MemeLens: Multilingual Multitask VLMs for Memes

Anonymous ACL submission

Abstract

Memes are a dominant medium for online communication and manipulation because meaning emerges from interactions between embedded text, imagery, and cultural context. Existing meme research is distributed across tasks (hate, misogyny, propaganda, sentiment, humour) and languages, which limits cross-domain generalization. To address this gap we propose MEMELENS, a unified multilingual and multitask explanation-enhanced Vision Language Model (VLM) for meme understanding. We consolidate 38 public meme datasets, filter and map dataset-specific labels into a shared taxonomy of 20 tasks spanning harm, targets, figurative/pragmatic intent, and affect. We present a comprehensive empirical analysis across modeling paradigms, task categories, and datasets. Our findings suggest that robust meme understanding requires multimodal training, exhibits substantial variation across semantic categories, and remains sensitive to over-specialization when models are fine-tuned on individual datasets rather than trained in a unified setting. We make the experimental resources publicly available to the community.[1]

1 Introduction

Memes are among the most widely shared forms of online content (Brody and Cullen, 2023; Barnes et al., 2024). By pairing an image with a small amount of overlaid text, memes communicate stance, sarcasm, in-group identity, or hostility. Their interpretation is compositional: meaning emerges from the *interaction* between the textual overlay and the visual scene, often mediated by background knowledge and cultural context, challenging purely textual or visual approaches.

This has motivated a growing body of work on multimodal meme understanding, spanning hate and abuse (Kiela et al., 2020), misogyny (Fersini et al., 2022), affect and humor (Sharma et al., 2020), *figurative language* (Xu et al., 2022), and persuasion techniques (Dimitrov et al., 2024; Alam et al., 2024b) across diverse datasets, tasks, and languages in practice.

Alongside these advances, the current state of the art remains largely task- and language-specific. Most efforts are organized around *a single dataset* and *a single phenomenon* (often via one shared task), which introduces at least three practical barriers: *(i)* limited transfer across label spaces, languages, and domains; *(ii)* evaluation protocols that are difficult to compare or reproduce across benchmarks; and *(iii)* missed opportunities for parameter sharing, where capabilities learned for one aspect of meme understanding (e.g., affect) could systematically benefit others (e.g., harassment detection).

In deployed settings (e.g., content moderation, real-time trend monitoring, and media-literacy), models must *simultaneously* reason about affect (sentiment and humor), pragmatic intent (e.g., sarcasm), and a broad spectrum of harm-related categories, while remaining robust to cultural and linguistic variation across platforms (Hee et al., 2024).

Existing resources for meme understanding remain uneven across languages. As shown in Figure 1, most datasets are English-centric, while non-English benchmarks are much more limited. Cross-lingual generalization is further hindered by mismatches in label definitions, annotation granularity, and guidelines across datasets (Bui et al., 2025). As a result, naive dataset mixing can be unreliable and may cause negative transfer in multilingual multitask training. At the same time, prior work argues that deeper meme understanding requires going beyond lexical and visual cues to infer latent intent and implied meaning (Hee et al., 2023).

