Bicocca_Unimib at EXIST 2026: Human-Centered Multimodal Sexism Detection in Memes with Physiological Signals

EXIST at CLEF 2026

Mattia Borgia*, Marco Viviani

Department of Informatics, Systems, and Communication (DISCo), University of Milano-Bicocca, Edificio U14 (ABACUS), Viale Sarca, 336 – 20126 Milan, Italy

Abstract

This paper describes the system submitted by the Bicocca_Unimib team to the EXIST 2026 shared task on sexism detection in memes, addressing Subtasks 2.1 (Sexism Identification), 2.2 (Source Intention), and 2.3 (Sexism Categorization). We propose a human-centered multimodal framework that extends the standard text-image paradigm by incorporating physiological signals collected from human observers, namely eye-tracking (ET), electroencephalography (EEG), and heart rate (HR), together with annotator demographic metadata. We develop and compare two neural architectures: a Late Fusion baseline, which concatenates representations from all modalities, and a Cross-Attention model, where physiological signals query the joint text-image representation to selectively emphasize relevant dimensions. Both models are trained under the Learning with Disagreements (LeWiDi) paradigm using soft labels, automatic multi-task loss weighting, 5-fold cross-validation, and early stopping. Official results show that the Late Fusion fold ensemble is the most consistent submission across subtasks, ranking highest among our runs for Sexism Identification and Source Intention, while the full-data refit performs best for Sexism Categorization. Overall, our findings suggest that physiological and subject demographic information can provide complementary signals for multimodal sexism detection, although their effectiveness remains constrained by data sparsity, class imbalance, and fold-specific variability.

Keywords

Sexism Detection, Memes, Multimodal Learning, Physiological Signals, Learning with Disagreements

1. Introduction

The *automatic detection of sexist content* in online multimedia is a challenging task at the intersection of Natural Language Processing (NLP), Computer Vision, and Social Computing. Memes, in particular, combine visual and textual elements in culturally grounded ways that require contextual knowledge to interpret correctly [4]. In this context, the EXIST 2026 shared task [1, 2] extends the long-running EXIST benchmark series by focusing exclusively on multimedia formats, including memes and videos, and by introducing a novel *human-centered* component. Selected subjects were exposed to each meme while their physiological and behavioral responses were continuously recorded. These signals, including *eye-tracking* (ET), *electroencephalography* (EEG), and *heart rate* (HR), are provided as additional features alongside the standard annotations, opening the possibility of leveraging human reactions as complementary information for automatic classifiers.

In this paper, we describe the system developed by the *Bicocca_Unimib* team for the three meme subtasks of EXIST 2026: Subtask 2.1, *Sexism Identification*; Subtask 2.2, *Source Intention*; and Subtask 2.3, *Sexism Categorization*. Our main contributions are:

- A multimodal pipeline integrating text (XLM-RoBERTa), image (CLIP), physiological signals (ET, EEG, HR aggregated per meme), and annotator demographic metadata;

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

*Corresponding author.

m.borgia4@campus.unimib.it (M. Borgia); marco.viviani@unimib.it (M. Viviani)

https://ikr3.disco.unimib.it/people/marco-viviani/ (M. Viviani)

0000-0002-2274-9050 (M. Viviani)

- A Cross-Attention mechanism in which physiological signals act as queries over the joint text-image representation;
- Training under the Learning with Disagreements (LeWiDi) paradigm [3] with soft labels and automatic task-loss weighting.

The remainder of the paper is organised as follows. Section 2 reviews related work, while Section 3 introduces the task and dataset. Section 4 describes the proposed methodology, and Section 5 reports the experimental results. Finally, Section 6 discusses the main findings and limitations, and Section 7 concludes the paper.

2. Related Work

This section situates our work within three relevant lines of research: the evolution of the EXIST shared tasks, recent approaches to multimodal meme understanding, and learning paradigms that explicitly account for annotator disagreement.

EXIST Shared Tasks. The EXIST lab was established in 2021 as the first multilingual benchmark for sexism detection, covering Twitter and Gab in English and Spanish. The 2023 edition introduced the source intention dimension and adopted the LeWiDi framework for handling annotator disagreement. The 2025 edition extended the task to memes, providing a large-scale benchmark for multimodal sexism detection. EXIST 2026 builds on these previous editions by adding physiological signals from human observers as a novel modality.

Multimodal Meme Understanding. Research on hateful and sexist meme detection has consistently shown that text-only or image-only models are insufficient, and that the way in which modalities are fused plays a central role [4]. Approaches combining CLIP [5] with domain-adapted text encoders, such as XLM-RoBERTa [6], have achieved strong results on sexism detection benchmarks. More recent work has explored graph-based inter-meme reasoning to capture batch-level relational context, as well as LLM-based captioning to supplement OCR text with semantic image descriptions [11]. Closely related to our work, [16] propose a human-centered multimodal approach combining eye-tracking, heart rate, and EEG signals for sexism detection in memes. Our work follows this multimodal line of research, but focuses on a systematic comparison of fusion architectures for integrating physiological signals within the EXIST 2026 benchmark. To the best of our knowledge, such signals had not been explored in automatic sexism detection before EXIST 2026.

