FourBytes at EXIST 2026: Cross-Attention Multimodal Fusion for Sexism Detection in Memes

Notebook for the EXIST Lab at CLEF 2026

Sujaudeen Nannai John[1], Thirumurugan Ravialagappan[2], Shrika Thota[3], Vishal Muralidharan[4] and Shanmugam Karthikeyen Shivaanee[5]

[1]*Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai 603110, Tamil Nadu, India*

Abstract

We describe our submission to EXIST 2026 Task 2.1, where systems must detect sexism in multilingual multimodal memes under both hard and soft evaluation settings. Our model combines XLM-T text representations with SigLIP image features through a cross-attention dual-encoder architecture and is trained on soft labels derived from annotator votes. The submitted run reaches an ICM score of -0.0188 (ranking 110th) on the Hard-hard evaluation over all instances, and an ICM score of -0.4376 (ranking 56th) on the Soft-soft evaluation leaderboard, with better results on English memes than on Spanish ones.

Keywords

multimodal sexism detection, meme classification, cross-attention fusion, multilingual transformers, EXIST 2026

1. Introduction

The EXIST lab at CLEF 2026 studies automated sexism detection in social media under multilingual and multimodal conditions. For the 2026 edition, the organizers extend the meme benchmark with physiological measurements in addition to the usual image, text, and label annotations [1, 2]. That makes the dataset useful for multimodal classification and for studying how people respond to potentially sexist content. Task 2.1 asks participants to decide whether a meme is sexist, both as a hard label and as a soft distribution over YES and NO.

This is not an easy binary classification setting. Meme text is short, noisy, and sometimes code-switched between English and Spanish. The image matters too, since the sexist meaning often comes from the interaction between caption and visual scene rather than from either modality by itself. On top of that, the soft evaluation rewards systems that produce usable probabilities instead of only good hard decisions. In practice, that calls for a multilingual text encoder, a strong visual backbone, and a fusion method that lets the two modalities interact at a finer level than simple pooled concatenation.

Our submission addresses that setting with a dual-encoder architecture in which the text stream attends to the visual stream through multi-head cross-attention. We use XLM-T multilingual tweet representations on the textual side [3], built on the XLM-R family [4], and SigLIP on the visual side [5]. The fused representation is pooled and passed to a binary classifier trained on soft labels from annotator votes. The implementation also includes an auxiliary head for the physiological variables released with the dataset, although the submitted run sets that auxiliary loss weight to zero.

Source code is available at: https://github.com/thirumuruganra/EXIST2026

The rest of the paper reviews related work, summarizes the Task 2.1 data interface, describes the model and training setup, and then reports the official results.

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

sujaudeenn@ssn.edu.in (S. N. John); thirumurugan2310277@ssn.edu.in (T. Ravialagappan); shrika2310548@ssn.edu.in (S. Thota); vishal2310253@ssn.edu.in (V. Muralidharan); shivaanee2310257@ssn.edu.in (S. K. Shivaanee)

2. Related Work

Research on harmful-content detection first focused on text alone and later moved toward multimodal meme understanding. The Hateful Memes Challenge made that shift explicit by showing how quickly strong text-only or image-only models fail when the dataset is built to block unimodal shortcuts [6]. SemEval-2022 Task 5 (MAMI) brought a similar setup to misogyny detection and treated meme understanding as both a binary problem and a fine-grained classification problem [7].

The EXIST campaigns push this line of work in a multilingual and more human-centered direction. EXIST 2023 introduced learning with annotator disagreement for multilingual tweets [8]. EXIST 2024 extended the benchmark from tweets to memes [9], and EXIST 2025 added TikTok videos [10]. In 2026, the organizers add physiological measurements on top of the meme and video benchmarks, while our submission focuses on Task 2.1 for memes [1, 2]. We follow the shared-task setup by training directly on soft labels and by keeping the implementation compatible with the released sensorial features.

Earlier multimodal baselines often relied on late fusion over pooled text and image embeddings. More recent systems usually work with token-level interaction instead. Our model follows that direction by combining multilingual XLM-R/XLM-T representations for noisy English and Spanish social-media text [4, 3] with SigLIP visual features [5] and joining them through cross-attention. That choice fits sexist memes reasonably well, since the label often depends on how a short caption changes the meaning of the image, or vice versa.

