Gated Multimodal Fusion with Neurophysiological Signals for Sexism Detection in Memes

Notebook for the EXIST Lab at CLEF 2026

Diego Zuniga[1,2,*]

[1]Université Côte d'Azur, Nice, France
[2]PRHLT Research Center, Universitat Politècnica de València, València, Spain

Abstract

We present Cloud-17's submission to the EXIST 2026 shared task on sexism identification in memes (Tasks 2.1–2.3, meme split only). Building on the human-centred multimodal framework of Arcos et al. [1], we introduce a *Gated Fusion* architecture that combines text (mDeBERTa-v3), image (ALIGN), EEG band-power features, and eye-tracking signals through learnable scalar gates conditioned on the text representation. We compare three fusion paradigms: early concatenation (*Early Fusion*), *Gated Fusion* (primary system), and *Cross Attention Gated* (an ablation), and submit three runs corresponding to *Gated Fusion*, an ensemble of *Gated Fusion*+*Cross Attention*, and *Cross Attention* alone. The gates provide per-meme modality importance scores without requiring post-hoc attribution. Across validation folds, our *Gated Fusion* consistently learns the hierarchy $\beta(\text{image}) \approx 0.98 \gg \alpha(\text{EEG}) \approx 0.481 \gg \lambda(\text{ET}) \approx 0.027$ for the non-textual modalities, with text serving as the residual anchor of the representation. This suggests that visual content dominates sexism perception in memes, EEG adds complementary neural signal, and eye-tracking contributes minimally once EEG is present. Our *Gated Fusion* model achieves macro F1 of 0.616 / 0.549 / 0.380 on Tasks 2.1/2.2/2.3 respectively, with *Cross Attention* leading on the categorisation task (F1 0.397).

Keywords

Sexism detection, Multimodal fusion, Gated fusion, EEG, Eye-tracking, Modality interpretability, EXIST 2026, Memes, Neurophysiological signals

1. Introduction

Memes have become a primary vehicle for sexist content on social media, combining text, images and cultural context in ways that have challenged purely textual classifiers [2]. The EXIST 2026 shared task [3, 4] is the first benchmark to pair memes with *neurophysiological signals*: Electroencephalography (EEG), eye tracking (ET) and heart rate (HR) recorded while human subjects viewed each meme.

Whether such physiological signals carry information beyond what text and image already provide is an open question. Arcos et al. [1] first showed that cross-attention over physiological signals can improve sexism categorization on EXIST 2025, especially on rare classes such as Misogyny. However, their fusion mechanism does not yield an explicit measure of *how much* each modality contributes to the prediction. Such a measure is valuable both for model interpretability and for guiding future multimodal designs.

We introduce a *Gated Fusion* architecture in which scalar gates, conditioned on the text representation, modulate the contribution of the other modalities (image, EEG, ET) at inference time. Trained on EXIST 2026, the model autonomously learns a clear modality hierarchy: $\beta(\text{Image}) \approx 0.98$, $\alpha(\text{EEG}) \approx 0.481$, $\lambda(\text{ET}) \approx 0.027$. Image content dominates, EEG provides a useful complementary signal, and eye tracking is consistently suppressed. This is a finding that the underlying cross-attention baseline does not show.

Contributions. This paper makes three contributions:

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
diego.zuniga-blassio@etu.univ-cotedazur.fr (D. Zuniga)

1. A gated multimodal fusion architecture for sexism detection in memes that integrates text, image, EEG and ET signals through learnable scalar gates conditioned on the textual representation.
2. An empirical ranking of non-textual modality relevance for the task: image $\gg$ EEG $\gg$ ET, with text as the residual anchor of the fused representation, robust across folds.
3. A structured comparison of three fusion paradigms: early concatenation, gated fusion, and cross-attention with gates across all three EXIST 2026 meme subtasks (2.1, 2.2, 2.3).

2. Related Work

Sexism detection in social media. The EXIST shared task [5, 6, 3] has progressively moved from tweet binary classification to multilabel categorization with soft, disagreement-aware labels [7]. The top models consistently rely on large pre-trained multilingual models (XLM-RoBERTa, mDeBERTa) with task-specific fine-tuning. Multimodality entered the task in EXIST 2024 with memes, combining textual content with visual context.

Multimodal hate speech detection. Fusing text and image for meme classification was popularised by the Hateful Memes Challenge [2]. Subsequent work has shown that vision-language pre-training (CLIP, ALIGN, Flamingo) [8, 9, 10] considerably improves meme understanding [11]. The most recent approaches fine-tune vision-language models on hateful content with prompt-based supervision. Across these methods, fusion is typically performed either by concatenation, by cross-attention, or by feeding both modalities into a unified transformer. More recently, Italiani et al. [12] proposed inter-meme graph reasoning to capture relational context across memes for sexism and misogyny detection.

