Aegis Athena at EXIST 2026: Multimodal Sexism Detection in Short Videos via Cross-Attention Fusion

Ruslan Pylypiv[1,*], Martina Szabóová[1,*]

[1]*Institute of Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 9, 042 00 Košice, Slovakia*

Abstract

This paper describes our participation in the EXIST 2026 shared task on sexism detection in short videos, where we address all three video subtasks: binary sexism identification (Task 3.1), source intention classification (Task 3.2), and fine-grained multi-label categorisation (Task 3.3). We propose a multimodal architecture that fuses four streams through cross-attention: structured scene descriptions generated by a vision–language model, automatic speech transcripts, raw video features, and physiological signals collected from viewers. The two textual streams share a single XLM-RoBERTa-large encoder and are disambiguated through learned type embeddings. We train under the learning-with-disagreement paradigm and systematically study eight modality configurations to identify which streams are complementary for each subtask. On the official EXIST 2026 test leaderboard, our team *Aegis Athena* reaches ICM-Hard-Norm of 0.6563, 0.5605 and 0.4549 on Tasks 3.1, 3.2 and 3.3 respectively (ICM-Soft-Norm of 0.5457, 0.3951 and 0.1297), corresponding to rankings of 4th, 17th and 3rd out of 76, 70 and 69 systems on the all-language hard evaluation, with first place on Task 3.3 and second place on Task 3.1 over English-only instances.

Keywords

sexism detection, multimodal learning, cross-attention fusion, vision-language models, short-video classification, EXIST 2026

1. Introduction

Sexism and online misogyny are an established and well-documented harm on social media, with short-video platforms emerging as a particularly problematic vector for younger audiences. Clinical research on online sexual harassment reports higher levels of depression, anxiety, and trauma in women who have experienced online sexual harassment [1]. In addition, the results of long-term studies identify verbal forms of sexism as the form of peer harassment most consistently associated with later depressive symptoms in adolescents [2]. Short-video recommender systems actively amplify this class of content: a controlled algorithmic study by Regehr et al. [3] found that on TikTok the share of recommended videos containing misogynistic content rose from 13% to 56% over only five days for accounts representing typologies of teenage boys. The EXIST organisers further describe TikTok's recommendation system as a mechanism that reinforces sexist ideologies via filter bubbles, with visual trends and content-moderation disparities contributing to the hypersexualisation of women [4]. Automatic detection of sexist content on these platforms is therefore a piece of moderation infrastructure with direct implications for users' well-being.

1.1. Why sexism detection in short videos is hard

While sexism detection has been studied extensively for textual content on social media, short-form video introduces qualitatively new challenges. First, sexist content in this medium is intrinsically *multimodal*: on-screen captions, spoken audio, visual context, gestures and music co-occur, and the sexist signal can be carried by any one of them, by their combination, or by the contrast between them. Arcos and Rosso [5], who introduced one of the first multimodal pipelines for sexism detection on

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

*Corresponding author.

ruslan.pylypiv@student.tuke.sk (R. Pylypiv); martina.szaboova@tuke.sk (M. Szabóová)

TikTok, find that combining textual, audio and visual features yields gains over text-only baselines that are particularly pronounced for source-intention detection rather than binary identification. Second, the modalities are frequently *semantically asynchronous*: a visually innocuous video may be accompanied by sexist voice-over, and an explicit visual scene may be paired with a critical, judgmental narration. This cross-modal inconsistency has been identified as a recurring failure mode of classifiers trained on a single modality [9], and motivates fusion mechanisms in which each modality can attend to the others rather than being aggregated only at the prediction level. Third, the informativeness of each modality is itself *task-dependent*. The overview of EXIST 2025 reports that among the input formats considered, textual input was the strongest single signal for TikTok sexism detection [4]. This finding deserves some unpacking, however, because the "textual" category in EXIST 2025 included both genuine transcripts and VLM-generated descriptions of the video; the latter, even though they are consumed by the classifier as text, are multimodal in origin. The same overview observes that top-performing systems in the video subtasks were those that combined such textual inputs with raw video or audio features through multimodal fusion. Together with Arcos and Rosso's earlier finding that multimodal gains are particularly pronounced for intention detection rather than for binary identification [5], this is evidence that different subtasks benefit from different combinations of input streams. Finally, a substantial fraction of sexism on social media is *implicit*, relying on stereotypes, objectification or ironic framing rather than overt slurs, and therefore cannot be captured by keyword matching but instead requires scene-level semantic understanding [4, 6]. Taken together, these properties make EXIST 2026 a problem where the architecture, the choice of modalities, and the way they are combined all matter, and where neither pure text-based nor pure end-to-end visual approaches are likely to be optimal.

1.2. The EXIST 2026 video subtasks

The EXIST campaign is a series of CLEF shared tasks dedicated to the detection and characterisation of sexism on social media, extending from text-only posts in 2021 to memes in 2024 and TikTok videos in 2025 [4]. The present work participates in the EXIST 2026 edition, which retains the bilingual (English / Spanish) video track introduced in 2025 and comprises three video subtasks organised hierarchically. Task 3.1 is a binary classification problem in which systems decide whether or not a short video contains sexist content. Task 3.2 is a hierarchical multiclass problem with three mutually exclusive output classes — NO, DIRECT and JUDGEMENTAL — where the two sexist subclasses distinguish between a video whose author directly expresses a sexist message and one that comments on or judges sexist behaviour without endorsing it. Task 3.3 is a hierarchical multi-label problem in which a video is assigned either NO or one or more of five fine-grained sexism categories: ideological and inequality, stereotyping and dominance, objectification, sexual violence, and misogyny and non-sexual violence [21].

All three subtasks are evaluated under the Learning with Disagreement paradigm with ICM-Soft-Norm as the primary metric [21, 4, 7]. For each video, the organisers provide the raw video file, automatically extracted text, and the human annotation labels used for LeWiDi evaluation. In addition, EXIST 2026 extends the video dataset with physiological signals collected: each stimulus was viewed by independent subjects, rather than by the task annotators themselves, while eye-tracking, heart-rate and EEG signals were recorded [21]. In the released data, each stimulus is associated with sensorial recordings from multiple subjects, with at least two subjects available for each recorded physiological modality. A full description of the dataset and metric is given in Section 3.

1.3. Our approach and contributions

Our architecture is designed in direct response to the challenges of Section 1.1. We ingest all four modalities the organisers release — the transcript, the raw video, and the sensorial signals, together with a fourth, structured VLM-generated scene description that we produce in preprocessing — and we combine them through cross-attention, so that any modality can attend to any other rather than being aggregated only at the prediction level.

Rather than committing to a single "best" modality combination, we systematically train eight

configurations of the same architecture: all six pairwise combinations of {VLM, transcript, video, sensorial}, plus the three-modality *trio* of transcript+VLM+video and the four-modality *quad* that adds the sensorial stream. This ablation is designed to test the broader hypothesis that the optimal modality subset is task-dependent. More specifically, we organise our analysis around three research questions:

- **RQ1:** Does the raw video stream add useful information beyond the textual streams alone?
- **RQ2:** Do structured VLM-generated scene descriptions add value beyond the combination of transcript and raw video?
- **RQ3:** Do physiological signals collected from independent viewing subjects provide complementary information beyond content-based modalities?

The full configuration table is given in Section 5, the architecture itself is detailed in Section 4, and the results are analysed with respect to these research questions in Section 6.

The decision to introduce structured VLM descriptions as a dedicated modality, rather than relying on the transcript alone, comes directly from inspection of the EXIST 2026 training data: a non-negligible fraction of videos consist of visual performances — for example, men cross-dressing to mock feminine behaviour or staging exaggerated re-enactments — in which the sexist signal lives in the visual scene and is largely absent from the spoken words. The same observation has been reported for the MuSeD Spanish TikTok corpus, where visual information was found to play a decisive role in recognising implicit or indirect sexism that is not evident from text alone [8].