Learning with Disagreements. The LeWiDi framework [3] advocates for preserving annotator disagreement rather than collapsing annotations into a single gold label. Training on soft probability distributions allows models to better reflect the inherent subjectivity of tasks such as hate speech and sexism detection. This paradigm is particularly relevant for EXIST 2026, where it serves as the official evaluation framework and motivates our use of soft labels during training.

3. Task and Dataset Description

This section describes the three EXIST 2026 meme subtasks addressed in this work and summarises the main characteristics of the training data. The dataset is multimodal: in addition to meme text and images, it includes crowd-worker annotations, annotator demographic information, and physiological signals recorded from human observers.

3.1. Subtasks

We participate in the three meme subtasks of EXIST 2026:

- **Subtask 2.1, *Sexism Identification*:** binary classification of whether a meme is sexist, with labels YES and NO;
- **Subtask 2.2, *Source Intention*:** classification of sexist memes according to the author's communicative intention, either DIRECT, when the meme is intentionally sexist, or JUDGEMENTAL, when it judges or reports on a sexist situation;
- **Subtask 2.3, *Sexism Categorization*:** multi-label categorization into one or more of six categories: IDEOLOGICAL-INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE, MISOGYNY-NON-SEXUAL-VIOLENCE, and NO.

3.2. Dataset

The training set contains 3,984 memes in English and Spanish, with an approximately balanced distribution between the two languages. The data are provided as a JSON file, where each meme is associated with a unique numeric identifier. For each meme, the dataset provides:

- **Textual content:** the OCR-extracted text overlaid on the meme image, together with the corresponding image filename;
- **Annotation labels:** six crowd-worker annotations per meme, from which soft probability distributions are derived for each subtask following the LeWiDi paradigm;
- **Annotator demographics:** per-annotator attributes, including gender, age group (18–22, 23–45, 46+), ethnicity, study level, and country of origin. Annotator profiles are heterogeneous: a single meme may be judged by annotators spanning multiple genders, age groups from 18–22 to 46+, education levels from less than high school to Master's degree, and multiple countries;
- **Physiological signals:** per-subject measurements stored under the `sensorial` field for the subjects who viewed each meme. Subjects are identified as EN1–EN7 and ES1–ES5, ES8, for a total of 13 subjects. Since not all subjects viewed every meme, the resulting physiological signal matrix is sparse.

Raw physiological signals. The physiological data are organised by modality and include the following raw features for each subject:

- ***Eye-Tracking* (ET):** 24 features per subject, including fixation duration, saccade amplitude, left and right pupil diameter, blink count, and reaction time;
- ***Heart Rate* (HR):** 4 features per subject, capturing the mean, variance, minimum, and maximum of the inter-beat interval;
- ***Electroencephalography* (EEG):** 80 features per subject, corresponding to spectral power across 16 electrode channels and 5 frequency bands (δ, θ, α, β, γ).

Table 1 reports the label distribution for each subtask. The dataset exhibits significant class imbalance, particularly in Subtask 2.3: the rarest category (MISOGYNY-NON-SEXUAL-VIOLENCE) accounts for only 1.5% of memes, compared to 44.4% for NO. This imbalance motivated the use of Multilabel Stratified K-Fold cross-validation [10], so that rare categories are proportionally represented in every fold.

4. Methodology

This section describes the methodology adopted for the three EXIST 2026 meme subtasks. Our system follows a multimodal pipeline that combines textual, visual, physiological, and demographic information. We first describe the preprocessing and feature extraction steps, then detail the construction of physiological and demographic features, the modality ablation used to guide the final system design, the two proposed architectures, and the training and submission strategy.

Table 1
Training set class distribution.

Subtask	Label	Count	%
2.1	YES	2,003	50.3
	NO	1,367	34.3
	Mixed	614	15.4
2.2	DIRECT	764	47.5
	JUDGEMENTAL	207	12.9
	NO	639	39.7
2.3	NO	1,367	44.4
	IDEOLOGICAL-INEQUALITY	280	9.1
	STEREOTYPING-DOMINANCE	282	9.2
	OBJECTIFICATION	351	11.4
	SEXUAL-VIOLENCE	144	4.7
	MISOGYNY-NON-SEXUAL-VIOL.	46	1.5

4.1. Text Preprocessing

Raw meme text was cleaned by removing URLs, user mentions, and HTML entities. Emojis were converted to their textual description using the `emoji` Python library, preserving semantic content (e.g., `:angry_face:`). A language prefix token (`[EN]` or `[ES]`) was prepended to each input to make the encoder explicitly language-aware.

