Team tai6le at EXIST 2026: System Description

Ji Wang[1,*], Shurong Chen[1,*], Hanqi Yang[1], Huanyu Dong[1] and Tian Du[1]

[1]Uppsala University, Department of Linguistics and Philology, Uppsala, Sweden

Abstract

This paper describes tai6le's participation in EXIST 2026, a shared task on sexism identification and categorization in multilingual multimedia social media content. We submitted systems for four subtasks: meme-based sexism identification (2.1), meme-based sexism categorization (2.3), video-based sexism identification (3.1), and video-based sexism categorization (3.3). Our submitted runs explored a set of pipelines, including XLM-RoBERTa-base [1] text-only fine-tuning, threshold and probability calibration for Learning with Disagreement [2] evaluation, caption-augmented meme inputs, an XLM-RoBERTa-large plus CLIP [3] late-fusion model for memes, and an XLM-RoBERTa-large text-only baseline. We also evaluated Gemma 4 [4] multimodal generative baselines internally, but did not include them in the final submission. Across the official all-language rankings, our strongest results were obtained in the soft-soft categorization setting: tai6le ranked 6th in Subtask 2.3 with ICM-Soft -4.3410 and ICM-Soft Norm 0.2699, and ranked 6th in Subtask 3.3 with ICM-Soft -5.7292 and ICM-Soft Norm 0.1583. For binary identification, our best all-language hard results were rank 24 in Subtask 2.1 and rank 33 in Subtask 3.1. The results indicate that calibrated text-centered systems can capture useful annotator-disagreement signals, especially for multi-label categorization, while hard-label performance remains sensitive to thresholding, class imbalance, and missing visual or audiovisual context.Despite exploring multimodal variants, the strongest official results were obtained by calibrated text-centered models.

Keywords

Sexism Identification, Multimodal, Social Media, Probability Calibration, XLM-RoBERTa

1. Introduction

Sexism detection in social media remains a difficult problem because sexist content is often implicit, culturally situated, and expressed through combinations of text, images, and short-form video conventions. The EXIST shared task series addresses this problem by evaluating systems for the identification and characterization of sexism in online content. EXIST 2026 extends this line of work to a fully multimedia setting, covering memes and TikTok-style videos in English and Spanish, and frames the task under a Learning with Disagreement paradigm in which annotator variability is treated as an informative signal rather than simple noise.

Team tai6le participated in four subtasks: meme-based sexism identification (Subtask 2.1), meme-based sexism categorization (Subtask 2.3), video-based sexism identification (Subtask 3.1), and video-based sexism categorization (Subtask 3.3). Subtasks 2.1 and 3.1 are binary classification tasks with labels NO and YES, while Subtasks 2.3 and 3.3 are hierarchical multi-label categorization tasks over NO and five sexism categories: IDEOLOGICAL AND INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE, and MISOGYNY AND NON-SEXUAL-VIOLENCE.

Our submitted systems and internal experiments were organized as several run families rather than a single architecture. While our core baseline fine-tunes XLM-RoBERTa-base on the official text fields, we also incorporated visual context through caption-augmented meme inputs and an XLM-RoBERTa-large plus CLIP late-fusion architecture. Additionally, we explored an XLM-RoBERTa-large text-only baseline and internal Gemma 4 multimodal generative baselines. For binary subtasks, the discriminative models were trained to match soft label distributions with a KL-divergence [5] objective. For multi-label categorization subtasks, they were trained with binary cross-entropy against the soft category distributions.

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

*Corresponding author.

ji.wang.2203@student.uu.se (J. Wang); shurong.chen.2448@student.uu.se (S. Chen); hanqi.yang.1110@student.uu.se (H. Yang); huanyu.dong.4827@student.uu.se (H. Dong); tian.du.2762@student.uu.se (T. Du)

Table 1
Subtasks Addressed

Subtask	Modality	Problem Type	Output
Task 2.1	Meme	Binary classification	YES / NO
Task 2.3	Meme	Multi-label classification	Sexism category labels
Task 3.1	Video	Binary classification	YES / NO
Task 3.3	Video	Multi-label classification	Sexism category labels

This paper describes our data preparation procedure, submitted system variants, and official results. The main finding is that our text-only approach is more competitive in soft-label settings than in hard-label settings, especially for the multi-label categorization subtasks. In the official all-language leaderboards, tai6le reached rank 6 in Subtask 2.3 soft-soft evaluation and rank 6 in Subtask 3.3 soft-soft evaluation. This suggests that the submitted probabilities captured useful disagreement patterns even when the corresponding hard-label decoding was less competitive.