Our official submissions, made under the team name *Aegis Athena*, achieve competitive results across all three EXIST 2026 video subtasks. On the all-language hard-evaluation leaderboards, our best runs rank 4th out of 76 systems on Task 3.1, 17th out of 70 on Task 3.2, and 3rd out of 69 on Task 3.3, with ICM-Hard-Norm scores of 0.6563, 0.5605 and 0.4549, respectively. On the corresponding soft evaluations, the same submissions obtain ICM-Soft-Norm scores of 0.5457, 0.3951 and 0.1297. The full official results, including English-only rankings are reported in Section 6.

2. Related Work

The EXIST campaign has driven recent work on sexism detection in social-media content, expanding from text-only posts in its first three editions to memes in EXIST 2024 [11] and TikTok videos in EXIST 2025 [4]. Across these editions, participants converged on a methodological core relevant to our work: fine-tuned transformer encoders such as XLM-RoBERTa, RoBERTa and DeBERTa, often in domain-adapted variants like HateBERT [12] or BERTweet, trained under the Learning with Disagreement paradigm against soft-label distributions [4]. For memes, a strategy particularly relevant to our work uses a vision–language model as a *translator* from image to text. The canonical formulation is Pro-Cap [23], which prompts a frozen VLM with hateful-content-related questions and uses the answers as image captions for a downstream BERT-based classifier, establishing that *what* a VLM is asked to describe matters as much as the choice of VLM itself. The same idea has driven the strongest EXIST meme systems: RoJiNG-CL at EXIST 2024 used GPT-4 to generate a one-sentence meme description, concatenated it with the OCR text, and fine-tuned standard text classifiers on the combined representation, finishing in the top three on the hard-evaluation leaderboard [13]; CogniCIC pushed this further at EXIST 2025 by replacing the downstream classifier with a multimodal LLM (Claude 3.7) consuming images and video frames directly [14]. The transition to video, however, introduces qualitatively different challenges — temporal dynamics, spoken audio, scene-level visual context — that the literature is only beginning to address.

2.1. Multimodal hate-speech and sexism detection in video

Multimodal hate detection in video is a young area shaped by a small number of benchmark datasets and a recent shift toward attention-based fusion. The field was anchored by HateMM [15], the first large-scale hate-video dataset (1,083 videos), and extended bilingually by MultiHateClip [16]. Early

systems relied on late fusion of pre-trained unimodal features; recent work has moved toward attention-based integration, with CMFusion [17] applying temporal cross-attention between video and audio, and MM-HSD [18] using Cross-Modal Attention as an *early* feature extractor across video, audio, transcripts and on-screen text. RAMF [19] extends the VLM-as-modality pattern from memes to videos, feeding three structured VLM-generated semantic streams (objective, hate-assumed and non-hate-assumed descriptions) through a Semantic Cross Attention module. These benchmarks remain small — on the order of 1,000–2,000 videos — which constrains end-to-end training and motivates parameter-efficient designs.

Within the EXIST setting specifically, the video track of EXIST 2025 produced the first published learnings from team systems applied to TikTok sexism detection, and these inform our design choices directly. ECA-SIMM-UVa, the top-ranked team on the hard evaluation of Task 3.1, adopted a segmentation-oriented approach that trained one classifier per channel (text, audio, video) and combined them through late-fusion rules including *One Is Enough*, Majority Voting and Probabilistic OIQ [20]. Their detailed analysis reported that the textual channel substantially outperformed both audio and video, and that late-fusion mechanisms ended up driven almost entirely by the text-channel decisions, with audio and video contributions effectively excluded by the calibrated thresholds [20]. DS@GT EXIST combined RoBERTa for transcripts, VideoMAE for video and CNN–MFCC pipelines for audio with a generative pipeline using Gemini to produce video summaries, and reported that combining the generative summaries with RoBERTa substantially improved performance over unimodal baselines [10]. CogniCIC used a multimodal LLM (Claude 3.7) end-to-end for the video subtasks and ranked first overall on the all-language hard evaluation of Task 3.2 [14]. Two structural observations follow. *First*, none of these EXIST systems uses cross-attention as the primary multimodal integration mechanism; fusion is performed either by late combination of unimodal predictions, by feature concatenation, or implicitly inside a single multimodal LLM. *Second*, ECA-SIMM-UVa themselves identify, as a direction for future work, the need for stronger multimodal integration that does not collapse onto the text channel [20]. Our design takes both observations as starting points.

2.2. Cross-attention fusion mechanisms

Cross-attention as a multimodal fusion mechanism originates in multimodal sentiment analysis, where the Multimodal Transformer (MulT) [24] introduced cross-modal attention modules operating on unaligned modality sequences. Subsequent work such as ALMT [25] introduced *language-as-anchor* variants in which text features guide the construction of an audio/visual representation, suppressing irrelevant or conflicting signals — a design consistent with the EXIST 2025 finding that the text channel is the most informative source for sexism detection in TikTok videos. Adoption of cross-attention for hate-content detection in videos is recent and limited to a handful of systems already mentioned above, namely CMFusion [17] and MM-HSD [18]. In the EXIST setting specifically, cross-attention has not yet been used as a primary fusion mechanism: integration is performed at the prediction level (ECA-SIMM-UVa [20], DS@GT [10]), by feature or text concatenation (RoJiNG-CL [13]), or implicitly inside a single multimodal LLM (CogniCIC [14]).

Taken together, the systems reviewed in this section reveal three recurring open challenges that motivate our approach. *First*, fusion across modalities in EXIST sexism-detection systems is predominantly performed late — at the prediction level — with cross-modal attention appearing only sporadically in adjacent hate-detection work and never, to our knowledge, applied to the full four-stream setting that the EXIST 2026 video track makes possible. *Second*, the text channel consistently dominates: as Fernández García et al. [20] report for the top-ranked EXIST 2025 system, late-fusion mechanisms tended to be driven almost entirely by the text channel, with audio and video contributions effectively excluded by calibrated thresholds. Vision–language descriptions partly mitigate this imbalance for memes but, in the few video systems that adopt them, the descriptions are used either as a direct LLM-classifier output or as a concatenated input feature, not as a separate semantic stream participating in attention-based fusion.

Third, the physiological signals newly introduced in EXIST 2026 are collected from independent viewing subjects rather than from the task annotators themselves, which raises an open question: whether such subject-level reactions can complement content-based video, transcript and VLM representations in sexism detection. Our work addresses all three points within a single unified architecture: cross-attention fusion as the primary integration mechanism, structured VLM-generated scene descriptions as a dedicated semantic modality alongside transcripts and raw video, and physiological signals as a fourth stream — all applied jointly to the three EXIST 2026 video subtasks.

3. Task and Dataset

This section completes the description of the EXIST 2026 video track previewed in Section 1.2, with a closer look at the data, the annotation protocol, and the evaluation metric used by the organisers.

The bilingual TikTok video dataset comprises 2,524 training and 674 test videos in Spanish and English, following the partitioning scheme established for EXIST 2025 [4, 20]. The videos were collected with the Apify TikTok Hashtag Scraper from a curated set of hashtags associated with potentially sexist content, and span the typical duration and resolution range of short-form TikTok posts. For each video the organisers release the raw video file, an automatic speech-to-text transcript, and a set of physiological signals to which we return below.

Labels are produced by a pool of six annotators, with each video labelled by a subset of two or three of them, and cover all three subtasks (sexism identification, source intention, fine-grained category) on a per-video basis. Rather than collapsing this evidence into a single ground-truth label, EXIST has followed the Learning with Disagreement paradigm since 2023 [4]: the full distribution of annotator votes is published and used directly as a soft label, while the corresponding majority-vote aggregation is provided as a complementary hard label for systems that prefer a single-class target.