4.2. Feature Extraction

Text Encoder. We used `cardiffnlp/twitter-xlm-roberta-base` [6] as the text encoder. This model is a multilingual RoBERTa variant pre-trained on a large corpus of multilingual tweets, making it particularly suitable for short, informal, and socially grounded textual content. Since meme text often contains abbreviations, emoji, non-standard spelling, and colloquial expressions, the Twitter-adapted encoder is better aligned with this register than standard multilingual encoders. For each meme, we used the resulting textual representation as the text component of our multimodal pipeline.

Image Encoder. We used CLIP ViT-B/32 [5] as the image encoder to obtain 768-dimensional visual embeddings for each meme. CLIP is well-suited to multimodal settings because it learns visual representations aligned with natural language, making it appropriate for content where image interpretation is closely tied to textual and semantic cues. All images were resized to 224×224 pixels and normalized according to the standard CLIP preprocessing pipeline.

4.3. Feature Engineering

Physiological Signal Processing. The raw per-subject physiological signals (24 ET, 4 HR, 80 EEG features) were transformed into per-meme feature vectors through two steps.

- First, the 80 raw EEG features (16 channels $\times$ 5 frequency bands) were aggregated into lobe-level representations by averaging the channels within each of four anatomically defined cortical lobes: Frontal (channels 0, 1, 8–11), Occipital (channels 6–7), Temporal (channels 12–13), and Parietal-Central (channels 2–5, 14–15). This reduces the EEG representation from 80 to 20 features per subject (4 lobes $\times$ 5 bands). The aggregation is motivated by the functional interpretation of cortical regions: the frontal lobe governs executive cognition and attention, the occipital lobe processes visual stimuli (particularly relevant for image-based content such as memes), the temporal lobe handles semantic comprehension and emotional response, and the parietal-central

region is associated with sensory integration, spatial attention, and motor planning, providing a complementary source of information alongside the other three regions.
- Second, for all three modalities, the per-subject values were aggregated across subjects by computing the *mean* and *standard deviation* across all subjects who viewed that meme. The mean captures the average physiological response to the meme; the standard deviation captures inter-subject variability, which may correlate with the perceptual ambiguity of the content. This yields the final feature counts: $24 \times 2 = 48$ ET features, $4 \times 2 = 8$ HR features, and $20 \times 2 = 40$ EEG features, for a total of 96 physiological signal features per meme.

Missing values arising from subjects who did not view a specific meme were handled differently depending on the nature of the feature. Discrete event-count features (e.g., blink count, fixation count) were imputed with zero, since their absence indicates that no event occurred. Continuous features (durations, amplitudes, spectral powers) were imputed using an iterative imputer fitted on the training partition of each fold. Non-negative quantities were additionally clipped to zero after aggregation to remove residual negative values introduced by the aggregation step.

Demographic Feature Computation. For each meme, 15 aggregated features were computed from the annotator pool: gender ratio (F/M), age group proportions (18–22, 23–45, 46+), education level proportions (high/mid/low, derived from annotator self-reported study levels), ethnicity proportions (White/Caucasian, Hispanic/Latino, other), and country-of-origin proportions (Spain, Latin America, Anglo countries, other). These features encode the demographic composition of the crowd-worker panel for each item, which may correlate with annotation patterns and sensitivity to different forms of sexism. It is worth noting that the annotation panel was designed to be balanced (50/50 male/female, equal thirds for age groups), resulting in near-constant gender and age feature values across all memes; education level and country of origin proved more variable and informative in our ablation.

Physiological Signal and Demographic Features. Both the 96-dimensional physiological signal vector and the 15-dimensional demographic vector are standardized using a `StandardScaler` fitted on the training partition of each cross-validation fold to prevent data leakage.

4.4. Modality Ablation: From Unimodal to Multimodal

Before developing the full multimodal pipeline, we conducted a progressive modality ablation on fold 0 to assess the individual and combined contribution of each input source. We trained a Late Fusion architecture in the following incremental configurations: (1) text only, (2) text and image, (3) text, image, and physiological signals, and (4) text, image, physiological signals, and demographics. Each modality addition improved the aggregate ICMSoftNorm sum on fold 0, with the largest gain coming from adding the image modality over text alone. Physiological signals and demographics provided smaller but consistent improvements. These results motivated the use of the full four-modality pipeline as the basis for all subsequent experiments.

4.5. Model Architectures

We propose and compare two multimodal architectures: a *Late Fusion* model (M1) and a *Cross-Attention* model (M2). The architectures of the two models differ in how they combine the available sources of information. The goal is to evaluate whether a straightforward fusion strategy is sufficient for this task, or whether a more interaction-based mechanism can better exploit the relationship between meme content and human physiological responses. Both architectures operate on the same set of input modalities and produce predictions for the three subtasks through task-specific output heads. Table 2 summarises their main differences.

Table 2
Comparison of the main architectural components of the Late Fusion and Cross-Attention models.

Component	Late Fusion	Cross-Attention
Text pooling	[CLS] token	Attention pooling
Physiol. integration	Concatenation	Cross-Attention
Demo integration	Concatenation	Late fusion (separate)
Output dimension	512	1,088

4.5.1. M1: Late Fusion Model

The Late Fusion model, whose architecture is illustrated in Figure 1, independently projects each modality and concatenates the resulting vectors, following the standard late fusion paradigm for multimodal learning [12].

