A Multimodal Method for Sexism Identification Incorporating Visual Descriptions and Physiological Signals

Notebook for the EXIST Lab at CLEF 2026

Yanfu Chen[1], Leilei Kong*

Foshan University, Foshan, China

Abstract

This paper presents the DiYiMing team's submission to the EXIST 2026 shared task on sexism detection.The identification of sexism within social media memes requires modeling complex multimodal inputs that exhibit multi-annotator disagreement alongside eye-tracking, heart rate, and electroencephalogram (EEG) signals. This task requires the identification of discriminatory expressions across text and images, the reconciliation of subjective labels, and the enforcement of cross-modal semantic alignment. To address these challenges, this paper introduces SGF-Multimodal, a classification framework that integrates text, images, and physiological signals through visual description enhancement and sensor gated fusion. The framework first leverages a visual description generation module to supplement non-textual semantics within images. It then extracts unimodal features via a text encoder and an image encoder, and encodes physiological signals (eye-tracking, heart rate, and EEG) into gating vectors to perform dimension-wise gated modulation on the joint text-image representations. Finally, multiple independent training runs combined with probability averaging are used to stabilize prediction outputs, thereby accommodating both hard-label classification and soft-label distribution prediction. By treating physiological signals as modulating information rather than additional feature blocks, the proposed method mitigates cross-modal semantic misalignment and training fluctuations.

Keywords

Sexism Identification, Visual Description Enhancement, Physiological Signals, Meme Classification

1. Introduction

The dissemination of sexist content on social media increasingly relies on multimodal forms, with memes serving as a primary vehicle [2]. Unlike traditional text-based posts, memes co-construct meaning through the interplay of image scenes, facial expressions, object symbols, posture relationships, and overlaid textual captions [3]. Consequently, expressions that appear lighthearted, humorous, or ambiguous on the surface often rely on visual metaphors, cultural consensus, and contextual clues to convey discriminatory attitudes [3, 4]. For such content, relying solely on literal text is insufficient to recover the underlying semantics, presenting a long-standing challenge in multimodal content moderation [1, 5].

This study addresses sexism identification in social media memes, aiming to determine whether a given meme contains sexist content, describes a sexist situation, or criticizes sexist behavior [1]. This task requires a system to recognize explicit and implicit sexist expressions at the text-image level, while simultaneously managing annotation disagreement and prediction uncertainty stemming from variations in subjective human interpretation. Furthermore, beyond standard text and image data, this task introduces eye-tracking (ET), heart rate (HR), and electroencephalogram (EEG) signals recorded while subjects were exposed to the stimulus samples. This addition expands the task from traditional text-image classification into a multimodal understanding problem that integrates behavioral and physiological responses [6, 7]. Consequently, the core scientific issue extends beyond the joint

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

*Corresponding author.

c2735066426@163.com (Y. Chen); kongleilei@fosu.edu.cn (L. Kong)

0009-0007-2332-0671 (Y. Chen); 0000-0002-4636-3507 (L. Kong)

modeling of text and images; it encompasses utilizing physiological signals to characterize individual perceptual differences regarding sexist content, mitigating supervisory noise caused by multi-annotator disagreement, and achieving effective alignment between text-image semantics and physiological responses to enhance identification robustness and interpretability.

Existing methods generally follow a "feature extraction - fusion - classification" pipeline [8, 9]. For meme scenarios, the most commonly used features comprise text and image encodings, or cross-modal interactive representations built upon them [9, 10]. Although these methods provide foundational baselines, they exhibit limitations under the current task configuration. First, Optical Character Recognition (OCR) only captures explicit text within images, failing to capture non-textual semantics such as posture, composition, visual irony, and character relationships; when the sexist meaning derives primarily from the image rather than the literal wording, OCR-only approaches discard critical discriminative clues [2, 11]. Second, even when additional physiological signals are available, utilizing them through simple feature concatenation fails to model or explain how these signals influence the comprehension of text-image content, thereby undermining their utility as indicators of human subjective reactions [7, 12].