A new component of the 2026 edition is the inclusion of physiological recordings. While each video is watched and labelled by a subset of the six annotators, only two designated subjects also contributed sensorial recordings during viewing. For each of these two subjects the dataset provides three signal streams — eye-tracking, heart rate and electroencephalography — yielding six physiological tokens per video in total [22]. To our knowledge this is the first edition of EXIST to release signals of this kind, and we treat them as a separate modality in our architecture (Section 4).

The primary metric reported by the organisers, and the one we optimise for, is ICM-Soft-Norm: the normalised, soft-label variant of the Information Contrast Measure (ICM) [7]. ICM generalises pointwise mutual information to hierarchical and multi-label classification by measuring how close a system's predicted distribution is to the ground-truth distribution. Its soft variant accepts both soft predictions and the soft annotator-vote labels described above, which makes it the natural fit for the Learning with Disagreement setting. The organisers additionally report the hard variant ICM-Hard-Norm, computed against the majority-vote labels, which we view as complementary — it measures categorical decision accuracy where the soft metric measures calibration under disagreement — and we report both throughout Section 6.

4. Methodology

Our architecture accepts up to four modality streams — VLM-generated structured scene descriptions (`vlm`), automatic speech transcripts (`trn`), raw video clips (`video`) and per-video physiological signals (`phys`) — and produces a single set of class logits for the current EXIST subtask. Each modality is first processed by a dedicated encoder, projected into a common $D{=}512$-dimensional space, and then combined through $N{=}2$ *fusion layers* that perform self-attention and cross-attention on a per-modality basis. The post-fusion representations are pooled per stream, concatenated, and passed through a small MLP classifier whose output dimension is the class count of the target subtask (2 for Task 3.1, 4 for Task 3.2 and 6 for Task 3.3). The overall layout is shown in Figure 1.

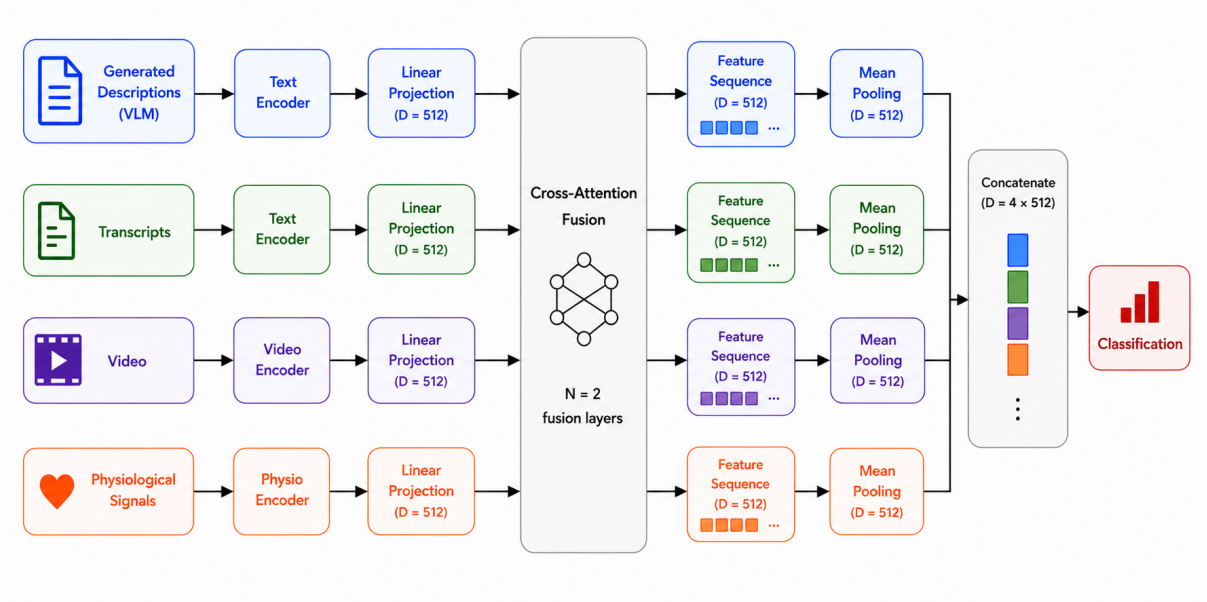

Figure 1: Architecture of our four modality model where streams are encoded independently, projected to a common dimension $D{=}512$, and combined through $N{=}2$ fusion layers performing per-modality self-attention followed by cross-attention. Pooled per-stream representations are concatenated and passed through an MLP classifier.

Text streams. Both textual streams — the VLM-generated scene description and the ASR transcript — are encoded by a *single* shared `xlm-roberta-large` backbone, whose 1024-dimensional contextual embeddings are projected to D by a linear layer. Sharing one encoder across the two streams is a deliberate choice: it nearly halves the text-side parameter count (saving on the order of 200M parameters relative to two independent large encoders) and, more importantly, it forces VLM-generated text and live transcripts to inhabit the same semantic space, which we found to stabilise multimodal training. To prevent the two streams from being mutually indistinguishable after encoding, we add a learned D-dimensional *type embedding* to every token of each stream, one per stream; the encoder is otherwise unmodified.

Video stream. Raw video is encoded by `MCG-NJU/videomae-base` (768-dimensional hidden states), with a linear projection to D. Each video is represented as K clips of 16 frames each; at training time $K{=}3$ clips are sampled randomly, providing a mild data-augmentation effect across epochs, while at inference time $K{=}4$ uniformly-spaced clips are used for a deterministic representation. On top of VideoMAE's intrinsic positional encoding we add two further learned positional embeddings: a temporal embedding indexed by the within-clip token position, and a clip-level embedding indexed by the clip's relative temporal position in the video rather than by its absolute ordinal, so that the same embedding space is shared consistently between the $K{=}3$ training clips and the $K{=}4$ inference clips. The clip-level embedding is important because the VideoMAE backbone has no native notion of *which* clip a token belongs to and would otherwise treat tokens from different clips as interchangeable.

Sensorial stream. The physiological signals released for EXIST 2026 are encoded by a purpose-built *PhysioEncoder*. Each video contributes six sensorial tokens — one for each combination of two viewers and three signal types (eye-tracking, heart rate, electroencephalography) — and each token is produced by projecting the raw per-token feature vector to D and adding four additive embeddings: a *modality* embedding (ET / HR / EEG), a *user* embedding (user 1 / user 2), a *position* embedding (one of six slots), and a *presence* embedding that distinguishes a genuinely zero-valued token from a token that is simply missing for this video. The presence embedding is the key element that lets the model learn a different attention behaviour for absent versus zero-valued signals, which we found necessary because missing

tokens are common in the released sensorial data.

4.1. Vision-language scene descriptions

The VLM-text stream is not provided by the organisers but generated ahead of time from the raw video as part of our preprocessing. We use `Qwen3-VL:30B`, served locally through Ollama, and prompt it to fill in five fixed semantic fields for each video: SCENE (visual setting and format), SITUATION (what is happening in context), PEOPLE (who appears and in what role), MESSAGE (what the video is saying or claiming) and TARGET (who or what the video is about). Asking the VLM for these specific fields, rather than a free-form caption, gives the downstream classifier explicit semantic anchors — most importantly TARGET and MESSAGE — that are directly relevant to sexism detection. The prompt is instructed to describe at two levels (literal content and situational context), to avoid moral or normative judgments, and to respect strict sentence limits per section. The full prompt text is given in Appendix A.

The frames that we show to the VLM are selected by a small content-aware pipeline rather than by uniform sampling. PySceneDetect first identifies scene boundaries, a small uniform grid of frames is then taken from within each scene, and finally a perceptual-hash step removes near-duplicate frames. This is cheap, deterministic, and avoids both the redundancy of uniform sampling on long static shots and the missing content of pure keyframe sampling.