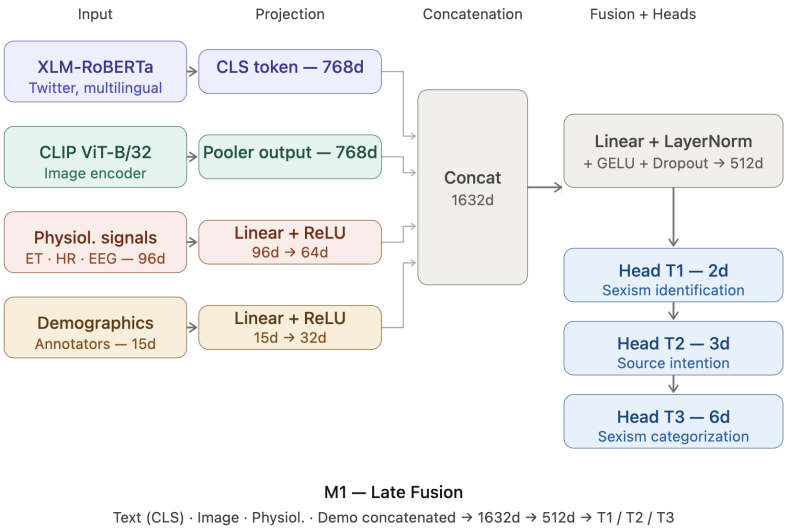

Figure 1: M1 Late Fusion architecture. All four modalities are independently projected and concatenated into a 1,632-dimensional vector, compressed to 512d via a linear layer with Layer Normalization and GELU activation. Three classification heads predict T1, T2, and T3.

Formally, let $\mathbf{t} \in \mathbb{R}^{768}$ be the [CLS] token produced by XLM-RoBERTa, $\mathbf{v} \in \mathbb{R}^{768}$ the pooled visual embedding produced by CLIP, $\mathbf{b}_{\text{raw}} \in \mathbb{R}^{96}$ the standardized physiological signal vector, and $\mathbf{d}_{\text{raw}} \in \mathbb{R}^{15}$ the standardized demographic vector. Each non-textual modality is first passed through an independent linear projection with ReLU activation:

$$\mathbf{b} = \text{ReLU}(W_b\,\mathbf{b}_{\text{raw}} + \beta_b) \in \mathbb{R}^{64}, \qquad \mathbf{d} = \text{ReLU}(W_d\,\mathbf{d}_{\text{raw}} + \beta_d) \in \mathbb{R}^{32} \tag{1}$$

where $W_b \in \mathbb{R}^{64\times 96}$, $W_d \in \mathbb{R}^{32\times 15}$ are learned weight matrices and β_b, β_d the corresponding bias terms. The four projected vectors are then concatenated and compressed into a joint multimodal representation $\mathbf{h} \in \mathbb{R}^{512}$:

$$\mathbf{h} = \text{GELU}(\text{LN}(W_f\,[\mathbf{t};\,\mathbf{v};\,\mathbf{b};\,\mathbf{d}] + \beta_f)) \tag{2}$$

where $[\mathbf{t};\,\mathbf{v};\,\mathbf{b};\,\mathbf{d}] \in \mathbb{R}^{1632}$ denotes the concatenation of the four vectors ($768+768+64+32 = 1632$), $W_f \in \mathbb{R}^{512\times 1632}$ and $\beta_f \in \mathbb{R}^{512}$ are the fusion layer parameters, $\text{LN}(\cdot)$ denotes Layer Normalization [7], and $\text{GELU}(\cdot)$ the Gaussian Error Linear Unit activation. Three independent classification heads are then applied to $\mathbf{h}$:

$$\hat{y}_{T1} = \text{softmax}(W_{T1}\,\mathbf{h} + \beta_{T1}), \quad W_{T1} \in \mathbb{R}^{2\times 512} \tag{3}$$

$$\hat{y}_{T2} = \text{softmax}(W_{T2}\,\mathbf{h} + \beta_{T2}), \quad W_{T2} \in \mathbb{R}^{3\times 512} \tag{4}$$

$$\hat{y}_{T3} = \sigma(W_{T3}\,\mathbf{h} + \beta_{T3}), \quad W_{T3} \in \mathbb{R}^{6\times 512} \tag{5}$$

where $\sigma(\cdot)$ denotes the element-wise sigmoid function, used for T3 since it is a multi-label task where the output probabilities are independent across categories.

4.5.2. M2: Cross-Attention Model

The Cross-Attention model, whose architecture is illustrated in Figure 2, introduces two improvements over Late Fusion: (1) attention pooling over all text token representations instead of using only the [CLS] token, and (2) a Cross-Attention mechanism where physiological signals query the joint meme representation.

