Through the Eyes of the Beholder: Biometric and Demographic Conditioning for Multimodal Sexism Detection

Notebook for the EXIST Lab at CLEF 2026

Ana-Maria Luisa Mocanu[1,†], Sebastian Mocanu[1,†], Ciprian-Octavian Truică[1,2] and Elena-Simona Apostol[1,*]

[1]National University of Science and Technology POLITEHNICA Bucharest, Splaiul Independenţei 313, Bucureşti 060042, Romania
[2]Academy of Romanian Scientists, Ilfov 3, Bucharest, 050044, Romania

Abstract

Detecting sexism on the internet is a fundamentally subjective task; our team, **VANGUARD**, addresses this challenge in the EXIST 2026 Task 2 by proposing a human-centered multimodal framework that analyses and incorporates the psychological and demographic characteristics of human annotators into the detection pipeline. We fuse five input modalities through a cross-attention architecture with Feature-wise Linear Modulation conditioning. Meme text is extracted and visually described with Gemma 4, then augmented by automatic translation between English and Spanish with NLLB-200. Text and image representations are produced by LoRA-adapted XLM-RoBERTa and CLIP encoders and fused with sensor features encoded by a pretrained autoencoder. To model annotator subjectivity, we frame Subtask 2.1 as a label distribution learning problem, optimizing a Kullback-Leibler divergence loss over the full annotator label distribution. At inference time, predictions are produced by soft-voting between the deep multimodal network and a complementary SVM trained on stylometric and physiological features. Our best submission ranks 29th out of 114 on Subtask 2.2 (source intention) under soft evaluation, and the normalized ICM scores remain above the baseline on Subtasks 2.1 and 2.2, indicating that annotator-centered conditioning contributes a usable signal. We release our full pipeline and analysis to support reproducible human-centered modeling.

Keywords

sexism detection, multimodal learning, meme analysis, biometric signals, label distribution learning, vision-language models, human-centered AI

1. Introduction

Meme culture has evolved from the early days of the internet into one of its dominant forms of communication. While often harmless, the internet's anonymity allows for the spread of controversial content [1, 2]; the subject of this paper is that the memes sometimes hide sexism [3], often as a response against feminism. Although mostly directed at women, sexism can affect men as well [4].

The main challenge with this type of content is its subtlety. Sexism in memes rarely appears as direct hate speech. Instead, it relies on irony (e.g., text formats like "SpongeBob mocking") and hidden visual meaning [5] that are hard to detect without a deep understanding of internet culture. Understanding memes remains highly subjective, as an image that one person finds offensive might be misunderstood or ignored by another, due to different personal biases and backgrounds.

CLEF 2026 Working Notes, September 21 – 24, 2026, Jena, Germany
*Corresponding author.
†These authors contributed equally.
ana_maria.mogoase@upb.ro (A. L. Mocanu); sebastian.mocanu@upb.ro (S. Mocanu); ciprian.truica@upb.ro (C. Truică); elena.apostol@upb.ro (E. Apostol)
https://sites.google.com/view/ciprian-octavian-truica (C. Truică); https://sites.google.com/view/elena-simona-apostol (E. Apostol)
0009-0005-2646-4086 (A. L. Mocanu); 0009-0007-0313-4724 (S. Mocanu); 0000-0001-7292-4462 (C. Truică); 0000-0001-6397-4951 (E. Apostol)

EXIST 2026 task 2 addresses this challenge. Instead of forcing a single label as the ground truth, the goal is to learn from human disagreement [6, 7]. By analyzing the multimodal dataset [8], we aim to "see" through the hidden online sexism by introducing a human-centered model. Rather than looking at the meme itself, our approach incorporates real-time reaction data from annotators who saw the memes as an important component.

Our specific contributions[1] are as follows:

- **Biometric and Demographic Conditioning:** We incorporate annotator eye-tracking, and heart rate variability features alongside demographic embeddings (gender, age, education level, ethnicity) directly into the neural model via FiLM conditioning [9], allowing the network to adjust its predictions based on the measurable physiological and social context of perception. Electroencephalography (EEG) bandpower features, while not individually significant for the neural branch, are additionally exploited by the classical SVM component of our ensemble. We report multi-seed ablations characterizing the empirical contribution of this conditioning in Section 5.
- **Cross-lingual Data Augmentation:** To improve cross-lingual robustness and expand the effective training set, we translate cleaned meme text and visual descriptions between English and Spanish, effectively doubling the training data while encouraging language-invariant learning.
- **VLM-based Text and Visual Enrichment:** We use the Gemma 4 [10] vision-language model to extract clean, uncensored embedded text from memes and to generate structured natural-language descriptions of their visual components, providing richer input representations than raw pixels alone.
- **Label Distribution Learning:** For Subtask 2.1, we frame prediction as a distribution learning problem, training the model to reproduce the full distribution of annotator opinions using a soft-label Kullback-Leibler (KL) divergence [11, 7], rather than optimizing for a majority-vote binary label.
- **Cross-Attention Multimodal Fusion:** Text representations from LoRA-adapted XLM-RoBERTa [12] and visual representations from LoRA-adapted CLIP [13, 14] are fused via cross-attention, with sensor embeddings injected through FiLM modulation [9].
- **Neural-Classical Ensemble:** Final predictions combine the deep multimodal model with an SVM trained on interpretable stylometric and physiological features via soft-voting, providing complementary coverage and acting as an algorithmic regularizer.
- **Ablation under Multiple-Comparison Control:** We retrain every architectural variant over five random seeds and test each against the baseline with Holm-Bonferroni correction [15] across the full family of 63 comparisons. No single component, including the biometric conditioning, survives the correction. We report this as a cautionary result.

The remainder of the paper is structured as follows. Section 2 reviews related work on sexism detection and multimodal meme analysis. Section 3 presents a statistical analysis of the EXIST 2026 dataset. Section 4 describes our full system architecture and training strategy. Section 5 reports experimental results and ablation studies. Section 6 concludes with a summary and the future directions.

2. Related Work

Sexism detection has evolved from text-based approaches using TF-IDF features with classical classifiers [16, 17] to deep learning with pre-trained embeddings [18], and more recently to transformer-based models that enable deeper contextual understanding [19]. The EXIST task series [20, 21, 22, 23] has been proven to advance the field, progressively expanding from text-only tweets to multimodal memes and videos. The 2026 edition introduces a fundamental shift from previous years. Beyond the meme content and annotator labels, the dataset includes physiological signals and demographic profiles of the annotators [8], making it possible to use human-centered modeling approaches.

[1] Code available at https://github.com/DS4AI-UPB/VANGUARD-CLEF2026-EXIST.

Memes pose a unique multimodal challenge, as their meaning arises from the interplay between text and image, often masked by humor or irony. The Hateful Memes Challenge [5] showed that state-of-the-art multimodal models struggled when offensiveness depended on the combination of individually benign modalities. Recent approaches have addressed the modality gap through VLM-based captioning [24], multimodal fusion and bias analysis [25], and graph-based reasoning [26]. We additionally condition the fusion on annotator biometrics through FiLM modulation [9], and use Gemma 4 [10] to both extract embedded meme text and generate structured visual descriptions, further augmenting the training data through cross-lingual translation with NLLB-200 [27].

Annotator disagreement in subjective tasks carries a genuine signal rather than noise [6]. Prior work has addressed this by training models to predict soft label distributions directly [7]. We adopt this framing, weighting each instance by its annotator entropy, and additionally employ Supervised Contrastive Learning [28] as an auxiliary loss to geometrically structure the embedding space alongside the primary KL divergence objective.

