ELiRF-UPV at EXIST 2026: Multimodal Fusion and Few-Shot LLMs for Sexism Detection in Memes and Videos

EXIST at CLEF 2026

Juan Alfonso García-Real[1,*], Vicent Ahuir[1,2,†] and María José Castro-Bleda[1,2,3,†]

[1]Universitat Politècnica de València, Camí de Vera, s/n, 46022 València, Spain

[2]ELiRF Research Group, VRAIN, Universitat Politècnica de València, Camí de Vera, s/n, Building 1F, 46022 València, Spain

[3]ValgrAI: Valencian Graduate School and Research Network of Artificial Intelligence, Universitat Politècnica de València, Camí de Vera, s/n, Building 1F, 46022 València, Spain

Abstract

This paper describes the participation of ELiRF-UPV in Tasks 2 and 3 of EXIST 2026 at CLEF 2026, covering all six subtasks of sexism identification, source-intention detection, and fine-grained sexism categorization in memes and TikTok videos. Following the official protocol, we report ICM and F1 in the hard–hard setting and ICM-Soft in the primary soft–soft setting, as computed by PyEvALL.

We submit two complementary families of systems, both applied to memes and videos and both relying on neutral Qwen3-VL-8B captioning to derive a visual description and an independent gender-relevance analysis that avoids confirmatory bias. The first, PhysioMeme-Fusion, is a trained multimodal system with three per-subtask heads optimized directly on soft labels—KL divergence for the mono-label heads and binary cross-entropy for the multi-label categorization—each with an explicit "No" class so that the soft metric drives learning. Its core pairs language-specific emotion-tuned encoders for English and Spanish with an enriched textual view of each instance (the on-screen text combined with the neutral visual description and the gender-relevance analysis, denoted `capsanocr`), and adds a physiological branch encoding the eye-tracking, heart-rate, and EEG signals released for the lab subjects, fused through a residual cross-attention block with modality dropout; an optional dense visual stream—SigLIP2 for memes and a VideoPrism video encoder for videos—is explored as an add-on. The second family is a training-free pipeline that prompts Qwen3.5-27B and Gemma-4-31B in a hierarchical cascade, passing meme images or sampled video frames directly to the model alongside the released text.

Our strongest result is the training-free few-shot pipeline: under the hard–hard evaluation it is competitive across both modalities and, in the official by-team ranking, places first among teams on video identification (T3.1), second on meme categorization (T2.3), and third on video intention (T3.2), which shows that a definition-rich hierarchical prompt is a powerful, data-free baseline for multimodal sexism detection. On the trained side, the soft-optimized fusion reaches the upper part of the primary soft–soft leaderboard on meme identification (T2.1, seventh among teams), and we find that the enriched textual view—building on the VLM-derived enrichment of Arcos et al.—carries the system, whereas the physiological branch yields no soft-metric gain (and is at times detrimental on the fine-grained categorization) and the dense visual stream degrades performance wherever the verbalized view is already present.

Keywords

sexism detection, multimodal learning, learning with disagreement, soft labels, physiological signals, vision-language models, few-shot learning, memes, TikTok videos, EXIST 2026

1. Introduction

Sexist discourse online is increasingly multimodal, expressed through memes and short videos whose meaning emerges from the interplay of image, text, and context and often turns on irony, stereotypes, and culturally situated knowledge. Humour and plausible deniability let it circulate while remaining ambiguous even to human readers, so sexism detection cannot be reduced to text classification [1].

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

*Corresponding author.

†Equal contribution.

jagarrea@etsinf.upv.es (J. A. García-Real); vahuir@upv.es (V. Ahuir); mcastro@dsic.upv.es (M. J. Castro-Bleda)

https://github.com/juangarcia83 (J. A. García-Real); https://www.upv.es/ficha-personal/vahuir (V. Ahuir); https://www.upv.es/ficha-personal/mjcastro (M. J. Castro-Bleda)

0009-0006-1840-6676 (J. A. García-Real); 0000-0001-5636-651X (V. Ahuir); 0000-0003-1001-8258 (M. J. Castro-Bleda)

The EXIST shared-task series has tracked this shift over six editions, from text-only tweets (2021–2023) to memes (2024) and TikTok videos (2025), under the Learning with Disagreement (LeWiDi) paradigm, which treats annotator disagreement as signal rather than noise [2, 3]. EXIST 2026 is exclusively multimodal—six subtasks covering binary identification, source-intention detection, and fine-grained categorization across memes (Task 2) and short videos (Task 3)—and enriches the corpus with physiological signals (eye-tracking, heart rate, and EEG) recorded while subjects viewed each stimulus [4]: auxiliary cues to implicit responses that people register physiologically even when they cannot articulate why [1].

This raises three challenges. First, the primary metric (ICM-Soft, computed with PyEvALL [5]) rewards predicted distributions that match the empirical annotator distribution. Second, the corpus is bilingual, and earlier EXIST systems report sizeable cross-lingual gaps under a single shared encoder [6, 7]. Third, vision-language models prompted to judge sexism directly are prone to confirmatory bias, over-assigning the positive class and inflating reported performance [8, 7].

We explore two families of systems, both applied to memes (Task 2) and short videos (Task 3). Our main contributions are: (i) *PhysioMeme-Fusion*, a trained multimodal architecture built on language-specific emotion-tuned encoders over an enriched textual view of each instance (on-screen text combined with a Qwen3-VL-8B [9] visual description and a gender-relevance analysis, denoted `capsanocr`) and a physiological branch, fused by a residual cross-attention block with modality dropout, plus an optional dense visual stream (SigLIP2 [10] for memes and a VideoPrism [11] video encoder for videos); (ii) three per-subtask heads trained directly on soft labels—KL divergence for the mono-label heads and binary cross-entropy for the multi-label categorization, each with an explicit "No" class—so that the LeWiDi metric drives learning; (iii) a training-free, hierarchical-conditional few-shot pipeline over Qwen3.5-27B [12] and Gemma-4-31B [13] that gates the fine-grained subtasks on the binary decision and is applied to both modalities; and (iv) a controlled comparison between the two families under the same official protocol.

2. Related Work

2.1. Multimodal Sexism in Memes and Videos

Memes spread prejudice under the guise of humour, and benchmarks such as MAMI [14] established misogynous meme detection as a multimodal task requiring both textual and visual evidence. Recent EXIST systems follow this direction: ArcosGPT [7] improves over an OCR-only baseline by enriching the text with BLIP captions and GPT-generated descriptions, with its best ViT+RoBERTa fusion reaching an F1-Macro of 0.83 (about +0.19 over text-only), while MemeWeaver [6] adds inter-meme graph reasoning over multimodal representations. For short videos, MuSeD [8] and FineMuSe [15] show that multimodal LLMs handle explicit sexism better than implicit cases and that textual cues remain critical even in visual content. This motivates our Task 3 design, which keeps both channels—sampled frames and the released transcription and on-screen text—available to the model.

2.2. Learning with Disagreement and Few-Shot MLLMs

LeWiDi treats annotation variability as signal rather than noise [3, 16]: soft-trained models do better under soft metrics and gold-trained ones under hard metrics, a tension EXIST resolves by reporting both. EXIST uses ICM and its soft extension ICM-Soft [5], which natively handle hierarchical, multi-label, and soft assignments, so predictions must be distributions and the loss must match the metric—which our per-subtask heads address directly. Separately, instruction-following multimodal LLMs classify with little or no training via definition-guided prompting [8], but risk confirmatory bias: prompted to judge sexism directly, a VLM rationalizes a positive label even for benign content, inflating recall [7]. We therefore use a two-stage neutral protocol (separate description and gender-relevance prompts, kept apart from the decision) and prompt Qwen3.5-27B and Gemma-4-31B in a hierarchical cascade.