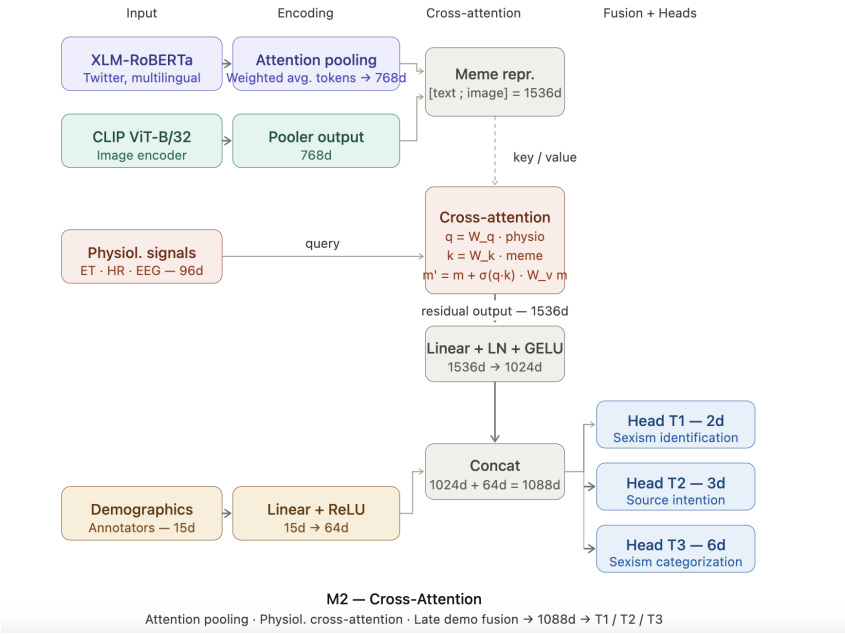

Figure 2: M2 Cross-Attention architecture. Text is encoded via attention pooling over all token representations. Physiological signals act as a query over the joint text-image representation via a Cross-Attention mechanism with residual connection, producing a modulated 1,536-dimensional meme representation. Demographic features are projected separately and fused at the end. The final 1,088-dimensional vector is passed to three classification heads for T1, T2, and T3.

Attention Pooling. From a formal point of view, let $H = [\mathbf{h}_1, \ldots, \mathbf{h}_L] \in \mathbb{R}^{L\times 768}$ be the sequence of L token-level hidden states produced by XLM-RoBERTa for an input of length L, where $\mathbf{h}_i \in \mathbb{R}^{768}$ is the hidden state of the i-th token. Instead of using only the first token ([CLS]), the text representation $\mathbf{t} \in \mathbb{R}^{768}$ is computed as a weighted sum of all token representations:

$$\mathbf{t} = \sum_{i=1}^{L} \alpha_i\,\mathbf{h}_i, \qquad \alpha_i = \frac{\exp(w^\top \mathbf{h}_i)}{\sum_{j=1}^{L}\exp(w^\top \mathbf{h}_j)} \tag{6}$$

where $w \in \mathbb{R}^{768}$ is a learned weight vector and $\alpha_i \in [0,1]$ are the normalized attention weights satisfying $\sum_{i=1}^{L}\alpha_i = 1$. This formulation allows the model to focus on the tokens most discriminative for the classification task, rather than relying on the [CLS] summary [13].

Cross-Attention. The joint meme representation, denoted as $\mathbf{m} = [\mathbf{t}; \mathbf{v}] \in \mathbb{R}^{1536}$, is modulated by the physiological signal vector via a scalar Cross-Attention mechanism [13]. Specifically, the

physiological vector $\mathbf{b}_{\text{raw}}$ is used to produce a query, while the meme representation $\mathbf{m}$ produces both a key and a value, which are formally defined as:

$$\mathbf{q} = W_q\,\mathbf{b}_{\text{raw}} \in \mathbb{R}^{64}, \quad \mathbf{k} = W_k\,\mathbf{m} \in \mathbb{R}^{64}, \quad a = \sigma\left(\frac{\mathbf{q}^\top \mathbf{k}}{\sqrt{d_a}}\right) \in [0,1] \tag{7}$$

$$\mathbf{m}' = \mathbf{m} + a \cdot W_v\,\mathbf{m} \in \mathbb{R}^{1536} \tag{8}$$

where $W_q \in \mathbb{R}^{64\times 96}$ and $W_k \in \mathbb{R}^{64\times 1536}$ are the query and key projection matrices, $d_a = 64$ is the attention dimension used for scaling, $\sigma(\cdot)$ is the sigmoid function producing a scalar attention score $a \in [0,1]$, and $W_v \in \mathbb{R}^{1536\times 1536}$ is the value projection matrix. The residual connection $\mathbf{m}' = \mathbf{m} + a \cdot W_v\,\mathbf{m}$ ensures the original representation is preserved when the physiological signal is uninformative (i.e., $a \approx 0$). The modulated representation $\mathbf{m}'$ is then passed through a fusion layer to obtain $\mathbf{f} \in \mathbb{R}^{1024}$, concatenated with the projected demographic vector $\mathbf{d}' \in \mathbb{R}^{64}$, and the resulting vector $\mathbf{o} = [\mathbf{f};\, \mathbf{d}'] \in \mathbb{R}^{1088}$ is fed to three classification heads identical in form to those of M1.