3. Task and Data

The proposed system uses the official JSON files released for the EXIST 2026 Memes Dataset [1, 2]. The lab guidelines describe a collection of more than 5,000 labeled memes in English and Spanish, with 3,984 training instances and 1,053 test instances. The language split is balanced, and the organizers note that a small number of duplicates carried over from the previous edition were removed in the 2026 release. Task 2.1 is a binary classification problem in which each meme must be labeled YES or NO depending on whether it contains sexist expressions, describes sexist behaviour, or criticizes such behaviour.

Each meme is represented by a JSON object that combines content fields and annotation metadata. The main content attributes used by our pipeline are the unique identifier `id_EXIST`, the language tag `lang`, the automatically extracted meme text `text`, the image filename `meme`, and the corresponding path `path_memes`. The released files also provide rich annotation metadata, including the number of annotators and demographic information such as gender, age group, ethnicity, study level, and country. For the three meme subtasks, the dataset contains `labels_task2_1`, `labels_task2_2`, and `labels_task2_3`, together with a `split` field identifying whether the instance belongs to the English or Spanish portion of the train or test partition.

For Task 2.1 specifically, the gold supervision is given as one label per annotator in `labels_task2_1`. In our implementation, these annotations are converted into a soft target by mapping YES to 1.0 and NO to 0.0 and then averaging over annotators. This gives a single probabilistic target per meme while preserving disagreement information, which is important because the shared task accepts both hard outputs and soft distributions. The guidelines also clarify that the higher-level subtasks use the placeholder label "-" for non-sexist instances and UNKNOWN for missing judgments, and that UNKNOWN is not a valid prediction label during evaluation. For the test partition, task labels are withheld, so systems must generate submission files directly from the provided meme identifiers and content.

The dataset interface further includes a `sensorial` block that stores physiological measurements collected under the Human-Centered AI setup of EXIST 2026. These signals are organized by subject and modality through `sensorial["users"]` and nested dictionaries for eye tracking (ET), heart rate (HR), and electroencephalography (EEG). The guidelines state that each stimulus was viewed by 2–4 subjects, and that whenever a modality is available for a meme it includes recordings from at least two subjects. Rather than raw time series, the release provides compact summary descriptors computed over the viewing interval from stimulus onset to user response. Concretely, ET contributes 24 variables

covering pupil diameter, blinks, fixations, saccades, and reaction time; HR contributes 4 summary statistics (mean, standard deviation, minimum, and maximum heart rate); and EEG contributes 80 bandpower features derived from 16 channels across the Delta, Theta, Alpha, Beta, and Gamma bands.

Our training code aggregates the available sensorial variables by averaging each feature over users and modalities, which yields one sensor vector per meme together with a validity mask. Missing variables are ignored when computing the auxiliary regression loss, so the setup still works when sensor coverage is incomplete. For validation, we do not use a purely random split. We stratify the training set by language and by annotator agreement, where agreement is bucketed according to whether the soft label is mostly negative, mixed, or mostly positive. In practice, this split gave us a more useful validation slice for threshold tuning and checkpoint selection.

4. System Architecture

4.1. Text and Vision Encoders

The textual backbone is `cardiffnlp/twitter-xlm-roberta-base`, corresponding to the XLM-T line of multilingual models for Twitter text [3]. We selected it because the task combines multilingual content with short, noisy social media text and still benefits from XLM-R pretraining [4]. The implementation keeps the full embedding layer trainable and unfreezes the top twelve transformer layers, which allows the encoder to adapt to the task while preserving the benefits of large-scale multilingual pretraining.

The visual backbone is the vision tower extracted from `google/siglip-base-patch16-256` [5]. This encoder produces a sequence of patch-level hidden states instead of a single pooled vector, which is important because the downstream fusion layer operates directly on token and patch sequences. When the hidden dimensionalities of the text and vision streams differ, a linear projection aligns the visual sequence with the textual hidden size before fusion.