To address these limitations, this paper proposes a multimodal identification framework centered on visual semantic supplementation and physiological signal regulation [11, 12, 13]. First, the system utilizes a vision-language model to transcribe key visual content from the meme image into a concise English description, which is then concatenated with the OCR text to compensate for semantics that OCR cannot directly encode. Second, the system models the enhanced text, original image, and physiological signals independently, feeding the aggregated ET, HR, and EEG features as a gated input to execute dimension-wise re-weighting on the joint text-image representations [15]. Third, given the limited sample scale and pronounced annotation disagreement inherent in shared tasks, we employ independent training across multiple random seeds and perform probability averaging to mitigate single-run training fluctuations, while maintaining architectural compatibility with both hard-output and soft-output evaluation modes.

The primary contributions of this paper are threefold:

1. We propose a visual description enhancement strategy tailored for meme scenarios, utilizing generated image descriptions to supplement visual semantics beyond OCR.
2. We design a gated fusion mechanism driven by physiological signals, enabling ET, HR, and EEG to participate in text-image modeling as regulatory variables rather than parallel feature blocks.
3. We accommodate both hard and soft evaluation requirements at both the methodological and experimental levels, and explicitly analyze the discrepancy between local label construction rules and official hard evaluation rules.

2. Related Work

2.1. Multimodal Hate Speech and Sexism Identification

Hate speech and discriminatory expression detection initially focused on the textual modality, aiming to identify offensive vocabulary, biased expressions, and harmful language targeting specific groups. As social media content has evolved from pure text to image-text mixtures, short videos, and memes, the research focus has shifted toward multimodal identification [1, 8]. In meme scenarios, text and images are rarely stacked in a parallel configuration; instead, they co-construct semantics through mutual reinforcement, inversion, irony, or ellipsis [3, 10]. Consequently, multimodal hate speech detection relies more heavily on cross-modal semantic alignment than standard text classification [9, 14].

Existing methods are broadly categorized into two paradigms. The first emphasizes explicit cross-modal interaction, commonly implemented via attention-based text-image alignment, cross-modal Transformer encoding, or structured relationship modeling between textual tokens and visual regions. These methods possess high expressive capacity and excel at capturing complex semantic relationships. The second paradigm relies on pre-trained encoders to extract text and image representations separately,

followed by classification via concatenation, weighting, or shallow fusion [10, 16]. The latter approach is more prevalent in shared tasks due to its training stability, low computational cost, and suitability for moderate data scales. This paper follows the second paradigm but optimizes the input construction and fusion approach to match the characteristics of the target task.

2.2. Application of Vision-Language Models in Content Moderation

Recent advancements in vision-language models have re-established image description generation as an effective auxiliary strategy in multimodal classification [13, 17]. Compared to directly feeding image encoding vectors into a classifier, converting image content into natural language descriptions to serve as additional text inputs provides two advantages. First, it textualizes visual clues—such as characters, actions, scenes, objects, and compositions—making it easier to align this information with the semantic space handled by text encoders [17]. Second, in multilingual scenarios, generating visual descriptions in a unified language acts as a bridge, mitigating language discrepancies present in the original OCR text.

For meme tasks, visual description enhancement carries significant practical relevance. The sexist nature of many memes is often embedded in character imagery, suggestive scenes, or stereotyped visual arrangements rather than the literal text on the image [4]. Under these conditions, if image description generation can supplement these non-textual semantics, it provides a more complete input for subsequent classification models without introducing complex graph structures or region-level alignment modules [13, 17]. This paper adopts this approach; however, unlike standard image description tasks, we explicitly require the generated descriptions to ignore text within the images to avoid redundancy and cross-contamination with OCR results.

