtasks, evaluated under both Hard and Soft label paradigms.

The inclusion of unedited ASR transcripts initially caused a performance drop compared to the text-only OCR baseline due to noise and music hallucinations. However, our confidence-filtering mechanism (`text_whisper_filtered`) gave the highest overall performance (Macro-F_1 0.821 for Hard Task 3.1), indicating that removing audio noise is effective for training transformers on short-form video content.

Table 2
Optuna hyperparameter search results for the top 3 configurations across Task 3.1, 3.2, and 3.3 for both Hard and Soft modes. For Task 3.1, the backbone is `twitter-xlm-roberta-large-2022`. For downstream tasks, the backbone indicates the corresponding checkpoint from Task 3.1.

Task	Rank	Backbone	Features	Mode	LR	Epochs	Macro-F_1
Task 3.1	1	-	`text_whisper_filtered`	Hard	2.3e-5	13	0.821
	2	-	`text_ocr`	Hard	1.6e-5	7	0.818
	3	-	`text_et_whisper_filtered`	Hard	1.4e-5	6	0.808
	1	-	`text_whisper_filtered`	Soft	1.9e-5	6	0.791
	2	-	`text_et_clip`	Soft	1.7e-5	11	0.789
	3	-	`text_et_whisper_filtered_clip`	Soft	1.8e-5	8	0.789
Task 3.2	1	Rank 1	`text_whisper_filtered`	Hard	1.3e-5	15	0.839
	2	Rank 1	`text_ocr`	Hard	1.6e-5	9	0.821
	3	Rank 3	`text_whisper_filtered`	Hard	1.3e-5	6	0.805
	1	Rank 1	`text_et_clip`	Soft	2.6e-5	5	0.833
	2	Rank 1	`text_et_whisper_filtered_clip`	Soft	1.3e-5	12	0.822
	3	Rank 3	`text_et_clip`	Soft	1.3e-5	10	0.815
Task 3.3	1	Rank 1	`text_ocr`	Hard	1.0e-5	13	0.824
	2	Rank 1	`text_whisper_filtered`	Hard	1.3e-5	10	0.821
	3	Rank 3	`text_whisper_filtered`	Hard	1.3e-5	9	0.799
	1	Rank 1	`text_whisper_filtered`	Soft	1.4e-5	9	0.794
	2	Rank 1	`text_et_clip`	Soft	1.4e-5	8	0.790
	3	Rank 1	`text_et_whisper_filtered_clip`	Soft	1.4e-5	8	0.788

Learning with Disagreement (Soft labels) guided the models to slightly favor human-centered signals like eye-tracking (`text_et_clip`), capturing the underlying subjectivity of the subjects.

5.2. Internal Test Set Evaluation

Following the optimization phase, we evaluated the sequential cascade models on our isolated 10% internal test set. Table 3 and Table 4 present the evaluation results under the Hard-Hard and Soft-Soft paradigms, respectively, tracking the performance across all three nested tasks for individual checkpoints (Rank-1, Rank-2, Rank-3) and their Ensemble. The Ensemble combines the three per-task checkpoints on a per-instance basis: for the Hard-Hard setting, the final label is obtained via majority voting over the three discrete predictions (with a $> 50\%$ agreement threshold for the multi-label Task 3.3, falling back to the most frequently predicted label when no majority is reached); for the Soft-Soft setting, the final probability distribution is the unweighted average of the three predicted soft distributions. Cascade gating is reapplied after aggregation in both settings, so that an Ensemble prediction of NO for Task 3.1 forces NO predictions for Tasks 3.2 and 3.3 on that instance.

Table 3
Hard-Hard pipeline — cascade evaluation on internal test split.

	Rank-1			Rank-2			Rank-3			Ensemble		
Task	ICM	ICM-N	F1	ICM	ICM-N	F1	ICM	ICM-N	F1	ICM	ICM-N	F1
3.1	**0.305**	**0.653**	**0.769**	0.188	0.594	0.730	0.285	0.643	0.762	0.266	0.633	0.756
3.2	-0.979	0.161	0.170	-1.017	0.147	0.157	**-0.943**	**0.173**	**0.173**	-0.979	0.161	0.165
3.3	-1.907	0.031	0.047	**-1.858**	**0.043**	**0.054**	-1.892	0.035	0.048	-1.904	0.031	0.047

In the Hard-Hard pipeline (Table 3), primary sexism identification (Task 3.1) is led by the Rank-1 configuration, achieving an ICM of 0.305, an ICM-N of 0.653, and a peak Macro-F_1 score of 0.769, closely

Table 4
Soft-Soft pipeline — cascade evaluation on internal test split.