When the resulting description is fed to the text encoder, we concatenate four of the five fields in a fixed order — `[TARGET] . . . [MESSAGE] . . . [SITUATION] . . . [PEOPLE]` — and drop the SCENE field. In preliminary runs we found SCENE to consist mostly of generic descriptions of camera angles, lighting and background objects, with very little signal for sexism; keeping it in the input only diluted the more useful fields. A representative example of a generated description (abbreviated) is:

> `[SCENE]` talking-head, home setting. `[SITUATION]` A woman shares a personal concern about dating at 40, questioning whether her age makes her unattractive to men. Her tone is earnest and reflective, and the static camera focuses on her upper body, emphasising her emotional delivery. `[PEOPLE]` a single woman, approximately 40 years old, as the main speaker. `[MESSAGE]` the speaker questions whether being 40 makes her unattractive to men, expressing vulnerability about being single and seeking validation. `[TARGET]` the speaker's own experience with age-related dating challenges, with implicit reference to societal perceptions of attractiveness for women in their 40s.

4.2. Cross-attention fusion

The fusion stack is the part of the model where the different modalities exchange information. As illustrated in Figure 2, the stack consists of $N{=}2$ identical fusion layers. Each layer updates every enabled modality in two stages: first, intra-modal self-attention refines the tokens of each stream independently; second, cross-attention lets each stream attend to the representations of the remaining active modalities.

Step 1 — intra-modal self-attention. For each modality m, we first apply a standard pre-norm self-attention block to its own token sequence. This operation is equivalent to the self-attention block of a Transformer encoder and allows the model to update each stream using only information internal to that modality:

$$\tilde{H}_m \;=\; H_m \;+\; \mathrm{SelfAttn}(\mathrm{LN}(H_m))\,. \tag{1}$$

Step 2 — cross-modal attention. For each modality m, the updated tokens $\tilde{H}_m$ are used to generate query representations. The remaining active modalities contribute key and value representations, which are aggregated into a single set of average keys $\bar{K}_m$ and average values $\bar{V}_m$. Cross-attention is then computed between the modality-specific queries and the aggregated cross-modal context:

$$H'_m = \tilde{H}_m + \mathrm{CrossAttn}\left(Q_m,\, \bar{K}_m,\, \bar{V}_m\right). \tag{2}$$

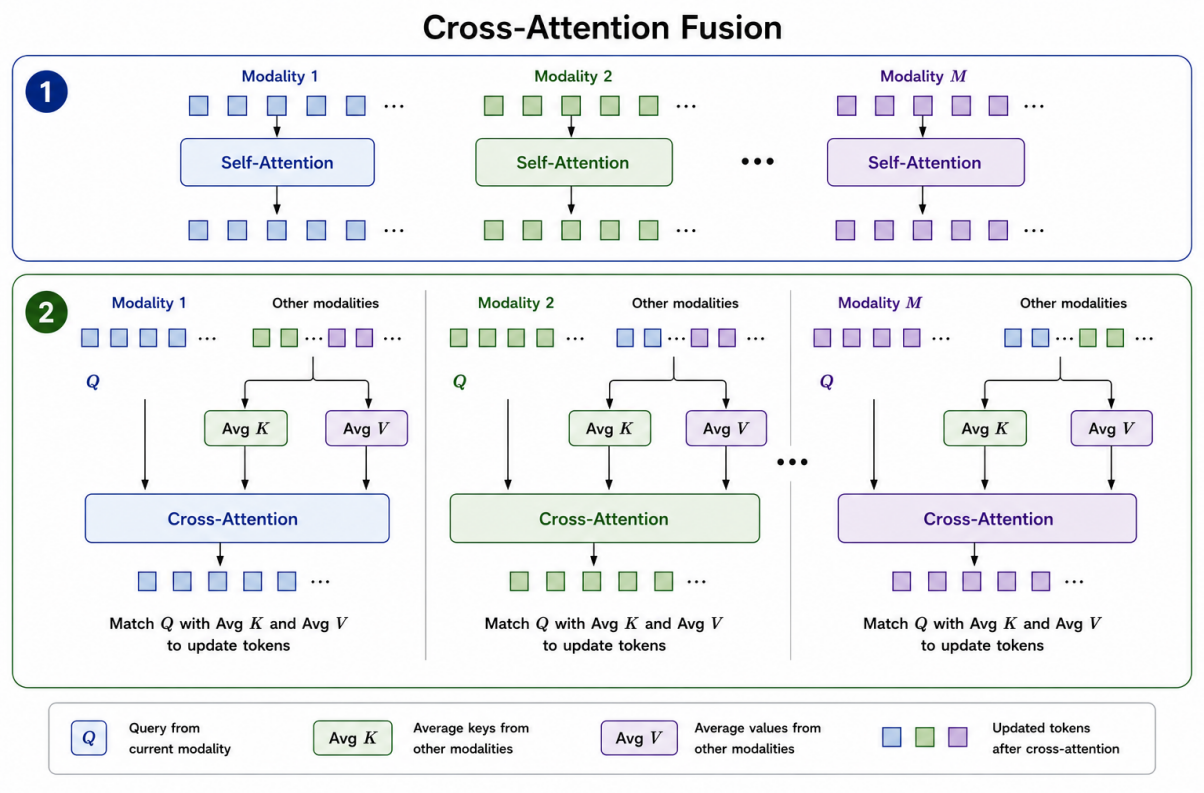

Figure 2: Symmetric cross-attention fusion mechanism. Each modality first undergoes self-attention and is subsequently updated through cross-attention, where its tokens serve as queries (Q) and the remaining modalities provide aggregated keys ($\bar{K}$) and values ($\bar{V}$). No modality is treated as a fixed anchor, allowing information exchange between all active modalities.

This operation injects information from all other modalities into the current modality while maintaining a symmetric fusion strategy, where no modality is treated as a privileged anchor.

From sequences to logits. After the final fusion layer, we collapse each stream into a single D-dimensional vector. For the two text streams, we use the post-fusion `[CLS]` token; for the video and physiological streams, we use a mask-aware mean over all present tokens. These per-modality vectors are concatenated, so the classifier input dimension scales with the number of enabled modalities. The resulting fused representation is passed through a three-layer MLP with GELU activations and dropout, producing the final logits.

4.3. Multi-task heads and losses

The whole architecture described above is shared across the three EXIST 2026 video subtasks; only the very last linear layer of the classifier changes its output dimension. Each subtask is trained independently, with the loss and output activation chosen to match the shape of the task, as summarised in Table 1.

Soft-label targets. For Tasks 3.1 and 3.2 the target is the annotator-vote distribution itself, used directly as a soft label, and the loss is the standard cross-entropy between this distribution and the model's softmax output. For Task 3.3 the soft label is per-class — each of the six labels has its own probability of applying to the video — and the loss is the sum of per-class binary cross-entropies. The `pos_weight` of each class is set to the inverse of its marginal positive rate on the training set, which compensates for the fact that the most severe categories (SEXUAL-VIOLENCE and MISOGYNY-NON-SEXUAL-VIOLENCE) are markedly under-represented.

Task	Outputs	Activation	Loss
3.1 (binary)	2	softmax	soft cross-entropy + Dirichlet perturbation
3.2 (multiclass)	3	softmax	soft cross-entropy + Dirichlet perturbation
3.3 (multi-label)	6	per-class sigmoid	weighted BCE + Beta perturbation

Table 1
Per-task output structure, activation, and loss function used by our model. Tasks 3.1 and 3.2 are mutually exclusive class problems; Task 3.3 is a multi-label problem in which any of the six labels may apply independently.