2.3. Physiological Signal Modeling and Human-Centered AI

Physiological signals such as eye-tracking, heart rate, and EEG are widely utilized in affective computing, cognitive modeling, and user research [12, 18]. Existing studies demonstrate that these signals reflect subjects' attention allocation, arousal levels, cognitive load, and emotional shifts. Distinct from "content modalities" like text and images, physiological signals act as a "response modality": they represent measurement results of a subject's reaction to a stimulus rather than the stimulus itself [7, 12]. Consequently, integrating these signals into content identification tasks has lacked a unified paradigm.

The target task introduces these signals within a multilingual, multimodal sexism shared task. This configuration implies that a system can either treat these signals as additional features or interpret them as proxy variables for users' subjective reactions. This paper adopts the latter approach; we do not require physiological signals to independently drive classification, but allow them to modulate the utilization of text-image representations. This design ensures that under conditions of limited sample scale and noise, allowing physiological signals to participate via regulation rather than dominance preserves model stability.

2.4. Disagreement Learning and Probabilistic Output Evaluation

This series of tasks preserves multi-annotator disagreement information, driving a transition in research from single hard-label training to detailed disagreement modeling. For these tasks, soft-label learning and probabilistic outputs are intimately linked with evaluation mechanisms [11, 16]. Diverging from traditional single-label classification, annotation disagreement implies that a portion of the samples are inherently borderline; forcing the system to compress these into a single categorical judgment discards crucial information regarding subjectivity. Thus, reporting both hard and soft results simultaneously in shared tasks has emerged as a standard experimental practice.

Against this backdrop, this paper retains both the hard-label training and soft-label modeling perspectives. This allows the system to compare directly against the official hard-hard evaluation while providing a natural interface for soft-soft evaluation and subsequent results analysis. More importantly, this dual setup helps reveal whether model performance stems from learning the majority opinion or from fitting the disagreement distribution.

3. Method

3.1. Task Definition

Given a multimodal input sample, let its image be denoted as I, text information as T, visual description as C, and physiological signal features as P. The task objective is to learn a discriminative function:

$$f(I, T, C, P) \rightarrow y$$

where y represents the sexism label of the sample. Under the hard-label setting, the model outputs discrete categories; under the soft-label setting, the model outputs a categorical distribution reflecting probabilistic information across differing annotation viewpoints. Since different modalities provide varying granularities of information, this paper adopts a step-by-step enhancement approach to construct inputs, evaluating the specific impact of additional semantic information on model performance.

3.2. Method Overview

The proposed system comprises four components: visual description generation, text-image encoding, physiological signal aggregation, and gated classification. First, the system offline generates a visual description for each meme image, explicitly requiring it to ignore text within the image while preserving purely visual semantics like characters, actions, expressions, objects, and scenes. Subsequently, the OCR text and visual description are concatenated and fed into a text encoder, while the original image is sent to a visual encoder. For sensor data, the `by_user` features of ET, HR, and EEG are aggregated into a fixed-length sample-level vector. Finally, a gating module modulates the joint text-image representation, and a classification head outputs the binary classification results. To reduce random fluctuations from a single training run, the inference stage takes the arithmetic mean of model output probabilities trained across multiple random seeds.

From the perspective of modeling logic, this framework does not construct an end-to-end, large unified cross-modal encoder. Instead, it decomposes the problem into two distinct sub-problems: first, how to represent the text-image semantics of the meme as completely as possible; second, how to enable physiological signals to modulate the utilization of text-image representations. The former is addressed by visual description enhancement, while the latter is realized through gated fusion. This division of labor maintains structural clarity and satisfies requirements for stability and reproducibility given the data scale. The overall workflow is illustrated in Figure 1.