	Rank-1			Rank-2			Rank-3			Ensemble		
Task	ICM-S	ICM-SN	CE	ICM-S	ICM-SN	CE	ICM-S	ICM-SN	CE	ICM-S	ICM-SN	CE
3.1	0.605	0.607	0.865	**0.780**	**0.638**	0.877	0.714	0.627	0.922	0.690	0.622	**0.719**
3.2	-3.757	0.070	0.123	-3.763	0.070	0.133	**-3.728**	**0.074**	0.127	-3.755	0.070	**0.111**
3.3	-7.868	**0.078**	–*	**-7.856**	**0.078**	–*	-8.286	0.055	–*	-8.027	0.069	–*

* CE not reported for Task 3.3: sigmoid outputs are not a valid probability distribution over classes, so cross-entropy in the multi-class sense is undefined.

followed by Rank-3 ($F_1 = 0.762$). The Ensemble performs competitively ($F_1 = 0.756$), although it remains slightly below the best single model. However, performance degrades at each stage of the cascade to downstream tasks. For Task 3.2 (Source Intention), the evaluation metrics drop significantly, with ICM values turning negative; the Rank-3 configuration performs best here, with an ICM of -0.943, an ICM-N of 0.173, and a Macro-F_1 of 0.173.

Task 3.3 (Sexism Type Categorisation) is the main performance bottleneck due to extreme class sparsity and multi-label complexity. All configurations show low metrics, with Rank-2 scoring the highest Macro-F_1 at 0.054 (ICM = -1.858, ICM-N = 0.043), while Rank-1 and the Ensemble drop to 0.047. This multi-label degradation reflects the compounded effect of error propagation within a sequential cascade, where misclassifications or low confidences at the initial gate stage propagate to downstream heads, particularly for complex underrepresented classes such as Misogyny-Non-Sexual-Violence.

In the Soft-Soft pipeline (Table 4), the models are evaluated on their ability to predict the continuous probability distribution of annotator disagreement. For Task 3.1, the Rank-2 checkpoint shows the tightest distribution alignment, with the highest ICM-S (0.780) and ICM-SN (0.638), yet the Ensemble attains the best Cross-Entropy (CE) of 0.719 against the higher values of the single models (e.g., 0.922 for Rank-3). This split reflects a calibration-versus-ranking trade-off between the two metrics: ICM-S measures how well the predicted distribution matches the *shape* of the annotator vote, whereas CE measures the *probability* assigned to the observed outcomes. A single checkpoint can thus track the disagreement shape most closely while the Ensemble, by averaging soft predictions, produces better-calibrated probabilities. The same pattern holds in Task 3.2, where the Ensemble obtains the top CE of 0.111 despite the general drop in ICM-S (-3.755) and ICM-SN (0.070). This shows that the Ensemble consistently minimises Cross-Entropy across tasks, and indicates that averaging soft predictions provides an effective calibration mechanism, smoothing overconfident distribution peaks.

For Task 3.3, Cross-Entropy is omitted because the multi-label sigmoid outputs do not sum to a valid probability distribution over a single choice, rendering standard multi-class cross-entropy undefined. Nonetheless, the alignment metrics reveal that Rank-1 and Rank-2 tie for the top position, both reaching an ICM-SN of 0.078, while the Ensemble scores 0.069 and Rank-3 lags behind at 0.055. This clear asymmetry between the hard and soft paradigms suggests that while features optimised for hard classification (such as filtered speech transcripts) excel at finding majority ground truth, alternative multi-modal configurations incorporating visual CLIP features and physiological signals are more adept at capturing subtle subject subjectivity and disagreement patterns.

5.3. Final Competition Test Set Results

Finally, we present the official competition results evaluated on the unseen test set provided by the EXIST 2026 organizers. Based on our internal cascade validation, we generated three submission runs for each task, wiring the top configurations from Task 3.1 to act as warm-started gates for Tasks 3.2 and 3.3.