Gated multimodal fusion. Arevalo et al. [13] introduced Gated Multimodal Units (GMU), which use sigmoid gates to learn the contribution of each modality at fusion time. This idea has since been adopted in audio-visual and multimodal learning more broadly. Compared to attention-based fusion, scalar gates offer two practical advantages: lower parameter count and direct interpretability of modality weights. To our knowledge, gated fusion has not been applied to multimodal sexism detection or to neurophysiology-augmented hate speech tasks.

Physiological signals for content moderation. EEG and eye-tracking have a long history in affective computing for estimating emotional response to multimedia stimuli [14]. Within the EXIST framework, Arcos et al. [1] were the first to integrate EEG, ET and HR into a multimodal sexism detection pipeline. They treat each subject's physiological response as a separate token and apply cross-attention between subject tokens and meme representations, reporting AUC improvements of up to 26% on Misogyny detection.

3. Dataset and Preprocessing

3.1. The EXIST 2026 meme dataset

The EXIST 2026 meme dataset comprises 3,984 training and 1,053 test memes in English and Spanish. Annotations follow a hierarchical scheme across three subtasks. **Task 2.1** asks whether a meme is sexist (binary). **Task 2.2**, applied only to sexist memes, distinguishes between *direct* sexism (explicit, intentional promotion of sexist ideas) and *judgemental* sexism (criticising, denouncing, or reporting sexist behaviour without endorsing it). **Task 2.3**, also restricted to sexist memes, assigns one or more categories: *Ideological Inequality*, *Stereotyping-Dominance*, *Objectification*, *Sexual Violence*, and *Misogyny-Non-Sexual-Violence*.

Because sexism perception is inherently subjective and depends on the context, each meme was independently annotated by six crowdsourced workers. This design explicitly acknowledges perspectivism:

workers may legitimately disagree on whether content is sexist, what its intent is, or which category applies. To preserve this disagreement, labels are provided in two forms: *hard labels* (majority vote) and *soft labels* (proportion of annotators selecting each option). Our models were trained on hard labels and evaluated using macro-averaged F1.

Beyond the image and the extracted OCR text, each meme is paired with physiological signals. Three modalities were recorded simultaneously: EEG, Eye Tracking, and Heart Rate.

Following preliminary ablation experiments, Heart Rate was excluded from our final models due to marginal contribution.

3.2. EEG Features

The EXIST 2026 dataset provides EEG recordings from a 16-electrode cap. Electrodes cover six anatomical regions: *frontal-polar* (Fp1, Fp2), *frontal* (F3, F4, F7, F8), *central* (C3, C4), *temporal* (T7, T8), *parietal* (P3, P4, P7, P8), and *occipital* (O1, O2). For each electrode, the dataset provides band-power estimates across five frequency bands (Delta, Theta, Alpha, Beta, and Gamma) averaged over the full viewing period rather than raw time-series. This yields 80 base features per subject (16×5) and is available for at least two participants who viewed each meme.

Preprocessing. Raw EEG power varies substantially across individuals due to anatomical factors unrelated to cognitive processing (e.g., skull thickness, electrode impedance). We therefore z-score each subject's 80 band-power features independently before any cross-subject aggregation. This normalisation ensures that the subsequent average reflects relative neural response patterns rather than individual baseline shifts. Then features are averaged across the participants, yielding a single EEG representation per meme.

Spatial feature engineering. Our raw channel powers capture local cortical activity but lack explicit spatial structure. Since sexism perception is a distributed cognitive process engaging prefrontal (social cognition), temporal (language), and occipital (visual) regions, we augmented the base features with representations that encode inter-regional relationships:

- **Regional means** (30 features): mean power per anatomical region $\times$ band;
- **Global means** (5): mean across all 16 channels per band.
- **Lateralization** (10): left- vs. right-hemisphere mean per band (8 electrodes each side).
- **Global band differences** (9): asymmetry indices between selected band pairs.
- **Regional differences** (19): cross-band differences within anatomical regions.
- **Log-ratios** (30): $\log\left(\frac{|\bar{B}_{\mathrm{num},r}|+\varepsilon}{|\bar{B}_{\mathrm{den},r}|+\varepsilon}\right)$ for five neurophysiologically motivated ratio pairs (e.g., θ/α, β/α) across the six regions, following standard biomarker conventions in affective EEG research [14].

These **handcrafted** spatial features provide the model with an explicit topological view of the brain's response. A natural extension, treating each of the 16 channels as an independent token with learned inter-channel attention, is left as future work (Section 8).

3.3. Eye Tracking Features

Eye tracking captures the visual attention where subjects directed their gaze and for how long. This provides content-specific behavioural signals that complement EEG's neurophysiological perspective. We retain four measures per subject:

- **Reaction time**: elapsed time from meme onset to first fixation. A longer delay suggests the content required more processing before engaging attention.
- **Fixation count**: number of discrete gaze fixations. How thoroughly the meme was visually explored.