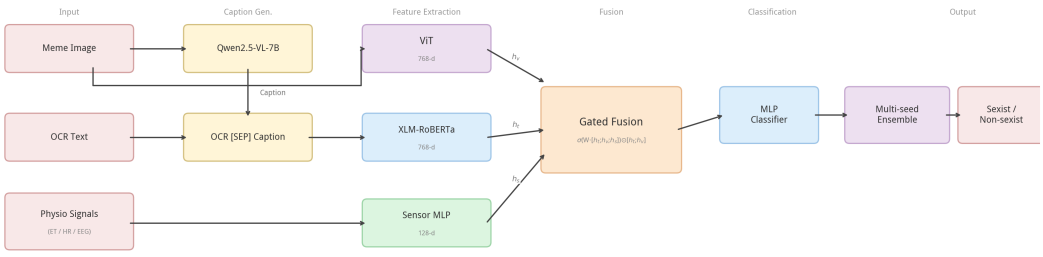

Figure 1: Overall framework of the proposed Sensor Gated Fusion (SGF-Multimodal) model.

3.3. Visual Description Enhancement and Data Construction

To overcome the limitation where OCR only recognizes explicit text within images, we designed a visual semantic enhancement module. In the target dataset, the `text` field originates from automatic OCR extraction, which reflects the literal content in the image but cannot encode visual clues such as

characters, actions, expressions, postures, and scene relationships. We utilize Qwen2.5-VL to generate a concise English description for each meme. The prompt explicitly instructs the model to focus on people's appearance, actions, gestures, body language, scenes, and visual symbols, while ignoring all text within the image. Descriptions generated this way serve to supplement visual semantics rather than duplicating OCR results. The expanded prompt is as follows:

```
PROMPT = (
    "Describe the visual content of this meme image in English. "
    "Focus on: people's appearance, expressions, actions, gestures,
       "
    "body language, objects, scenes, and any visual symbols. "
    "Ignore all text in the image. "
    "Write one concise sentence."
)
```

Subsequently, we concatenate the OCR text and the visual description as:

OCR_text [SEP] visual_description

where `[SEP]` is used to distinguish the two information sources. This design yields three advantages. First, the text encoder simultaneously receives literal text and language-translated visual clues from the image. Second, since the visual descriptions are uniformly written in English, they serve as an intermediary bridge under a mixed English-Spanish corpus. Third, compared to directly using the image as a second modality input, this design allows visual information to gain an extra layer of explicit expression within the text space, reducing the pressure on the model to learn cross-modal mappings entirely on its own.

3.4. Feature Extraction

The text branch employs a pre-trained XLM-RoBERTa to encode the concatenated text input, extracting sentence-level representations. Considering that visual descriptions increase input length, we adopt a longer maximum input length in the primary experimental setup to mitigate the risk of truncating useful information. The image branch uses a ViT to extract global image representations. We do not introduce additional object detection or region-level features here; instead, we prioritize global visual encoding results to keep the training workflow clean and stable.

For sensor data, the `sensorial` field of each sample contains features organized by modality and user. We extract the `by_user` features of ET, HR, and EEG from `sensorial.modalities`. Missing values for any user are padded with zeros, and an average is taken across the user dimension to derive modality-level sample representations. Finally, these are concatenated into a fixed-length 108-dimensional vector according to 24 dimensions for ET, 4 dimensions for HR, and 80 dimensions for EEG. Although this approach discards finer-grained temporal structures, it enables all samples to obtain a uniform sensor input representation without substantially increasing complexity.

3.5. Physiological Signal Gated Fusion

The primary model contains three encoding branches. The text branch outputs a 768-dimensional sentence-level representation h_t, the image branch outputs a 768-dimensional global visual representation h_v, and the sensor branch maps the 108-dimensional vector into a 128-dimensional hidden space via a two-layer Perceptron. The text and image representations are first concatenated to form the base text-image representation:

$$h_{tv} = [h_t; h_v]$$

Subsequently, the sensor representation is concatenated with the base text-image representation and fed into a Sigmoid gating network to generate a gating vector g of the same dimension as the text-image

representation:

$$g = \sigma(W[h_{tv}; h_s] + b)$$

Next, dimension-wise re-weighting is performed on the base text-image representation:

$$h_{\text{fused}} = h_{tv} \odot g$$

Finally, h_{fused} is fed into the classification head to obtain the binary classification prediction. The core mechanism does not treat physiological signals as a third parallel semantic modality, but leverages them to conduct "preservation-suppression" style regulation on the text-image representation. Intuitively, sensor information is treated as a control variable tied to the subject's reaction, used to influence which dimensions within the joint text-image representation the model should prioritize.