For a detailed comparison, we separate the results into the *Hard-Hard* (standard majority vote consensus) and *Soft-Soft* (learning with disagreement) evaluation paradigms across all three tasks for the combined language set (*ALL*). In each setting, we report the top 4 performing institutional systems,

the organizers' baseline, and our highest-ranked run to contextualize our performance within the final leaderboard. The leaderboards are sorted according to the raw ICM value, ICM-Hard or ICM-Soft, depending on the evaluation paradigm.

5.3.1. Hard-Hard Evaluation Paradigm

Tables 5, 6, and 7 show the leaderboards for Tasks 3.1, 3.2, and 3.3 under the traditional discrete consensus framework. In these leaderboard tables, our team's row is both bolded and shaded in light gray, to distinguish it from the plain bold used in Tables 3 and 4 to mark the best score per row.

Table 5
Official EXIST 2026 Leaderboard: Task 3.1 (Sexism Identification) – Hard-Hard Evaluation (*ALL* instances).

System	Ranking	ICM-Hard	ICM-Hard Norm	F1 YES
EXIST2025-test-gold	0	0.9907	1.0000	1.0000
ELiRF-UPV_2	1	0.3307	0.6669	0.7627
German&Blai_3	2	0.3284	0.6657	0.7487
ELiRF-UPV_1	3	0.3195	0.6613	0.7504
Aegis-Athena_2	4	0.3096	0.6563	0.7528
NeverChorizoInMyPaella_3	**38**	**0.1312**	**0.5662**	**0.6852**
EXIST2025-test_majority-class	74	-0.4244	0.2858	0.0000
EXIST2025-test_minority-class	75	-0.6036	0.1954	0.6117

Table 6
Official EXIST 2026 Leaderboard: Task 3.2 (Source Intention) – Hard-Hard Evaluation (*ALL* instances).

System	Ranking	ICM-Hard	ICM-Hard Norm	F1
EXIST2025-test-gold	0	1.3244	1.0000	1.0000
AIdonnow_2	1	0.3344	0.6262	0.7065
AIdonnow_3	2	0.3165	0.6195	0.7001
XORIK_3	3	0.2876	0.6086	0.6966
ELiRF-UPV_1	4	0.2843	0.6073	0.6901
NeverChorizoInMyPaella_2	**37**	**-0.0470**	**0.4823**	**0.5746**
EXIST2025-test_majority-class	69	-0.7537	0.2155	0.2375
EXIST2025-test_minority-class	70	-2.4749	0.0000	0.0586

Table 7
Official EXIST 2026 Leaderboard: Task 3.3 (Sexism Categorisation) – Hard-Hard Evaluation (*ALL* instances).

System	Ranking	ICM-Hard	ICM-Hard Norm	F1
EXIST2025-test-gold	0	1.5453	1.0000	1.0000
AIdonnow_1	1	-0.1162	0.4624	0.3750
AIdonnow_2	2	-0.1392	0.4549	0.3999
Aegis-Athena_2	3	-0.1394	0.4549	0.4193
German&Blai_3	4	-0.1743	0.4436	0.4285
NeverChorizoInMyPaella_3	**21**	**-0.3523**	**0.3860**	**0.4097**
EXIST2025-test_majority-class	52	-0.9530	0.1916	0.1188
EXIST2025-test_minority-class	68	-6.7467	0.0000	0.0025

Under the Hard-Hard paradigm, our systems achieved moderate overall standings. In Task 3.1 (Table 5), Run 3 obtained an ICM-Hard Norm of 0.5662 (ranking 38th), lagging behind the leading submission by *ELiRF-UPV_2* (0.6669 Norm, 0.7627 F_1). This performance gap suggests that aggregating multi-annotator perspectives via simple majority votes shifts the problem back to a rigid surface textual matching space, penalising our physiological representation.

A similar outcome can be observed in Task 3.2 (Table 6), where our OCR-only setup (Run 2) ranked 37th with an ICM-Hard Norm of 0.4823, compared to the top score of 0.6262 set by *AIdonnow_2*. The gap shrinks significantly in Task 3.3 (Table 7), where our multimodal ensemble (Run 3) climbed up to the 21st position with an ICM-Hard Norm of 0.3860 and an F_1 score of 0.4097, suggesting that combining filtered transcripts and visual frames stabilizes cascading errors even when evaluating hard targets.

5.3.2. Soft-Soft Evaluation Paradigm

By contrast, our approach performed considerably better under the *Learning with Disagreement* evaluation setup. Tables 8, 9, and 10 detail the soft distribution results, again with our team's row bolded and shaded in light gray.