2.3. Physiological Signals in NLP

Eye-tracking, heart rate, and EEG have long been used to capture implicit affective and cognitive responses, but their integration into NLP-oriented classification pipelines is still relatively recent. Annotator gaze has been shown to predict subjective hatefulness ratings and to improve text-only hate speech detection models [17], while EEG and eye-tracking have also been combined with LLM-derived word relevance to model word-level neural states during reading [18]. Most directly, Arcos et al. [1] recorded ET, HR, and EEG from 16 subjects over the 3,984 EXIST memes and found significant sexist/non-sexist processing differences in eye-tracking and EEG features, but no significant heart-rate effect; their fusion model reached an AUC of 0.794, a 3.4% relative improvement over a strong VLM-based baseline.

2.4. Positioning with Respect to Arcos et al. (2026)

Our work is most closely related to Arcos et al. [1], with whom we share the use of physiological signals, VLM-derived enriched textual representations, and attention-based fusion on the EXIST meme corpus. Their work introduces a human-centered multimodal resource by recording eye-tracking, heart rate, and EEG responses while subjects viewed EXIST memes, and proposes a fusion model that combines VLM-generated textual descriptions and preliminary sexism analyses with XLM-RoBERTa [19] text representations and physiological features through cross-attention. We differ in four respects.

First, whereas their fusion system is trained with a weighted binary cross-entropy objective and evaluated mainly under an AUC/F1-oriented regime, with binary sexism detection as the headline result and intention and categorization also reported, we train three per-subtask heads directly against the official soft-label objective: KL divergence for the mono-label subtasks and binary cross-entropy for the multi-label categorization, each with an explicit "No" class. This makes the optimization more directly aligned with the LeWiDi setting and the official soft evaluation.

Second, although both approaches enrich OCR text with VLM-generated visual and sexism-related descriptions, our `capsanocr` view (Section 4) is built from two separate neutral prompts and is encoded with language-specific emotion-tuned encoders, rather than the single XLM-RoBERTa backbone used in their fusion classifier. This choice is motivated by the language-dependent performance gaps reported in previous EXIST systems [7]. We also evaluate an optional dense visual stream—SigLIP2 for memes and a VideoPrism video encoder for videos—as an extension rather than as part of the core architecture.

Third, while Arcos et al. obtain text-aware physiological representations through cross-attention and concatenate them with the main textual representation, our fusion keeps the text representation as a residual anchor and injects physiological evidence as a residual cross-attention term. In our formulation, the physiological descriptor acts as a query over the textual token sequence, so the physiological signal can reweight textual evidence without forming an independent decision pathway. Modality dropout is further used to handle missing or unreliable physiological signals.

Fourth, we extend the human-centered multimodal setting beyond memes to short videos (Task 3) and pair the trained soft-label system with a training-free hierarchical few-shot family. This enables a controlled comparison between soft-optimized multimodal fusion and definition-guided large-VLM prompting under the same official protocol.

3. Tasks, Data and Evaluation

3.1. Tasks 2 and 3: Memes and Videos

Task 2 (memes) and Task 3 (short videos) share the same three-subtask structure; videos additionally carry temporal, acoustic, and overlay-text cues. The subtasks are: **x.1 — Sexism identification**, a binary decision (YES/NO) on whether the instance is sexist, depicts a sexist situation, or criticizes one; **x.2 — Source intention**, classifying a sexist instance as DIRECT (sexist by itself or inciting sexism) or JUDGEMENTAL (reporting or condemning it), dropping the REPORTED class of earlier tweet editions; and **x.3 — Sexism categorization**, a multi-label assignment over five categories (IDEOLOGICAL-

INEQUALITY, STEREOTYPING-DOMINANCE, OBJECTIFICATION, SEXUAL-VIOLENCE, and MISOGYNY-NON-SEXUAL-VIOLENCE).

3.2. Dataset

The bilingual (EN/ES) EXIST 2026 corpus builds on previous EXIST editions, reusing the EXIST 2024 memes and the EXIST 2025 TikTok videos within a new human-centered experimental setting enriched with physiological signals (eye-tracking, heart rate, and EEG) [4]: 3,984 training / 1,053 test memes and 2,524 training / 674 test videos. The two languages are nearly balanced for memes, while the video training partition is skewed toward Spanish. Each instance provides the extracted `text`, the image or video, per-annotator labels, and annotator metadata, enabling both hard-label supervision and the soft labels—the empirical label distribution—required by LeWiDi.

3.3. Physiological Resource

Distinct from the annotation layer, EXIST 2026 releases physiological recordings from a separate pool of subjects who viewed the stimuli under instrumentation [4]: eye-tracking (fixations, saccades, pupil diameter, reaction time), heart rate, and EEG band power. Each stimulus was seen by only 2–4 lab subjects, so the physiological branch must aggregate across few subjects and tolerate missing modalities. We use the three modalities as complementary, content-level evidence rather than labels and ablate their individual contributions (Section 4); prior work on the same memes found no heart-rate effect while eye-tracking and EEG carried discriminative signal [1], a pattern our ablations re-examine.

3.4. Evaluation

Following the official protocol, all subtasks are evaluated with PyEvALL using ICM and its soft extension ICM-Soft [5], an information-theoretic similarity that natively supports hierarchical, multi-label, and (in the soft case) probabilistic assignments, rewarding predictions that match the empirical annotator distribution. We report two settings per subtask: *hard–hard* (also reporting F1 and normalized ICM) and the primary *soft–soft*. The hierarchy is enforced at evaluation time through PyEvALL rather than during training, keeping our per-subtask heads decoupled. Hard labels follow per-task vote thresholds: for Task 2 memes, T2.1 requires more than three of the six votes, T2.2 the intention chosen by more than two annotators, and T2.3 a category selected by more than one; Task 3 videos use a threshold of more than one annotator for all three subtasks.

4. System A: PhysioMeme-Fusion for Memes and Videos

PhysioMeme-Fusion is our trained multimodal system, applied to both memes (Task 2) and short videos (Task 3) with a single shared architecture. It encodes each instance through two streams: a language-specific emotion-tuned text stream over an enriched textual view—a neutral Qwen3-VL-8B [9] visual description, an independent gender-relevance analysis, and the literal on-screen text, a representation we denote `capsanocr` (Section 4.1)—and a physiological stream. What changes between tasks is only the composition of the textual view (overlay text for memes; transcription plus on-screen text for videos) and the number of lab subjects and annotators per instance; the encoders, fusion, and heads are identical. The two streams are fused by a residual cross-attention block in which the physiological representation attends over the textual features, and three per-subtask heads are trained on soft labels with an explicit "No" class, so that non-sexism is a learned outcome rather than a thresholded residual. Modality dropout keeps the model robust when part of the physiological descriptor is unavailable. An optional dense visual stream—SigLIP2 for memes and a VideoPrism video encoder for videos—with a bilinear text–image interaction is evaluated as an add-on rather than a core component, and the text-only baseline—the same encoder and heads without the cross-attention block—isolates the contribution of the lab signals as a clean ablation. Figure 1 gives an overview.