Eye-tracking and EEG signals have been previously used in affective computing for emotion recognition [29, 30], but their integration into sexism detection introduced in EXIST 2026 [8] is unique. We encode the retained physiological features through a pretrained sensor autoencoder before joint training and inject them via FiLM modulation [9] into a cross-attention architecture.

3. Data Analysis

The EXIST 2026 dataset [8] consists of 3984 memes which were viewed by 16 subjects. Additionally, it contains human data composed of recorded eye movements, heart rate, and EEG of the annotators.

3.1. Demographic and Physiological Analysis

To model the multimodal indicators of sexism, we fused textual, visual, demographic, and physiological data. The ground truth was established via majority voting, where a meme was classified as sexist (i.e., $y = 1$) if $\geq 50\%$ of annotators agreed. A soft-label target (p_{yes}) was also retained to model annotator disagreement through the soft KL Divergence objective during training. We additionally implemented an uncertainty-weighted variant of this objective, applying a per-instance factor $\exp\big(-p_{yes}(1-p_{yes})\big)$ that emphasizes high-consensus samples. This variant was explored in preliminary single-task experiments and is not used in the multitask runs reported in this paper.

Demographic and Physiological Selection. Pearson's χ^2 tests confirmed that demographic traits (i.e., gender, age, study level, and ethnicity) significantly influence labeling behavior ($p < 0.001$), leading to their inclusion via trainable embedding layers. For physiological telemetry, we aggregated Eye Tracking (ET), Heart Rate (HR), and EEG data [29, 30] by computing the mean across users per instance. Following t-tests and point biserial correlation analysis represented in Table 1, we retained only the features showing significant variance ($p < 0.05$).

Feature Analysis. The distribution of predictive features is visualized in Figure 1. The analysis of the boxplots indicates that sexist content triggers significantly higher reaction times, fixation counts, and increased heart rate variability (HR Std.). This suggests that sexist memes generate greater cognitive friction than non-sexist ones.

Despite their predictive power, the Spearman correlation heatmap represented in Figure 2 reveals extreme multicollinearity among eye-tracking metrics, with Fixations and Saccades yielding $\rho \approx 1.0$. While this indicates functional redundancy, all significant features were retained to allow the neural network to optimize weighting internally.

Network Integration. To prevent features with large numerical scales, such as reaction time, from dominating the gradients, the 4-dimensional sensor vector was standardized using a `StandardScaler` fitted on the training split and applied to the validation and test data. Reaction time was additionally log-transformed prior to scaling to compress its heavy right tail. This ensures that subtle fluctuations in heart rate variance are weighed equitably alongside high-magnitude eye-tracking durations during multimodal fusion.

Table 1
Statistical Significance of Demographic and Physiological Features

Feature Category	Variable	Test Statistic	p-value
Demographics (χ^2)	Gender	$\chi^2 = 79.51$	$< 0.001^{**}$
	Age	$\chi^2 = 187.77$	$< 0.001^{**}$
	Study Level	$\chi^2 = 45.77$	$< 0.001^{**}$
	Ethnicity	$\chi^2 = 41.13$	$< 0.001^{**}$
Physiological (t-test)	Reaction Time	$t = 8.17$	$< 0.001^{**}$
	Fixations Count	$t = 5.52$	$< 0.001^{**}$
	Saccades Count	$t = 5.44$	$< 0.001^{**}$
	HR Std.	$t = 3.45$	$< 0.001^{**}$
	Mean Pupil Diameter	$t = 0.00$	0.9962
	Mean Heart Rate	$t = -1.16$	0.2466
	EEG Alpha Power	$t = -1.01$	0.3123

** Indicates statistical significance at $\alpha = 0.05$ level.

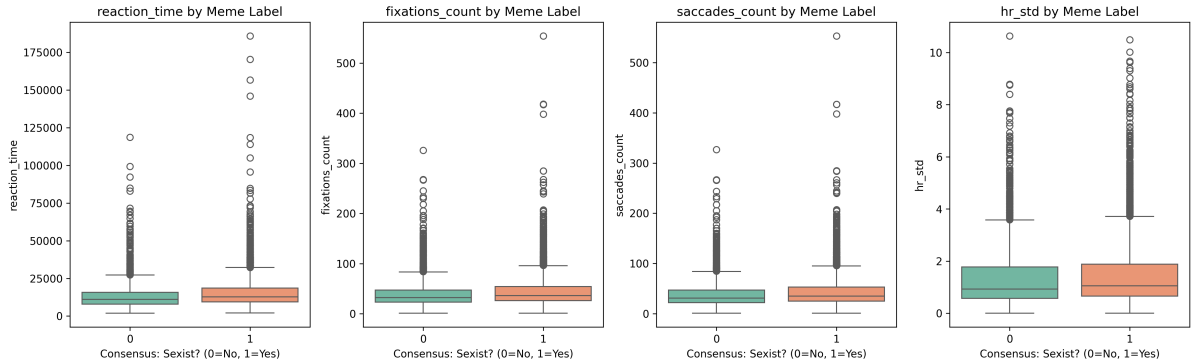

Figure 1: Distribution of predictive physiological features across meme classes.

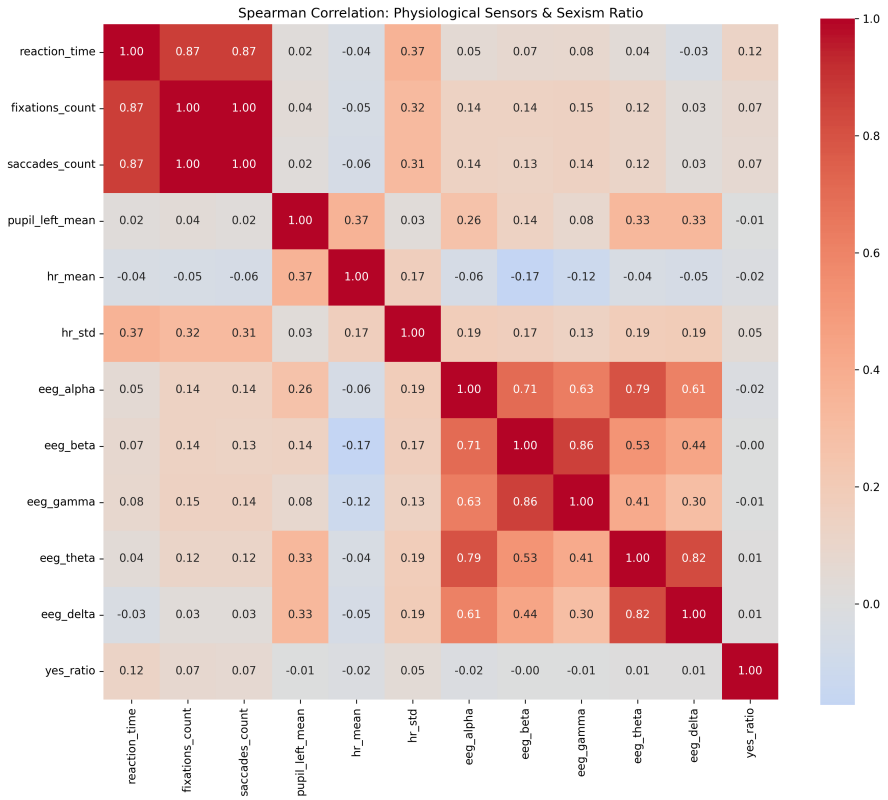

Figure 2: Spearman correlation heatmap of aggregated physiological sensors.

3.2. Text

To capture the linguistic nuances of the sexist memes, we extracted metrics from the embedded text across $N = 3,984$ valid records before preprocessing.

Linguistic Distribution. The dataset presents a highly balanced bilingual composition, with the records split almost perfectly between English ($n = 2,005$) and Spanish ($n = 1,979$). This is crucial for training, ensuring the network does not falsely correlate a specific language with the presence of sexism.