- **Fixation duration**: mean duration of each fixation. Longer fixations indicate more sustained attention to specific regions of the meme.
- **Saccade count**: number of rapid gaze shifts.

These measures together encode the temporal dynamics and spatial breadth of how subjects processed each meme, reflecting underlying cognitive processes relevant to content perception.

Preprocessing. Following the same procedure as EEG, features are z-scored per subject to remove inter-individual baseline differences, then averaged across the participants per meme. Unit conversions were applied before z-scoring (reaction time: ms $\rightarrow$ s; fixation duration: ns $\rightarrow$ ms). Both the mean and standard deviation across participants are retained, as inter-subject variability in viewing behaviour is itself informative: high disagreement in reaction time or fixation count may reflect ambiguity in the meme's visual significance. This yields an 8 dimensional ET representation per meme.

4. System Description

We evaluate three multimodal architectures of increasing complexity, all built on the same encoders and task heads for fair comparison.

1. *Early Fusion:* It serves as our baseline, concatenation with no explicit modality weighting.
2. *Gated Fusion:* It is our primary contribution: a gated architecture where the model learns modality importance scores.
3. *Cross Attention Gated:* We added bidirectional cross-attention before gating, serving as an ablation experiment.

Training details are described in Section 5.

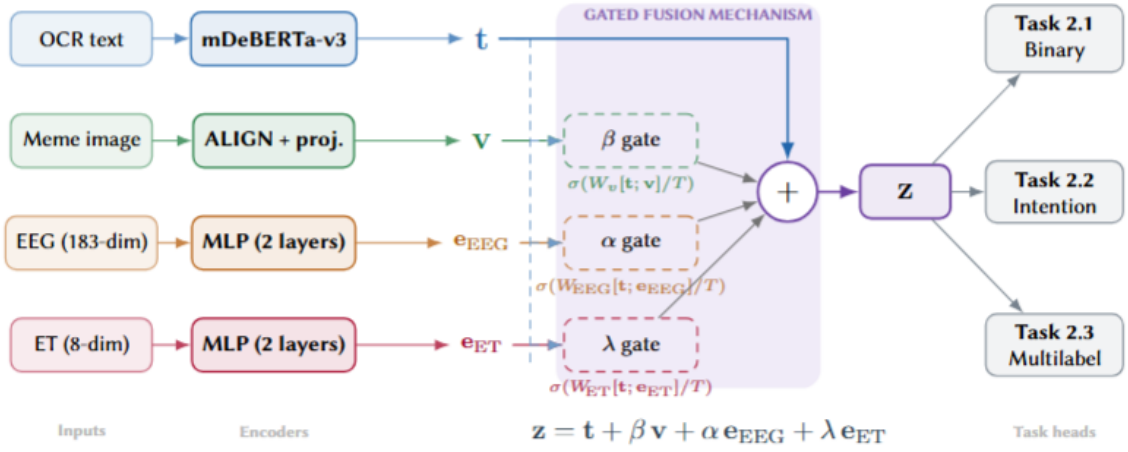

Figure 1: Gated Fusion architecture. Text (mDeBERTa-v3) serves as the *residual anchor* and conditions three sigmoid gates (dashed lines) that control the contribution of each non-textual modality. Temperature *T=0.3* sharpens gate outputs; gate values $(\beta, \alpha, \lambda) \in (0, 1)$ are directly interpretable as modality importance scores.

4.1. Shared Components

Text encoder. `microsoft/mdeberta-v3-base` [15] encodes the OCR-extracted meme text. We fine-tune all layers and extract the `[CLS]` representation ($d_t = 768$).

Vision encoder. `kakaobrain/align-base` [9] encodes the meme image. We unfreeze the last three transformer blocks and project the pooled visual representation to $d_v = 768$ via a linear layer.

Physio encoders. Separate two-layer MLPs with LayerNorm and ReLU encode the EEG features ($d_{\text{EEG}} = 183$) and ET features ($d_{\text{ET}} = 8$), both projected to $d = 768$.

Multi-task heads. Three task-specific classifiers operate on the fused representation: $\mathcal{H}_{2.1}$ (binary), $\mathcal{H}_{2.2}$ (binary, masked for non-sexist memes), $\mathcal{H}_{2.3}$ (5-class multilabel, masked for non-sexist memes). The joint loss is:

$$\mathcal{L} = 0.40\,\mathcal{L}_{2.1} + 0.35\,\mathcal{L}_{2.2} + 0.25\,\mathcal{L}_{2.3}$$

$\mathcal{L}_{2.1}$ and $\mathcal{L}_{2.2}$ use binary cross-entropy; $\mathcal{L}_{2.3}$ uses BCE with logits for multilabel classification. The weights were set manually to reflect task difficulty and the hierarchical dependency between tasks: Task 2.1 receives the highest weight as the primary binary decision upon which Tasks 2.2 and 2.3 are conditioned; Task 2.3 receives the lowest weight as the most fine-grained and sparse multilabel subtask.