4.6. Training Objective

Since the task follows the LeWiDi paradigm, we trained the models on soft label distributions using task-specific losses combined through automatic multi-task weighting.

Soft Label Losses. For T1 and T2 (single-label tasks) we used *soft cross-entropy*, formally defined as:

$$\mathcal{L}_{\text{CE}}(p,\, q) = -\sum_{c=1}^{C} q_c \log p_c \tag{9}$$

where $p_c \in [0,1]$ is the predicted probability for class c, $q_c \in [0,1]$ is the corresponding soft target derived from the annotator distribution (with $\sum_c q_c = 1$), and $C = 2$ for T1 and $C = 3$ for T2.

For T3 (multi-label) we used *element-wise binary cross-entropy with soft targets*, formally defined as:

$$\mathcal{L}_{\text{BCE}}(p,\, q) = -\sum_{c=1}^{C} \big[q_c \log p_c + (1 - q_c)\log(1 - p_c)\big] \tag{10}$$

where $C = 6$ is the number of categories, $p_c \in [0,1]$ is the predicted probability for category c produced by a sigmoid function (independent per category, so $\sum_c p_c$ need not equal 1), and $q_c \in [0,1]$ is the proportion of annotators assigning that category.

Automatic Task Weighting. We adopted the Automatic Weighted Loss (AWL) as defined in [8], which frames each task loss as a Gaussian likelihood with a learnable homoscedastic uncertainty parameter:

$$\mathcal{L} = \sum_{i=1}^{3} \frac{1}{2\sigma_i^2}\,\mathcal{L}_i + \log \sigma_i \tag{11}$$

where $\mathcal{L}_1 = \mathcal{L}_{\text{CE}}^{T1}$, $\mathcal{L}_2 = \mathcal{L}_{\text{CE}}^{T2}$, $\mathcal{L}_3 = \mathcal{L}_{\text{BCE}}^{T3}$ are the per-task losses, and $\sigma_i > 0$ ($i = 1, 2, 3$) are learnable scalar parameters initialised to 1. A larger σ_i down-weights task i, while the $\log \sigma_i$ regularization term prevents the trivial solution $\sigma_i \to \infty$.

4.7. Training Setup and Hyperparameter Search

Both architectures were trained with AdamW using differential learning rates: encoder parameters at 5×10^{-6} and head/projection parameters at 5×10^{-4} (Late Fusion) or 8×10^{-4} (Cross-Attention). A cosine learning rate schedule with linear warmup was applied. Gradient norms were clipped at 1.0, and early stopping was applied with patience = 2 (max 6 epochs).

Successive Halving. Hyperparameters were identified via Successive Halving [9]: 16 randomly sampled configurations were trained for 2 epochs (Bracket 1); the top 4 by validation loss advanced to 4 epochs (Bracket 2); the top 2 were trained to completion up to 5 epochs (Bracket 3). The search was conducted on fold 0 for each architecture separately, exploring encoder learning rate, head learning rate, dropout, batch size, weight decay, and warmup ratio. The selected configurations are reported in Table 3.

Table 3
Hyperparameter configurations selected by Successive Halving.

Hyperparameter	Late Fusion	Cross-Attention
Encoder LR	5×10^{-6}	5×10^{-6}
Head LR	5×10^{-4}	8×10^{-4}
Dropout	0.2	0.1
Batch size	8	4
Weight decay	0.02	0.05
Warmup ratio	0.05	0.15

4.8. Cross-Validation and Ensemble Strategy

We applied Multilabel Stratified 5-Fold cross-validation [10], stratifying on the binarised T3 labels to ensure rare categories are represented in every fold. Three systems were submitted:

- **Run 1:** Average of the Late Fusion and Cross-Attention fold ensembles (10 models).
- **Run 2:** Late Fusion fold ensemble (5 models).
- **Run 3:** Late Fusion retrained on all 3,984 training samples (no OOF estimate available).

Hard labels are derived from soft predictions via argmax for T1 and T2; T3 applies a per-category threshold of 0.5, falling back to NO if no category exceeds the threshold.

5. Experiments and Results

This section reports the experimental evaluation of the proposed systems. We first analyse the effect of the main design choices through an ablation study conducted on a single validation fold. We then report the out-of-fold (OOF) performance of the submitted systems on the training data and, finally, present the official test set results released by the task organizers.

5.1. Ablation Study

We conducted a systematic ablation on fold 0, measuring the ICMSoftNorm across the three subtasks.