Text Length Analysis. Memes inherently rely on brief, punchy text and prominent visual cues, occasionally exhibiting longer blocks of text. As detailed in Table 2, the analysis of the extracted textual fields reveals an average character count of 124.3 ($\sigma = 104.72$) and a mean word count of 21.6 words ($\sigma = 18.12$). The word distributions are right-skewed. The maximum word count reaches 299 words, while 75% of the dataset contains 26 words or fewer.

Table 2
Descriptive Statistics of Textual Features ($N = 3,984$)

Feature	Mean	Std. Dev.	Min	25%	Median	Max
Character Count	124.30	104.72	9.00	64.00	96.00	1777.00
Word Count	21.60	18.12	2.00	11.00	17.00	299.00

3.3. Image

For the visual modality, exploratory data analysis was conducted on the structural and pixel-level properties of the raw image.

Dimension and Formatting. The resolution of the memes varies significantly, with widths ranging from 173 to 4744 pixels and heights from 124 to 6000 pixels. The aspect ratio distribution shown in Table 3 exhibits a distinct median exactly at 1 and a mean of 1.03. As represented in Figure 3, the dominance of the 1:1 square format is characteristic of meme formats. Standardizing the input via center-cropping or padding to a fixed square resolution results in minimal information loss. However, this affects taller memes, typically those with a storytelling format, which may benefit from being split into multiple sub-images before being fed as input.

Visual Tones. An analysis of the average grayscale pixel brightness, ranging from $0 =$ Black to $255 =$ White, revealed an average brightness of $\mu = 138.31$ ($\sigma = 43.55$), centered perfectly in mid-tones, with the interquartile range falling between 108.35 and 168.68.

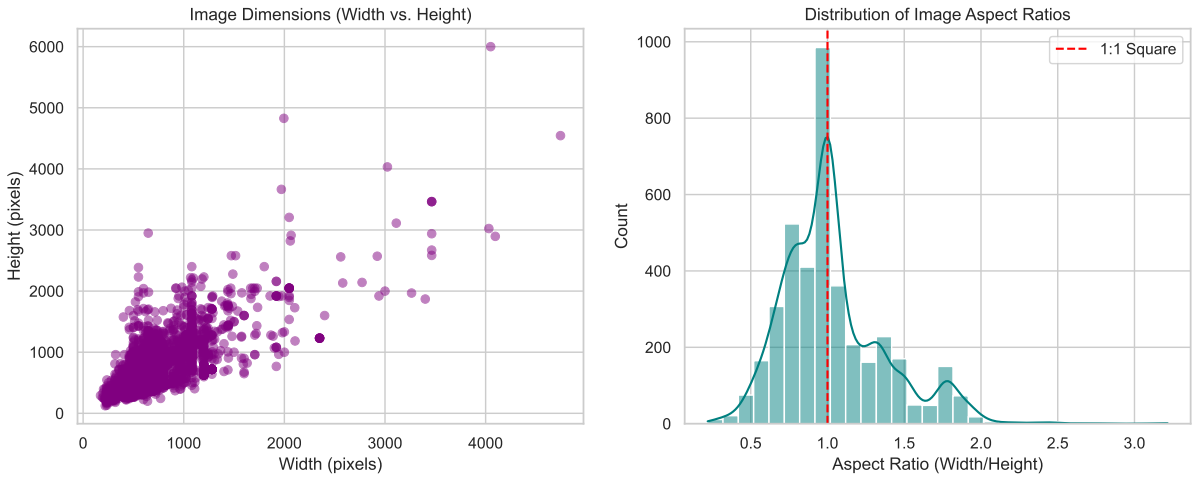

Figure 3: Image dimensions (width vs. height) and the resulting aspect ratio distribution, highlighting a dominant $1:1$ structural format.

Table 3
Descriptive Statistics of Visual Features ($N = 3984$)

Feature	Mean	Std. Dev.	Min	25%	Median	Max
Image Width (px)	713.41	355.18	173.00	500.00	640.00	4744.00
Image Height (px)	745.33	405.74	124.00	480.00	650.00	6000.00
Aspect Ratio (W/H)	1.03	0.34	0.22	0.80	1.00	3.22
Pixel Brightness	138.31	43.55	16.19	108.35	136.03	251.79

4. Methodologies

Our system is a multi-stage pipeline. We first enrich each meme with clean embedded text and a structured visual description extracted by a vision-language model (see Subsection 4.1), then expand the training set through cross-lingual translation (see Subsection 4.2). The enriched data feeds a deep multimodal network which fuses five input streams through cross-attention and FiLM conditioning (see Subsection 4.4), whose physiological branch is warm-started by a pretrained autoencoder (Subsection 4.3) and whose embedding space is shaped by an auxiliary contrastive objective (Subsection 4.5). At inference time, the network is combined with a feature-based SVM through soft-voting (Subsection 4.6). An overview of the complete pipeline is shown in Figure 4.

4.1. Text Enhancer

The textual data and the visual descriptions generation were extracted using the Gemma 4 [10] vision-language model via the Ollama framework. To ensure efficient processing, we used the 4-bit quantized version Effective-4B (e4b) of the model. The inference settings were kept strict, with a 0.0 temperature and a 0.1 top_p, to be able to extract the literal meaning of the text and avoid hallucinations. A fallback re-extraction pass for the items whose JSON output failed to parse used the same strict settings for text transcription but slightly relaxed sampling with a temperature of 0.4 for the visual descriptions.

4.2. Data Augmentation

To address the limitation of the training set size and to improve the cross-lingual robustness, we performed translation-based data augmentation using the NLLB-200 distilled model. For each meme, the cleaned text and the visual description were translated into the other language (e.g., English to Spanish), and the translated copies were appended to the training set with an "_aug" suffix on their identifiers. This approach doubles the available training data and helps the encoders learn language-invariant representations of sexist content. To prevent leakage, training and validation splits were assigned by parent ID on the original de-duplicated dataset, with each augmented copy placed in the same split as its source meme [27].

4.3. Sensor Auto Encoder

The four retained physiological features (i.e., log reaction time, fixation count, saccade count, and heart-rate standard deviation) are low-dimensional and noisy, which makes them prone to being overwhelmed by the high-capacity text and vision encoders during joint training. To obtain a more robust and informative representation, we pretrain a sensor autoencoder in an unsupervised manner prior to the main multimodal training stage.

The autoencoder is a symmetric multilayer perceptron which maps the 4-dimensional standardized sensor vector to a 32-dimensional latent representation through a $4 \to 64 \to 32$ stack with Layer Normalization, GELU activations [31], and dropout. The decoder mirrors this as $32 \to 64 \to 4$ to reconstruct the input. The network is trained for 50 epochs with the Adam optimizer [32] and a mean-squared-error reconstruction objective. Once pretrained, the encoder weights are transferred into the corresponding sensor branch of the multimodal model, where they continue to be finetuned end-to-end. This warm start provides the fusion network with a meaningful 32-dimensional physiological