4.2. Run 1: Early Fusion

As a concatenation baseline, all four modality representations are combined into a single vector and passed through a fusion MLP before the task heads:

$$\mathbf{z} = \text{MLP}_{\text{fuse}}([\mathbf{t};\, \mathbf{v};\, \mathbf{e}_{\text{EEG}};\, \mathbf{e}_{\text{ET}}])$$

where MLP_{fuse} consists of a linear projection, LayerNorm, and ReLU activation, projecting the $4d$-dimensional concatenation back to $d = 768$. This architecture imposes no content-adaptive modality weighting: every meme contributes equally from all four modalities regardless of which signals are informative for that specific instance. Therefore *Early Fusion* serves as a strong baseline that quantifies the benefit of the adaptive gating introduced in Run 2.

4.3. Run 2: Gated Fusion

The core limitation of *Early Fusion* is that it assigns fixed, equal weight to all modalities regardless of meme content. *Gated Fusion* addresses this by learning adaptive scalar gates that modulate each modality's contribution at inference time. Note that neither EarlyFusion nor GatedFusion performs cross-modal interaction between modalities, each representation remains independent before fusion. This is a deliberate design choice in GatedFusion: by keeping representations independent, the gates can attribute contribution cleanly to each modality, preserving interpretability. CrossAttnGated (Section 4.4) relaxes this constraint by introducing cross-attention before gating, at the cost of gate stability.

Gate computation. Each gate is a scalar in $(0, 1)$ produced by a sigmoid over a learned linear projection of the text embedding *jointly* with the corresponding modality embedding:

$$\beta = \sigma\left(\frac{W_v\,[\mathbf{t};\, \mathbf{v}]}{T}\right), \quad \alpha = \sigma\left(\frac{W_{\text{EEG}}\,[\mathbf{t};\, \mathbf{e}_{\text{EEG}}]}{T}\right), \quad \lambda = \sigma\left(\frac{W_{\text{ET}}\,[\mathbf{t};\, \mathbf{e}_{\text{ET}}]}{T}\right)$$

where $W_v, W_{\text{EEG}}, W_{\text{ET}} \in \mathbb{R}^{1\times 1536}$, $T = 0.3$ is a temperature parameter, and σ is the sigmoid function.

Gated fusion. Text serves as the most reliable signal, the anchor of the representation, since OCR always yields some output, while the remaining modalities contribute additively, scaled by their gates:

$$\mathbf{z} = \mathbf{t} + \beta\,\mathbf{v} + \alpha\,\mathbf{e}_{\text{EEG}} + \lambda\,\mathbf{e}_{\text{ET}}$$

This additive formulation keeps the text representation intact regardless of gate values: even if all gates collapse to zero, $\mathbf{z} = \mathbf{t}$ and the model degrades gracefully to a text-only baseline.

Interpretability. The gate values (β, α, λ) are directly interpretable as modality importance. Crucially, they are *instance-level*: the same model can assign high β to a visually explicit meme and low β to one where the sexism resides entirely in the text. We analyse the empirical distribution of gate values across the 1,053 test memes in Section 7.

4.4. Run 3: Cross-Attention Gated

In *Gated Fusion*, each gate assesses modality relevance from static, independent representations: text and the target modality have not yet interacted when the gate is computed. *Cross Attention Gated* solves this constraint by introducing bidirectional cross-attention between text and each non-textual modality before gating, allowing representations to incorporate cross-modal context before the fusion decision is made.

Cross-modal enrichment. Each modality attends to, and is attended by, the text representation. Let $\mathrm{CA}(\mathbf{q}, \mathbf{k})$ denote multi-head cross-attention (8 heads) with query $\mathbf{q}$ and key-value $\mathbf{k}$, and LN layer normalisation with a residual connection. The enriched representations are:

$$\mathbf{t}' = \mathrm{LN}(\mathbf{t} + \mathrm{CA}(\mathbf{t},\, \mathbf{v}) + \mathrm{CA}(\mathbf{t},\, [\mathbf{e}_{\mathrm{EEG}};\, \mathbf{e}_{\mathrm{ET}}]))\,, \tag{1}$$

$$\mathbf{v}' = \mathrm{LN}(\mathbf{v} + \mathrm{CA}(\mathbf{v},\, \mathbf{t}))\,, \tag{2}$$

$$[\mathbf{e}'_{\mathrm{EEG}};\, \mathbf{e}'_{\mathrm{ET}}] = \mathrm{LN}([\mathbf{e}_{\mathrm{EEG}};\, \mathbf{e}_{\mathrm{ET}}] + \mathrm{CA}([\mathbf{e}_{\mathrm{EEG}};\, \mathbf{e}_{\mathrm{ET}}],\, \mathbf{t}))\,. \tag{3}$$

Each modality is treated as a single-token sequence for the cross-attention operations.