3.6. Training Objectives and Output Strategy

The current experiments maintain two training routes: hard labels and soft labels. In the soft setup, the target distribution is directly constructed from the voting ratios of six annotators:

$$P(\text{YES}) = \text{YES votes}/6$$

$$P(\text{NO}) = \text{NO votes}/6$$

This route naturally applies to soft-soft evaluation and aligns closer with the task design of "learning from disagreement". In the hard setup, the system uses single-label supervision compressed via voting. It is necessary to clarify that current experimental implementations utilize the rule YES ≥ 3, whereas the official hard-hard evaluation requires a category to be chosen by "more than 3 annotators" to enter the hard gold standard; meaning it corresponds to YES > 3 for the positive class, and $3/6$ tie-vote samples are removed in the official hard evaluation [1].

Regarding training stability, the system adopts a freeze-unfreeze training strategy, linear warmup, gradient clipping, and an early stopping mechanism. Specifically, the text and image backbones are frozen during the first few epochs, training only the fusion layer and classification head. Once the classification layer achieves initial convergence, the entire model is unfrozen. Given the limited data scale, the high proportion of category-boundary samples, and the inherently volatile nature of multimodal training, we further utilize independent training with multiple random seeds and average the softmax probabilities during inference to enhance the robustness of the final output.

4. Experiments

4.1. Dataset

This paper uses the officially released EXIST 2026 Memes Dataset. According to the task guidelines, Task 2.1 is a binary classification task within a meme scenario, with target categories designated as YES and NO [1]. The training set contains 3,984 samples, and the test set contains 1,053 samples. The language distribution of the training set is relatively balanced, containing 2,005 English samples and 1,979 Spanish samples; the test set comprises 513 English samples and 540 Spanish samples [6]. Each training sample is independently annotated by 6 annotators, making it suitable for either constructing a single hard label or directly forming a soft label distribution. To concentrate the data description, Table 1 summarizes the basic statistical information most relevant to the experiments in this paper.

4.2. Data Preprocessing and Input Construction

On the text side, we retain the original OCR output and append the generated visual description behind the OCR text as supplementary information. The text input employs a uniform [SEP] separation format to facilitate the model in distinguishing between the two information sources. For extremely short OCR

Table 1
EXIST 2026 Task 2.1 Data Statistics

Item	Value	Explanation
Training set samples	3984	Total number of meme training samples
Test set samples	1053	Total number of meme test samples
Training set English samples	2005	lang = en
Training set Spanish samples	1979	lang = es
Test set English samples	513	lang = en
Test set Spanish samples	540	lang = es
Total YES votes	13286	Aggregated votes from 6 annotators across the training set
Total NO votes	10618	Aggregated votes from 6 annotators across the training set
Samples with YES = 3/6	614	Hard/soft boundary samples
Proportion of YES = 3/6	15.41%	Reflects the proportion of disagreement samples
Samples with ET, HR, and EEG signals	3984	Available across all training samples

texts, visual descriptions frequently provide a supplementing effect; for samples with longer OCR texts where image clues remain crucial, visual descriptions offer an additional visual semantic reference.

On the image side, lightweight data augmentation is applied during the training phase, including random horizontal flips, mild rotations, brightness perturbations, and contrast perturbations. This is intended to alleviate overfitting while preventing overly aggressive augmentation from destroying the layout and semantic clues of the meme. On the sensor side, we do not perform additional complex normalization procedures, prioritizing length consistency across different modalities in sample-level representations and robustness against missing values.

4.3. Experimental Setup

Visual descriptions are generated offline prior to training. The system calls Qwen2.5-VL-7B to generate English descriptions sample by sample for both training and test set images, writing the results back to JSON files. The text and image branches of the primary model utilize pre-trained XLM-RoBERTa-base and ViT-base, respectively. Due to the text length increase after concatenating visual descriptions, the maximum input length of the main model is set to 128; comparative setups omitting visual descriptions typically use a shorter length as a reference for input enhancement efficacy.