Fusion Strategy. We compared four integration strategies for physiological signals: (i) simple concatenation (Late Fusion), (ii) multiplicative gating, where physiological signals generate a sigmoid gate that modulates the joint text-image representation element-wise, (iii) a Mixture-of-Experts (MoE) approach [15], and (iv) Cross-Attention. In the MoE configuration, two configurations were evaluated: 8 experts with hidden dimension 128, and 4 experts with hidden dimension 256; a learned router selected the top-2 experts per input for each modality. The gating mechanism proved a robust baseline, offering a strong improvement over simple concatenation. Cross-Attention yielded the highest fold-0 performance among all tested strategies; however, this advantage did not replicate consistently across all five folds in the OOF evaluation (see Section 5.3), suggesting that the Cross-Attention hyperparameter configuration is more sensitive to fold-specific variation, likely due to the aggressive learning rate and small batch size selected during the search.

Loss Function Variants. We tested three alternatives to the standard soft CE/BCE losses. *Focal loss* ($\gamma = 2$) consistently degraded T1 and T2 performance. *Class-weighted BCE* was motivated by the severe T3 class imbalance but similarly hurt T1 and T2 without improving T3. *Label smoothing* ($\epsilon = 0.1$) was ineffective, which we attribute to the soft labels already serving as a natural regulariser. We also experimented with hierarchical conditioning—passing the predicted T1 soft output as an additional input to the T2 and T3 heads—which provided marginal improvement on T2 at the cost of T3.

Caption Augmentation. We evaluated augmenting the text input with visual descriptions of the meme image generated by Qwen2.5-VL-3B [14]. For each meme, the model was prompted to describe the visual content of the image without reading the overlaid text. This caption was concatenated to the OCR-extracted meme text using a `[IMG]` separator token, and the full combined string was passed to XLM-RoBERTa. The motivation was to provide the text encoder with visual context that the CLIP image embedding encodes only implicitly. Contrary to expectations, caption augmentation consistently degraded performance across all model variants. The generated descriptions appeared to introduce redundant or noisy information on top of the CLIP image embedding, without adding discriminative signal.

Physiological Signal Feature Engineering. We explored per-channel EEG features (without lobe aggregation), lobe-aggregated EEG features, and different normalization strategies. Lobe-aggregated EEG features consistently outperformed raw per-channel features. Robust scaling and additional feature engineering degraded performance in all tested configurations.

5.2. Training Set Results

Table 4 reports out-of-fold performance for the submitted systems, measured with ICMSoftNorm.

Table 4
Out-of-fold ICMSoftNorm results. N/A = not available (full refit).

System	T1	T2	T3	Sum
Run 1 – Ensemble	0.449	0.377	0.238	**1.064**
Run 2 – Late Fusion	0.444	0.371	0.241	1.056
Run 3 – Full Refit	N/A	N/A	N/A	N/A

The ensemble (Run 1) improves over the Late Fusion model (Run 2) by 0.008 points. Subtask 2.3 (T3) is consistently the most challenging, reflecting the severe class imbalance: the rarest category (MISOGYNY-NON-SEXUAL-VIOLENCE) has only 46 dominant training instances compared to 1,367 for NO.

5.3. Official Test Set Results

Table 5 reports the official rankings and scores obtained by our three submitted runs on the EXIST 2026 test set, evaluated by the task organizers.

Notably, the relative ordering of runs differs across subtasks. For T2.1 and T2.2, the Late Fusion fold ensemble (Run 2) consistently ranks highest among our submissions, confirming the OOF evaluation trend. For T2.3, however, the full refit (Run 3) performs best, suggesting that training on the complete dataset provides additional benefit specifically for the most difficult and imbalanced subtask. The ensemble (Run 1) ranks lower than expected on T2.3, which may reflect that averaging predictions from the Cross-Attention model — which showed weaker OOF performance on T3 — reduced the signal from the stronger Late Fusion predictions.

Table 5
Official test set results for the three submitted runs across all subtasks. Rankings are relative to all participating systems (T2.1: ≈140 systems; T2.2 and T2.3: ≈115 systems each). ICM-SN = ICMSoftNorm. CE = Cross Entropy (T2.3 not reported by organizers).

Subtask	Run	Ranking	ICM-Soft	ICM-SN	CE
T2.1 — Sexism ID	Run 2 (Late Fusion)	29	−0.2804	0.4549	0.9078
	Run 1 (Ensemble)	31	−0.2925	0.4530	0.8937
	Run 3 (Full Refit)	34	−0.3043	0.4511	0.9336
T2.2 — Source Intention	Run 2 (Late Fusion)	12	−1.1752	0.3750	1.4233
	Run 1 (Ensemble)	13	−1.1770	0.3748	1.4036
	Run 3 (Full Refit)	15	−1.2031	0.3721	1.4580
T2.3 — Categorization	Run 3 (Full Refit)	12	−4.8823	0.2412	—
	Run 2 (Late Fusion)	13	−4.9132	0.2396	—
	Run 1 (Ensemble)	17	−5.1125	0.2290	—

6. Discussion

This section discusses the main findings emerging from the experimental results. We focus on the relative behaviour of the proposed architectures, the contribution and limitations of physiological and demographic information, and the effect of model complexity under the constraints of the available dataset and computational resources.