4.2. Cross-Attention Fusion

The core modeling decision is to fuse the two modalities through cross-attention instead of through pooled concatenation. Let $T \in R^{n \times d}$ denote the sequence of textual hidden states and $V \in R^{m \times d}$ the projected sequence of visual hidden states. After layer normalization and dropout, the model applies multi-head attention with text tokens as queries and image patches as keys and values. This lets each textual token gather information from the parts of the meme image that are most relevant to its semantics.

The attention output is added back to the textual sequence through a residual connection, followed by a feed-forward block with its own residual pathway. We then apply masked mean pooling over the updated text sequence using the attention mask from the tokenizer. The resulting fused vector is normalized, passed through a GELU activation, and sent to the main classifier head, which outputs a single logit for the YES class.

We chose this fusion block because it gives token-level text–image interaction without adding a large multimodal stack on top of the encoders. That matters for memes, where the same image can lead to different labels once the overlaid text changes. Figure 1 summarizes the training and inference pipeline used in our submission.

4.3. Auxiliary Sensor Head

In addition to the primary classifier, the model includes an auxiliary head that predicts the aggregated physiological feature vector. The head is a small multilayer perceptron with layer normalization, GELU, dropout, and a final linear layer. During training, the corresponding loss is computed as a masked mean squared error so that only available sensor variables contribute.

Figure 1: End-to-end methodology of our Task 2.1 system. The model encodes meme text and image separately, fuses them through cross-attention, optionally predicts aggregated sensorial variables with an auxiliary head, and produces both hard and soft submissions after validation-based threshold selection.

Although this mechanism is implemented in the model, the final submission disables the auxiliary objective by setting the loss weight α to 0.0. We still describe it here because it is part of the actual system code and because it remains an obvious extension for future experiments with the extra signals released in EXIST 2026.

5. Experimental Setup

Table 1
Main training configuration used by FourBytes_1.

Component	Setting
Text encoder	twitter-xlm-roberta-base
Vision encoder	siglip-base-patch16-256
Batch size	16
Max text length	128
Epochs	5
Learning rates	1e-5 / 5e-6 / 5e-5
Warmup ratio	0.1
Weight decay	0.01
Early stopping patience	2
Threshold grid	0.30 to 0.70
Auxiliary loss weight	0.0

5.1. Optimization and Regularization

The training pipeline uses Hugging Face Accelerate with mixed-precision computation. Both the text and vision encoders enable gradient checkpointing, which reduces memory pressure and makes the dual-encoder configuration easier to train on commodity GPUs. The optimizer is AdamW with differential learning rates: 1×10^{-5} for the text encoder, 5×10^{-6} for the vision encoder, and 5×10^{-5} for the task-specific heads. We train for up to five epochs with a warmup ratio of 0.1, dropout of 0.3 in the auxiliary head, weight decay of 0.01, and early stopping patience of two epochs.

The primary objective is `BCEWithLogitsLoss` computed against the soft label for each meme. When enabled, the auxiliary loss is a masked mean squared error between predicted and aggregated physiological features. The total loss is the sum of the primary loss and α times the auxiliary loss. In the final configuration, $\alpha = 0.0$, so model selection depends exclusively on the primary task.

5.2. Inference and Submission Formatting

Validation uses soft binary cross-entropy as the checkpointing criterion and performs a grid search over thresholds from 0.30 to 0.70 in order to maximize hard-label F1 on the validation subset. At test time, the model outputs probabilities for the YES class, and the final hard submission uses the selected threshold instead of a fixed 0.5 decision boundary.

For the soft submission, the implementation applies a piecewise linear rescaling that maps the chosen hard threshold to 0.5. The motivation is to preserve the decision boundary while still emitting a symmetric YES/NO distribution in the JSON format expected by PyEvALL. The official evaluation scripts used in our workflow compute ICM, normalized ICM, F1 for the Hard-hard evaluation, and Soft-soft ICM, normalized Soft-soft ICM, and cross-entropy for the Soft-soft evaluation.

6. Results and Discussion

Table 2 reports the official leaderboard results for the submitted run FourBytes_1. The system obtains an overall Hard-hard F1 of 0.7323 and ranks 110 on the global Hard-hard evaluation. On the Soft-soft evaluation, the rank improves to 56, although the Soft-soft ICM score remains negative.