During training, the dataset is split internally into training and validation sets according to a fixed ratio to guarantee that local parameter tuning and variant comparisons proceed on a unified split. The main model is trained using a small batch size paired with gradient accumulation to balance memory constraints and effective batch sizes; concurrently, learning rate warmup, early stopping, and freeze-unfreeze strategies are deployed to boost stability. The final inference stage averages the probabilities of models trained using multiple random seeds, exporting hard and soft outputs respectively.

4.4. Evaluation Metrics

In accordance with the EXIST task setup, results are reported from two perspectives:

- **Official Metrics**: ICM, ICM-Norm, ICM-soft, ICM-soft-Norm.
- **Auxiliary Metrics**: F1(YES), Cross-Entropy.

The official guidelines explicitly distinguish between two evaluation modes. The first is *hard-hard evaluation*, where the system outputs a hard label, and the ground truth is also compressed into a hard label via annotator voting; for Task 2.1, the officials designate "the category chosen by more than 3 annotators" as the hard gold standard, and if no category exceeds this threshold, the sample is removed from hard-hard evaluation. The second is *soft-soft evaluation*, where the system outputs a probability distribution for each category and compares it with the soft gold standard formed by the annotator distribution. Under these two settings, the official core metrics are calculated based on ICM

and ICM-soft respectively, while positive class F1 is also reported in the hard-hard results of Task 2.1. In current local experiments, model selection primarily relies on F1(YES), Accuracy, or validation set loss.

4.5. Comparative Methods and Baseline Settings

To systematically analyze the contribution of individual components to overall performance, this paper divides comparative methods into several progressive tiers. The first category consists of text baselines, including a pure text model utilizing only OCR text, and a text-enhanced model that adds visual descriptions to the OCR text, aiming to evaluate the utility of input enhancement itself. The second category comprises standard text-image models, which directly concatenate text and image representations for classification, observing whether merely adding the image modality can stably boost performance without introducing physiological signals. The third category encompasses fusion models incorporating physiological signals, which include both a non-gated version that directly concatenates sensor vectors with text-image representations, and the gated fusion model proposed in this paper. Finally, multi-random-seed probability averaging is introduced atop the main model to form a gated fusion ensemble model, evaluating gains brought by prediction stability.

Under this comparative framework, discrepancies among methods primarily manifest in three aspects: first, whether the input contains visual descriptions; second, whether physiological signals are incorporated into the fusion process; third, whether physiological signals engage in fusion via direct concatenation or gated modulation. Such a setup effectively isolates the "contribution of visual enhancement", the "contribution of physiological signals", and the "contribution of the gating mechanism".

4.6. Experimental Results

Tables 2 and 3 present the experimental results of the model on the hard-label and soft-label tasks, respectively. According to the best results returned by the official evaluation platform, our final submitted model achieved relatively stable classification performance on the hard-label task and maintained strong distribution-fitting capabilities on the soft-label task. We focus on comparing the effects of different input constructions and fusion strategies on system performance, specifically addressing three questions: first, whether visual description enhancement stably outperforms pure OCR inputs; second, whether physiological signal gating outperforms simple text-image concatenation; third, whether multi-random-seed probability averaging can further elevate output stability, particularly in soft evaluation scenarios.