Configuration	`vlm`	`trn`	`video`	`phys`
`vlm + trn`	✓	✓		
`vlm + video`	✓		✓	
`vlm + phys`	✓			✓
`trn + video`		✓	✓	
`trn + phys`		✓		✓
`video + phys`			✓	✓
`trio`	✓	✓	✓	
`quad`	✓	✓	✓	✓

Table 2
The eight modality configurations evaluated under the same architecture and training protocol. The same set is trained independently for each of the three EXIST 2026 video subtasks.

Soft-label perturbation. On top of the soft-label cross-entropies we apply a simple data-augmentation trick on the label side. Instead of using the raw soft label p as the target directly, at each training step we draw a slightly noisy version $\hat{p}$ from a distribution centred on p and use $\hat{p}$ as the target. For Tasks 3.1 and 3.2 we draw $\hat{p}$ from $\mathrm{Dir}(\alpha \cdot p)$ over the class simplex; for the per-class Bernoulli targets of Task 3.3 we draw each class independently from a Beta distribution centred on the same p. The concentration parameter α controls how aggressive the perturbation is: a large α keeps $\hat{p}$ close to p, a small α allows it to drift further. The motivation is that in EXIST 2026 each video is annotated by only two to three annotators, so the soft label itself is already a noisy estimate of the underlying distribution; resampling it at each step gives the model exposure to this annotator-level uncertainty rather than letting it overfit to a single point estimate. A small label smoothing of $\varepsilon{=}0.05$ is additionally applied to the cross-entropy losses for Tasks 3.1 and 3.2; it is disabled for Task 3.3, where the per-class weighting already counteracts over-confidence on the dominant negative class.

5. Experimental Setup

To answer the research questions formulated in Section 1.3, we train and evaluate eight configurations of the same architecture, listed in Table 2. These configurations cover all six pairwise combinations of the four streams {`vlm`, `trn`, `video`, `phys`}, plus the three-modality *trio* (`vlm + trn + video`) and the four-modality *quad* (`trio + phys`). The same set of configurations is trained independently for each of the three EXIST 2026 video subtasks, resulting in 24 internal model-selection runs. From these experiments, we selected three official submissions per subtask, following the competition limit.

This design separates model analysis from official submission constraints. The pairwise configurations allow us to test whether raw video, structured VLM descriptions, and physiological signals provide complementary information when combined with a single additional stream. The `trio` configuration represents the strongest content-only setting, while the `quad` configuration tests whether physiological signals from independent viewing subjects add value on top of the full content-based model.

5.1. Training and evaluation

All models are trained on a single NVIDIA H200 GPU on the PERUN supercomputer of the Technical University of Košice, Slovakia. We use AdamW with a single linear warm-up phase over the first 10%

	Hard evaluation			Soft evaluation	
Configuration	**ICM-Hard**	**ICM-Hard-Norm**	**Macro-F_1**	**ICM-Soft**	**ICM-Soft-Norm**
`vlm + trn`	0.2582	0.6303	0.7526	0.1877	0.5329
`vlm + video`	0.2864	0.6446	0.7596	0.2554	0.5448
`vlm + phys`	0.2081	0.6050	0.7359	-0.0865	0.4848
`trn + video`	0.2433	0.6228	0.7477	0.0351	0.5062
`trn + phys`	0.1951	0.5985	0.7344	-0.1274	0.4776
`video + phys`	-0.1568	0.4215	0.6117	-0.5908	0.4015
`trio`	0.2901	0.6464	0.7631	0.0072	0.5013
`quad`	**0.3096**	**0.6563**	**0.7695**	**0.2602**	**0.5457**

Table 3
Task 3.1 (binary) results on the official EXIST 2026 test set across the eight modality configurations. ICM-Soft-Norm is the primary metric. Best per column in **bold**.

of training steps, followed by cosine decay to zero. Differential learning rates are applied per parameter group: 1×10^{-5} for the text encoder, 2×10^{-5} for VideoMAE, 5×10^{-5} for the PhysioEncoder, and 1×10^{-4} for the classifier head. We additionally apply layer-wise learning-rate decay with factor 0.9 inside the two large backbones. The bottom 18 layers of XLM-RoBERTa-large and the bottom 8 layers of VideoMAE are frozen throughout training, which we found to be the best speed/accuracy trade-off in preliminary runs.

The batch size is 8 with gradient accumulation over 2 steps, and training runs for 10–15 epochs depending on the subtask. We apply three regularisation strategies. First, *modality dropout* zeroes out one whole modality with probability 0.15 per training step, so that the model does not over-rely on any single stream. Second, we use label smoothing with $\varepsilon{=}0.05$ for the soft cross-entropy losses of Tasks 3.1 and 3.2; this is disabled for Task 3.3, where the objective is multi-label binary cross-entropy. Third, we use early stopping based on validation loss. In preliminary experiments, selecting checkpoints directly by the competition metric ICM-Soft-Norm often led to overly confident predictions, which degraded generalisation on the hidden test set. Validation loss produced better-calibrated probability estimates and more stable downstream results.

Evaluation uses 5-fold stratified cross-validation, with folds stratified jointly by language (EN / ES) and class. Ambiguous samples with no majority label are kept only on the training side of each fold. For each configuration, we select the best checkpoint per fold by validation loss. At inference time, we ensemble the five fold checkpoints by averaging the per-class softmax probabilities for Tasks 3.1 and 3.2, and the per-class sigmoid probabilities for Task 3.3. The resulting probabilities are used as soft predictions, while hard predictions are obtained from the same outputs and submitted in PyEvALL-compatible JSON format.

6. Results and Analysis

This section reports the numerical results of all eight modality configurations on the three EXIST 2026 video subtasks. For each subtask we report both the hard-evaluation metrics (ICM-Hard, ICM-Hard-Norm, macro-F_1) and the soft-evaluation metrics (ICM-Soft, ICM-Soft-Norm).

6.1. Results on Task 3.1 (binary sexism identification)

On Task 3.1 the four-modality `quad` configuration achieves the best Aegis-Athena score on every single metric, both hard and soft, with ICM-Hard-Norm 0.6563 and ICM-Soft-Norm 0.5457. On the official EXIST 2026 all-language hard-evaluation leaderboard this places our system 4th out of 76 ranked submissions, and on the English-only evaluation it ranks 2nd (ICM-Hard-Norm 0.6909). For historical context, the best-ranked all-language system on the comparable EXIST 2025 Task 3.1 hard evaluation reached ICM-Hard-Norm 0.6001 [20]; the cross-year comparison is not formally valid since the test sets differ, but the magnitude of the gap is indicative of progress on this subtask. The runner-up across

Configuration	Hard evaluation			Soft evaluation	
	ICM-Hard	ICM-Hard-Norm	Macro-F_1	ICM-Soft	ICM-Soft-Norm
`vlm + trn`	0.1553	0.5586	**0.6512**	**-0.9847**	**0.3951**
`vlm + video`	**0.1603**	**0.5605**	0.6483	-1.2488	0.3670
`vlm + phys`	-0.1110	0.4581	0.5661	-1.7552	0.3131
`trn + video`	0.0399	0.5151	0.6236	-1.4412	0.3465
`trn + phys`	-0.0086	0.4968	0.5948	-1.6852	0.3205
`video + phys`	-0.6672	0.2671	0.5073	-1.9352	0.3036
`trio`	0.1075	0.5406	0.6326	-1.1649	0.3759
`quad`	0.1191	0.5450	0.6364	-1.0841	0.3845

Table 4
Task 3.2 (source intention) results on the official EXIST 2026 test set across the eight modality configurations. ICM-Soft-Norm is the primary metric. Best per column in **bold**.