2. Task and Dataset

2.1. Task Overview

EXIST 2026 is a shared task on sexism identification and characterization in social media. The 2026 edition focuses on multimedia content, including memes and TikTok videos, and introduces physiological data for multimodal sexism characterization. The lab defines six subtasks around three main objectives: sexism identification, source intention detection, and sexism categorization, applied to two types of multimodal data: memes and videos [6, 7]. In this system description, we report our submissions for four subtasks: sexism identification in memes, sexism categorization in memes, sexism identification in videos, and sexism categorization in videos. We did not submit systems for the source intention subtasks.

2.2. Subtasks Addressed

We participated in four subtasks: Task 2.1, Task 2.3, Task 3.1, and Task 3.3.

Task 2.1 requires systems to determine whether a meme is sexist itself, describes a sexist situation, or criticizes sexist behavior, using the labels YES and NO. Task 3.1 follows the same binary identification setting for videos. Tasks 2.3 and 3.3 require fine-grained sexism categorization for memes and videos, respectively. These two categorization subtasks are formulated as multi-label classification problems, since an instance may be associated with more than one sexism category. The category inventory consists of five labels: IDEOLOGICAL AND INEQUALITY, STEREOTYPING AND DOMINANCE, OBJECTIFICATION, SEXUAL VIOLENCE, and MISOGYNY AND NON-SEXUAL VIOLENCE [7].

2.3. Dataset

The EXIST 2026 dataset contains 8,235 multimedia instances in English and Spanish. The meme subset contains 5,037 instances, including 3,984 training instances and 1,053 test instances. The video subset contains 3,198 instances, including 2,524 training instances and 674 test instances. The dataset extends previous EXIST meme and TikTok video resources and adds human-centered signals collected during annotation, including eye tracking, heart rate, and EEG [6, 7]. Different system variants used their own internal development splits derived from the official training data. We therefore report split sizes together with each experiment when needed.

Table 2
Submitted Prediction Files and Corresponding Models

File	Model Description
task2_1_hard_tai6le_1.json	XLM-RoBERTa-base text-only
task2_1_hard_tai6le_2.json	XLM-RoBERTa-large + CLIP multimodal
task2_1_hard_tai6le_3.json	XLM-RoBERTa-base text-only (threshold-calibrated)
task2_1_soft_tai6le_1.json	XLM-RoBERTa-base text-only
task2_1_soft_tai6le_2.json	XLM-RoBERTa-large + CLIP multimodal
task2_1_soft_tai6le_3.json	XLM-RoBERTa-base text-only (probability-calibrated)
task2_3_hard_tai6le_1.json	XLM-RoBERTa-base text-only (threshold-calibrated)
task2_3_hard_tai6le_2.json	XLM-RoBERTa-large text-only
task2_3_hard_tai6le_3.json	XLM-RoBERTa-base text-only
task2_3_soft_tai6le_1.json	XLM-RoBERTa-large text-only
task2_3_soft_tai6le_2.json	XLM-RoBERTa-base text-only (probability-calibrated)
task2_3_soft_tai6le_3.json	XLM-RoBERTa-base text-only
task3_1_hard_tai6le_1.json	XLM-RoBERTa-base text-only
task3_1_hard_tai6le_2.json	XLM-RoBERTa-base text-only (threshold-calibrated)
task3_1_hard_tai6le_3.json	XLM-RoBERTa-base text-only
task3_1_soft_tai6le_1.json	XLM-RoBERTa-base text-only (probability-calibrated)
task3_1_soft_tai6le_2.json	XLM-RoBERTa-base text-only
task3_1_soft_tai6le_3.json	XLM-RoBERTa-base text-only
task3_3_hard_tai6le_1.json	XLM-RoBERTa-base text-only (threshold-calibrated)
task3_3_hard_tai6le_2.json	XLM-RoBERTa-large text-only
task3_3_hard_tai6le_3.json	XLM-RoBERTa-base text-only
task3_3_soft_tai6le_1.json	XLM-RoBERTa-base text-only (probability-calibrated)
task3_3_soft_tai6le_2.json	XLM-RoBERTa-large text-only
task3_3_soft_tai6le_3.json	XLM-RoBERTa-base text-only