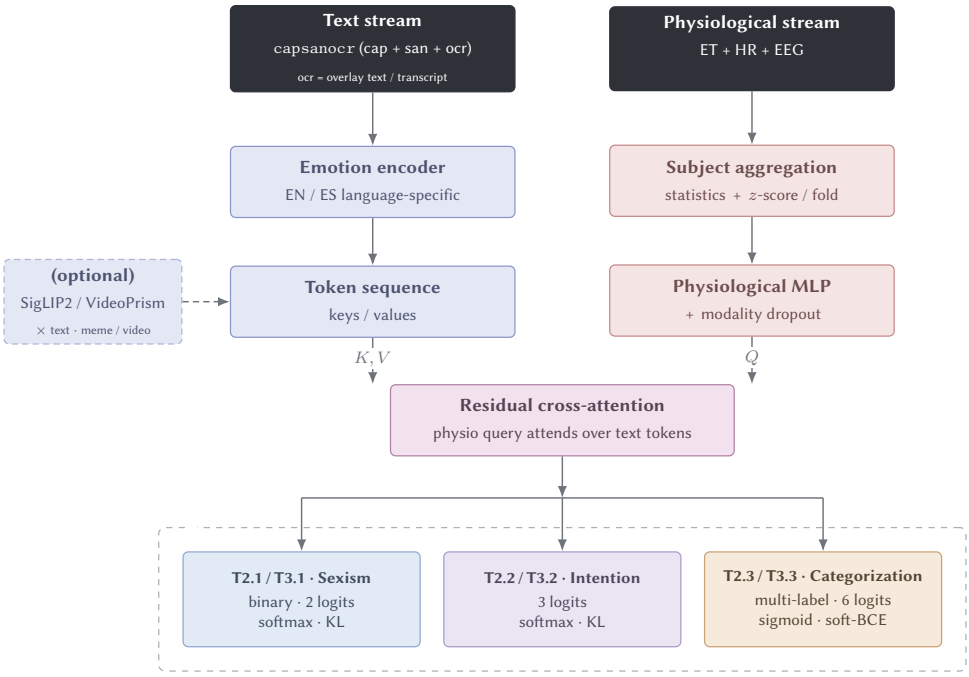

Figure 1: PhysioMeme-Fusion architecture, shared by Tasks 2 and 3. A language-specific emotion text stream over `capsanocr`—whose `ocr` component is the meme overlay text for Task 2 and the released transcription and on-screen text for Task 3—and a subject-aggregated physiological stream are fused by a residual cross-attention block (a physiological query attends over the text tokens) and read by three soft-label heads. The optional dense visual stream (dashed) couples a raw visual encoder to the text through a bilinear interaction—SigLIP2 over the meme image for Task 2 and a VideoPrism video encoder over the clip for Task 3; modality dropout makes the model robust when the physiological signal is absent.

4.1. Textual Encoding and Optional Visual Stream

The core of the system is a textual stream that consumes an enriched textual view of the instance, which we denote `capsanocr`. The idea of enriching the on-screen text with a VLM-generated visual description and a preliminary sexism analysis follows Arcos et al. [1], who combine OCR with a VLM caption and analysis as their enriched content representation; our contribution here is the per-language, two-prompt construction described below and the `capsanocr` naming we adopt for it. The view is the concatenation, with a `[SEP]` separator and always in the language of the instance, of three complementary sources:

- `cap` — a neutral *visual description* produced by Qwen3-VL-8B, which verbalizes what the content depicts (people, gestures, setting, juxtapositions) without judging it; for memes it describes the image and for videos it describes the sampled frames;
- `san` — an independent *gender-relevance analysis*, also produced by Qwen3-VL-8B through a separate prompt, that reasons about whether and how the content relates to sexism;
- `ocr` — the literal *on-screen text*: the overlay text of the meme for Task 2 (the `text` field of the dataset), and the transcription together with the on-screen text released with the corpus for Task 3.

The name `capsanocr` is our shorthand for the concatenation `cap` + `san` + `ocr`. Generating the visual description and the gender-relevance analysis with two *separate* prompts—rather than a single combined one—is a deliberate departure that keeps the sexism reasoning apart from the final classification decision, which mitigates the confirmatory bias discussed in Section 2. The construction is tolerant to missing sources, and the simpler views—on-screen text alone (`ocr`) and description-plus-analysis without it (`capsan`)—are retained as ablations. Because the description and the analysis already verbalize the visual content, the text stream alone provides the main semantic grounding of the instance.

As an optional extension, a dense visual stream encodes the raw visual content and couples it to the text representation through the bilinear interaction of Section 4.4. For memes this stream is SigLIP2 over the raw meme image; for videos it is a VideoPrism video encoder (`google/videoprism-lvt-large-f8r288`) that embeds the clip directly rather than per individual frame. Conceptually, this stream is an *alternative route* to the same visual signal that `cap` already carries: whereas `cap` verbalizes the image (or the sampled frames) as text through Qwen3-VL-8B, SigLIP2 / VideoPrism encode the raw pixels directly as a dense visual representation, so the two paths differ only in how the visual content reaches the model (as a description versus as an embedding) and not in what they aim to capture. We treat it as an add-on whose contribution is measured against the text-only core rather than as a default component.

4.2. Language-Specific Emotion Encoders

Because both corpora are bilingual, the textual stream uses an emotion-tuned encoder per language: `j-hartmann/emotion-english-roberta-large` [20] for English and `daveni/twitter-xlm-roberta-emotion-es` [21] for Spanish. These encoders produce affective representations over the `capsanocr` text that help capture sarcasm, aggressiveness, and emotionally charged stereotypes. Using language-specific encoders rather than a single multilingual backbone is motivated by the recurrent cross-lingual gap reported in the EXIST literature [6, 7].

4.3. Physiological Encoding

The physiological stream encodes the three released modalities for the lab subjects per instance (2–4 per meme, 2 per video): eye-tracking (ET: pupil diameter, blink, fixation and saccade descriptors, and reaction time), heart rate (HR), and EEG band power. Per modality, the per-subject values of each variable are pooled across the available subjects into a fixed statistical vector—its mean, standard deviation, maximum, and minimum, i.e. $[\text{mean}, \text{sd}, \text{max}, \text{min}]$—so that every variable yields the same four summaries regardless of how many subjects viewed the instance. Any variable that is not populated for a given instance—most commonly the eye-tracking blink descriptors, which can be empty for memes when a subject does not blink during the short viewing interval—is zero-filled, so that instances with full and partial physiological coverage can be mixed in the same batch. The resulting per-variable summaries are concatenated into a single physiological descriptor, which is z-scored using normalization statistics estimated on the training partition of each fold only—never the validation partition—and passed through a small MLP. During training, modality dropout randomly suppresses the physiological stream, encouraging the model to remain accurate when the signals are unavailable at test time.