Table 8
Official EXIST 2026 Leaderboard: Task 3.1 (Sexism Identification) – Soft-Soft Evaluation (*ALL* instances).

System	Ranking	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2025-test-gold	0	2.8488	1.0000	0.1962
German&Blai_2	1	0.5358	0.5940	1.2302
Dominican-Papi_2	2	0.4586	0.5805	0.8305
German&Blai_3	3	0.4549	0.5798	0.7784
AIdonnow_1	4	0.4505	0.5791	0.8436
NeverChorizoInMyPaella_3	**7**	**0.2675**	**0.5469**	**1.3183**
EXIST2025-test_majority-class	57	-1.2877	0.2740	4.4285
EXIST2025-test_minority-class	60	-2.0051	0.1481	5.5402

Table 9
Official EXIST 2026 Leaderboard: Task 3.2 (Source Intention) – Soft-Soft Evaluation (*ALL* instances).

System	Ranking	ICM-Soft	ICM-Soft Norm	Cross Entropy
EXIST2025-test-gold	0	4.6948	1.0000	0.2550
AIdonnow_1	1	-0.3847	0.4590	1.0930
Dominican-Papi_2	2	-0.5080	0.4459	1.2472
Dominican-Papi_3	3	-0.6813	0.4274	1.1624
Dominican-Papi_1	4	-0.7860	0.4163	1.2824
NeverChorizoInMyPaella_2	**17**	**-1.3753**	**0.3535**	**2.6836**
EXIST2025-test_majority-class	46	-3.1337	0.1663	4.4354
EXIST2025-test_minority-class	50	-15.4368	0.0000	8.8286

Table 10
Official EXIST 2026 Leaderboard: Task 3.3 (Sexism Categorisation) – Soft-Soft Evaluation (*ALL* instances).

System	Ranking	ICM-Soft	ICM-Soft Norm
EXIST2025-test-gold	0	8.3833	1.0000
NeverChorizoInMyPaella_3	**1**	**-5.4930**	**0.1724**
thebocadillos_1	2	-5.5814	0.1671
nodicus_3	3	-5.5964	0.1662
NeverChorizoInMyPaella_2	**4**	**-5.6924**	**0.1605**
NeverChorizoInMyPaella_1	**5**	**-5.6935**	**0.1604**
EXIST2025-test_majority-class	25	-6.8222	0.0931
EXIST2025-test_minority-class	41	-11.6668	0.0000

Under Soft-Soft evaluation, our cascade methodology was notably more robust. For Task 3.1 (Table

8), our fully multimodal Run 3 secured 7th place[2] overall with an ICM-Soft Norm of 0.5469, placing it among the top-ranked systems. In Task 3.2 (Table 9), our multimodal framework reached 17th place[3] (0.3535 ICM-Soft Norm).

Our strongest result is in Task 3.3 (Table 10). In this complex, fine-grained multi-label scenario, our configuration Run 3 (OCR + Whisper + CLIP + ET) achieved the **1st place overall**[4] across the competition, registering the highest ICM-Soft (-5.4930) and ICM-Soft Norm (0.1724) among all international participants. Our auxiliary runs (Run 2 and Run 1) also ranked highly, taking the 4th and 5th place respectively.

This result supports our hypothesis: while simple textual features (OCR) may suffice to capture the flat consensus of a majority vote, tracking the subjective boundaries of nested sexism categories benefits from human-centered physiological signals. Fusing speech-aligned visual tokens with eye-tracking (ET) features effectively captures the underlying annotator disagreement profiles, which makes it better suited for modeling soft label distributions.

Table 11
Comparison of Official EXIST 2026 Soft-Soft evaluation rankings across EN and ES language subsets, showing cross-lingual stability after back-translation.