2.4. Labels and Evaluation Setting

The official evaluation includes both hard-hard and soft-soft settings. In the hard setting, systems provide conventional discrete predictions. In the soft setting, systems provide probabilities for each category. For binary identification subtasks, hard outputs assign each instance to either YES or NO, while soft outputs provide a probability distribution over the two labels. For categorization subtasks, hard outputs assign one or more sexism categories to each instance when applicable, while soft outputs provide category-level confidence scores. In our submitted systems, we produced both hard and soft predictions for all four addressed subtasks. For the internal development experiments, we used the official training data with our own development split for model selection and threshold calibration. The official test set was only used for generating the final submitted predictions.

3. Submitted Systems

Table 2 summarizes the submitted prediction files and their corresponding model sources. The run IDs should therefore be interpreted per subtask and per evaluation type rather than as one globally identical model. In particular, `tai6le_1`, `tai6le_2`, and `tai6le_3` can refer to different model families depending on the task: XLM-RoBERTa-base text-only, threshold-calibrated and probability-calibrated XLM-RoBERTa-base text-only, XLM-RoBERTa-large text-only, and XLM-RoBERTa-large + CLIP multimodal. This mapping is important when interpreting the official leaderboard results.

4. Data Processing

We processed the official EXIST 2026 meme and video datasets for the four subtasks in which we participated. The meme subtasks used the EXIST 2026 Memes Dataset, which contains 3,984 training

Table 3
Data distributions and internal splits across tasks

Subtask	Task type	Official train	Internal train	Internal dev	Test
2.1	Binary	3,984	3,586	398	1,053
2.3	Multi-label	3,984	3,580	404	1,053
3.1	Binary	2,524	2,271	253	674
3.3	Multi-label	2,524	2,259	265	674

instances and 1,053 test instances. The video subtasks used the EXIST 2026 Videos Dataset, which contains 2,524 training instances and 674 test instances. Both datasets include English and Spanish examples.

For each subtask, we aligned the official metadata with the corresponding soft and hard gold files from the training split. The soft labels were used as the main training targets. For binary subtasks, the label vector was ordered as [NO, YES] and validated to sum to one. For multi-label subtasks, the label vector was ordered as [NO, IDEOLOGICAL AND INEQUALITY, STEREOTYPING-DOMINANCE, MISOGYNY AND NON-SEXUAL-VIOLENCE, SEXUAL-VIOLENCE, OBJECTIFICATION]. Label names were normalized to handle spelling and separator variants before vector construction.

The official training data was split into an internal training set and development set with a 0.1 development ratio and random seed 42. The split was stratified by language and label information using scikit-learn [8]. For binary subtasks, stratification used language and the soft-label argmax. For multi-label subtasks, stratification used language, the number of positive categories, and the highest-probability label. This was intended to preserve both language balance and approximate label balance in the internal development set.

Most submitted runs were text-centered and relied on the official text/OCR/transcript fields. Some meme runs additionally used generated visual captions or CLIP visual features. We did not use physiological signals in the final submitted systems. We normalized line breaks and surrounding whitespace. For meme instances, we also constructed a conservative cleaned text field that removes URLs, platform-like links, social media handles, repeated whitespace, and line breaks. The text-only submissions reported here used the original text field rather than additional image, audio, video-frame, or sensor-derived features. This design was chosen for stability and reproducibility under the submission deadline, but it also limits the system: sexism in memes and videos may depend on visual cues, speaker tone, interaction between modalities, or contextual references that are not fully recoverable from text alone.