4.4. Fusion

In the core configuration, the language-specific encoder produces a sequence of L contextual text tokens $\mathbf{H} \in \mathbb{R}^{L \times d}$, retained rather than collapsed to a single vector, and a residual cross-attention block injects the physiological evidence into this textual representation. The aggregated physiological descriptor is projected by the MLP of Section 4.3 to a single query token $\mathbf{q} \in \mathbb{R}^{d}$, while the text tokens $\mathbf{H}$ supply the keys and values. With h heads and per-head width $d_h = d/h$, the block computes scaled dot-product attention,

$$\mathbf{a} = \text{softmax}\left(\frac{(\mathbf{q}W_Q)(\mathbf{H}W_K)^\top}{\sqrt{d_h}} + M \right) \mathbf{H}W_V, \tag{1}$$

where W_Q, W_K, W_V are the learned projections and M is an additive mask that assigns $-\infty$ to padding positions so they receive zero weight. The attention weights in Eq. (1) are a probability distribution over the L text tokens; consequently the physiological signal contributes only through *where* it attends, and the resulting context vector $\mathbf{a}$ is by construction a convex combination of the textual values, never an independent additive term. The block closes with a residual connection and layer normalization around the textual anchor,

$$\mathbf{z} = \text{LayerNorm}\big(\mathbf{h}_{\text{cls}} + \mathbf{a}\big), \tag{2}$$

where $\mathbf{h}_{\text{CLS}}$ is the pooled text token. The fused representation $\mathbf{z} \in \mathbb{R}^d$ is shared by the three prediction heads. When the optional dense visual stream is enabled (SigLIP2 for memes, VideoPrism for videos), a bilinear text–image interaction (in the spirit of Hate-CLIPper) first couples the visual features with the text representation, and the resulting joint tokens replace $\mathbf{H}$ as the keys and values of Eq. (1); the rest of the block is unchanged.

The assignment of roles in Eq. (1)—text as keys and values, the physiological descriptor as the query—is a deliberate design choice rather than an arbitrary one. The text stream is the densely populated, primary carrier of meaning: `capsanocr` already verbalizes the visual content and the gender-relevance reasoning, and it is present for every instance. The physiological descriptor, by contrast, is sparse, aggregated from only two to four subjects, and absent for most of the corpus. Placing the rich, always-present textual evidence as the keys and values and the sparse physiological signal as the query lets the lab signal *modulate* which textual tokens are emphasized, instead of injecting independent content of its own; the prediction therefore remains anchored in the text and the model cannot be driven by a noisy physiological vector. Were the roles reversed—text querying a physiological sequence—the output would be a combination of physiological values and would collapse whenever those signals are missing, which is precisely the regime that dominates this corpus.

This residual formulation makes the physiological contribution controllable: since the attention output of Eq. (1) is a convex combination of the projected text values, the physiological query can only reweight evidence already present in the text stream, and the residual of Eq. (2) preserves the textual anchor $\mathbf{h}_{\text{CLS}}$. When the signal is uninformative—as is often the case under the sparse coverage of these corpora—the cross-attention term reduces to a near-constant offset absorbed by the residual and the subsequent layer normalization, so the network degrades gracefully toward the text-only model; an informative query instead sharpens attention toward the relevant tokens. Combined with the modality dropout of Section 4.3, this makes the contribution of the lab signals a clean, isolable ablation (cross-attention versus plain concatenation).

4.5. Prediction Heads

Following the Learning with Disagreement protocol, supervision targets are the empirical annotator distributions rather than hard labels, and we use three per-subtask heads, shared in form across both tasks (memes and videos):

- **Identification (T2.1 / T3.1)** — a 2-dim distribution $[P(\text{No}), P(\text{Yes})]$, softmax output, KL-divergence loss;
- **Intention (T2.2 / T3.2)** — a 3-dim distribution $[P(\text{No}), P(\text{Direct}), P(\text{Judgemental})]$, softmax output, KL-divergence loss;
- **Categorization (T2.3 / T3.3)** — a 6-dim multi-label vector $[P(\text{No}), P(c_1), \ldots, P(c_5)]$ with independent sigmoids and a soft binary cross-entropy loss.

UNKNOWN annotations are ignored. Each head includes an explicit "No" class, so non-sexism is predicted directly rather than inferred as a residual; the taxonomic hierarchy linking the subtasks is restored only at evaluation time through PyEvALL, which avoids the soft-evaluation collapse observed when a multi-class objective is forced onto the multi-label task (Section 2). Training uses the full corpus of each task, including instances with no clear majority, since a target such as $[0.5, 0.5]$ is itself a valid soft target for KL and BCE. This is consistent with the official evaluation: under the primary soft–soft setting, the empirical annotator distribution—including such ambiguous items—*is* the ground truth, so learning these distributions directly serves ICM-Soft rather than merely adding noise. In the secondary hard–hard setting, by contrast, PyEvALL removes items with no majority class from the gold standard, so these instances cannot distort the hard metric either.

4.6. Configurations and Ablations

The design induces a compact ablation ladder that isolates each contribution, and the same ladder is run for both tasks. We build up from the textual core and then add the physiological modalities

incrementally: on-screen text only (`ocr`); the full enriched view `capsanocr`, which adds the neutral Qwen3-VL-8B visual description and the gender-relevance analysis; `capsanocr+EEG`; and the complete `capsanocr+EEG+HR+ET`. The optional dense visual stream is evaluated as an additional configuration on top of the text core—SigLIP2 for memes and a VideoPrism video encoder for videos. Because physiological coverage exists only for a fraction of instances, we expect the physiological gain to be modest and concentrated on the most ambiguous fine-grained categories; a carefully measured small or null effect, with leakage controlled per fold and an otherwise identical baseline, is itself a valid empirical finding that the architecture is designed to isolate.

5. System B: Few-Shot Classification with Qwen3.5-27B and Gemma-4-31B

System B is a training-free family that casts the three subtasks of Task 2 and Task 3 as a single *hierarchical-conditional* few-shot prompt: an instruction-following multimodal model receives a small set of labelled exemplars and a query instance, and predicts the subtask labels in a cascade. Unlike System A, it produces *hard* labels and is therefore evaluated in the hard–hard setting; it serves as a strong large-VLM reference against which the trained, soft-optimized fusion model can be contrasted under the same official protocol. The backend is interchangeable through a single switch: Qwen3.5-27B, a natively unified vision–language model, or Gemma-4-31B; decoding is greedy (`do_sample=False`), so the pipeline is deterministic. In essence, each prediction combines three ingredients: a definition-rich system prompt, a small leakage-free pool of labelled exemplars selected to match the subtask, and a cascade that runs the binary subtask first and conditions the fine-grained ones on its decision. The same recipe applies to memes (with the image) and to videos (with sampled frames), so a single pipeline covers both tasks. Figure 2 gives an overview.

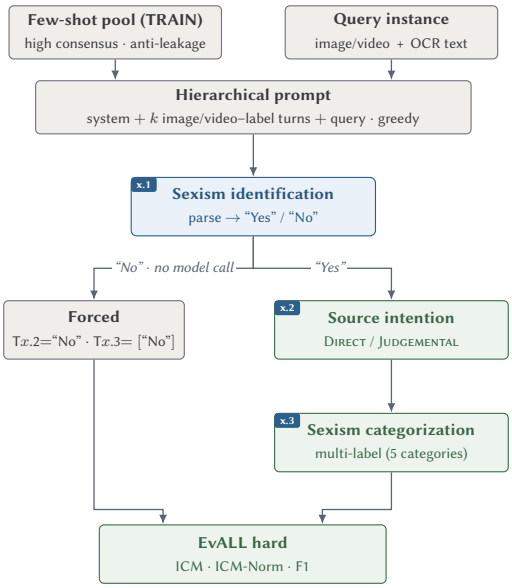

Figure 2: System B hierarchical-conditional few-shot pipeline (shown for Task 2; Task 3 is identical, with sampled video frames in place of the image). The binary subtask Tx.1 runs on every instance; only "Yes" instances reach the model for Tx.2 and Tx.3, while "No" instances are forced to non-sexist labels without a model call, ensuring hierarchical consistency.