Task	Submission (Run)	Rank (EN)	Rank (ES)
Task 3.1	Run 3	11	6
	Run 2	9	12
	Run 1	13	9
Task 3.2	Run 3	22	7
	Run 2	18	4
	Run 1	20	8
Task 3.3	Run 3	1	16
	Run 2	2	18
	Run 1	3	19

Furthermore, a breakdown of EXIST 2026 individual language performance confirms that our back-translation pipeline reduced cross-lingual imbalance and, in turn, the risk of overfitting towards one language. As shown in Table 11, despite the dataset's initial Spanish dominance, our models exhibited balanced degradation. In other words, performance is balanced across languages for Task 3.1, but the downstream tasks show a task-dependent skew, for example, the 1st-place Task 3.3 result is driven mainly by the English subset (ranks 1-3), whereas the Spanish subset places mid-table (16-19); Task 3.2 shows the opposite pattern. We attribute the Task 3.3 English-side advantage to the back-translated samples providing more balanced per-class coverage for English, while the original Spanish annotations retain their natural class distribution. Therefore, no single language collapses, confirming the back-translation pipeline prevented the mono-lingual overfitting that the original 60/40 ES/EN split would have caused.

6. Conclusions

In this work we showed that by using a heavily pre-trained social-media transformer backbone alongside confidence-based ASR filtering, and structuring the three tasks as a logical hierarchy to minimise redundant downstream computations, we addressed the challenge of detecting sexism in short-form videos. Our ranking 1st overall in Task 3.3 (Soft-Soft) shows that fusing speech-aligned visual tokens with physiological eye-tracking signals is effective for modeling subject subjectivity and disagreement

[2]Task 3.1, ranked 7th, was composed of a combination of OCR text, whisper transcription filtered by whisper, clip images extracted from timestamps of the transcripts and ET physiological signals.
[3]Task 3.2, ranked 17th, was similarly composed of OCR + Whisper + CLIP + ET, as in Task 3.1.
[4]Task 3.3, ranked 1st, had an equivalent configuration to Tasks 3.2 and 3.1.

in multi-label natural language scenarios. However, the significant performance drop in the Hard-Hard paradigm suggests that while these features are better suited for capturing nuanced disagreement patterns, they may not be as effective for majority-vote classification, which relies more heavily on clear textual cues. A practical finding from our pipeline is that filtering ASR output by token confidence is a simple, model-agnostic preprocessing step that removes background-music hallucinations and is the main driver of our best Task 3.1 result. The cascade design, however, carries a structural cost: errors at the Task 3.1 gate propagate to the conditioned downstream heads, which partly explains the degradation observed on Tasks 3.2 and 3.3 under hard evaluation.

7. Future Work

Future work will focus on the following main areas: (i) implementing class-targeted data augmentation (e.g., synthetic text generation via LLMs) specifically aimed at the most severely underrepresented minority categories of Task 3.3; (ii) incorporating annotator demographic metadata directly into the network architecture to better model the origins of the disagreement signals; (iii) exploring deeper fusion techniques for physiological features, moving beyond basic MLPs to temporal architectures capable of capturing the sequential dynamics of human physiological responses; (iv) skipping the normalization phase entirely to allow the models to process raw text, preserving the natural language style typically used in social media contexts; (v) experimenting with speech-agnostic image extraction approaches such as SIFT [23]; (vi) investigating the Hard-Hard/Soft-Soft performance asymmetry through explicit probability calibration (e.g., temperature scaling or fine-tuned decision thresholds) [24], leaving open whether the gap reflects genuine noise in the physiological signal or it requires a calibration correction.

Acknowledgements

Our participation in the EXIST 2026 competition was carried out as part of the Applications of Computational Linguistics subject of the MSc in Artificial Intelligence, Pattern Recognition and Digital Imaging at Universitat Politècnica de València. As part of this course, we thank both Paolo Rosso and David Gimeno-Gómez, from Universitat Politècnica de València, for their insightful feedback and suggestions for this work. We also thank the lab organizers of EXIST 2026 and CLEF for providing the multimodal datasets and coordinating this evaluation framework.

Declaration on Generative AI

During the preparation of this work, the authors used Anthropic Claude Sonnet to assist with code debugging and table formatting, and used both Claude Sonnet and Grammarly for English grammar correction and academic stylistic refinement to improve clarity or conciseness. All outputs were reviewed and edited by the authors, who take full responsibility for the publication's content.

References

[1] M. E. Moloney and T. P. Love. *Assessing Online Misogyny: Perspectives from Sociology and Feminist Media Studies*. Sociology Compass, 12(5):e12577, 2018. https://doi.org/10.1111/soc4.12577

[2] F. Barbieri, L. Espinosa Anke & J. Camacho-Collados. *XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond*. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC), pp. 258–266, Marseille, France, 2022. https://aclanthology.org/2022.lrec-1.27/

[3] I. Arcos. *ArcosGPT at EXIST 2025: Identifying Sexism in Memes with Multimodal Deep Learning: Fusing Text and Visual Cues*. In Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Vol. 4038, CEUR-WS.org, 2025. https://ceur-ws.org/Vol-4038/paper_141.pdf

[4] I. Arcos, P. Rosso, and E. Gomis. *Human-Centered Multimodal Fusion for Sexism Detection in Memes with Eye-Tracking, Heart Rate, and EEG Signals*. In *Proceedings of LREC 2026*, 2026.