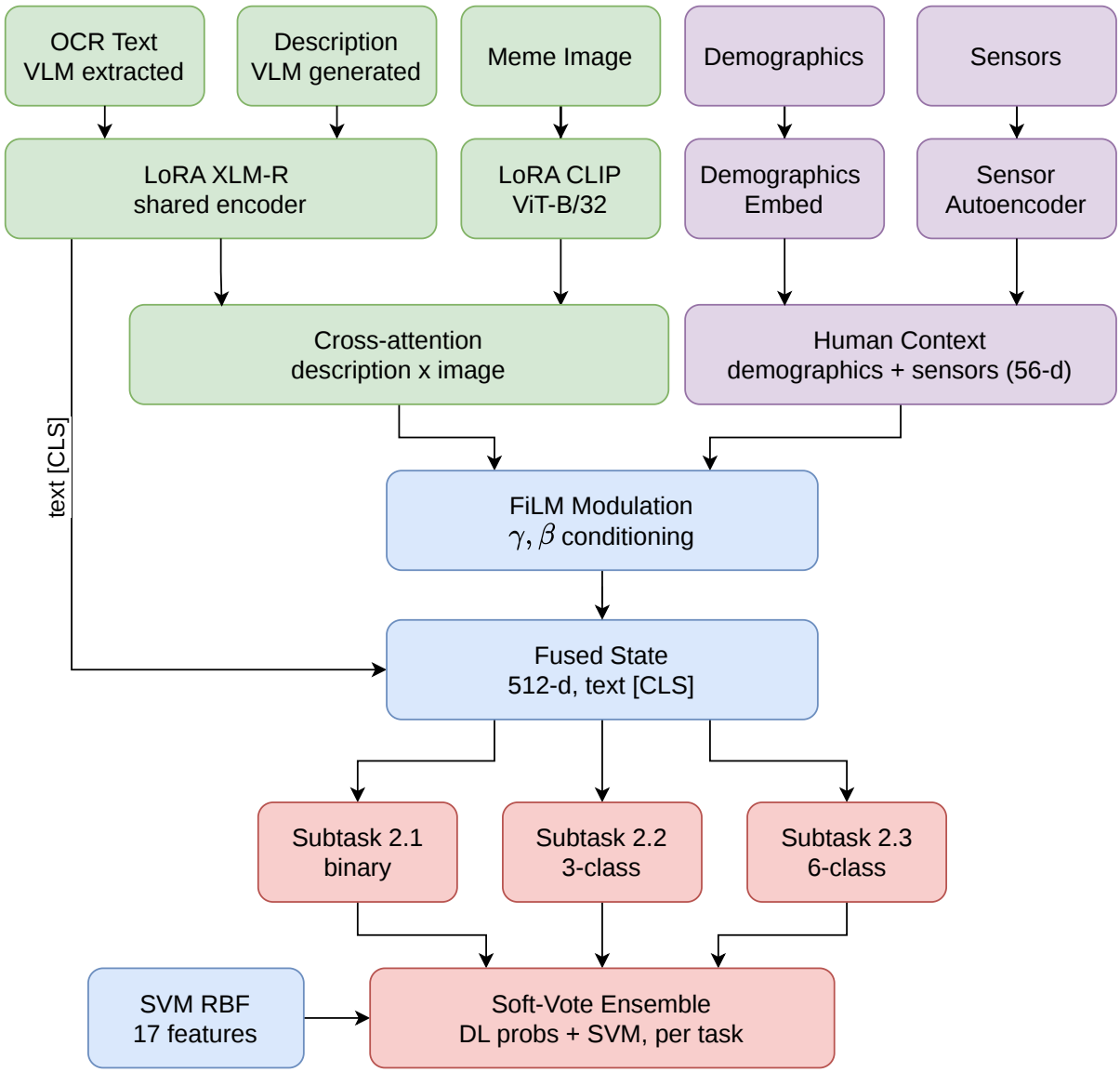

Figure 4: Overview of our architecture. Five input streams, the VLM-extracted embedded text, the VLM-generated visual description, the meme RGB image, the annotator demographics, and the physiological sensor vector, are encoded by a shared LoRA XLM-RoBERTa for embedded text and description, a LoRA CLIP vision encoder for images, demographic embedding layers, and a pretrained sensor autoencoder. The projected description and image-patch sequences are fused by multi-head cross-attention. The resulting representation is modulated through FiLM by the 56-dimensional human-context vector formed from demographics concatenated with the sensor embedding, and is then concatenated with the text [CLS] token to form the 512-dimensional fused state shared across subtasks. Three task heads (binary for Subtask 2.1, three-class for 2.2, six-class multi-label for 2.3) produce per-task predictions, which at inference are combined with a feature-based SVM through soft-voting. Color denotes the meme-content path (green), the annotator-context path (purple), and the task outputs (red).

embedding from the first training step, rather than forcing it to learn one from scratch alongside the much larger language and vision components.

4.4. Cross-Attention Fusion

Our model fuses five input streams: the meme embedded text (extracted via VLM), the VLM-generated visual description, the meme image, annotator demographics, and the physiological sensor vector.

Encoders. The embedded text and the visual description are independently encoded by a shared

LoRA-adapted XLM-RoBERTa [12], while the image is encoded by a LoRA-adapted CLIP ViT-B/32 vision model [13, 14]. LoRA adapters (rank 16, $\alpha = 32$) are inserted into the attention projections of both encoders, keeping the pretrained backbones frozen and training only a small number of additional parameters. The token sequences from each encoder are linearly projected to a common 256-dimensional space with Layer Normalization [33] and dropout.

Grounding the description in the image. To resolve the meaning of a meme, which often emerges from the interplay between what is written and what is shown, we ground the textual description in the visual content through a multi-head cross-attention layer [34]. The projected description sequence serves as the query, and the projected image patch sequence serves as the keys and values. A residual connection followed by Layer Normalization [33] yields a fused vision-language sequence, from which we take the [CLS] position as the fused representation $h_{\text{fused}} \in \mathbb{R}^{256}$.

Human conditioning via FiLM. Annotator demographics are passed through dedicated embedding layers (i.e., gender, age, study level, ethnicity) and averaged across valid annotations of each meme to form a demographic vector. This is concatenated with the 32-dimensional physiological embedding produced by the pretrained sensor encoder (Subsection 4.3), giving a 56-dimensional human-context vector. Following FiLM [9], two linear layers map this vector to per-feature scale γ and shift β parameters, which modulate the fused representation as $h_{\text{mod}} = h_{\text{fused}} \odot (1+\gamma) + \beta$. Both FiLM projections are zero-initialized, so the model begins training as an unconditioned multimodal classifier and progressively learns how much to let annotator context reshape its predictions.

Classification Heads. The modulated representation is concatenated with the text [CLS] embedding to form the final 512-dimensional fused state, which is shared across all subtasks. Each active subtask has its own classification head: a binary head for Subtask 2.1, a three-class head for Subtask 2.2, and a six-class multi-label head for Subtask 2.3. For Subtask 2.1, an auxiliary binary sexism head is attached to h_{fused} to provide an additional training signal, and a 128-dimensional projection of the final fused state feeds the supervised contrastive objective.

4.5. Contrastive Learning

In addition to the primary KL divergence objective, we employ Supervised Contrastive Learning (SupCon) [28] as an auxiliary loss to shape the embedding space. This operates on a 128-dimensional projection of the final fused state as in Equation (1), where z_i are L2-normalized embeddings, $P(i)$ denotes the set of positive pairs, and $\tau = 0.07$ is the temperature parameter.