Configuration	Hard evaluation			Soft evaluation	
	ICM-Hard	ICM-Hard-Norm	Macro-F_1	ICM-Soft	ICM-Soft-Norm
`vlm + trn`	-0.4748	0.3464	0.4100	-6.6937	0.1008
`vlm + video`	-0.2581	0.4165	0.3918	-7.3756	0.0601
`vlm + phys`	-0.4258	0.3622	0.3241	-7.3978	0.0588
`trn + video`	-0.3154	0.3979	0.3710	-6.9094	0.0879
`trn + phys`	-0.4934	0.3404	0.3832	-7.4713	0.0544
`video + phys`	-0.8564	0.2229	0.2027	-8.1381	0.0146
`trio`	-0.1568	0.4215	0.4106	-6.2080	**0.1297**
`quad`	**-0.1394**	**0.4549**	**0.4193**	**-6.2080**	**0.1297**

Table 5
Task 3.3 (multi-label fine-grained categorisation) results on the official EXIST 2026 test set across the eight modality configurations. Macro-F_1 is computed as the unweighted mean of per-class F_1 scores over the six output labels. ICM-Soft-Norm is the primary metric. Best per column in **bold**.

most columns is the `trio`, indicating that the physiological stream contributes a small but consistent gain on top of the three content-based modalities. Two patterns at the bottom of the table are worth noting. First, the content-free pair `video + phys` collapses to ICM-Hard-Norm 0.4215, confirming that at least one strong textual stream is necessary. Second, configurations that include only one textual stream (`trn + phys`, `vlm + phys`) are noticeably weaker than those that include both, suggesting that the two textual streams are complementary rather than redundant.

6.2. Results on Task 3.2 (source intention)

Task 3.2 produces the most interesting picture in our ablation. The best configuration under *hard* evaluation is `vlm + video` (ICM-Hard-Norm 0.5605, ranking 17th out of 70 on the official all-language hard leaderboard), while the best configuration under *soft* evaluation is the different pair `vlm + trn` (ICM-Soft-Norm 0.3951, ranking 7th out of 49). Both top configurations are pairs rather than the full `quad`, and both phys-containing pairs are markedly weaker. The `video + phys` configuration collapses to ICM-Hard-Norm 0.2671, the lowest score in our entire ablation, and adding the physiological stream on top of the `trio` produces only a marginal gain on hard evaluation while losing slightly on soft. The divergence between hard and soft optima within a single task, together with the disproportionate cost of including physiological signals, is the cleanest signal in our ablation that modality complementarity is task-dependent; we return to this in Section 7.

6.3. Results on Task 3.3 (fine-grained categorisation)

On Task 3.3 the strongest configuration is the four-modality `quad`, which achieves the best score on every hard metric (ICM-Hard-Norm 0.4549, macro-F_1 0.4193) and ties with the `trio` on the primary soft metric at ICM-Soft-Norm 0.1297. This places our submission 3rd out of 69 ranked systems on

the official EXIST 2026 all-language hard-evaluation leaderboard, and **1st out of 69** on the English-only hard evaluation (ICM-Hard-Norm 0.4719). For historical context, the best-ranked all-language EXIST 2025 system on this subtask reached ICM-Hard-Norm 0.3765 [4]; while the test sets differ and a strict comparison is not meaningful, the substantial absolute gap indicates that the multi-label categorisation subtask has seen visible progress across editions. Adding the physiological stream on top of the `trio` brings a further +3.3 ICM-Hard-Norm points despite the well-known difficulty of fine-grained multi-label classification on small datasets, indicating that the sensorial signal carries information that the content-only model does not fully capture even on the hardest of the three subtasks. The absolute scores on this subtask are nevertheless substantially lower than on Tasks 3.1 and 3.2 for every configuration, reflecting the much harder nature of multi-label categorisation on a small, severely imbalanced dataset.

7. Conclusion

We have proposed a multimodal cross-attention architecture for short-video sexism detection that fuses up to four streams — structured VLM-generated scene descriptions, automatic speech transcripts, raw video features, and physiological viewer signals — within a single unified model. The same architecture is trained independently for the three EXIST 2026 video subtasks, with only the classifier head differing in output dimension. Our team *Aegis Athena* ranks 4th out of 76 ranked systems on the all-language hard evaluation of Task 3.1 (ICM-Hard-Norm 0.6563), 17th out of 70 on Task 3.2 (ICM-Hard-Norm 0.5605), and 3rd out of 69 on Task 3.3 (ICM-Hard-Norm 0.4549). On the English-only hard evaluation we rank *first* on Task 3.3 (ICM-Hard-Norm 0.4719) and second on Task 3.1 (ICM-Hard-Norm 0.6909).The corresponding soft-evaluation scores — ICM-Soft-Norm 0.5457, 0.3951 and 0.1297 — rank 8th, 7th and 11th on the all-language leaderboards of their respective subtasks. Beyond these headline results, the systematic ablation across eight modality configurations provides clear answers to the three research questions posed in this work.

RQ1: does raw video add value over textual streams alone? On Task 3.1 the answer is clearly yes: `vlm + video` beats `vlm + trn` on every metric (ICM-Soft-Norm 0.5448 versus 0.5329), and adding the `video` stream on top of the textual pair to form the `trio` adds another 1.6 ICM-Hard-Norm points. On Task 3.2 the role of video is even more pronounced, in that `vlm + video` is the best configuration on hard evaluation, beating both `trio` and `quad`. This is consistent with the intuitive view that source intention is conveyed visually through tone, framing and performance, often more directly than through the words actually spoken. On Task 3.3 the pattern is consistent but weaker: configurations with `video` dominate their video-free counterparts on most metrics, but the absolute gains are small because the task itself is hard for every configuration.

RQ2: do VLM descriptions add value over transcript+video? On all three subtasks, configurations that include `vlm` are the strongest. The cleanest test is the direct comparison `trn + video` versus `vlm + video`: the latter is better on Task 3.1 (ICM-Hard-Norm 0.6446 versus 0.6228), and the gap is much larger on Task 3.2 (0.5605 versus 0.5151). A similar pattern holds when adding `vlm` on top of the `trn + video` pair to form the `trio`. We interpret this as evidence that the structured TARGET and MESSAGE fields generated by the VLM contain semantic information that neither the raw video nor the transcript supplies on its own, particularly for intention detection where the TARGET field directly encodes who the message is aimed at.

RQ3: do physiological signals add value beyond content-based modalities? On all three subtasks, the four-modality `quad` configuration performs at least as well as the content-only `trio` under the primary metric: ICM-Soft-Norm rises from 0.5013 to 0.5457 on Task 3.1 (a gain of 0.044), from 0.3759 to 0.3845 on Task 3.2 (a small but consistent gain), and stays unchanged at 0.1297 on

Task 3.3. The picture is similar on hard evaluation, where adding `phys` yields a clear $+1.0$ ICM-Hard-Norm point on Task 3.1 and $+3.3$ ICM-Hard-Norm points on Task 3.3, and a small $+0.4$ point gain on Task 3.2. We therefore answer RQ3 in the affirmative: when the three content modalities are already in place, the physiological stream is genuinely complementary information rather than redundant noise. The magnitude of the gain, however, scales with how strong the content-only baseline already is and with how much calibration headroom the metric leaves — the largest improvement appears on the Task 3.1 soft metric, where the `trio` under-performs its hard score and the additional stream supplies information the content modalities cannot recover on their own.

Why pairs with physiological signals struggle. The picture changes when `phys` is paired with only a single content modality. Three such pairs appear in our ablation, and the relative ranking between them is informative about *which* content anchor the physiological signals can be combined with most usefully.

`video + phys` is by a clear margin the weakest configuration on every subtask — ICM-Hard-Norm 0.4215, 0.2671 and 0.2229 on Tasks 3.1, 3.2 and 3.3 respectively, in each case roughly twenty points below the best configuration for that subtask. Raw video features alone carry little of the contextual information needed to adjudicate sexism: who the addressee is, whether the speaker is performing or sincere, whether a scene is being parodied or shown straight. The physiological stream is itself a fairly weak signal, recorded from only two subjects per video, and pairing two such weak streams without a strong textual anchor leaves the model with no reliable way to ground the prediction.