[5] 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 CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Vol. 4038, CEUR-WS.org, 2025. https://ceur-ws.org/Vol-4038/paper_151.pdf

[6] 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 CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Vol. 4038, CEUR-WS.org, 2025. https://ceur-ws.org/Vol-4038/paper_161.pdf

[7] L. Plaza, J. Carrillo-de-Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, and 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.

[8] L. Plaza, J. Carrillo-de-Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, and D. Spina. *Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media (Extended Overview)*. In *CLEF 2026*. Working Notes, E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, and J. M. Struß (Eds.), 2026.

[9] P. Italiani, D. Gimeno-Gómez, L. Ragazzi, G. Moro, P. Rosso. *MemeWeaver: Inter-Meme Graph Reasoning for Sexism and Misogyny Detection*. In *Findings of the Association for Computational Linguistics: EACL 2026*, Rabat, Morocco, pp. 2120–2134. https://doi.org/10.18653/v1/2026.findings-eacl.111

[10] P. Italiani, F. Maqbool, D. Gimeno-Gómez, E. Fersini & C. D. Martínez-Hinarejos. *TrankilTwice at EXIST 2025: Detecting Sexism in Memes under Multi-Lingual Settings*. In Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, Vol. 4038, CEUR-WS.org, 2025. https://ceur-ws.org/Vol-4038/paper_156.pdf

[11] OpenAI Research, openai/clip-vit-base-patch32, https://huggingface.co/openai/clip-vit-base-patch32, 2021. Accessed: 2026-03-20.

[12] Google Research, google/siglip-base-patch16-224, https://huggingface.co/google/siglip-base-patch16-224, 2023. Accessed: 2026-03-23.

[13] OpenAI Research, whisper-large-v3, https://huggingface.co/openai/whisper-large-v3, 2023. Accessed: 2026-03-15.

[14] CardiffNLP Research, twitter-xlm-roberta-large-2022, https://huggingface.co/cardiffnlp/twitter-xlm-roberta-large-2022, 2022. Accessed: 2026-04-01.

[15] 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), PMLR Vol. 202, pp. 28492–28518, 2023. arXiv:2212.04356. https://arxiv.org/abs/2212.04356

[16] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. *Sigmoid Loss for Language Image Pre-Training*. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11975–11986, 2023. arXiv:2303.15343. https://arxiv.org/abs/2303.15343

[17] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. *Unsupervised Cross-lingual Representation Learning at Scale*. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 8440–8451, 2020. https://aclanthology.org/2020.acl-main.747/

[18] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. *Attention Is All You Need*. In Advances in Neural Information Processing Systems (NeurIPS) 30, 2017. arXiv:1706.03762. https://arxiv.org/abs/1706.03762

[19] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. *Optuna: A Next-generation Hyperparameter Optimization Framework*. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), pp. 2623–2631, 2019. https://arxiv.org/abs/1907.10902

[20] R. Sennrich, B. Haddow, and A. Birch. *Improving Neural Machine Translation Models with Monolingual Data*. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 86–96, 2016. https://aclanthology.org/P16-1009/

[21] M. Siino, F. Lomonaco, P. Rosso *Backtranslate what you are saying and I will tell who you are*. In Expert Systems, 2024. https://doi.org/10.1111/exsy.13568

[22] D. Hendrycks and K. Gimpel. *Gaussian Error Linear Units (GELUs)*. arXiv:1606.08415, 2016. https://arxiv.org/abs/1606.08415

[23] P. De. *Key Frame Extraction from Videos Based on SIFT and Structural Similarity*. In Communication and Intelligent Systems, Lecture Notes in Networks and Systems, Springer Nature Singapore, 2023. https://doi.org/10.1007/978-981-99-2100-3_29

[24] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. *On Calibration of Modern Neural Networks*. In Proceedings

of the 34th International Conference on Machine Learning (ICML), 2017. https://arxiv.org/abs/1706.04599