For inference, the trained model was applied to the official test metadata for each subtask. Soft outputs were written in the official PyEvALL [9, 10] JSON format as label-probability dictionaries. Hard outputs were derived from the same probabilities using the decoding rule for each run: either argmax or a tuned YES threshold for binary subtasks, and thresholding for multi-label subtasks. All files were validated for the expected naming convention and JSON structure before being compressed into the final submission package.

5. Models and Experiment Setups

5.1. Text-Only Models

For the text-only setting, we used XLM-RoBERTa-base with the official textual fields provided by the dataset. For meme instances, the input was the official OCR text. For video instances, the input was the official textual content. This setting was used in the text-only XLM-RoBERTa-base run for the binary identification subtasks. It was also used in the selected formal XLM-RoBERTa-base run for the video subtasks, where Tasks 3.1 and 3.3 used the official original video text only.

5.2. Multimodal

For the multimodal setting, we used generated visual captions as textual descriptions of meme images. The final classifier was still XLM-RoBERTa-base, but the input combined information from two sources: the official meme OCR text and the generated visual caption. For Task 2.1, we used a single-template format combining OCR and caption text. For Task 2.3, we used a RoBERTa-style [11] separated format. This setting corresponds to the meme part of the selected formal XLM-RoBERTa-base run.

5.3. Calibration

The calibrated text-only variant was derived from XLM-RoBERTa-base probability outputs and did not introduce a new trained backbone. Instead, it focused on tuning and finalizing the model outputs for the two official evaluation contexts: threshold selection for hard submissions and probability calibration for soft submissions.

For hard submissions, predicted probabilities were decoded with task-specific thresholds selected on the internal development split using the official ICM metric. For binary subtasks, this involved selecting a single threshold for the YES probability. For multi-label subtasks, label-specific thresholds were selected, and empty hard-label predictions were post-processed as NO.

For soft submissions, probabilities were calibrated separately on the internal development split using ICMSoft as the target metric. Binary subtasks used logit-based calibration, while multi-label subtasks used label-wise probability scaling. The calibrated outputs were then converted into the official PyEvALL [9, 10] JSON format and validated before submission.

5.4. XLM-RoBERTa-large + CLIP multimodal

For the multimodal configurations, we developed a late-fusion architecture integrating text and visual features. Specifically, we employed XLM-RoBERTa-Large [1] to process the conservatively cleaned text alongside OpenAI's CLIP-ViT-Base-Patch32 [3] to extract visual representations from the meme images. To analyze the model's overconfidence toward the minority class and its sensitivity to the task's primary metric, we evaluated decision thresholds beyond the default 0.5 on the development set. This sweep suggested an optimal classification boundary around $\tau = 0.53$ based on the argmax of the Information Contrast Model (ICM) score, but we report it as a development-set observation rather than as a tuned property of the final submission.

5.5. Gemma 4 MLLM (Zero-Shot 26B and LoRA-Tuned E4B)

For the internal generative multimodal binary baseline, we evaluated two configurations from Google's Gemma 4 family on the sexism identification subtasks. The first is a zero-shot baseline using gemma-4-26B-A4B-it [4]. The second is a PEFT LoRA[12] fine-tune of the smaller dense sibling gemma-4-E4B-it [4], where the base weights are frozen and a low-rank update is injected into the attention and FFN projections. Classification is performed not by free-form generation but by first-token logit probing over the YES/NO token IDs at the assistant-response position. To counteract the soft-label collapse observed in the 26B zero-shot run, we layer two calibration mechanisms: label smoothing during training, and post-hoc temperature scaling applied to logits at inference.

5.6. XLM-RoBERTa-large Text-Only Models

In addition to the base architectures, we trained an XLM-RoBERTa-large text-only baseline across all four subtasks. This run family relied exclusively on the official text column for both memes and videos and was initialized with a random seed of 42. For the binary identification subtasks (Task 2.1 and Task 3.1), the models were optimized using KL divergence [5]. On the internal development split, the model reached a KL divergence [5] of 0.2378 for memes and 0.4757 for videos. For the multi-label categorization subtasks (Task 2.3 and Task 3.3), the models were trained using Binary Cross-Entropy

(BCE) loss. Unlike the calibrated runs that tuned decision boundaries, the hard predictions for this large baseline retained the default decision threshold of 0.5 across all categories. On the development set, this setup achieved a hard macro F1 of 0.1799 for Subtask 2.3 and 0.3099 for Subtask 3.3, alongside soft macro Mean Absolute Error (MAE) scores of 0.1498 and 0.1582, respectively.