`trn + phys` performs noticeably better but is still clearly behind the textual pair `vlm + trn` and the textual-visual pair `vlm + video` on every subtask. The transcript does provide a content anchor, but on TikTok content this anchor is often thin: many videos contain only short utterances, song lyrics, or voice-over fragments, and the transcript alone does not capture the visual performance through which sexism is frequently delivered. Combined with the noisy physiological signal, this leaves `trn + phys` able to recover the easy cases but not the implicit or visually-mediated ones.

`vlm + phys` is the strongest of the three phys-paired configurations on Task 3.1 (ICM-Hard-Norm 0.6050), which we attribute to the VLM-text stream being a stronger content anchor than either the raw video or the transcript alone: its TARGET and MESSAGE fields make the discriminative information explicit and text-based, which then leaves the physiological stream with something concrete to complement. Even here, however, the configuration is weaker than `vlm + video` and `vlm + trn`. Taken together, this confirms the picture that emerges from the RQ3 analysis above: physiological signals are at their most useful as a *fourth* modality added on top of a strong multi-stream content backbone, rather than as a substitute for any single content stream.

Limitations. Three further limitations of the present work deserve mention. First, the VLM-text stream is produced offline by a single 30B-parameter vision–language model; a smaller VLM or a different prompt structure might trade quality for inference cost and would be worth studying. Second, soft labels in EXIST 2026 are estimated from only two to three annotators per video, so even our soft-label perturbation scheme is bounded by the underlying label noise. Third, our ablation deliberately excludes the raw audio stream, following prior EXIST findings; whether a strong audio encoder (e.g. Wav2Vec 2 or a foundation audio model) could match or exceed the gain from the VLM description is an open question.

Future work. Two architectural extensions seem particularly promising. The first is the combination of cross-attention fusion with a stronger audio modality: if audio provides intention-bearing cues currently funnelled implicitly through the VLM description, an explicit audio stream should be able to do this more directly. The second is to strengthen the physiological branch in two complementary ways. On the data side, denoising or multi-annotator alignment preprocessing should reduce the subject-level noise that we believe currently limits the contribution of `phys`. On the model side, our *PhysioEncoder* is trained from scratch on the small EXIST 2026 sensorial set; replacing it with a pre-trained backbone

— for example, a foundation model for EEG such as LaBraM or BIOT, an ECG-pretrained encoder for the heart-rate stream, or a domain-specific eye-tracking transformer — would let the physiological stream inherit representations learned from orders of magnitude more recording hours than EXIST itself provides, which we expect to grow the current gain from a modest complement to a more substantial driver of performance. Finally, the structured VLM-description pattern we use here is independent of the downstream classifier and should transfer naturally to other multimodal harm-detection tasks — hate, harassment, bullying — where the visual scene carries information that the spoken transcript does not.

Acknowledgments

This research was funded by the Slovak Research and Development Agency under the contract No. APVV 22-0414

Declaration on Generative AI

During the preparation of this work, the author(s) used `Qwen3-VL-30B` to generate structured vision–language scene descriptions from the raw video data used as one of the input modalities in the proposed architecture. In addition, ChatGPT (OpenAI) was used to produce the architecture and fusion diagrams, and Claude (Anthropic) was used for grammar checking and language polishing.

After using these tools, the author(s) reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Iroegbu, M., O'Brien, F., Muñoz, L.C., Parsons, G.: Investigating the psychological impact of cyber-sexual harassment. Journal of Interpersonal Violence **39**(15–16), 3424–3445 (2024). https://doi.org/10.1177/08862605241231615

[2] Dahlqvist, H.Z., Landstedt, E., Young, R., Gådin, K.G.: Dimensions of peer sexual harassment victimization and depressive symptoms in adolescence: A longitudinal cross-lagged study in a Swedish sample. Journal of Youth and Adolescence **45**(5), 858–873 (2016). https://doi.org/10.1007/s10964-016-0446-x

[3] Regehr, K., Shaughnessy, C., Minzhu, Z., Shaughnessy, N.: Safter scrolling: How algorithms popularise and gamify online hate and misogyny for young people Technical report, Association of School and College Leaders, University College London, University of Kent (2024) https://www.alignplatform.org/resources/safter-scrolling-how-algorithms-popularise-and-gamify-online-hate-and-misogyny-young

[4] Plaza, L., Carrillo-de-Albornoz, J., Arcos, I., Rosso, P., Spina, D., Amigó, E., Gonzalo, J., Morante, R.: Overview of EXIST 2025: Learning with disagreement for sexism identification and characterization in tweets, memes, and TikTok videos (extended overview). In: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 4038. CEUR-WS.org (2025). https://ceur-ws.org/Vol-4038/paper_135.pdf

[5] Arcos, I., Rosso, P.: Sexism identification on TikTok: A multimodal AI approach with text, audio, and video. In: Goeuriot, L., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2024. LNCS, vol. 14958, pp. 61–73. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-71736-9_2

[6] Arora, A., Nakov, P., Hardalov, M., Sarwar, S.M., Nayak, V., Dinkov, Y., Zlatkova, D., Dent, K., Bhatawdekar, A., Bouchard, G., Augenstein, I.: Detecting harmful content on online platforms: What platforms need vs. where research efforts go. ACM Computing Surveys **56**(3), article 72, 1–17 (2023). https://doi.org/10.1145/3603399