[1] anonymous.com

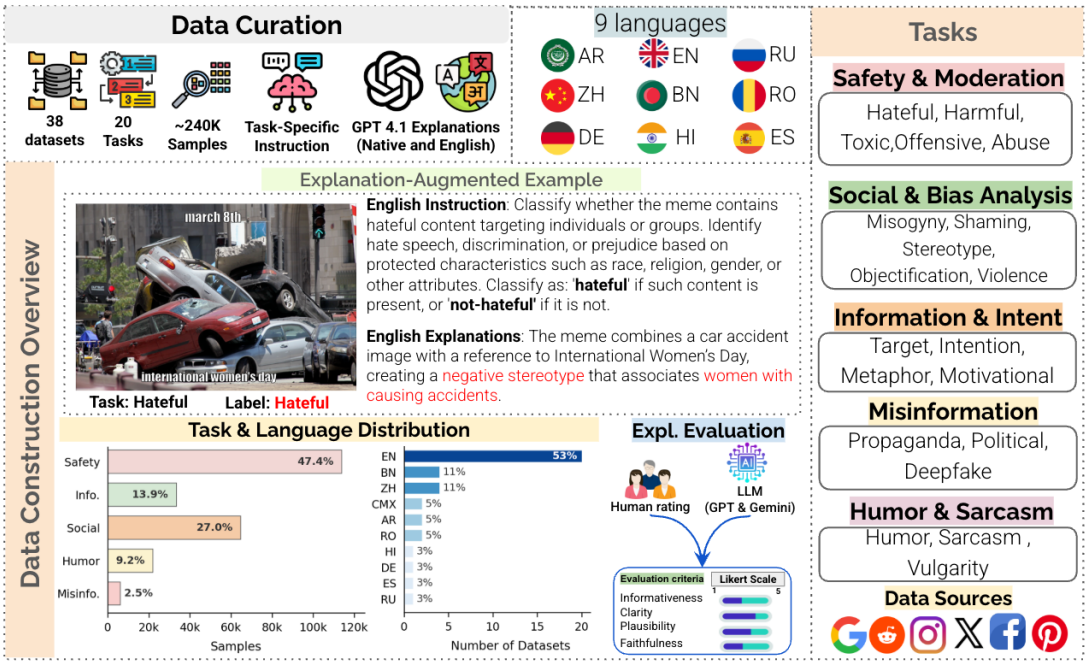

Figure 1: **Overview of tasks and datasets in MEMELENS.** Dataset-specific labels are mapped into a unified taxonomy for multi-task training and cross-dataset evaluation. Task categories include Safety, Info., Social, Humor, and Misinfo. Language abbreviations denote EN, BN, ZH, CMX (Hindi-English), AR, RO, HI, DE, ES, and RU.

Motivated by these gaps, we frame meme understanding as a problem of *multilingual, multimodal, multitask learning*, where models learn shared representations while harmonizing heterogeneous taxonomies across datasets to enable robust generalization across domains and settings, improving transfer, consistency, and performance. Concretely, we compile an extensive collection of meme datasets (38 sources in our current collection), spanning multiple languages and tasks.

This unification also raises an empirical question: *which* modeling paradigm best supports multilingual multitask meme understanding under realistic heterogeneity? To address it, we benchmark *(i)* unimodal baselines (text-only, image-only), *(ii)* multimodal fusion and sequence-based architectures, and *(iii)* causal/instruction-style VLMs, analyzing not only aggregate accuracy but also cross-task and cross-lingual transfer. *Finally*, we situate our study relative to emerging efforts that emphasize robustness and generalization in meme understanding (Liu et al., 2025; Chen et al., 2025).

We summarize our main contributions below:

- We curate and consolidate 38 publicly available meme datasets, applying rigorous filtering and a unified annotation mapping to create a coherent, multilingual, multi-task resource.
- We introduce a unified taxonomy spanning 9 languages and 20 tasks, together with a consistent training and evaluation protocol for cross-dataset benchmarking.
- We formulate meme understanding as **multilingual, multimodal and multitask learning** and present a unified training setup over a large dataset mixture curated under a consistent *text-over-image* meme definition.
- We provide explanation-enhanced supervision and a systematic empirical study across modeling paradigms, including unimodal baselines, multimodal sequence models, and large VLMs, evaluated under a unified training and benchmarking framework.

Our empirical analysis yields several key observations. We find that *robust meme understanding consistently benefits from multimodal training*, but exhibits substantial variability across semantic task categories and datasets. Task families involving implicit or rhetorical meaning, such as humor and sarcasm, remain challenging across all paradigms. Additionally, we observe that models fine-tuned on individual datasets tend to over-specialize, *motivating unified multitask training when broad coverage and cross-dataset robustness are required.*

2 Related Work

Meme benchmarks. A major driver of progress in computational meme understanding has been curated datasets and shared tasks that frame meme semantics as supervised prediction (Sharma et al.,

2022). The Kiela et al. (2020) Hateful Memes Challenge showed the limitations of unimodal models, motivating approaches that integrate visual evidence with overlaid text. Beyond hate, SemEval expanded meme understanding to affective and pragmatic phenomena. Memotion (Sharma et al., 2020) targets sentiment and emotion (e.g., humor, sarcasm, offensiveness, motivation), while MAMI (Fersini et al., 2022) focuses on misogyny and fine-grained traits. SemEval also formalized persuasion and propaganda in Task 6 (2021) (Dimitrov et al., 2021) and later extended it to multilingual settings in Task 4 (2024) (Dimitrov et al., 2024).

Multilingual and multicultural meme understanding. While early meme benchmarks centered on English, subsequent work has emphasized the multilingual nature of meme culture and the limits of monolingual generalization (Alam et al., 2022). For instance, MUTE (Hossain et al., 2022) targets Bengali and code-mixed hateful memes, and Multi^{3}Hate (Bui et al., 2025) offers a parallel multilingual meme dataset to study cross-cultural annotation differences and cross-lingual VLM behavior.