Gated fusion. Because $\mathbf{t}'$ already encodes cross-modal context, the gates need only be conditioned on the enriched text ($W \in \mathbb{R}^{1\times 768}$, single-input), unlike the dual-input gates of *Gated Fusion*:

$$\beta = \sigma\!\left(\frac{W_v\, \mathbf{t}'}{T}\right), \quad \alpha = \sigma\!\left(\frac{W_{\mathrm{EEG}}\, \mathbf{t}'}{T}\right), \quad \lambda = \sigma\!\left(\frac{W_{\mathrm{ET}}\, \mathbf{t}'}{T}\right).$$

The fused representation follows the same additive structure as our *Gated Fusion* model, replacing all embeddings with their attended counterparts:

$$\mathbf{z} = \mathbf{t}' + \beta\, \mathbf{v}' + \alpha\, \mathbf{e}'_{\mathrm{EEG}} + \lambda\, \mathbf{e}'_{\mathrm{ET}}.$$

Our *Cross Attention* model introduces approximately 9.4M additional parameters over our *Gated Fusion* from the four cross-attention layers. We treat it as an experiment to assess whether explicit cross-modal interaction before gating improves over the simpler content-adaptive weighting of our *Gated Fusion* model.

5. Experimental Setup

5.1. Training

All models are trained with AdamW ($lr = 2 \times 10^{-5}$, warmup ratio 0.1), batch size 16, and bfloat16 automatic mixed precision on a single NVIDIA A100 GPU. We use 3-fold stratified cross-validation with early stopping (patience 5) monitored on a combined validation F1:

$$F1_{\mathrm{val}} = 0.40\, F1_{2.1} + 0.35\, F1_{2.2} + 0.25\, F1_{2.3},$$

same task weights as in the training loss. The best checkpoint per fold is saved.

5.2. Submitted Runs

Run selection followed two criteria. First, **analytical**: fold average validation F1 showed that *Gated Fusion* and *Cross Attention* outperformed *Early Fusion* on Tasks 2.2 and 2.3, with *Gated Fusion* leading on intention detection and *Cross Attention* on categorisation (Table 1). *Early Fusion* achieved marginally higher Task 2.1 F1, but at the cost of substantially lower agreement with the other two architectures (67% vs. 75% predicted sexist), suggesting a different and less consistent decision boundary. Second, **scientific positioning**: since simple concatenation-based fusion had already been explored in related work [1], submitting EarlyFusion would replicate an existing approach rather than advance the state of the art. Submitting the two novel architectures and their ensemble instead allowed us to demonstrate the added value of content-adaptive modality weighting over a previous baseline.

- **Run 1** (*Gated Fusion*): fold ensemble of GatedFusion checkpoints. Best overall validation F1 on Tasks 2.1 and 2.2.
- **Run 2** (Ensemble GatedFusion+CrossAttnGated): average of both models' softmax probabilities, combining complementary strengths across tasks.
- **Run 3** (*Cross Attention*): fold ensemble of CrossAttnGated checkpoints. Best validation F1 on Task 2.3.

5.3. Prediction Formatting

Task 2.1. Hard label: $\hat{y} = \mathbf{1}[p(\text{sexist}) \geq 0.5]$. Soft label: raw sigmoid probability pair $(p_{\text{yes}}, p_{\text{no}})$.

Task 2.2. Hard label: non-sexist memes receive No; sexist memes are assigned Direct or Judgemental via argmax over conditional probabilities. Soft label: hierarchical distribution $p_{\text{no}} = 1 - p_{2.1}$, $p_{\text{direct}} = p_{2.1} \cdot p(\text{direct} \mid \text{sexist})$, $p_{\text{judg}} = p_{2.1} \cdot p(\text{judg} \mid \text{sexist})$, normalised to sum to 1.

Task 2.3. Hard labels used category specific thresholds calibrated to match the training set label frequency distribution, rather than a fixed $\tau = 0.5$: for a category with training prevalence f among sexist memes, the threshold is set to the $(1 - f)$-th percentile of predicted scores on the test set. This prevents systematic suppression of rare categories (e.g., Misogyny-NSV, $f \approx 2.5\%$, $\tau \approx 0.20$). Soft labels are submitted as raw sigmoid outputs per category, consistent with the task format where probabilities need not sum to 1.

6. Results

Table 1 reports macro-averaged F1 and AUC on held-out validation folds. Official ICM and ICM-soft scores on the test set will be added in the camera-ready version following the release of EXIST 2026 official results.

Table 1
Macro F1 and $\text{AUC}_{2.1}$ on held-out validation folds. Tasks 2.2 and 2.3 are evaluated on sexist memes only. EarlyFusion is included as a baseline reference but was not submitted. Best result per column in **bold**.