6. Experiments and Internal Validation

Our experiments were conducted as a set of parallel system variants implemented in PyTorch [13] and the Hugging Face Transformers library [14], rather than as a single shared model. Different team members explored different modeling choices, including XLM-RoBERTa-base text-only models, calibrated text-only variants, XLM-RoBERTa-large text-only models, caption-augmented meme models, and an XLM-RoBERTa-large + CLIP multimodal model. This design allowed us to compare conservative text-only systems with larger or multimodal variants under the same task setting.

For internal validation, we created our own development split from the official training data, since the official test labels were not available during system development. For meme subtasks, the official training set contained 3,984 instances, which were split using a 0.1 development ratio; exact counts vary slightly by subtask due to stratification and are reported in Table 3. For video subtasks, the official training set contained 2,524 instances, which were split using a 0.1 development ratio; exact counts vary slightly by subtask due to stratification and are reported in Table 3.

Each model family was developed and selected by its corresponding contributor using the internal development split. Depending on the model, this involved choosing the best checkpoint, input format, random seed, decision threshold, or probability calibration parameters. For hard outputs, we treated ICM as the main development metric when threshold selection was needed. For soft outputs, we used ICMSoft-oriented calibration when probability calibration was applied. Because the submitted runs came from different model families, we did not force all systems to use the same architecture or the same selection procedure.

Before submission, candidate outputs from the different model families were converted into the official PyEvALL [9, 10] JSON format and checked for label validity. For multi-label hard outputs, empty predictions were replaced with NO, and predictions containing concrete sexism categories were kept mutually exclusive with NO. For multi-label soft outputs, the NO probability was included so that the output matched the expected label inventory.

For the final paper, we report the submitted variants as a team-level comparison. When discussing final test performance, we compare all submitted runs for each subtask and evaluation context, and highlight the strongest submitted result among them. This is important because the best run can differ across subtasks and between hard-hard and soft-soft evaluation.

7. Results and Analysis

Overall, our official results show a clear difference between hard-label and soft-label behavior. The strongest positions were achieved in soft-soft multi-label categorization, where probability calibration and category-level confidence scores are rewarded for matching annotator disagreement rather than only discrete majority labels.

Table 4 summarizes our best run per subtask and evaluation setting. Three points stand out. First, our strongest positions are all in soft-soft categorization: rank 6 on both Subtask 2.3 and Subtask 3.3 (and rank 4 on the English-only Subtask 3.3 leaderboard), where calibrated category probabilities are rewarded for matching annotator disagreement. Second, hard-label and binary identification are consistently mid-pack (ranks 22–33), indicating that converting probabilities into discrete labels—especially for rare sexism categories—is the main bottleneck. Third, added model capacity did not consistently help: the XLM-RoBERTa-large+CLIP multimodal run did not outperform the calibrated XLM-RoBERTa-base text-only models, and on Subtask 2.1 it was our weakest submitted run. The internal Gemma 4 configurations showed a similar difficulty with soft-label calibration (Section 7.4).

Table 4
Best tai6le run per subtask on the official all-language leaderboard. For each row, ICM-Norm is ICM-Soft Norm on soft rows and ICM-Hard Norm on hard rows; rank on the all-language leaderboard is shown in parentheses.