Multitask meme modeling. A smaller body of work studies *multi-task meme understanding*. Early work by Chauhan et al. (2020) proposed a multimodal multi-task architecture spanning humor, sarcasm, offensiveness, motivational content, and sentiment. More recent efforts introduce datasets and models that jointly model multiple facets within a unified setting (e.g., hate, targets, stance, and humor) (Shah et al., 2024), and benchmarks that evaluate VLMs across diverse meme tasks (e.g., humor and sarcasm) (Gavit et al., 2025). Despite these advances, most studies still train and report results *separately per dataset and per phenomenon*.

Explanations, roles, and reasoning. Meme semantics is often implicit and context dependent, motivating approaches beyond surface classification, including connotation modeling and explanation generation (Martinez Pandiani et al., 2025). Prior work also shows that harmfulness can be satirical and highly context sensitive (Pramanick et al., 2021). Several datasets use explanations as supervision, including hateful reasons for conditional rationale generation (Hee et al., 2023) and explanations for entity and role understanding (Sharma et al., 2023). More recent resources elicit richer meaning representations using multimodal QA (Agarwal et al., 2024), interpretation-augmented captioning (Park et al., 2025), and bilingual detection-plus-explanation setups (Kmainasi et al., 2025; Gu et al., 2025). In parallel, model-centric work distills or structures reasoning traces to improve prediction and interpretability (Lin et al., 2023, 2024; Hee and Lee, 2025; Bayan Kmainasi et al., 2026). At the same time, explanations are not always helpful, as rationales can drift from meme evidence and *explain-then-detect* pipelines may underperform without reasoning-aware objectives (Lu et al., 2025; Mei et al., 2025b).

VLMs and meme-specific adaptation. With the rise of strong VLMs, meme research has increasingly focused on adapting and prompting these models. Prompt-based approaches leverage the implicit knowledge of language models for hateful meme detection (Cao et al., 2022). Retrieval-based methods target out-of-domain generalization in evolving meme ecosystems (Mei et al., 2024, 2025a). Agentic and multi-agent paradigms explore zero-shot or adaptive evaluation for harmful memes (Liu et al., 2025; Chen et al., 2025). Finally, meme-centered safety benchmarks for VLMs improve ecological validity by using real memes to probe harmful outputs (Lee et al., 2025).

3 Dataset

3.1 Dataset Curation

Language and task coverage. We curate a multilingual, multimodal collection of 38 publicly available meme datasets spanning nine languages such as English, Arabic, Bengali, Chinese, German, Spanish, Hindi, Romanian, and Russian, including both monolingual and code-mixed settings. These datasets cover 20 tasks as shown in Figure 1.

Filtering. The source datasets vary substantially in annotation schemes, label schema, and even in what is considered a "meme." Following prior work, we operationalize memes as images with embedded/overlaid text (image–text pairs) (Fersini et al., 2022; Sharma et al., 2020; Kiela et al., 2020). Accordingly, we formulate all tasks as *multimodal* and require an explicit text modality (either released by the dataset or obtained via OCR) to ensure consistent cross-dataset processing and reliable downstream modeling. In our curated set, we observed that some datasets include a non-trivial fraction of images that do not conform to this definition (e.g., images without embedded text; MMHS (Gomez et al., 2020)) and/or do not release extracted textual content (e.g., BanglaAbuseMeme (Das and Mukherjee, 2023), MIMIC Islam-

ophobia (Singh et al., 2024)). We therefore identify and remove samples with empty text by extracting text with a language-specific OCR pipeline based on EasyOCR.[2] We choose EasyOCR due to its public availability and its use in closely related prior work (Alam et al., 2024b). Filtering is performed independently within each dataset split (train/validation/test). Across datasets, we remove ∼92K samples with empty textual content, mostly from the MMHS dataset, which includes many images without embedded text. This filtering improves visual textual alignment while preserving the original dataset splits for remaining instances.

Label taxonomy. We design a taxonomy that *(i)* spans the task inventory observed across meme datasets, *(ii)* accommodates heterogeneous label structures (binary, multi-class, multi-label), and *(iii)* enables *semantic alignment* across datasets through explicit label definitions. The taxonomy serves as a *supervision schema* rather than a claim that a fixed ontology can exhaustively capture meme meaning.

Unified taxonomy and label harmonisation. Our taxonomy aligns heterogeneous benchmarks by defining a shared task inventory with 20 tasks and by normalising dataset-specific label encodings and surface forms such as `0/1`, yes/no, naming variants, and dataset-specific subclasses into canonical labels. This reduces spurious variation introduced by annotation conventions and supports reliable multilingual multitask training and cross-benchmark comparison.

Task families. We group the 20 tasks into five semantic categories.