This addition complements the KL divergence objective by explicitly structuring the embedding space. Memes with the same consensus label are pulled together while those with differing labels are pushed apart. The dual optimization encourages representations that are simultaneously distributionally calibrated and geometrically well-separated.

$$\mathcal{L}_{\text{SupCon}} = -\frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\text{sim}(z_i, z_p)/\tau)}{\sum_{a \neq i} \exp(\text{sim}(z_i, z_a)/\tau)} \tag{1}$$

4.6. Ensembling

The final prediction system combines two complementary classifiers through soft-voting. The first is our deep multimodal model, a LoRA-adapted neural network described in Subsection 4.4 that processes all five input modalities through cross-attention and FiLM conditioning [9]. The second is a feature-based Support Vector Machine (SVM) with an RBF kernel, trained on a 17-dimensional feature vector comprising text stylometry (i.e., length, word count, capitalization ratio, punctuation ratio), demographic ratios (e.g., female and older-annotator proportions), and log-scaled physiological signals (i.e., eye-tracking, heart rate, and EEG bandpower). For Subtask 2.1 the SVM is a binary classifier, for Subtask 2.2 a multiclass classifier, and for Subtask 2.3 a One-vs-Rest classifier producing one binary decision per category. All classifiers use balanced class weights. The ensemble prediction is computed in Equation (2), where the blending weight α and the decision threshold τ are jointly optimized via grid search on the

validation set to maximize the macro F1-score. The SVM acts as an algorithmic regularizer, operating on interpretable handcrafted features and providing a complementary signal that can correct neural model errors.

$$\hat{p} = \alpha \cdot p_{\text{DL}} + (1 - \alpha) \cdot p_{\text{SVM}} \tag{2}$$

5. Experiments

5.1. Experimental Setup

All the experiments were conducted on a single NVIDIA GPU with PyTorch 2.x. For the text augmentation and meme visual interpretation, Gemma 4 [10] was used. The text encoder is XLM-RoBERTa-base, and the image encoder is CLIP ViT-B/32, both adapted with LoRA, using rank 16, α 32, and a 0.1 dropout. Training used AdamW with differential learning rates [35], 3×10^{-5} for fusion heads and 8×10^{-6} for LoRA parameters, with weight decay of 0.1. Gradients were accumulated over 4 steps and clipped to norm 1.

5.2. Results

Tables 4 and 5 report our official EXIST 2026 results under the soft and hard evaluation protocols, respectively, broken down by subtask and language split, alongside the corresponding validation scores computed on the held-out 20% of the training set with PyEvALL. For reference, each table also lists the organizers majority and minority class baselines on the All split.

Our system is most competitive on Subtask 2.2 (i.e., source intention), where it ranks 29th of 114 systems under soft evaluation, indicating that the fused multimodal representation captures the distinction between direct and judgmental sexism reasonably well. Performance on Subtask 2.1 (i.e., binary sexism detection) is mid-range, with the English split ranked 86th of 214 with hard ICM-Norm 0.5496 and F1 0.7336 outperforming the Spanish split, which is ranked 141st of 214 with hard ICM-norm 0.4306 and F1 0.6697 on test. This English-Spanish gap is much less pronounced on validation, where the two splits perform similarly under hard evaluation (ICM-Norm 0.5609 vs. 0.4679), and Spanish even exceeds English on the YES-class F1 (0.7732 vs. 0.7556). The widening of the gap from validation to test suggests that the Spanish portion of the test distribution diverges more from the training distribution than the English portion does, plausibly because the multilingual pretraining of XLM-RoBERTa and CLIP captures English internet sexism cues more idiomatically than Spanish ones. Subtask 2.3 (i.e., fine-grained categorization) is the hardest setting, reflecting the difficulty of the six-way multi-label problem under severe class imbalance, where rare categories receive few positive annotations.

Relative to the organizers baselines, our submissions clear the majority-class system by a wide margin on Subtasks 2.1 and 2.2 under both protocols (e.g., 0.4861 vs. 0.2947 hard ICM-Norm on Subtask 2.1, and 0.3612 vs 0.1369 on Subtask 2.2), confirming that the human-centered conditioning contributes a usable signal rather than noise. Subtask 2.3 is the exception, although the soft score is 0.1568 still exceeds the majority class baseline which is 0, the hard scores collapses onto that baseline exactly with 0.0703 ICM-Norm.

5.3. Augmentation and Ensemble Ablation

Table 6 compares four configurations on Subtask 2.1, isolating the contributions of cross-lingual augmentation and the SVM ensemble. Three of these correspond to our three official submissions. The augmentation-only variant without ensemble was retained for ablation purposes and was not submitted due to limited submission numbers.

The pure deep learning model trained on the non-augmented corpus achieves the highest overall scores on validation with ICM-Norm 0.5210 and F1 Yes 0.7651, outperforming every variant that adds

Table 4
Soft evaluation results for Tasks 2.1, 2.2, and 2.3. For each language split the best submitted run is reported. The test and validation columns are each selected for their own best score and may therefore originate from different runs.

Task	Split	Rank	ICM-Soft	ICM-Soft Norm	CE	Val. ICM-Soft	Val. ICM-Soft Norm	Val. CE
2.1	All	75	−0.5545	0.4109	0.9389	−0.5584	0.4126	0.9414
	EN	75	−0.4360	0.4292	0.9200	−0.4013	0.4372	0.9255
	ES	70	−0.6843	0.3909	0.9568	−0.7408	0.3842	0.9570
2.2	All	29	−1.5152	0.3389	1.4575	−1.4622	0.3475	1.4564
	EN	30	−1.3087	0.3572	1.4479	−1.1936	0.3734	1.4360
	ES	31	−1.7624	0.3169	1.4667	−1.7647	0.3188	1.4764
2.3	All	41	−6.4755	0.1568	–	−6.6050	0.1550	–
	EN	41	−6.2800	0.1607	–	−6.4860	0.1590	–
	ES	38	−6.6920	0.1525	–	−6.7264	0.1505	–
Organizer baselines (test set, All split)								
2.1	Maj.-class	140	−2.3568	0.1212	4.4015	–	–	–
2.1	Min.-class	141	−3.5089	0.0000	5.5672	–	–	–
2.2	Maj.-class	112	−5.0745	0.0000	5.5565	–	–	–
2.2	Min.-class	114	−18.9382	0.0000	8.0245	–	–	–
2.3	Maj.-class	74	−9.8173	0.0000	–	–	–	–
2.3	Min.-class	114	−50.0353	0.0000	–	–	–	–

Table 5
Hard evaluation results for Tasks 2.1, 2.2, and 2.3. For each language split the best submitted run is reported. The test and validation columns are each selected for their own best score and may therefore originate from different runs.

Task	Split	Rank	ICM-Hard	ICM-Hard Norm	F1	Val. ICM-Hard	Val. ICM-Hard Norm	Val. F1
2.1	All	117	−0.0274	0.4861	0.6920	0.0410	0.5210	0.7651
	EN	86	0.0977	0.5496	0.7336	0.1208	0.5609	0.7556
	ES	141	−0.1362	0.4306	0.6697	−0.0610	0.4679	0.7732
2.2	All	71	−0.3992	0.3612	0.3719	−0.3581	0.3709	0.4040
	EN	60	−0.3303	0.3854	0.3825	−0.3588	0.3714	0.4017
	ES	69	−0.4367	0.3479	0.3644	−0.3662	0.3655	0.4033
2.3	All	138	−2.0711	0.0703	0.0919	−1.8981	0.1159	0.1361
	EN	143	−2.0015	0.0747	0.0938	−1.8276	0.1164	0.1389
	ES	130	−2.1173	0.0667	0.0901	−1.9635	0.1148	0.1332
Organizer baselines (test set, All split)								
2.1	Maj.-class	202	−0.4038	0.2947	0.6821	–	–	–
2.1	Min.-class	214	−0.6468	0.1711	0.0000	–	–	–
2.2	Maj.-class	167	−1.0445	0.1369	0.1839	–	–	–
2.2	Min.-class	183	−2.0637	0.0000	0.0697	–	–	–
2.3	Maj.-class	140	−2.0711	0.0703	0.0919	–	–	–
2.3	Min.-class	174	−3.3135	0.0000	0.0318	–	–	–

either augmentation or the SVM ensemble. The same configuration also gives the best test ICM-Norm on the All split with 0.4861.