Subtask	Eval	Best run	Model	ICM-Norm	F1 / CE
2.1	soft	`tai6le_1`	XLM-R-base text-only	0.4525 (r32)	CE 0.9275
2.1	hard	`tai6le_3`	XLM-R-base, thresh.-calibrated	0.5919 (r24)	F1-YES 0.7279
2.3	soft	`tai6le_2`	XLM-R-base, prob.-calibrated	0.2699 (r6)	—
2.3	hard	`tai6le_1`	XLM-R-base, thresh.-calibrated	0.3255 (r27)	F1 0.4069
3.1	soft	`tai6le_2`	XLM-R-base text-only	0.5188 (r19)	CE 1.1215
3.1	hard	`tai6le_3`	XLM-R-base text-only	0.5737 (r33)	F1-YES 0.6679
3.3	soft	`tai6le_3`	XLM-R-base text-only	0.1583 (r6)	—
3.3	hard	`tai6le_2`	XLM-R-large text-only	0.3825 (r22)	F1 0.2656

Table 5
Selected XLM-RoBERTa-base runs. (internal development split).

Run type	Task	Input	Development result
Text-only baseline	2.1	OCR text	Macro-F1 = 0.7134; KL = 0.1979
Text-only baseline	3.1	video text	Macro-F1 = 0.7297; KL = 0.4436
Selected formal run	2.1	OCR + visual caption	Best seed MCC = 0.3313; Macro-F1 = 0.6618
Selected formal run	2.3	OCR + visual caption	Best seed Macro-F1 range = 0.2080-0.2272
Selected formal run	3.1	video text	Best seed MCC range = 0.3972-0.4763
Selected formal run	3.3	video text	Best seed Macro-F1 range = 0.2488-0.3084

These results support two interpretations. First, the textual fields and transcripts contain enough information for competitive disagreement-aware probability estimates in categorization tasks. Second, converting those probabilities into discrete hard labels is difficult, especially when minority sexism categories are rare and when meme or video meaning depends on visual, audio, or contextual cues absent from the text field.

7.1. Selected XLM-RoBERTa-base Runs

Text-only models handle instances with explicit lexical cues reliably, but tend to over-flag gender-related content that is not actually sexist. The development figures in this subsection come from different team members' pipelines, each with its own split and evaluation script, and are not directly comparable across rows or with the calibrated variant in Section 7.2.

For the text-only baseline, overt hostile wording (e.g. the OCR text "CHUPENME LA PIJA!!!") is easy to flag, whereas politically or socially framed statements that merely mention women and pregnancy trigger false positives. Adding visual captions gives at most a modest and task-dependent benefit on memes: short or socially sensitive OCR such as "ABUSO" is still mislabeled, and fine-grained categorization remains limited by sparse textual evidence. The video subtasks are more stable because the official transcripts carry more explicit context, though noisy transcripts still hurt minority categories such as MISOGYNY AND NON-SEXUAL VIOLENCE. The calibrated text-only variant separates hard-label thresholding from soft-label probability calibration, so its contribution is output finalization rather than a new backbone architecture.

7.2. XLM-RoBERTa-base Text-Only Fine-Tuning and Calibration

Hard thresholding and soft calibration are two separate decisions, and treating them separately is what drives the gap between our soft and hard results. On top of the text-only model we did not change the backbone; we only tuned how its probabilities are converted into the two official submission formats.

Table 6
Selected hard-label thresholds

Subtask	Selected hard-label threshold(s)
2.1	YES threshold = 0.52
2.3	[0.5, 0.4, 0.19, 0.28, 0.5, 0.5]
3.1	YES threshold = 0.61
3.3	[0.5, 0.5, 0.32, 0.5, 0.15, 0.5]

Table 7
Calibration parameters for soft probability submissions

Subtask	Calibration method	Selected parameters
2.1	binary logit calibration	temperature = 0.5, bias = -0.25
2.3	multi-label scaling	scales = [1.15, 1.0, 1.5, 1.3, 1.0, 1.3]
3.1	binary logit calibration	temperature = 0.5, bias = -1.0
3.3	multi-label scaling	scales = [2.0, 1.0, 2.0, 0.25, 0.25, 0.25]

Table 8
Internal development hard-hard results of the calibrated text-only variant

Subtask	ICM	ICMNorm	FMeasure / Macro F1	Accuracy
2.1	-0.122235	0.436985	0.383867	0.604790
2.3	-0.733666	0.347671	0.373989	N/A
3.1	0.133815	0.567017	0.688514	0.690476
3.3	-0.226480	0.431380	0.361393	N/A