The English results are consistently stronger than the Spanish results. Even with a multilingual encoder, the model seems to adapt less well to the Spanish portion of the task. There are several plausible reasons for that gap, including lexical variation in captions, lower effective coverage of Spanish training examples, or differences in how sexist content is expressed across the two languages. We did not run

Table 2
Official leaderboard results for FourBytes_1 on EXIST 2026 Task 2.1.

Split	Hard-hard Rank	Hard-hard ICM	Hard-hard F1 YES	Soft-soft Rank	Soft-soft ICM	Soft-soft Cross Entropy
ALL	110	-0.0188	0.7323	56	-0.4376	0.9524
EN	68	0.1357	0.7633	47	-0.2789	0.9112
ES	151	-0.1732	0.7050	61	-0.6107	0.9915

controlled language-specific ablations, so we leave these as working explanations rather than firm claims.

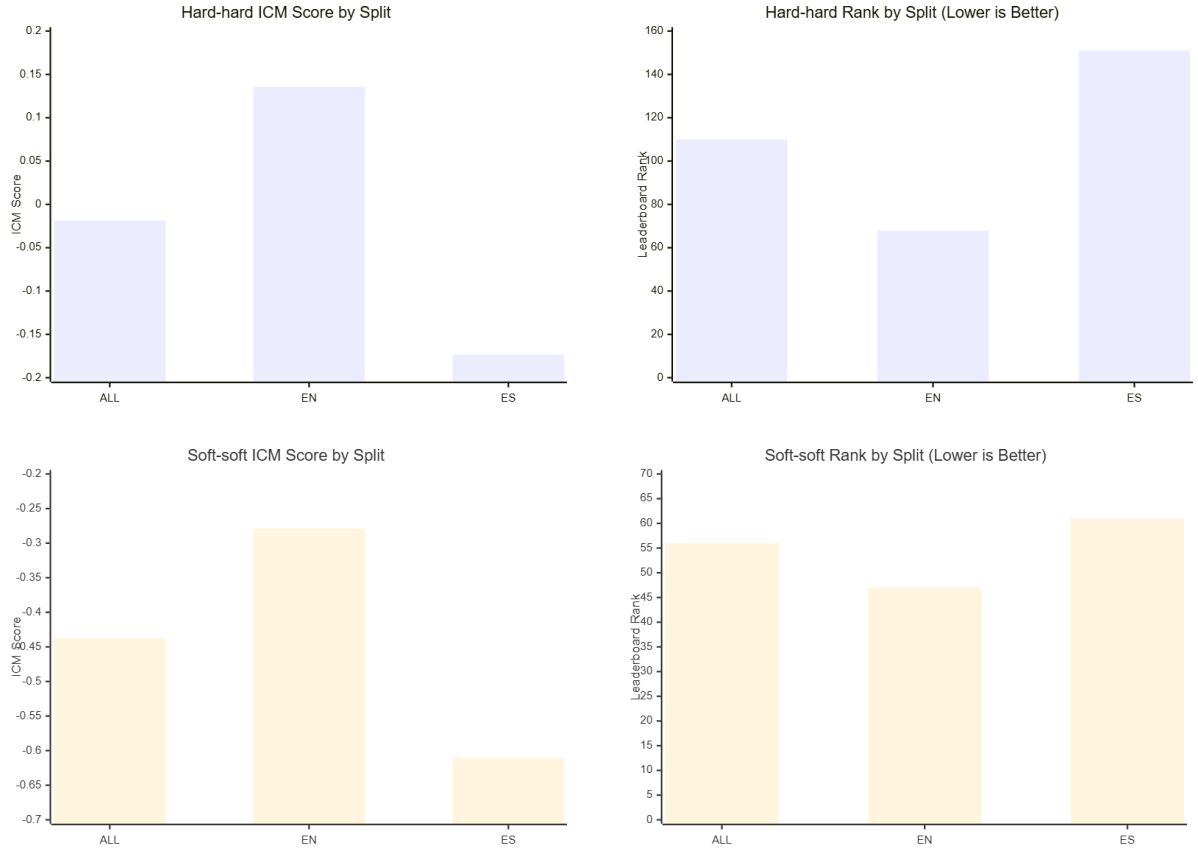

Figure 2: Leaderboard trends for FourBytes_1 across the ALL, EN, and ES splits. The left column shows the official ICM scores for the Hard-hard (top) and Soft-soft (bottom) evaluations, while the right column shows the corresponding leaderboard ranks. English is consistently the strongest split, while Spanish remains the most difficult condition.