Cross-lingual augmentation produces a clear language trade-off. Comparing the non-augmented ensemble against the augmented ensemble on validation, augmentation lifts Spanish performance (i.e., $0.4249 \rightarrow 0.4925$ ICM-Norm) while degrading English (i.e., $0.5494 \rightarrow 0.5001$). The same pattern repeats on test, where the augmented ensemble achieves our best English score with ICM-Norm 0.5496 and F1 0.7336 while the non-augmented variants are stronger on Spanish. Augmentation therefore acts as a transfer mechanism, lifting the weaker language at the cost of the stronger one.

The SVM ensemble does not provide a uniform improvement, its effect on validation depends on the training regime. Adding the ensemble to the pure deep learning system lowers performance in the no-augmentation setup ($0.5210 \rightarrow 0.4963$ ICM-Norm) but raises it in the augmentation setup ($0.4561 \rightarrow 0.4963$ ICM-Norm). However, on test the effect partially reverses, where the ensemble paired with augmentation yields the best English score, suggesting that the SVM contributes complementary signal precisely when the deep model has been exposed to noisier translated data. The ensemble is therefore best understood as a regularizer that helps in the noisier training regime but adds variance

Table 6
Ablation study on Subtask 2.1 validation (i.e., subset of training set) and test (i.e., results from the submitted runs) sets with hard evaluation. All three configurations correspond to officially submitted runs. *Aug.* indicates whether cross-lingual translation augmentation was used during training and *Ens.* indicates whether the SVM ensemble was applied at inference. The augmented without ensemble configuration was not submitted for official evaluation.

	Configuration		ICM-Hard Norm			F1 Yes		
Dataset	**Aug.**	**Ens.**	All	EN	ES	All	EN	ES
Validation	–	–	**0.5210**	**0.5609**	0.4679	**0.7651**	**0.7556**	**0.7732**
	✓	✓	0.4963	0.5001	**0.4925**	0.7395	0.7409	0.7382
	–	✓	0.4963	0.5494	0.4249	0.7418	0.7277	0.7532
	✓	–	0.4561	0.4540	0.4582	0.7463	0.7449	0.7477
Test	–	–	**0.4861**	0.5414	**0.4306**	0.6920	0.7167	**0.6697**
	✓	✓	0.4712	**0.5496**	0.3929	**0.6924**	**0.7336**	0.6545
	–	✓	0.4657	0.5235	0.4079	0.6766	0.7012	0.6540
	✓	–	–	–	–	–	–	–

when the underlying DL system is already well fitted. This combination of effects explains why no single submission dominates across all language splits, and why the best test scores per split in Tables 4 and 5 come from different submitted runs.

5.4. Architectural Component Ablation

To assess the contribution of the individual neural-network components, we retrained the system from scratch with each component independently disabled, repeating the procedure with five distinct random seeds per configuration. We probe the contribution of FiLM-based human conditioning, the visual description and image streams both individually and jointly (i.e., "text-only" variant), the supervised contrastive auxiliary loss, the auxiliary sexism head, and the multi-task formulation itself with single-task variants trained on a single subtask. All ablation runs use the non-augmented training corpus and the deep model alone without the SVM ensemble, matching the configuration of our best-performing official submission on the All split. Table 7 reports the mean and standard deviation of hard-evaluation scores across the five seeds, while the Table 8 reports the soft evaluation.

Table 7
Component ablation on the validation set, hard evaluation on both English and Spanish split. Each cell shows mean ± standard deviation over five random seeds. All configurations trained on the non-augmented dataset using only the deep learning model.

Configuration	Task 2.1		Task 2.2		Task 2.3	
	ICM-Norm	F1	ICM-Norm	F1	ICM-Norm	F1
Full system (baseline)	0.4877 ± 0.0066	0.6607 ± 0.0054	0.3517 ± 0.0158	0.3916 ± 0.0105	0.1174 ± 0.0018	0.1339 ± 0.0024
w/o FiLM conditioning	0.4773 ± 0.0209	0.6557 ± 0.0130	0.3438 ± 0.0200	0.3882 ± 0.0132	0.1174 ± 0.0018	0.1339 ± 0.0024
w/o visual description	0.4863 ± 0.0223	0.6571 ± 0.0153	0.3565 ± 0.0193	0.4120 ± 0.0218	0.1206 ± 0.0031	0.1395 ± 0.0060
w/o image	0.4863 ± 0.0890	0.6508 ± 0.0703	0.3255 ± 0.0924	0.3635 ± 0.0832	0.1174 ± 0.0018	0.1339 ± 0.0024
text only	0.4957 ± 0.0210	0.6678 ± 0.0134	0.3405 ± 0.0250	0.3848 ± 0.0143	0.1174 ± 0.0018	0.1339 ± 0.0024
w/o SupCon loss	0.4876 ± 0.0094	0.6618 ± 0.0058	0.3539 ± 0.0143	0.3917 ± 0.0099	0.1174 ± 0.0018	0.1339 ± 0.0024
w/o auxiliary head	0.4827 ± 0.0093	0.6566 ± 0.0093	0.3567 ± 0.0067	0.3967 ± 0.0058	0.1174 ± 0.0018	0.1339 ± 0.0024
Single-task (2.1 only)	0.4850 ± 0.0172	0.6581 ± 0.0122	–	–	–	–
Single-task (2.2 only)	–	–	0.3648 ± 0.0214	0.4035 ± 0.0151	–	–
Single-task (2.3 only)	–	–	–	–	0.1346 ± 0.0246	0.1605 ± 0.0375

We applied paired Welch t-tests against the baseline configuration for every ablation-metric pair across both hard and soft evaluation, yielding a family of 63 tests, and correct for multiple comparisons using the Holm-Bonferroni step-down procedure at $\alpha = 0.05$. No ablation produced an effect that survives this correction. The closest case is the increase in soft ICM-Soft Norm on Subtask 2.3 when the visual description is removed with uncorrected Welch $p = 0.001$, but this fails the Holm threshold

Table 8
Component ablation on the validation set, soft evaluation on both English and Spanish split. Each cell shows mean ± standard deviation over five random seeds. All configurations are trained on the non-augmented dataset, using only the deep learning model. Cross-Entropy is not reported for Task 2.3.

Configuration	Task 2.1		Task 2.2		Task 2.3	
	ICM-Soft Norm	CE	ICM-Soft Norm	CE	ICM-Soft Norm	CE
Full system (baseline)	0.3935 ± 0.0095	0.9636 ± 0.0024	0.3342 ± 0.0078	1.4779 ± 0.0098	0.1464 ± 0.0039	–
w/o FiLM conditioning	0.3893 ± 0.0077	0.9641 ± 0.0043	0.3306 ± 0.0067	1.4791 ± 0.0112	0.1450 ± 0.0034	–
w/o visual description	0.4031 ± 0.0084	0.9589 ± 0.0079	0.3415 ± 0.0100	1.4710 ± 0.0116	0.1635 ± 0.0060	–
w/o image	0.3884 ± 0.0450	0.9477 ± 0.0256	0.3319 ± 0.0273	1.4610 ± 0.0279	0.1542 ± 0.0174	–
text only	0.3837 ± 0.0132	0.9495 ± 0.0062	0.3278 ± 0.0095	1.4633 ± 0.0092	0.1534 ± 0.0090	–
w/o SupCon loss	0.3944 ± 0.0099	0.9628 ± 0.0025	0.3347 ± 0.0077	1.4773 ± 0.0099	0.1465 ± 0.0038	–
w/o auxiliary head	0.3990 ± 0.0045	0.9610 ± 0.0045	0.3371 ± 0.0061	1.4767 ± 0.0105	0.1481 ± 0.0034	–
Single-task (2.1 only)	0.3889 ± 0.0024	0.9606 ± 0.0060	–	–	–	–
Single-task (2.2 only)	–	–	0.3406 ± 0.0099	1.4701 ± 0.0122	–	–
Single-task (2.3 only)	–	–	–	–	0.1553 ± 0.0407	–

($0.05/63 \approx 0.00079$). We therefore cannot conclude that any individual architectural component carries a statistically detectable benefit at our seed count. The system gains appear to arise from the integration of components rather than any single contribution we can isolate.