Model	$\text{F1}_{2.1}$	$\text{F1}_{2.2}$	$\text{F1}_{2.3}$	$\text{AUC}_{2.1}$
EarlyFusion (baseline, 3 folds)	**0.627**	0.532	0.370	0.669
GatedFusion (Run 1, 3 folds)	0.616	**0.549**	0.380	**0.680**
GatedFusion+CrossAttnGated (Run 2, ensemble)	0.613	0.543	0.389	0.673
CrossAttnGated (Run 3, 3 folds)	0.611	0.537	**0.397**	0.666

Table 2
Gate value statistics for the test set (1,053 memes). Gates are sigmoid outputs conditioned on text and modality embeddings.

Gate	Modality	Mean	Median	Std
β	Image (ALIGN)	0.982	0.990	0.022
α	EEG	0.481	0.474	0.091
λ	Eye Tracking	0.027	0.020	0.020

Task 2.1: Sexism identification. Our *Early Fusion* model achieved the highest binary F1 (0.627), which is expected: concatenation with equal weights is a strong prior when all modalities are roughly equally informative at the binary level. However, *Early Fusion's* higher F1 on this task did not translate into better performance on the hierarchically harder tasks, nor into consistency with the other architectures. This was one of the motivations to exclude it from submission (Section 5). On the other hand, *Gated Fusion* achieved the highest AUC (0.680), suggesting its probability estimates are better calibrated even if its hard-label F1 is marginally lower.

Task 2.2: Source intention. *Gated Fusion* leads on Task 2.2 (F1 0.549), the task requiring the model to distinguish DIRECT from JUDGEMENTAL sexism. This is consistent with the role of adaptive gating: intention detection likely benefits from selectively suppressing modalities that add noise at this level of distinction, precisely what the gates are designed to do.

Task 2.3: Sexism categorisation. Our *Cross Attention Gated* model leads on Task 2.3 (F1 0.397), the most challenging subtask. Category prediction requires distinguishing between semantically overlapping classes (e.g., Stereotyping-Dominance vs. Objectification), a task that benefits from the richer cross-modal representations produced by bidirectional cross-attention before gating. The ensemble (Run 2) interpolates between *Gated Fusion* and *Cross Attention Gated*, achieving intermediate performance across all three tasks.

Complexity Vs. Interpretability. These results prompt a clear observation: architectural complexity does not guarantee performance gains. Our *Cross Attention Gated* model introduces $\sim$9.4M additional parameters over our *Gated Fusion* model through four cross-attention layers, yet underperforms on Tasks 2.1 and 2.2 and leads only on the category subtask. *Gated Fusion*, by contrast, adds only three lightweight linear projections over *Early Fusion*, the gate weight matrices, yet achieves competitive or superior performance across all tasks while producing directly interpretable modality weights at no additional inference cost. This imbalance implies that architectural decisions may be more effective than deeper or attention mechanisms: while cross-attention provides expressive ability without a correspondingly strong prior, the gates impose a structured inductive bias (text modality weighting) that is consistent with the semantics of the task. Furthermore, the multi-task formulation with hierarchical masking adds a form of structural complexity, by conditioning Tasks 2.2 and 2.3 on Task 2.1 predictions, which consistently improves performance on the finer subtasks without increasing model parameters. The gated fusion thus occupies a position in the complexity-interpretability trade-off: richer than a concatenation baseline, more transparent than a cross-attention model, and competitive across all three evaluation tasks.

7. Analysis: Gate Interpretability

The central advantage of our *Gated Fusion* over both its concatenation baseline and prior multimodal approaches is that the gate values (β, α, λ) are not quantities requiring post-hoc attribution. They are explicit scalars that the model exposes by construction. This section examines what the model learned

about modality importance for sexism detection, using gate distribution over the 1,053 test memes (Table 2).

A stable modality hierarchy emerges. The model autonomously learns the ordering $\beta \gg \alpha \gg \lambda$, consistent across all three training folds (gate std below 0.12 for every modality). This consistency is itself a result: the hierarchy is not an artefact of a particular initialisation but a stable property the model converges to. To our knowledge, this is the first quantitative, learned estimate of relative modality importance for sexism perception in memes.

Image dominates ($\beta \approx 0.98$). The image gate indicates the model relies almost entirely on visual content. This is expected: in memes, sexism is frequently expressed through the image, the depicted scene, the juxtaposition of figures, visual stereotypes rather than the text alone [2]. The low variance ($\sigma = 0.022$) shows this reliance is consistent across memes, not driven by a small subset.

EEG carries complementary signal ($\alpha \approx 0.481$). The EEG gate is high but not saturated, and notably more variable ($\sigma = 0.091$) than the image gate. Two observations follow. First, the model assigns substantial weight to neural signals, indicating they encode information not recoverable from text and image alone, consistent with Arcos et al. [1], who report an EEG-only AUC of 0.717 rivalling text-only models. Second, the higher variance means the EEG gate is the one that adapts most across memes: the model modulates its reliance on brain signals depending on content, which is precisely the behaviour the gating mechanism was designed to enable.