A separate issue is calibration. The Soft-soft evaluation rank is notably better than the Hard-hard evaluation rank, yet the negative Soft-soft ICM values and relatively high cross-entropy show that the predicted distributions are still not well calibrated. That matches the way the system was trained: it is optimized mainly for classification quality and only adjusted afterward through threshold-based rescaling. A stronger calibration step, such as temperature scaling or direct optimization of a distributional objective, would probably help here.

Figure 2 shows these trends more directly, demonstrating that while the ranking gaps are large, the absolute performance distributions across splits overlap significantly.

Even so, the run is not without useful signals. The cross-attention layer gives a clear way for textual and visual evidence to interact, and the validation split takes both language and annotator agreement into account. The implementation also leaves room for future use of the physiological side information released by the organizers. In the submitted configuration, however, that branch is inactive, so we

cannot claim any benefit from those variables yet.

7. Conclusion

We presented our cross-attention dual-encoder system for EXIST 2026 Task 2.1. The model combines a multilingual text encoder, a SigLIP vision encoder, and a fusion block that lets text representations attend to image patches before pooled classification. The official results show that this setup is competitive enough to serve as a useful baseline, especially on the Soft-soft evaluation, but they also make the main weaknesses clear: the model is less robust on Spanish memes and its output probabilities are not calibrated well enough.

The next steps follow directly from those weaknesses. We want to test calibration methods that target the Soft-soft evaluation more explicitly, examine language-aware fine-tuning for the English–Spanish gap, and turn the auxiliary physiological branch back on to see whether it provides useful supervision for multimodal representations.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools in the preparation of this work.

References

[1] L. Plaza, J. C. de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social Media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), 2026.

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

[3] 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, European Language Resources Association, Marseille, France, 2022, pp. 258–266. URL: https://aclanthology.org/2022.lrec-1.27.

[4] 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: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.

[5] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid Loss for Language Image Pre-Training, 2023. URL: https://arxiv.org/abs/2303.15343. arXiv:2303.15343.

[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, 2020. URL: https://arxiv.org/abs/2005.04790. arXiv:2005.04790.

[7] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, SemEval-2022 Task 5: Multimedia Automatic Misogyny Identification, in: G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, S. Ratan (Eds.), Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Association for Computational Linguistics, Seattle, United States, 2022, pp. 533–549. URL: https://aclanthology.org/2022.semeval-1.74/. doi:10.18653/v1/2022.semeval-1.74.

[8] L. Plaza, J. C. de Albornoz, R. Morante, E. Amigó, J. Gonzalo, D. Spina, P. Rosso, Overview of EXIST 2023: Learning with Disagreement for Sexism Identification and Characterization, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2023, volume 14163 of *Lecture Notes in Computer Science*, Springer, 2023. URL: https://link.springer.com/chapter/10.1007/978-3-031-42448-9_23. doi:10.1007/978-3-031-42448-9_23.

[9] L. Plaza, J. C. de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, D. Spina, Overview of EXIST 2024: Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2024, volume 14662 of *Lecture Notes in Computer Science*, Springer, 2024. URL: https://link.springer.com/chapter/10.1007/978-3-031-71908-0_5. doi:10.1007/978-3-031-71908-0_5.

[10] L. Plaza, J. C. de Albornoz, I. Arcos, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2025: Learning with Disagreement for Sexism Identification and Characterization in Tweets, Memes, and TikTok Videos, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2025, volume 16089 of *Lecture Notes in Computer Science*, Springer, 2025. URL: https://link.springer.com/chapter/10.1007/978-3-032-04354-2_16. doi:10.1007/978-3-032-04354-2_16.