[7] Amigó, E., Delgado, A.D.: Evaluating extreme hierarchical multi-label classification. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 5809–5819 (2022). https://doi.org/10.18653/v1/2022.acl-long.399
[8] De Grazia, L., Martín-Lara, D., Albacete, S., Plaza-del-Arco, F.M., Plaza, L.: MuSeD: A multimodal Spanish dataset for sexism detection in social media videos. arXiv preprint arXiv:2504.11169 (2025). https://arxiv.org/abs/2504.11169
[9] Mazhar, A.A., Zada, I., Aldhayan, M., Alsalamah, S., Asiri, M.M., Ayadi, M., Alshahrani, A., Anwar, M.S.: AI-powered detection of cyberbullying in short-form video content: A hybrid deep learning framework. PLOS ONE **21**(2), e0338799 (2026). https://doi.org/10.1371/journal.pone.0338799
[10] Ali, M., Yendapalli, L., Tawfik, B., Winzenried, M.: Tackling sexism in multimodal social media: Exploring hybrid generative-transformer models. In: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 4038. CEUR-WS.org (2025). https://ceur-ws.org/Vol-4038/paper_139.pdf
[11] Plaza, L., Carrillo-de-Albornoz, J., Ruiz, V., Maeso, A., Chulvi, B., Rosso, P., Amigó, E., Gonzalo, J., Morante, R., Spina, D.: Overview of EXIST 2024 — Learning with Disagreement for Sexism Identification and Characterization in Tweets and Memes. In: Goeuriot, L., et al. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2024. LNCS, vol. 14958, pp. 93–117. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-71908-0_5
[12] Caselli, T., Basile, V., Mitrović, J., Granitzer, M.: HateBERT: Retraining BERT for abusive language detection in English. In: Proceedings of the 5th Workshop on Online Abuse and Harms, pp. 17–25 (2021). https://arxiv.org/abs/2010.12472
[13] Ma, J., Li, R.: RoJiNG-CL at EXIST 2024: Leveraging large language models for multimodal sexism detection in memes. In: Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 3740. CEUR-WS.org (2024). https://ceur-ws.org/Vol-3740/paper-100.pdf
[14] Alcántara, T., Garcia-Vazquez, O., Calvo, H., Valdez-Rodríguez, J.E.: CogniCIC at EXIST 2025: Identifying sexist content in text and visual media using transformers and generative AI models. In: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 4038. CEUR-WS.org (2025). https://ceur-ws.org/Vol-4038/paper_138.pdf
[15] Das, M., Raj, R., Saha, P., Mathew, B., Gupta, M., Mukherjee, A.: HateMM: A multi-modal dataset for hate video classification. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 17, pp. 1014–1023 (2023). https://arxiv.org/abs/2305.03915
[16] Wang, H., Tan, R.Y., Naseem, U., Lee, R.K.-W.: MultiHateClip: A multilingual benchmark dataset for hateful video detection on YouTube and Bilibili. In: Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), pp. 7493–7502 (2024). https://doi.org/10.1145/3664647.3681521
[17] Zhao, et al.: Enhanced multimodal hate video detection via channel-wise and modality-wise fusion (CMFusion). arXiv preprint arXiv:2505.12051 (2025). https://arxiv.org/abs/2505.12051
[18] Cespedes-Sarrias, et al.: MM-HSD: Multi-modal hate speech detection in videos. arXiv preprint arXiv:2508.20546 (2025). https://arxiv.org/abs/2508.20546
[19] Yang, S., Chen, T., Yue, J., Cheng, G., Jiao, J., Fu, Z.: Reasoning-Aware Multimodal Fusion for Hateful Video Detection. arXiv preprint arXiv:2512.02743 (2025). https://arxiv.org/abs/2512.02743
[20] Fernández García, D., Amigó Cabrera, E., Cardeñoso Payo, V.: ECA-SIMM-UVa at EXIST 2025: A segmentation oriented approach to sexism detection in TikTok videos based on a "One Is Enough" paradigm. In: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, vol. 4038. CEUR-WS.org (2025). https://ceur-ws.org/Vol-4038/paper_151.pdf
[21] Plaza, L., Carrillo-de-Albornoz, J., Arcos, I., Aloy-Mayo, M., Rosso, P., García-Arias, E., Spina, D.: 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).
[22] Plaza, L., Carrillo-de-Albornoz, J., Arcos, I., Aloy-Mayo, M., Rosso, P., García-Arias, E., Spina, D.: Overview of EXIST 2026: Physiological Data for Multimodal Sexism Characterization in Social

Media (Extended Overview). In: Sánchez Salido, E., Barrón-Cedeño, A., García Seco de Herrera, A., MacAvaney, S., Struß, J.M. (eds.) Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings. CEUR-WS.org (2026).

[23] Cao, R., Hee, M.S., Kuek, A., Chong, W.H., Lee, R.K.-W., Jiang, J.: Pro-Cap: Leveraging a frozen vision–language model for hateful meme detection. In: Proceedings of the 31st ACM International Conference on Multimedia (MM '23), pp. 5244–5252 (2023). https://doi.org/10.1145/3581783.3612498

[24] Tsai, Y.-H.H., Bai, S., Liang, P.P., Kolter, J.Z., Morency, L.-P., Salakhutdinov, R.: Multimodal Transformer for unaligned multimodal language sequences. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 6558–6569 (2019). https://doi.org/10.18653/v1/P19-1656

[25] Zhang, H., Wang, Y., Yin, G., Liu, K., Liu, Y., Yu, T.: Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis (ALMT). In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 756–767 (2023). https://arxiv.org/abs/2310.05804

A. VLM scene-description prompt

This appendix gives the full text of the prompt used to generate the structured VLM scene descriptions described in Section 4. The prompt is bilingual; the English version is reproduced verbatim below, and a Spanish-language version following the same structure is used for Spanish videos.

System prompt (English)

```
You are a careful multimodal describer for short-form videos.
Your job is to explain what is happening in the video - the
situation, the people, and what the video is about - so that a
downstream text model can reason about its social content.

Describe at two levels:
1. What is literally shown and said (people, actions, words, framing).
2. What is happening in context (what kind of situation this is, who
   is the subject of attention or commentary, what the speaker's point
   or message seems to be, and the tone of the interaction).

You MAY interpret the situation at the level of "what is going on
here" - e.g. "a man in a wig parodying stereotypical feminine
behavior", "a woman giving a serious critique of a political idea",
"two friends joking with each other". This situational framing is
exactly what is needed.

You MUST NOT output moral or normative judgments - do not say whether
something is sexist, misogynistic, degrading, offensive, empowering,
respectful, or harmful. Do not output yes/no labels for any such
property. Describe the situation richly and let the reader judge.

Other rules:
- Do not guess the identity of real people unless a name is clearly
  shown.
- When describing people, note apparent gender presentation,
  approximate age range, and role in the scene. Mention visible
  ethnicity or skin tone only when it is contextually relevant to what
  is being discussed.
- Note whether a person appears to be performing, parodying, sincerely
  speaking, reacting, narrating, or being talked about by someone else.
- If the speaker's tone is notable (sarcastic, mocking imitation,
  deadpan, shouting, exaggerated), say so.
- If there is a notable contrast between what is said and what is
  shown, mention it.
- If the camera lingers on, zooms into, or repeatedly cuts to specific
  body parts or specific people in a notable way, mention it.
- Be compact. Respect the sentence limits given for each section. Do
  not repeat information across sections.
- Every requested section must be filled. Never leave a section empty.
- Output plain text using the exact section headers requested. No
  JSON, no markdown fences, no preamble, no closing remarks.
- Write in English.
```

Section instructions (English)

```
Produce a compact description using the following five sections in
this exact order. Respect the sentence limits strictly. Every section
must be present and non-empty.

[SCENE]
One sentence only. Name the format and the visual setting in a few
words. Common formats include: talking head, GRWM (get-ready-with-me
where someone does makeup/hair while speaking), storytime, skit,
dance, meme with text overlay, gameplay with voiceover, slideshow,
reaction video, street interview, vlog, animation, cooking/activity
with voiceover, or similar. If the speaker is doing a parallel
activity while talking (makeup, cooking, driving, walking), say so
explicitly - do not call it a plain talking head.

[SITUATION]
The most important section. In 2-4 sentences, explain what is
actually going on in this video in context: is someone telling a
story, performing a character, parodying a type of person, giving an
opinion, reacting to something, being interviewed, joking with a
friend, lip-syncing, dancing, demonstrating something? If someone is
playing a character that is not themselves (e.g. a man dressed as a
woman, or exaggerating a stereotype), state that clearly. If the
speaker's tone is notable (sarcastic, mocking imitation, deadpan,
shouting), mention it here. If there is a notable contrast between
what is said and what is shown, or if the camera lingers or zooms in
a notable way, mention it here in one short clause.

[PEOPLE]
1-3 sentences. For each distinct person: apparent gender presentation,
approximate age range, and role in the scene (main speaker, target of
commentary, silent background, interviewer, interviewee, etc.). Note
voiceover-only or on-screen-only participants. Note whether anyone is
performing, parodying, or being talked about rather than speaking.
Mention ethnicity or skin tone only when the video's content makes it
contextually relevant.

[MESSAGE]
1-3 sentences. What the video is saying, claiming, joking about, or
reacting to. What is the speaker's apparent point, or the joke's
apparent punchline? Stay descriptive: "the speaker argues that X",
"the joke is that Y", not "this is good/bad". If there is no clear
message (e.g. pure dance or aesthetic video), say so in one sentence.

[TARGET]
1-2 sentences. Who or what the video is about, focused on, commenting
on, reacting to, or making fun of. This may or may not be someone
present on screen - it could be an absent group, a public figure, a
stereotype, an idea, a movement, or the main subject themselves. If
the video criticizes, comments on, or jokes about an idea, a movement,
or a subset of a group, name the target specifically and, when useful,
clarify who is NOT the target. If the video is not about anyone or
anything in particular, say so.
```