5.1. Models

We instantiate System B with two open-weight, instruction-following multimodal LLMs: Qwen3.5-27B [12], a natively unified vision–language model, and Gemma-4-31B [13]. Both ingest interleaved

image–text turns—meme images for Task 2 and sampled video frames for Task 3—and are multilingual, covering the English and Spanish of both corpora, which is what lets a single, training-free pipeline run unchanged over the two modalities and under the two architectures. Their structured textual outputs are parsed into the official label set, and the backend is toggled by a single switch with greedy decoding (`do_sample=False`) keeping each run deterministic, so the two models yield a clean backend comparison (Section 7).

5.2. Few-Shot Pool Construction

The exemplar pool is always built from the *training* partition—the only instances with labels—which keeps it stable across runs and, crucially, prevents leakage. The pool mirrors the cascade rather than the flat label space, so the exemplars a subtask sees never contradict the constraints its prompt imposes. Within each subtask, candidates are ranked by annotator consensus—the fraction of annotators agreeing with the majority label—so that the most prototypical cases are shown first:

- **Tx.1** sees the full binary problem, with an equal number of "Yes" and "No" exemplars per class (two each for memes, three each for videos).
- **Tx.2** receives only instances the upstream stage already marked sexist, so its pool *excludes* "No": three DIRECT and three JUDGEMENTAL exemplars, concentrating the demonstrations on the harder intention boundary.
- **Tx.3** likewise excludes "No" and prefers *single-label* exemplars—a meme or video tagged with a single category teaches that category more cleanly than a multi-tagged one— selecting two exemplars per sexist category through a score that combines label purity and consensus.

This conditional construction removes the contradiction that arises when a system prompt forbids "No" in Tx.2/Tx.3 while the exemplars still contain it, and it is cached on disk and validated on load so that a pool generated under an earlier (non-conditional) design is detected and regenerated rather than silently reused.

5.3. Hierarchical Prompting

Each subtask is driven by an extensive, structured system prompt that encodes the operational EXIST definition together with anti-error rules distilled from the literature: the binary prompt explicitly flags benevolent sexism as a frequent false-negative risk and counter-speech as a frequent source of ambiguity, and supplies an arbiter test (*does the joke require a specific gender to land?*); the intention prompt frames the decision as *whose side is the creator on?* and treats irony, implicit sexism, and multimodal contradiction as ambiguity cues, instructing the model to distinguish content that endorses sexist stereotypes from content that denounces, reports, or criticizes sexist experiences [8, 15]; the categorization prompt reproduces the five guideline definitions plus disambiguation rules for the most-confused category pairs. Crucially, the decision rules are not ad hoc: the category definitions and the operational criteria (including the treatment of benevolent sexism and counter-speech) are taken from the official EXIST 2026 Lab Guidelines [4] and the task definitions of the EXIST overview [2], rather than from our own interpretation, so that the prompts and the gold labels of the exemplar pool share the same labelling criteria; ranking exemplars by annotator consensus (Section 5.2) further reduces the risk that a demonstration contradicts the rules its prompt states. On top of the system prompt, each pool exemplar is materialized as a genuine multimodal turn pair—a `user` turn carrying the image and OCR text and an `assistant` turn carrying the gold label—so the model performs true few-shot learning with images, not text-only demonstrations. The verbatim system prompts for the three meme subtasks (T2.1, T2.2, T2.3) will be released in our public repository (Appendix A); the Task 3 video subtasks use analogous prompts on the video setting of Section 5.5.

5.4. Cascade Inference

Inference follows the taxonomic hierarchy. The binary subtask Tx.1 is run over every instance; only instances predicted "Yes" are passed through the model for Tx.2 and Tx.3, while instances predicted "No" are assigned Tx.2 = "No" and Tx.3 = ["No"] without a model call. This halves the number of calls on the expensive fine-grained subtasks and guarantees hierarchical consistency by construction: a category can never be assigned to an instance deemed non-sexist. The hierarchy is reimposed as a hard override, so the binary decision always governs the downstream labels.

5.5. Video Processing

Task 3 reuses the same hierarchical-conditional pipeline but treats the video as a first-class modality. Each exemplar and the query are passed as a genuine multimodal turn carrying the `.mp4` itself: the Qwen backend samples frames at $\mathrm{fps} = 0.5$ with a per-frame budget of $\approx$307K pixels ($\approx$512 visual tokens, $\approx$40 frames for an 80 s clip), while Gemma uses its processor's default video sampling. Following the Gemma guidance, the video block is placed *before* the text (modality-first), an ordering Qwen tolerates without penalty. The textual channel is the transcription and on-screen text released in the `text` field, appended after the video; no separate transcription is computed.

The exemplar pool adds three video-specific refinements over the meme pool of Section 5.2: it is *bilingually stratified* (each class half Spanish, half English) to avoid a language bias; T3.1 and T3.2 use *consensus diversity*, drawing roughly 70% of exemplars from high-consensus and 30% from medium-consensus instances (agreement in $[0.5, 0.7]$) to expose borderline cases; and T3.3 appends a few *co-occurrence* exemplars with two or more active categories, after the two single-label exemplars per category, so the model learns that categories combine. Optionally, each exemplar can be followed by a short rule-grounded rationale, which we treat as an ablatable choice.

The cascade, decoding, and parsing are identical to the meme case (Section 5.4), with reasoning prefixes stripped before parsing and out-of-memory recovery added for T3.3 (pool reduction, memory cleanup, and checkpoint resumption), since video inference is far heavier than image inference. The verbatim video prompts—which add medium-specific rules absent from the meme prompts (POV/scenario framing, performance of sexist audio, and multichannel cues such as facial expression, editing, and reply context)—will be released in the same repository (Appendix A).

6. Experimental Setup

6.1. Hyperparameter Optimization

For model selection and ablation we use 5-fold stratified cross-validation, stratified with respect to the target label distribution and language when possible. We tune PhysioMeme-Fusion with Optuna [22] (TPE sampler, MedianPruner pruner), preferring model-based search over a dense grid at equal budget [23]. Each trial is scored by the mean ICM-Soft-Norm over the five validation folds—the primary LeWiDi metric itself rather than a surrogate loss—and the search never observes the test set: the per-fold z-scoring statistics are re-estimated inside each fold (Section 4.3) and the selected configuration queries the official test set exactly once. The space covers the peak learning rate, the AdamW weight decay, the OneCycleLR warmup fraction `pct_start`, the number of cross-attention heads, and the modality-dropout rate; batch size, epochs, sequence length, and gradient clipping are fixed by convention.

7. Results

We report the official EXIST 2026 leaderboard results for the three subtasks of Tasks 2 and 3, evaluated by PyEvALL over *all* instances (both languages). The soft–soft setting is the primary one under LeWiDi;

we also report the hard–hard setting. For each of our runs we give the official metrics together with the leaderboard rank out of the total number of participating systems.

7.1. Task 2 Results

Table 1 reports the soft–soft evaluation of our trained System A under its three configurations: the `capsanocr` text core (Run 1), the same model with the physiological branch added through cross-attention (Run 2), and the variant that additionally enables the SigLIP2 visual stream (Run 3). Table 2 reports the hard–hard evaluation, where Run 1 is the hard adaptation of the physiological soft model and Runs 2 and 3 are the few-shot System B with the Gemma-4-31B and Qwen3.5-27B backends, respectively.