Eye tracking is consistently suppressed ($\lambda \approx 0.027$). The ET gate is uniformly low across the test set. We interpret this as evidence of *redundancy* rather than *uselessness*: eye-tracking and EEG both index attentional and cognitive processing, and given access to both, the model concentrates its reliance on the richer EEG representation (80 features vs. 8). This does not imply ET is uninformative in isolation, only that it adds little *marginal* signal once EEG is present. The finding directly justifies our exclusion of an even weaker physiological signal, Heart Rate, from the final models.

Cross-attention obscures the hierarchy. The same gates in our *Cross Attention Gated* are far less stable: the image gate spans $\beta \in [0.07, 1.00]$ across folds, although the model achieved comparable task performance. We attribute this to the OCR-induced redundancy intrinsic to memes. The text input is extracted directly from the image, so text and image representations share substantial semantic content. Cross-attention entangles these representations before gating; once entangled, the gate can no longer attribute contribution cleanly to one modality or the other, thus its value becomes unstable. *Gated Fusion* avoids this by gating over independent representations, and this preserves interpretability. This is the core argument for preferring *Gated Fusion*: it attains competitive performance *and* yields stable, meaningful modality attributions.

Per-meme gate variability. Figure 2 illustrates three representative cases from the test set that highlight the adaptive behaviour of the gates. The meme with the highest EEG gate (α=0.700) contains explicit objectifying visual content, consistent with a stronger neural response beyond what text and image encode alone. The meme with the highest eye-tracking gate (λ=0.163) is characterised by dense text, suggesting that fixation and saccade patterns capture extended reading behaviour. Finally, the meme receiving the lowest image gate (β=0.744) in the test set is a cartoon where all sexist content is textual so the model correctly suppresses a visually uninformative channel. These examples demonstrate that the gates respond to content in a semantically coherent manner.

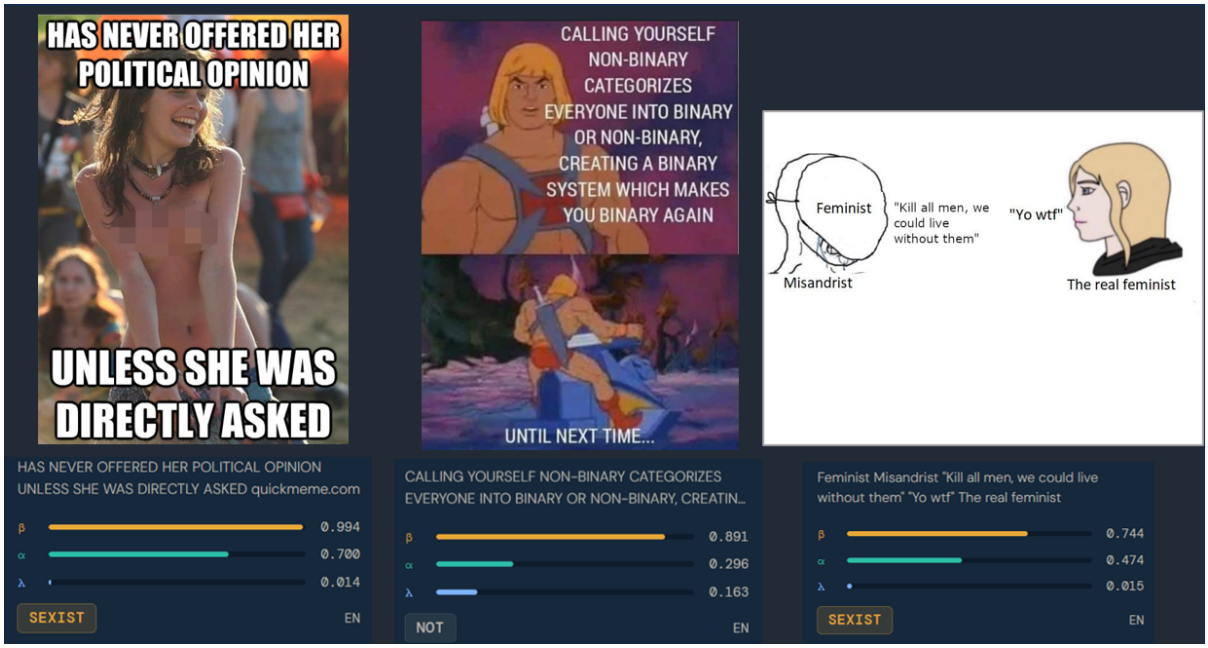

Figure 2: Three memes illustrating content-adaptive gate behaviour. *Left*: objectifying visual content elicits the highest EEG gate (α=0.700), indicating neural signal beyond text and image alone. *Centre*: dense OCR text yields the highest eye-tracking gate (λ=0.163), consistent with extended visual scanning of written content. *Right*: a cartoon meme where sexism is entirely textual receives the lowest image gate (β=0.744).