Architecture Comparison. The OOF evaluation shows that the performance gap between Late Fusion and the ensemble is modest (0.008 points). The official test set results reported in Section 5.3 are consistent with this observation: Run 2, corresponding to the Late Fusion fold ensemble, ranks highest among our submissions on T2.1 and T2.2, while the ensemble (Run 1) provides no clear advantage over the simpler model. Although Cross-Attention achieved stronger results in the single-fold evaluation on fold 0, this advantage did not consistently replicate across all five folds or on the official test set. We attribute this behaviour to the hyperparameter configuration selected for Cross-Attention, namely a high head learning rate, low dropout, and small batch size, which favoured fast convergence on the fold used during hyperparameter search but generalized less well to unseen data.

Value of Physiological Signals. Physiological signals contributed positively in the progressive modality ablation and appeared more effective when integrated through Cross-Attention rather than simple concatenation. However, with only 2 subjects per meme, the aggregated physiological statistics are inherently noisy and may not fully capture stable population-level responses. A larger subject pool would be required to assess the true potential of physiological signals for this task.

Demographic Features. The annotation panel was designed to be demographically balanced, with a 50/50 male/female split and equal thirds for age groups. As a result, gender and age features are nearly constant across memes, limiting their discriminative power. Education level and country of origin, instead, showed greater variability and proved more informative in the ablation.

Mixture of Experts. We evaluated two MoE configurations: 8 experts with a hidden dimension of 128, and 4 experts with a hidden dimension of 256. Both configurations underperformed Late Fusion, confirming the expected limitation that, with only 3,984 training memes, the routing mechanism cannot acquire meaningful expert specialisations. Each expert sees, on average, fewer than 1,000 training examples, which is likely insufficient for specialisation to emerge. This finding is nonetheless informative, as it suggests that architectural complexity alone does not compensate for limited data, and that simpler but better-motivated mechanisms, such as Cross-Attention and gating, are preferable

at this dataset scale. This is consistent with observations in the broader MoE literature [15], where specialisation benefits typically require datasets orders of magnitude larger.

Limitations. The absence of dedicated GPU infrastructure constrained the experimental scope more broadly: larger text encoders, more extensive hyperparameter searches, and more complex architectures could not be evaluated within the available session time limits on Kaggle. Finally, single-fold performance estimates substantially overestimate true generalisation performance, highlighting the importance of multi-fold evaluation even at the exploratory stage.

7. Conclusion

In this paper, we presented a human-centered multimodal system for sexism detection in memes, integrating textual, visual, physiological, and demographic information. A progressive modality ablation showed that each input source contributed positively, with the image modality providing the largest individual gain. We compared two fusion architectures, Late Fusion and Cross-Attention, both trained under the LeWiDi paradigm using soft labels and automatic multi-task loss weighting.

Overall, the Late Fusion fold ensemble proved to be the most consistent submission across subtasks, outperforming the Cross-Attention ensemble on the official test set despite slightly lower out-of-fold scores. Future work should further investigate the use of larger vision-language models for structured semantic feature extraction, together with a systematic comparison of prompting strategies.

Declaration on Generative AI

The author used a large language model (Claude, Anthropic) as an assistant during the preparation of this paper. Specifically, the AI tool was used to support the development and debugging of the experimental pipeline, and to review and improve the clarity and structure of the manuscript. All scientific content, experimental design, results, and conclusions are the sole responsibility of the author.

References

[1] 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.

[2] 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.

[3] E. Leonardelli, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, A. Uma, and M. Poesio. SemEval-2023 Task 11: Learning with Disagreements (LeWiDi). In *Proceedings of SemEval-2023*, pages 2304–2318, 2023.

[4] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, and D. Testuggine. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of ICML*, 2021.

[6] F. Barbieri, L. Espinosa Anke, and J. Camacho-Collados. XLM-T: Multilingual Language Models in Twitter for Sentiment Analysis and Beyond. In *Proceedings of LREC*, 2022.

[7] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[8] A. Kendall, Y. Gal, and R. Cipolla. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *Proceedings of CVPR*, 2018.

[9] K. Jamieson and A. Talwalkar. Non-Stochastic Best Arm Identification and Hyperparameter Optimization. In *Proceedings of AISTATS*, 2016.

[10] K. Sechidis, G. Tsoumakas, and I. Vlahavas. On the Stratification of Multi-Label Data. In *Proceedings of ECML PKDD*, 2011.

[11] P. Italiani, D. Gimeno-Gómez, L. Ragazzi, G. Moro, and P. Rosso. MEMEWEAVER: Inter-Meme Graph Reasoning for Sexism and Misogyny Detection. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 2120–2134, 2026.

[12] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2019.

[13] 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)*, volume 30, 2017.

[14] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang, T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025.

[15] X. Han, H. Nguyen, C. Harris, N. Ho, and S. Saria. FuseMoE: Mixture-of-Experts Transformers for Fleximodal Fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

[16] 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.