Table 1
Task 2 **soft–soft** official results (all instances), trained System A. ICM-Soft and ICM-Soft-Norm are the PyEvALL LeWiDi metrics; CE is the cross-entropy between the predicted and the empirical annotator distributions, computed by us as an auxiliary calibration diagnostic (lower is better). Rank is the official leaderboard position; *Team rank* is our position in the by-team ranking.

Subtask	Configuration	ICM-Soft	ICM-Soft$_N$	CE	Rank	Team rank
T2.1	`capsanocr` (text core)	−0.036	0.494	0.907	13th	7th/69
	+ physiological	−0.034	0.495	0.913	12th	
	+ SigLIP2	−0.710	0.386	1.070	94th	
T2.2	`capsanocr` (text core)	−1.359	0.355	1.476	23rd	9th/57
	+ physiological	−1.316	0.360	1.491	19th	
	+ SigLIP2	−1.885	0.300	1.624	53rd	
T2.3	`capsanocr` (text core)	−4.949	0.238	–	15th	8th/56
	+ physiological	−5.240	0.222	–	18th	
	+ SigLIP2	−5.789	0.193	–	27th	

Table 2
Task 2 **hard–hard** official results (all instances). Run 1 is the hard adaptation of the physiological System A; Runs 2–3 are the few-shot System B with the Gemma-4-31B and Qwen3.5-27B backends. ICM-Hard is the primary hard metric; F1 is the positive-class / macro F1. Rank is the official leaderboard position; *Team rank* is our position in the by-team ranking.

Subtask	System	ICM-Hard	ICM-Hard$_N$	F1	Rank	Team rank
T2.1	System A (hard)	0.178	0.590	0.742	25th	4th/101
	Few-shot Gemma-4-31B	0.260	0.632	0.776	16th	
	Few-shot Qwen3.5-27B	0.327	0.667	0.779	**8th**	
T2.2	System A (hard)	−0.292	0.399	0.455	37th	5th/87
	Few-shot Gemma-4-31B	0.073	0.525	0.552	9th	
	Few-shot Qwen3.5-27B	0.088	0.531	0.550	**8th**	
T2.3	System A (hard)	−0.735	0.348	0.446	21st	2nd/87
	Few-shot Gemma-4-31B	−0.191	0.460	0.541	6th	
	Few-shot Qwen3.5-27B	−0.107	0.478	0.552	**3rd**	

Three patterns are consistent across the three subtasks. First, the dominant source of improvement on the textual side is the enrichment of the input rather than the physiological branch. Moving from the literal overlay text (`ocr`) to the full `capsanocr` view—which adds the neutral Qwen3-VL-8B visual description (`cap`) and the independent social/contextual analysis (`san`)—is where the large gain is concentrated: on our cross-validation split it averages an improvement of about 17% across subtasks. Verbalizing the visual content and the gender-relevance reasoning gives the encoder the context that the raw overlay text lacks. By contrast, adding the physiological branch on top of `capsanocr` produces only a marginal change, and the increments shrink by an order of magnitude at each step: still on the validation split, adding EEG improves over `capsanocr` by about 1.8% on average, and adding heart

rate and eye-tracking on top of that (capsanocr+EEG → capsanocr+EEG+HR+ET) adds a further 0.65% only. Consistent with this diminishing return, Runs 1 and 2 are indistinguishable on T2.1 and T2.2 (differences in ICM-Soft-Norm of one thousandth and five thousandths, respectively) and the text-only core is in fact slightly better on T2.3. This is the small-or-null physiological effect anticipated in Section 4.4, consistent with the sparse physiological coverage and the near-zero correlation between physiological variability and annotator disagreement; we report it as a genuine negative finding rather than omit it. Second, the optional SigLIP2 visual stream *degrades* performance markedly in every soft subtask—dropping, for example, from rank 13 to rank 94 on T2.1—which is why we treat it as an exploratory add-on rather than a core component; we attribute the drop to the capsanocr captions already verbalizing the visual content, so the raw visual stream adds parameters and noise without complementary signal. Third, in the hard–hard setting the training-free few-shot System B clearly outperforms the trained network on all three subtasks, with Qwen3.5-27B the strongest backend throughout: it ranks 8th, 8th, and 3rd on T2.1, T2.2, and T2.3, and the advantage widens on the harder fine-grained categorization, where the soft-trained network struggles. The gap between the few-shot runs and the hard adaptation of System A is largest on T2.2 and T2.3, suggesting that the large instruction-following models exploit the rich category definitions in the prompt more effectively than the small trained heads recover them from the soft targets.

7.2. Task 3 Results

Tables 3 and 4 report the official soft–soft and hard–hard results for our two video runs. In the hard–hard setting (Table 4) both runs are the training-free few-shot System B, one with the Gemma-4-31B backend and one with the Qwen3.5-27B backend. In the soft–soft setting (Table 3), both runs are the trained System A fusion applied to video: the first is the capsanocr text core with the full physiological branch (the video counterpart of the +physiological meme run), and the second adds a dense visual stream that, for video, replaces SigLIP2 with a VideoPrism video encoder (google/videoprism-lvt-large-f8r288).

Table 3
Task 3 **soft–soft** official results (all instances), the trained System A applied to video. The two runs are the capsanocr text core with the full physiological branch and the same model with an added visual stream that replaces the SigLIP2 used for memes with a VideoPrism video encoder. ICM-Soft and ICM-Soft-Norm are the PyEvALL LeWiDi metrics; CE is the cross-entropy between the predicted and the empirical annotator distributions (lower is better; not released for T3.3). Rank is the official leaderboard position; *Team rank* is our position in the by-team ranking.

Subtask	Configuration	ICM-Soft	ICM-Soft$_N$	CE	Rank	Team rank
T3.1	capsanocr + physio	−0.971	0.330	1.117	51st	23rd/27
	+ VideoPrism	−0.982	0.328	1.060	52nd	
T3.2	capsanocr + physio	−2.879	0.193	3.643	40th	17th/23
	+ VideoPrism	−2.917	0.189	3.503	42nd	
T3.3	capsanocr + physio	−6.724	0.099	–	22nd	11th/23
	+ VideoPrism	−6.924	0.087	–	26th	

Two observations stand out. First, on video the few-shot System B is highly competitive in the hard–hard setting: the Qwen3.5-27B backend tops the T3.1 leaderboard and the Gemma-4-31B backend is third; both rank in the top tier on intention (T3.2, 4th and 7th); and the advantage narrows on the fine-grained categorization (T3.3, 19th and 35th)—the same subtask that is hardest in Task 2. Second, the backend ordering inverts relative to memes: whereas Qwen3.5-27B dominated every Task 2 subtask, on video it leads only on identification (T3.1), while Gemma-4-31B is the stronger backend on both intention (T3.2, 0.607 vs. 0.601 ICM-Hard$_N$) and categorization (T3.3, 0.390 vs. 0.350). In the soft–soft setting both System A video runs sit in the lower half of the leaderboard—ranks in the 40s–50s on T3.1 and T3.2 and 22nd/26th on T3.3—with the high cross-entropy on T3.2 ($\approx$3.5) indicating predicted distributions poorly calibrated against the annotator distributions; System A's soft fusion, which is

Table 4
Task 3 **hard–hard** official results (all instances), the few-shot System B with the Gemma-4-31B and Qwen3.5-27B backends. ICM-Hard is the primary hard metric; F1 is the positive-class (T3.1) / macro (T3.2–T3.3) F1 reported by PyEvALL. Rank is the official leaderboard position; *Team rank* is our position in the by-team ranking.