Table 9
Internal development soft-soft results after ICMSoft calibration

Subtask	ICMSoft	ICMSoftNorm	CrossEntropy
2.1	-0.933458	0.353556	0.968475
2.3	-4.878031	0.250633	2.156901
3.1	0.150383	0.526535	1.097378
3.3	-6.889815	0.145921	1.811297

For hard submissions, thresholds were selected on the internal development split using ICM: a single YES threshold for binary subtasks, and label-specific thresholds for multi-label subtasks, with empty predictions mapped to NO (Table 6). For soft submissions, probabilities were calibrated separately against ICMSoft—logit calibration for binary subtasks and label-wise scaling for multi-label subtasks (Table 7). For multi-label subtasks, the threshold and scaling arrays follow the label-vector order described in Section 4, including NO as the first entry.

The two settings reward different things: hard-hard needs a good discrete decision boundary, soft-soft needs probabilities that track annotator disagreement. Decoding choices matter most for the multi-label subtasks, where the system must separate NO from several concrete categories, including low-frequency ones such as SEXUAL VIOLENCE and MISOGYNY AND NON-SEXUAL VIOLENCE.

Table 8 reports the internal development hard-hard results of the calibrated text-only variant after threshold tuning.

Table 9 reports the internal development soft-soft results after ICMSoft calibration.

7.3. XLM-RoBERTa-large + CLIP multimodal (Subtask 2.1)

Late-fusion with CLIP did not improve performance on the official test set and was our weakest Subtask 2.1 submission. The numbers in this subsection are measured on our internal development split; this configuration was submitted only for Subtask 2.1 (`tai61e_2`), combining XLM-RoBERTa-Large text features with CLIP-ViT-Base-Patch32 visual features. On development it reaches a cross-entropy of 0.9598, comparable to the text-only model rather than clearly lower, so the visual features added no distinct soft-label advantage.

The more informative finding is the gap between probabilistic behavior and discrete decisions. Under the default threshold $\tau = 0.50$, the system behaves as an aggressive detector on dev: YES-class recall reaches 81.59%, but NO-class recall drops to 47.37%; global accuracy is 67.96%, yet the hard ICM is slightly negative at -0.04997. Because ICM penalizes "uninformative or safe" over-predictions, falling back on the majority class under high-variance annotations yields little information contrast relative to a naive baseline, even at a respectable raw accuracy.

A threshold grid sweep from $\tau = 0.40$ to 0.55 on the development set shows that the default $\tau = 0.50$ was strongly penalized, whereas $\tau = 0.53$ would have maximized both Macro-F1 (0.6474) and ICM. We report this as a development-set observation about boundary sensitivity, not as a tuned property of the final submission.

On the official test set, `tai61e_2` ranked 114th in hard-hard (ICM-Hard -0.0237, ICM-Hard Norm 0.4879, F1-YES 0.6966) and 113th in soft-soft (ICM-Soft -0.8586, ICM-Soft Norm 0.3620, Cross Entropy 0.9470), well below our text-only runs (`tai61e_3`, rank 24 hard; `tai61e_1`, rank 32 soft). The development behavior did not transfer to unseen test data, and the simpler text-only models generalized better under the official metric.

7.4. Gemma 4 MLLM Configurations

Gemma 4 produced over-confident soft distributions that calibration could soften but not fix, so we did not submit either configuration. Both were evaluated only on the internal development split.

The zero-shot 26B-A4B-it baseline (first-token YES/NO probing, $T = 2.0$) reaches a hard accuracy of 76.8% on Subtask 3.1 and 71.6% on Subtask 2.1, but its soft quality is poor: mean per-instance KL against the annotator-consensus gold is 2.916 on 2.1 and 1.677 on 3.1. The logits routinely collapse to near-0/1 on items whose gold consensus is close to {0.5, 0.5}, so a single temperature can soften the family but cannot recover the bimodal disagreement structure. The LoRA-tuned E4B stage targets exactly this: on Subtask 3.1 a post-hoc temperature fit ($T \approx 3.68$) cuts mean KL from 1.677 to 0.463 (a 72% reduction), at the cost of hard accuracy dropping from 76.8% to 65.6% (best checkpoint 67.2%), reflecting that the E4B base is a materially weaker discriminator than the 26B-A4B MoE even after seven-projection LoRA adaptation. We retain this line as a generative baseline for future work where soft-distribution fidelity is the primary target.