Two qualitative observations remain. First, the no-image configuration exhibits dramatically higher variance than every other configuration with standard deviation of 0.089 on Subtask 2.1 ICM-Norm against ≤ 0.023 for the others. This suggests that the image stream provided training stability even when its mean contribution is small. Second, the multi-task baseline and most ablations of it produce essentially identical performance on Subtask 2.3 with 0.1174 ± 0.0018 on hard ICM-Norm. This indicates that the Subtask 2.3 head is collapsing to a near-trivial solution across configurations, only single-task training on 2.3 with ICM-Norm 0.1346 ± 0.0246 breaks this pattern, suggesting that multi-task inference, not architectural choices, is the binding constraint on Subtask 2.3 performance.

5.5. Discussions

Our results partially support the central hypothesis of this work: annotator physiological and demographic context is statistically associated with the perceived sexism of meme content, and a system that conditions on it stays above the trivial baseline on Subtasks 2.1 and 2.2. At the same time, our multi-seed ablation conditions tempers the claim, when each component is isolated, none produces an effect distinguishable from run-to-run variance after correction. We therefore frame the human-centered conditioning as a useful component of an integrated system rather than as an independently decisive module, and we report the negative ablation openly as a finding.

Comparing the same configuration on validation and test for the All split (i.e., both English and Spanish) with hard ICM-Norm 0.5210 vs. 0.4861, the model exhibits modest validation overfitting but no catastrophic distribution shift. The English-Spanish gap on test with 0.5496 vs 0.4306 hard ICM-Norm is substantially wider than on validation with 0.5609 vs 0.4679, suggesting that the Spanish portion of the test distribution diverges from training more than English portion does, consistent with the multilingual pretraining of our text and vision backbones favoring English idioms of online sexism. The strong relative standing on Subtask 2.2 which is ranked 29th of 114 overall for soft and 71st of 183 for hard evaluation indicates that the cross-attention grounding of the visual description in the image is helpful for reasoning about communicative intent, where both modalities are jointly informative.

The gap between hard and soft performance is also informative. Because we explicitly optimize a soft KL divergence over the annotator label distribution, the model is trained to reproduce disagreement rather than collapse to a majority vote, which is better aligned with the soft evaluation protocol. This is consistent with the learning from disagreement framing [6, 7] that motivated our approach.

5.6. Limitations

Our approach has several limitations. First, physiological and demographic signals are aggregated by averaging across all annotators of a meme, which discards individual-level variation. A viewer-specific model could better capture the subjectivity the dataset was designed to expose.

Second, the eye-tracking features exhibit extreme multicollinearity (e.g., Fixations and Saccades with $\rho \approx 1.0$), so the four-dimensional sensor vector likely carries less independent information than its dimensionality suggests.

Third, EEG signals were not individually significant in our statistical analysis and therefore do not inform the neural branch, entering only through the classical SVM, leaving open the question of whether richer temporal EEG modeling, rather than aggregated bandpower features, might unlock signal that our current setup leaves unused.

Fourth, our multi-seed ablation (Tables 7 and 8) finds that no individual architectural component produces an effect distinguishable from run-to-run variance at $n = 5$ seeds after Holm-Bonferroni corrections. While this rules out large isolated contributions, smaller real effects could exist below our detection threshold and would require either larger seed counts or a less noisy evaluation regime to confirm.

Fifth, the model class coverage is uneven on the multi-class subtasks, on validation the per-class F1 for the JUDGEMENTAL category from Task 2.2 and for every minority category in Task 2.3 is 0.0, meaning the model never predicts these categories despite their presence in the training set, broken only by single-task training on 2.3 alone, which indicates that multi-task interference rather than architectural design is the binding constraint on fine-grained categorization performance. This is a symptom of severe class imbalance combined with a soft objective that does not explicitly penalize ignoring rare classes, and explains why our overall Subtask 2.3 ICM-Norm remains close to the trivial baseline.

Sixth, our cross-lingual augmentation strategy lifts Spanish performance but degrades English seen in Table 6, a more selective approach, such as translation-quality filtering or back-translation consistency checks, might preserve the benefit without the cost.

Seventh, our handling of tall, storytelling-format memes relies on square padding, which compresses their content. Splitting such memes into sub-images was not explored and could improve both OCR and visual description quality extraction. Finally, the deep model and the SVM are optimized separately and combined using grid search, rather than trained jointly end-to-end.

6. Conclusions

We presented a human-centered multimodal system for sexism detection in memes that combines textual, visual, demographic, and physiological modalities through a FiLM-conditioned [9] cross-attention architecture. Our approach treats sexism identification as a distribution learning problem, using a soft-label KL divergence to capture the inherent subjectivity of human annotators. The system is most competitive on Subtask 2.2 (source intention), where it ranks 29th of 114 overall under soft evaluation, and it stays above the organizers baseline on Subtasks 2.1 and 2.2, confirming that the human-centered conditioning contributes signal rather than noise.

The statistical analysis performed on the EXIST 2026 dataset revealed that annotator physiological responses are significantly associated with the perceived sexism of meme content. By incorporating these signals through a pretrained sensor autoencoder with FiLM modulation, the model learns to adjust its predictions based on the measured cognitive friction that sexist content produces in human viewers. At the same time, our multi-seed ablation under multiple comparison control found that no individual component can be shown to carry a statistically detectable benefit at our seed count. The system signal appears to arise from the integration of the components rather than from any single module. We view this as a useful, honestly reported result for future work that adds physiological and demographic signals to subjective NLP tasks.

In future work, we would like to explore individual-level sensor modeling rather than meme-level aggregation, attention between the Gemma extracted text and image modalities, larger seed counts and

lower noise evaluation regimes to resolve the small per-component effects, and adjusting a universal model for all of EXIST Task 2 subtasks that learns from all the labels before being specialized for each task.

Acknowledgments

The research presented in this paper is supported in part by The Academy of Romanian Scientists, through the funding of the project "NetGuardAI: Intelligent system for harmful content detection and immunization on social networks" (AOȘR-TEAMS-IV).

Declaration on Generative AI

During the preparation of this work, the authors used Claude Opus 4.6 for rephrasing in order to improve clarity and style. 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. Isaac, Reinforcement of sexism through memes, National Dialogue on Gender-Based Cyber Violence (2018) 1–13.

[2] S. Dutta, D. Singh, Reinforcement of sexism through memes: Harassment and the current digital culture, Name Page No. Analyzing Trends in Variations of Dow Jones Stocks and Cryptocurrency Prices (2021) 165–182.

[3] C. Argüello-Gutiérrez, A. Cubero, F. Fumero, D. Montealegre, P. Sandoval, V. Smith-Castro, I'm just joking! perceptions of sexist humour and sexist beliefs in a latin american context, International Journal of Psychology 58 (2023) 91–102. doi:https://doi.org/10.1002/ijop.12884.

[4] M. Duggan, Online harassment 2017 (2017). URL: http://hdl.handle.net/20.500.11990/10.

[5] 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://proceedings.neurips.cc/paper/2020/hash/1b84c4cee2b8b3d823b30e2d604b1878-Abstract.html.