Subtask	System	ICM-Hard	ICM-Hard$_N$	F1	Rank	Team rank
T3.1	Few-shot Gemma-4-31B	0.320	0.661	0.750	3rd	
	Few-shot Qwen3.5-27B	0.331	0.667	0.763	**1st**	1st/32
T3.2	Few-shot Gemma-4-31B	0.284	0.607	0.690	**4th**	3rd/30
	Few-shot Qwen3.5-27B	0.268	0.601	0.679	7th	
T3.3	Few-shot Gemma-4-31B	−0.339	0.390	0.416	**19th**	8th/30
	Few-shot Qwen3.5-27B	−0.463	0.350	0.410	35th	

mid-table on memes, thus transfers less well to video, where our submissions are competitive only in the hard setting.

7.3. Qwen3.5-27B vs. Gemma-4-31B

Both backends run the identical few-shot pipeline, so Table 2 isolates the effect of the model. Qwen3.5-27B outperforms Gemma-4-31B on all three subtasks under ICM-Hard$_N$, and the margin grows from the binary task (T2.1, 0.667 vs. 0.632) to the fine-grained categorization (T2.3, 0.478 vs. 0.460); the two are closest on intention (T2.2). Both backends nonetheless land in the top decile of the hard leaderboard on the fine-grained subtasks, confirming that a definition-rich hierarchical prompt is a strong, training-free baseline. A more detailed comparison of hierarchical consistency, invalid-output rate, and qualitative error patterns—particularly where visual evidence contradicts the surface text—is left to the error analysis.

8. Discussion

8.1. What Worked

The clearest positive result is the value of enriching the textual input. The `capsanocr` view—overlay text augmented with a neutral Qwen3-VL-8B visual description and an independent gender-relevance analysis—is what places the trained System A in the upper part of the soft leaderboard: on memes it reaches 13th on identification (T2.1), 23rd on intention (T2.2), and 15th on the fine-grained categorization (T2.3), all from a single text core without any raw-image stream. This confirms that verbalizing the visual content and reasoning explicitly about gender relevance gives the text encoder the context that the literal overlay text lacks, and that a description-based route to the image (Section 4) is sufficient to be competitive on the soft metric. Using language-specific emotion-tuned encoders for English and Spanish, rather than a single multilingual backbone, is consistent with this outcome: it lets each language be modeled in its stronger semantic space and avoids the cross-lingual interference repeatedly reported for shared encoders in the EXIST literature [6, 7].

On the hard–hard setting the picture is even sharper: the training-free few-shot pipeline (System B) is the stronger system on all three Task 2 subtasks and on Task 3 identification. With the Qwen3.5-27B backend it ranks 8th, 8th, and 3rd on T2.1–T2.3 and *first* on T3.1, while the Gemma-4-31B backend is close behind and leads the video intention and categorization subtasks. A definition-rich, hierarchical prompt is therefore a strong baseline that needs no task-specific training and is especially useful where labelled data is scarce or where instances demand broad world knowledge—irony, cultural references, and implicit stereotypes that small trained heads recover only weakly. That the same pipeline is competitive across two modalities and two backbones, with only a single switch changed, also indicates that most of its accuracy comes from the prompt design rather than from any one model.

8.2. What Did Not Work

Two design choices did not pay off, and we report both as genuine negative results. First, the physiological branch produced only a marginal change over the text-only core: as detailed in Section 7.1, it leaves T2.1 and T2.2 numerically almost unchanged and is slightly *detrimental* on the multi-label T2.3 (0.222 vs. 0.238 ICM-Soft-Norm), with the per-step gains across the ablation ladder shrinking by an order of magnitude. This matches the sparse physiological coverage of the corpus—only a fraction of instances carry lab signals, recorded from a handful of subjects each—and the near-zero correlation we observed between physiological variability and annotator disagreement. It is also consistent with the modality-specific pattern reported on the same memes by Arcos et al. [1], where heart rate carried no significant signal; aggregating sparse, partly uninformative channels into a single descriptor appears to add more variance than discriminative content under the soft metric.

Second, the optional SigLIP2 visual stream degraded performance markedly in every soft subtask—most strikingly collapsing T2.1 from 13^{th} to 94^{th}, with smaller but consistent drops on T2.2 and T2.3. Because `capsanocr` already verbalizes the visual content as text, the raw visual stream is largely redundant and adds parameters and noise without complementary signal; the video counterpart, in which SigLIP2 is replaced by a VideoPrism video encoder, shows the same lack of benefit. Together these two findings point in one direction: for this corpus, a single well-constructed textual view is a stronger and more robust representation than either dense visual embeddings or sparse physiological summaries bolted on top of it.

A subtler negative concerns calibration rather than ranking. In the soft–soft setting the video runs sit in the lower half of the leaderboard despite competitive hard performance, with a high cross-entropy on T3.2 (≈ 3.5) that signals predicted distributions poorly matched to the annotator distributions; the soft fusion that is mid-table on memes thus transfers imperfectly to video. The remaining recurring failure modes are over-reliance on text overlays, missed irony, and occasional unstable label formatting in the generative outputs.

8.3. Error Analysis

We analyze errors by language, modality, and subtask, paying particular attention to instances with high annotator disagreement, which are the most influential under the ICM-Soft metric. The fine-grained categorization (Tx.3) is the hardest subtask for every system: it carries the most negative ICM-Soft scores and the largest gap between the trained model and the few-shot pipeline, indicating that the small soft-trained heads recover the category boundaries less effectively than the large models exploit the explicit category definitions supplied in the prompt. Within T2.3, errors concentrate on the categories that are both rarer and more implicit—stereotyping/dominance and the non-sexual forms of misogyny—rather than on the more lexically marked sexual-violence class.

A second pattern is that the better backend changes with the modality. On memes, Qwen3.5-27B dominates Gemma-4-31B on all three subtasks, whereas on video the ordering inverts: Qwen3.5-27B leads only on identification (T3.1), while Gemma-4-31B is the stronger backend on both video intention (T3.2) and categorization (T3.3). This suggests the two models differ in how they integrate sampled video frames with the released transcript and on-screen text, and cautions against selecting a single backbone from meme results alone. Finally, cases where the visual evidence contradicts the surface text remain a notable source of error for both system families, echoing the OCR-versus-image tension documented on MuSeD [8]; this is precisely the regime in which a single fused textual view, rather than separate channels competing at inference time, appears to help.

8.4. Limitations

Three caveats bound our conclusions. First, on the physiological data: the signals cover only a fraction of the corpus, are recorded from very few subjects per instance, and are aggregated into a fixed per-variable summary that discards their temporal structure; our null result should therefore be read as specific to this sparse, low-coverage regime rather than as evidence that physiological cues are uninformative in

principle. The cross-attention fusion is also weakly interpretable, as it does not reveal *which* features mattered. Second, on the few-shot pipeline: it is sensitive to prompt wording and output-format constraints, and—producing only hard labels—it cannot be scored under the primary soft metric, so its strength shows only in the hard–hard setting. Third, on scope: the trained encoders are computationally heavy, the video pipeline samples frames rather than modeling motion, and all findings are confined to the bilingual EN/ES EXIST 2026 data and the official metrics, so they may not transfer to other languages, platforms, or definitions of sexism.