8. Conclusion and Future Work

We presented a Gated Fusion architecture for multimodal sexism detection in memes that integrates text, image, EEG, and eye-tracking signals through learnable, text-conditioned scalar gates. Evaluated on EXIST 2026 Tasks 2.1, 2.2, and 2.3, *Gated Fusion* achieves competitive macro F1 while producing directly interpretable modality importance scores as a byproduct of inference.

The gate analysis reveals a stable hierarchy: $\beta(\text{image}) \approx 0.98 \gg \alpha(\text{EEG}) \approx 0.481 \gg \lambda(\text{ET}) \approx 0.027$, which is consistent across folds and constitutes, to our knowledge, the first quantitative, *learned* estimate of modality importance for this task.

We further show that increasing architectural complexity through cross-attention does not uniformly improve performance and degrades gate stability, motivating the simpler gated design as the better trade-off between performance and interpretability.

Limitations. The physiological signals were averaged over the full meme-viewing period and across participants, which limits the representativeness of the learned EEG and ET representations. Thus, the model learns to capture the average response of the individuals rather than a person level signal. Furthermore, this average discards the temporal dynamics that may carry additional information about when and how cognitive processing of sexist content unfolds.

Future work. The MLP-encoded EEG representation used here treats all 16 channels as a flat feature vector, losing the spatial structure of the brain's response. We plan to extend this approach with a spatial EEG Transformer that treats each channel as an independent token with topographic positional embeddings derived from electrode coordinates, enabling the model to learn inter-region dependencies directly from data. Beyond this, we plan to implement per-channel gates conditioned on text and image. This architecture would enable anatomically grounded analysis of *which brain regions* respond most to sexist content, a natural bridge between computational sexism detection and cognitive neuroscience. We will also move to a per-subject dataset (instead of per-meme) formulation to preserve individual physiological responses and enable subject-level cross-validation, providing a more rigorous

experimental design for evaluating the contribution of neurophysiological signals.

Acknowledgments

This work was carried out during a research internship at the PRHLT Research Center, Universitat Politècnica de València. Thanks to the EXIST 2026 organizers for providing the dataset and evaluation framework.

Declaration on Generative AI

The author used Claude (Anthropic) as a writing assistance tool to support manuscript drafting and editing. The author takes full responsibility for the scientific content, experimental design, results, analysis, and conclusions presented in this work.

References

[1] I. Arcos, P. Rosso, E. Gomis-Vicent, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and EEG signals, in: Proceedings of the 15th Language Resources and Evaluation Conference (LREC-COLING 2026), Torino, Italy, 2026.

[2] D. Kiela, H. Firooz, A. Mohan, V. Goyal, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, in: Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020, pp. 2611–2624.

[3] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), 2026.

[4] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, Overview of EXIST 2026: Physiological data for multimodal sexism characterization in social media (extended overview), in: E. Sánchez Salido, A. Barrón-Cedeño, A. García Seco de Herrera, S. MacAvaney, J. M. Struß (Eds.), CLEF 2026 Working Notes, 2026.

[5] F. Rodríguez-Sánchez, J. Carrillo-de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, C. Largeron, E. Fersini, EXIST 2023: sEXism Identification in Social neTworks shared task report, in: Working Notes of CLEF 2023, CEUR Workshop Proceedings, 2023.

[6] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, E. García-Arias, D. Spina, EXIST 2025: Learning with disagreement for sexism identification in social networks, in: Working Notes of CLEF 2025, CEUR Workshop Proceedings, 2025.

[7] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, SemEval-2023 task 11: Learning with disagreements (Le-Wi-Di), in: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023), 2023, pp. 2304–2318.

[8] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR, 2021, pp. 8748–8763.

[9] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 4904–4916.

[10] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, K. Simonyan, Flamingo: A visual language model for few-shot learning, Advances in Neural Information Processing Systems (NeurIPS) 35 (2022) 23716–23736.

[11] S. Pramanick, S. Sharma, D. Bhatt, M. S. Akhtar, T. Chakraborti, T. Chakraborty, MOMENTA: A multimodal framework for detecting harmful memes and their targets, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4439–4455.

[12] P. Italiani, D. Gimeno-Gómez, L. Ragazzi, G. Moro, P. Rosso, MEMEWEAVER: Inter-meme graph reasoning for sexism and misogyny detection, in: Findings of the Association for Computational Linguistics: EACL 2026, 2026.

[13] J. Arevalo, T. Solorio, M. Montes-y Gómez, F. A. González, Gated multimodal units for information fusion, in: ICLR Workshop Track, 2017.

[14] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Pun, T. Ebrahimi, I. Patras, L. Rothkrantz, DEAP: A database for emotion analysis using physiological signals, IEEE Transactions on Affective Computing 3 (2012) 18–31.

[15] P. He, X. Liu, J. Gao, W. Chen, DeBERTa: Decoding-enhanced BERT with disentangled attention, in: International Conference on Learning Representations (ICLR), 2021.