[6] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, J. Artif. Intell. Res. 72 (2021) 1385–1470. URL: https://doi.org/10.1613/jair.1.12752. doi:10.1613/JAIR.1.12752.

[7] B. Wu, Y. Li, Y. Mu, C. Scarton, K. Bontcheva, X. Song, Don't waste a single annotation: improving single-label classifiers through soft labels, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5347–5355. URL: https://aclanthology.org/2023.findings-emnlp.355/. doi:10.18653/v1/2023.findings-emnlp.355.

[8] I. Árcos, P. Rosso, E. Gomis-Vicent, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and EEG signals, Proceedings of the 2026 Joint International Conference on Computationa Linguistics, Language Resources and Evaluation (LREC 2026) abs/2602.23862 (2026). URL: https://doi.org/10.48550/arXiv.2602.23862. doi:10.48550/ARXIV.2602.23862. arXiv:2602.23862.

[9] E. Perez, F. Strub, H. de Vries, V. Dumoulin, A. C. Courville, Film: Visual reasoning with a general conditioning layer, in: S. A. McIlraith, K. Q. Weinberger (Eds.), Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, AAAI Press, 2018, pp. 3942–3951. URL: https://doi.org/10.1609/aaai.v32i1.11671. doi:10.1609/AAAI.V32I1.11671.

[10] Google DeepMind, Gemma 4 model card, https://ai.google.dev/gemma/docs/core/model_card_4, 2026. Accessed: 2026-06-03.

[11] S. Kullback, R. A. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 79–86. URL: https://api.semanticscholar.org/CorpusID:120349231. doi:10.1214/aoms/1177729694.

[12] 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.

[13] 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: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, Proceedings of Machine Learning Research, PMLR, 2021, pp. 8748–8763. URL: http://proceedings.mlr.press/v139/radford21a.html.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, W. Chen, Lora: Low-rank adaptation of large language models, CoRR abs/2106.09685 (2021). URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[15] S. Holm, A simple sequentially rejective multiple test procedure, Scandinavian Journal of Statistics 6 (1979) 65–70. URL: https://api.semanticscholar.org/CorpusID:122415379.

[16] A. Jha, R. Mamidi, When does a compliment become sexist? analysis and classification of ambivalent sexism using twitter data, in: D. Hovy, S. Volkova, D. Bamman, D. Jurgens, B. O'Connor, O. Tsur, A. S. Doğruöz (Eds.), Proceedings of the Second Workshop on NLP and Computational Social Science, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 7–16. URL: https://aclanthology.org/W17-2902/. doi:10.18653/v1/W17-2902.

[17] R. P. De Pelle, V. P. Moreira, Offensive comments in the brazilian web: a dataset and baseline results, in: Brazilian Workshop on Social Network Analysis and Mining (BRASNAM), SBC, 2017, pp. 510–519. doi:https://doi.org/10.5753/brasnam.2017.3260.

[18] F. Gasparini, I. Erba, E. Fersini, S. Corchs, Multimodal classification of sexist advertisements, in: C. Callegari, M. van Sinderen, P. Novais, P. G. Sarigiannidis, S. Battiato, Á. S. S. de León, P. Lorenz, M. S. Obaidat (Eds.), Proceedings of the 15th International Joint Conference on e-Business and Telecommunications, ICETE 2018 - Volume 1: DCNET, ICE-B, OPTICS, SIGMAP and WINSYS, Porto, Portugal, July 26-28, 2018, SciTePress, 2018, pp. 565–572. URL: https://doi.org/10.5220/0006859405650572. doi:10.5220/0006859405650572.

[19] P. Parikh, H. Abburi, N. Chhaya, M. Gupta, V. Varma, Categorizing sexism and misogyny through neural approaches, ACM Trans. Web 15 (2021) 17:1–17:31. URL: https://doi.org/10.1145/3457189. doi:10.1145/3457189.

[20] L. Plaza, J. Carrillo-de-Albornoz, I. Árcos, M. Aloy-Mayo, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2026: Learning with disagreement for sexism identification and characterization in memes and tiktok videos (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[21] L. Plaza, J. Carrillo-de-Albornoz, I. Árcos, M. Aloy-Mayo, P. Rosso, D. Spina, E. Amigó, J. Gonzalo, R. Morante, Overview of EXIST 2026: Learning with disagreement for sexism identification and characterization in memes and tiktok videos (extended overview), in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[22] L. Plaza, J. Carrillo-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: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. D. Nunzio, L. Soulier, P. Galuscáková, A. G. S. de Herrera, G. Faggioli, N. Ferro

(Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II, Lecture Notes in Computer Science, Springer, 2024, pp. 93–117. URL: https://doi.org/10.1007/978-3-031-71908-0_5. doi:10.1007/978-3-031-71908-0_5.

[23] L. Plaza, J. Carrillo-de-Albornoz, I. Árcos, 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: J. Carrillo-de-Albornoz, A. G. S. de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Lecture Notes in Computer Science, Springer, 2025, pp. 266–289. URL: https://doi.org/10.1007/978-3-032-04354-2_16. doi:10.1007/978-3-032-04354-2_16.

[24] R. Cao, M. S. Hee, A. Kuek, W. Chong, R. K. Lee, J. Jiang, Pro-cap: Leveraging a frozen vision-language model for hateful meme detection, in: A. E. Saddik, T. Mei, R. Cucchiara, M. Bertini, D. P. T. Vallejo, P. K. Atrey, M. S. Hossain (Eds.), Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023, ACM, 2023, pp. 5244–5252. URL: https://doi.org/10.1145/3581783.3612498. doi:10.1145/3581783.3612498.

[25] G. Rizzi, F. Gasparini, A. Saibene, P. Rosso, E. Fersini, Recognizing misogynous memes: Biased models and tricky archetypes, Inf. Process. Manag. 60 (2023) 103474. URL: https://doi.org/10.1016/j.ipm.2023.103474. doi:10.1016/J.IPM.2023.103474.

[26] P. Italiani, D. Gimeno-Gómez, L. Ragazzi, G. Moro, P. Rosso, Memeweaver: Inter-meme graph reasoning for sexism and misogyny detection (2026) 2120–2134. URL: https://aclanthology.org/2026.findings-eacl.111/.

[27] M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Y. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, CoRR abs/2207.04672 (2022). URL: https://doi.org/10.48550/arXiv.2207.04672. doi:10.48550/ARXIV.2207.04672. arXiv:2207.04672.

[28] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Advances in neural information processing systems 33 (2020) 18661–18673. doi:https://doi.org/10.48550/arXiv.2004.11362.

[29] J. Z. Lim, J. Mountstephens, J. Teo, Emotion recognition using eye-tracking: Taxonomy, review and current challenges, Sensors 20 (2020) 2384. URL: https://doi.org/10.3390/s20082384. doi:10.3390/S20082384.

[30] B. Fu, C. Gu, M. Fu, Y. Xia, Y. Liu, A novel feature fusion network for multimodal emotion recognition from eeg and eye movement signals, Frontiers in Neuroscience 17 (2023) 1234162. doi:10.3389/fnins.2023.1234162.

[31] D. Hendrycks, K. Gimpel, Gaussian error linear units (gelus), arXiv preprint arXiv:1606.08415 (2016). doi:https://doi.org/10.48550/arXiv.1606.08415.

[32] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, in: Y. Bengio, Y. LeCun (Eds.), 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings, 2015. URL: http://arxiv.org/abs/1412.6980.

[33] L. J. Ba, J. R. Kiros, G. E. Hinton, Layer normalization, CoRR abs/1607.06450 (2016). URL: http://arxiv.org/abs/1607.06450. arXiv:1607.06450.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, CoRR abs/1706.03762 (2017). URL: http://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[35] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.