9. Conclusion and Future Work

This paper presents two complementary strategies for EXIST 2026, both applied to memes (Task 2) and short videos (Task 3): a trained multimodal fusion architecture, PhysioMeme-Fusion, optimized directly for the soft-label objective with an explicit "No" class per head, and a training-free few-shot pipeline driven by definition-rich hierarchical prompts over two instruction-following multimodal LLMs. Evaluated under the official PyEvALL protocol, the two systems occupy clearly different regimes. In the soft–soft setting, the trained fusion ranks in the upper part of the meme leaderboard (seventh among teams on identification), and its accuracy is driven almost entirely by the enriched textual view (`capsanocr`)—a description-based, neutrally prompted route to the image that proves sufficient without any raw-image stream. In the hard–hard setting, the few-shot pipeline is the stronger system on every Task 2 subtask and tops the video identification leaderboard; in the official by-team ranking it places first among teams on video identification (T3.1), second on meme categorization (T2.3), and third on video intention (T3.2), confirming that a carefully designed prompt is a competitive, data-free baseline across modalities and backbones.

Equally informative are our two negative results, reported in full rather than omitted: the physiological branch yields at best a marginal and sometimes detrimental change under the soft metric, and the dense SigLIP2 / VideoPrism visual stream degrades performance wherever a verbalized textual view is already present. Taken together, they indicate that for this corpus a single well-constructed textual representation is more robust than additional dense or sparse modalities layered on top of it—a finding we believe is useful to the community precisely because the auxiliary signals were expected to help.

Future work follows directly from these limitations. On the physiological side, we plan coverage-aware and more interpretable fusion—learning per-instance how much to trust the lab signals and identifying which features carry the discriminative content—together with acquisition denser than the current two-to-four subjects per instance. On the modeling side, we will work on calibrating the soft-label predictions (the weak point of the video runs), on stronger temporal modeling for video beyond frame sampling, and on combining the complementary strengths of the two families—for instance, distilling the few-shot models' definition-grounded reasoning into the soft-trained heads so that a single system is competitive under both the soft and the hard metric.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033 and ERDF "ERDF A way of making Europe" (project PID2024-155948OB-C55), and by the grant PDC2025-164848-I00 funded by MICIU/AEI/10.13039/501100011033.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Claude, and Gemini in order to: Drafting content, Improve writing style, Generate literature review, and Content enhancement (including support with code development and debugging, and refining design ideas). After using these tools and services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] I. Arcos, P. Rosso, E. Gomis-Vicent, Human-centered multimodal fusion for sexism detection in memes with eye-tracking, heart rate, and EEG signals, arXiv preprint arXiv:2602.23862, 2026. arXiv:2602.23862.

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

[3] L. Aroyo, C. Welty, Truth is a lie: Crowd truth and the seven myths of human annotation, AI Magazine 36 (2015) 15–24.

[4] L. Plaza, J. Carrillo-de Albornoz, I. Arcos, M. Aloy-Mayo, P. Rosso, D. Spina, EXIST 2026: Physiological data for multimodal sexism characterization in social media, in: Advances in Information Retrieval — 48th European Conference on Information Retrieval (ECIR 2026), Lecture Notes in Computer Science, Springer, 2026, pp. 297–305. doi:10.1007/978-3-032-21321-1_40.

[5] E. Amigó, A. D. Delgado, Evaluating extreme hierarchical multi-label classification, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 5809–5819.

[6] P. Italiani, D. Gimeno-Gómez, L. Ragazzi, G. Moro, P. Rosso, MemeWeaver: Inter-meme graph reasoning for sexism and misogyny detection, in: Findings of the Association for Computational Linguistics: EACL 2026, Association for Computational Linguistics, 2026. Preprint: arXiv:2601.08684.

[7] I. Arcos, ArcosGPT at EXIST 2025: Identifying sexism in memes with multimodal deep learning: Fusing text and visual cues, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025.

[8] L. De Grazia, P. Pastells, M. Vázquez Chas, D. Elliott, D. Sánchez Villegas, M. Farrús, M. Taulé, MuSeD: A multimodal spanish dataset for sexism detection in social media videos, in: Proceedings of the Conference on Language Modeling (COLM 2025), 2025.

[9] Qwen Team, Qwen3-VL technical report, 2025. arXiv:2511.21631.

[10] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, X. Zhai, SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. arXiv:2502.14786.

[11] L. Zhao, N. B. Gundavarapu, L. Yuan, et al., VideoPrism: A foundational visual encoder for video understanding, in: Proceedings of the 41st International Conference on Machine Learning (ICML), 2024.

[12] Qwen Team, Qwen3.5: Towards native multimodal agents, https://qwen.ai/blog?id=qwen3.5, 2026.

[13] Gemma Team, Google DeepMind, Gemma 4 Technical Report, Technical Report, Google DeepMind, 2026.

[14] 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), Association for Computational Linguistics, 2022, pp. 533–549.

[15] L. De Grazia, D. Sánchez Villegas, D. Elliott, M. Farrús, M. Taulé, Beyond binary classification: Detecting fine-grained sexism in social media videos, 2026. doi:10.48550/arXiv.2602.15757. arXiv:2602.15757, fineMuSe.

[16] 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.

[17] Ö. Alaçam, S. Hoeken, S. Zarrieß, Eyes don't lie: Subjective hate annotation and detection with gaze, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), Association for Computational Linguistics, 2024, pp. 187–205.

[18] Y. Zhang, Q. Li, S. Nahata, T. Jamal, S.-K. Cheng, G. Cauwenberghs, T.-P. Jung, Integrating

large language model, EEG, and eye-tracking for word-level neural state classification in reading comprehension, IEEE Transactions on Neural Systems and Rehabilitation Engineering 32 (2024) 3465–3475.
[19] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 8440–8451.
[20] J. Hartmann, Emotion English RoBERTa-large, https://huggingface.co/j-hartmann/emotion-english-roberta-large, 2022.
[21] D. Vera, O. Araque, C. A. Iglesias, GSI-UPM at IberLEF 2021: Emotion analysis of spanish tweets by fine-tuning the XLM-RoBERTa language model, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), 2021.
[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD), ACM, 2019, pp. 2623–2631. doi:10.1145/3292500.3330701.
[23] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, Journal of Machine Learning Research 13 (2012) 281–305.

A. System Prompts and Reproducibility

For reproducibility, the system prompts used by the few-shot family of Section 5—for the meme subtasks T2.1–T2.3 and their video counterparts T3.1–T3.3—will be released, together with the code and the few-shot exemplar pools, in our public repository: https://github.com/juangarcia83/EXIST-2026. The same prompts are shared by both backends (Qwen3.5-27B and Gemma-4-31B). In addition to the prompts, the notebooks implementing both system families, along with their full configurations and hyperparameter settings, will also be added to the repository, so that all reported experiments can be reproduced end to end.

The model checkpoints used in this work are publicly available on the Hugging Face Hub. For System A: the captioning VLM Qwen3-VL-8B (https://huggingface.co/Qwen/Qwen3-VL-8B-Thinking); the language-specific emotion encoders `j-hartmann/emotion-english-roberta-large` (https://huggingface.co/j-hartmann/emotion-english-roberta-large) and `daveni/twitter-xlm-roberta-emotion-es` (https://huggingface.co/daveni/twitter-xlm-roberta-emotion-es); and the optional dense visual streams SigLIP2 (https://huggingface.co/google/siglip2-so400m-patch14-384) and VideoPrism (https://huggingface.co/google/videoprism-lvt-large-f8r288). For System B: the few-shot backends Qwen3.5-27B (https://huggingface.co/Qwen/Qwen3.5-27B) and Gemma-4-31B (https://huggingface.co/google/gemma-4-31B).