Table 2
Results for Task 2.1 — Hard-Hard Setting

Method	ICM-Hard	ICM-Hard Norm	F1 YES
Image + Text	0.0630	0.5321	0.6527
Image + Text + Visual Description	0.0835	0.5425	0.6591
Image + Text + Visual Description + Physiological Signals	0.1017	0.5517	0.6662

Table 3
Results for Task 2.1 — Soft-Soft Setting

Method	ICM-Soft	ICM-Soft Norm	Cross Entropy
Image + Text	-0.3123	0.4498	1.0866
Image + Text + Visual Description	-0.2894	0.4535	0.9879
Image + Text + Visual Description + Physiological Signals	-0.2230	0.4642	0.9932

5. Conclusion

This paper introduces SGF-Multimodal, a multimodal sexism identification system designed for EXIST 2026 Task 2.1. To leverage the implicit guidance offered by physiological responses, we propose a Sensor Gated Fusion (SGF) mechanism that utilizes signals such as eye-tracking and EEG to dynamically adjust the interaction weights of text-image features. Concurrently, integrating Qwen2.5-VL to generate explicit visual descriptions significantly enhances the model's comprehension of contextual and non-textual image semantics. Experimental results demonstrate that this framework achieves competitive performance on the EXIST 2026 test set, with ablation studies validating the efficacy of each component. Future work will explore advanced sequential sensor modeling architectures to better capture the dynamic temporal progression of physiological signals.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62276064).

Declaration on Generative AI

During the preparation of this work, the author(s) used Qwen2.5-VL-7B in order to generate concise visual descriptions of meme images to augment input semantics. After using this tool, the author(s) reviewed, curated, and formatted the content as needed and take full responsibility for the publication's content.

References

[1] L. Plaza, J. C. de Albornoz, I. Arcos, M. Aloy-Mayo, E. Gomis-Vicent, 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, Lecture Notes in Computer Science, 2026. doi:10.1007/978-3-032-21321-1_40.

[2] E. Fersini, F. Gasparini, G. Rizzi, A. Saibene, B. Chulvi, P. Rosso, A. Lees, J. Sorensen, Semeval-2022 task 5: Multimedia automatic misogyny identification, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 533–549.

[3] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR, 2022, pp. 12888–12900.

[4] D. Kiela, H. Firooz, A. Mohan, V. Goswami, A. Singh, P. Ringshia, D. Testuggine, The hateful memes challenge: Detecting hate speech in multimodal memes, Advances in neural information processing systems 33 (2020) 2611–2624.

[5] R. Gomez, J. Gibert, L. Gomez, D. Karatzas, Exploring hate speech detection in multimodal publications, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2020, pp. 1470–1478.

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

[7] M. Wu, C. Zhao, A. Su, D. Di, T. Fu, D. An, M. He, Y. Gao, M. Ma, K. Yan, et al., Hypergraph multimodal large language model: Exploiting eeg and eye-tracking modalities to evaluate heterogeneous responses for video understanding, in: Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 7316–7325.

[8] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework, in: International conference on machine learning, PMLR, 2022, pp. 23318–23340.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[10] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, S. C. H. Hoi, Align before fuse: Vision and language representation learning with momentum distillation, Advances in neural information processing systems 34 (2021) 9694–9705.

[11] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.

[12] Z. Ahmad, N. Khan, A survey on physiological signal-based emotion recognition, Bioengineering 9 (2022) 688.

[13] A. N. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, Learning from disagreement: A survey, Journal of Artificial Intelligence Research 72 (2021) 1385–1470.

[14] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, J. Liu, Uniter: Universal image-text representation learning, in: European conference on computer vision, Springer, 2020, pp. 104–120.

[15] K. Zhang, L. He, X. Jiang, W. Lu, D. Wang, X. Gao, Cognitioncapturer: Decoding visual stimuli from human eeg signal with multimodal information, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 14486–14493.

[16] J. Lan, D. Frassinelli, B. Plank, Mind the uncertainty in human disagreement: Evaluating discrepancies between model predictions and human responses in vqa, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 39, 2025, pp. 4446–4454.

[17] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, S. Hoi, Instructblip: Towards general-purpose vision-language models with instruction tuning, Advances in neural information processing systems 36 (2023) 49250–49267.

[18] A. Saavedra, R. Chocarro, M. Cortinas, N. Rubio, Exploring unconscious user responses to affective computing in interactive prototypes: a consumer neuroscience study, Behaviour & Information Technology 44 (2025) 5028–5047.