8. Conclusion and Future Work

In this paper, we described our participation in EXIST 2026, covering four subtasks: sexism identification in memes (2.1), sexism categorization in memes (2.3), sexism identification in videos (3.1), and sexism categorization in videos (3.3). We submitted both hard and soft predictions for each addressed subtask. Our systems centered on two complementary backbones: an XLM-RoBERTa-base pipeline using text-only and caption-augmented inputs together with threshold and probability calibration for official hard and soft outputs, and an XLM-RoBERTa-large + CLIP late-fusion architecture for binary meme identification.

Our analysis surfaced three consistent patterns. First, text-based models handle instances with explicit lexical cues reliably but tend toward false positives on gender-related yet non-sexist content, particularly political or social commentary that mentions women. Second, adding visual captions provides at most a modest and task-dependent benefit on memes and does not resolve cases where the

OCR is short, ambiguous, or socially sensitive without being overtly sexist; fine-grained categorization in particular remains limited by sparse textual evidence. Third, the video subtasks were comparatively more stable because the official transcripts often contain explicit contextual cues, although noisy or fragmented transcripts continue to hurt minority categories such as MISOGYNY AND NON-SEXUAL VIOLENCE. The XLM-RoBERTa-large + CLIP development results further showed that the gap between probabilistic alignment (low cross-entropy of 0.9598) and discrete decision quality (a slightly negative ICM under the default threshold) is sensitive to threshold choice, with $\tau = 0.53$ yielding the empirical peak in both Macro F1 and ICM within that architecture on the development split.

Several directions remain open. First, we did not participate in the source-intention subtasks (2.2 and 3.2); extending our pipelines to the hierarchical DIRECT / JUDGEMENTAL distinction is a natural next step, especially under the hierarchy-aware ICM evaluation. Second, we did not exploit the physiological signals (eye tracking, heart rate, and EEG) released as part of the EXIST 2026 human-centered framework; incorporating these features as auxiliary inputs or as priors over annotator disagreement is a promising direction for the soft-label setting, where the disagreement signal itself is the prediction target. Third, stronger vision-language encoders, multimodal pretraining on meme-style image-text composites, and per-category (rather than a single global) threshold calibration are likely to address the recall imbalance on minority sexism categories. Finally, since both languages and both modalities exhibit different failure modes, modality- and language-conditional calibration may yield further gains under the ICM-soft objective.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] 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, Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[2] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, SemEval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 338–347. URL: https://aclanthology.org/2021.semeval-1.41/. doi:10.18653/v1/2021.semeval-1.41.

[3] 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, volume 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[4] Google DeepMind, Gemma 4 model card, 2026. URL: https://ai.google.dev/gemma/docs/core/model_card_4, accessed: June 2026.

[5] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html.

[6] 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.

[7] 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.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: https://jmlr.org/papers/v12/pedregosa11a.html.

[9] J. Carrillo-de Albornoz, PyEvALL: The python library to evaluate all, https://pypi.org/project/PyEvALL/, 2025. Python evaluation library for classification, ranking, and Learning with Disagreement evaluation contexts.

[10] E. Amigo, J. Carrillo-de Albornoz, M. Almagro-Cadiz, J. Gonzalo, J. Rodriguez-Vidal, F. Verdejo, EvALL: Open access evaluation for information access systems, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Association for Computing Machinery, 2017, pp. 1301–1304. URL: https://doi.org/10.1145/3077136.3084145. doi:10.1145/3077136.3084145.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.

[12] 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, arXiv preprint arXiv:2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685.

[13] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems, volume 32, 2019, pp. 8024–8035. URL: https://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.

[14] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/. doi:10.18653/v1/2020.emnlp-demos.6.