1. **Safety & Moderation** hateful, harmful, toxic, offensive, abuse, vulgarity
2. **Social & Bias** misogyny, shaming, stereotype, objectification, violence
3. **Information & Intent** target, intention, metaphor, motivational
4. **Misinformation** propaganda, political, deepfake
5. **Humor & Sarcasm** humor, sarcasm

Multi-label normalisation. For multi-label settings like the misogyny categories in MIMIC, we normalise composites by decomposing them into shared primitives, for example `prejudice+humiliation` → {`prejudice`, `humiliation`}. We retain an `unspecified` label when the source annotation is explicitly underspecified. This preserves the original signal without forcing incompatible granularity into a single label space.

Mapping policy. Table 5 summarises the harmonisation used to construct MEMELENS. Since constituent datasets adopt heterogeneous schemes ranging from binary indicators such as `0/1` and yes/no to fine-grained taxonomies, we map each original label to a human-readable canonical form. When the original label space is reduced, marked with †, we either collapse rare or ambiguous subclasses into a canonical task label, for example MMHS 41-way → `hateful` and `not-hateful`, or remove under-represented residual categories such as *other* in MET-Meme En and *unspecified* in HarMeme. Canonical labels follow a consistent `positive` and `not-positive` convention to support unified evaluation across datasets and tasks.

Label cardinality and task formulation. Our taxonomy supports binary, multi-class, and multi-label supervision (reported in Table 8), so we do not convert all tasks into binary form. When the original label space is semantically meaningful and stable, we preserve it. Examples include the 3-way ROMEMES (Păiş et al., 2024) deepfake labels `real`, `fake`, and `deepfake`, the multi-class harmfulness levels and structured target categories in HARMEME (Pramanick et al., 2021), and the 7-way misogyny-category annotations in MIMIC. For multi-label datasets, we retain the signal using a standard one-vs-rest formulation implemented as per-label binary heads.

For the remaining tasks, we adopt a binary canonical form to support capability-oriented benchmarking across heterogeneous sources. Many datasets differ only in label encodings such as yes/no and `0/1`, or use incompatible granularities for the same phenomenon. Mapping to a shared binary form such as `hateful` and `not-hateful` removes mismatches and enables fair cross-dataset training and evaluation. In addition, multi-class schemes reflect dataset-specific intensity scales or annotation artifacts not consistently defined across benchmarks. Collapsing these to a coarse binary decision tests whether a model detects the core phenomenon rather than fitting a grading rubric.

We also normalize semantically equivalent label variants such as *no_harmful*, *not_harmful*, and *non-hateful* into a single canonical label to reduce

[2] https://github.com/JaidedAI/EasyOCR

variation driven by naming conventions rather than underlying phenomena.

3.2 Dataset Statistics

After preprocessing and filtering, the benchmark comprises approximately 178K, 22K and 40K instances for train/validation/test, respectively. Most datasets use binary labelings, while a smaller subset uses multi-class schemes; the number of labels ranges from 2 to 7, depending on the task and original annotation design.

Table 8 gives detailed dataset statistics. Figure 2 shows how instances are distributed across source languages and downstream tasks. The collection exhibits strong variation in language coverage. English accounts for the largest portion of instances at roughly 180K and spans most task categories, with a large portion in hate-related tasks at roughly 79K. Other languages are more task-focused. Bengali, Arabic, and Russian primarily map to offensive, misogyny, and toxic tasks, respectively, while Chinese is largely task-focused on propaganda, metaphor, and intention across datasets with varying annotation schemes and domain-specific characteristics. Hate- and harm-related tasks constitute a substantial portion of the benchmark across languages. The largest single source is `MMHS` with approximately 41K, 6K, and 12K instances in train, validation, and test, and several other datasets provide multilingual coverage, including `FHM`, `MIMIC Islamophobia`, `MUTE`, and `Multi3Hate`. The diversity in datasets and label structures supports analyses of cross-dataset transfer and taxonomy mismatch, while preserving the original annotation schemes aside from label-name normalization.

3.3 Explanation Augmentation

Explanation generation. We additionally developed explanations for all 38 datasets in *MemeLens*. Each augmented instance pairs the original label with a short natural-language rationale/explanation that justifies the decision and explicitly grounds it in *both* the visual content and the overlaid/extracted text, enabling evaluation under standard label prediction as well as explanation-aware settings. Following Kmainasi et al. (2025), we generate an explanation e by sampling from the conditional model $e \sim p_\theta\big(e \mid \mathbf{x}^{(I)}, \mathbf{x}^{(T)}, \mathbf{g}_t\big)$, where $\mathbf{x}^{(I)}$ is the meme image, $\mathbf{x}^{(T)}$ is its text modality, and $\mathbf{g}_t$ denotes task-specific annotation guidelines. To support multilingual and cross-cultural analysis, we provide explanations in English and, for

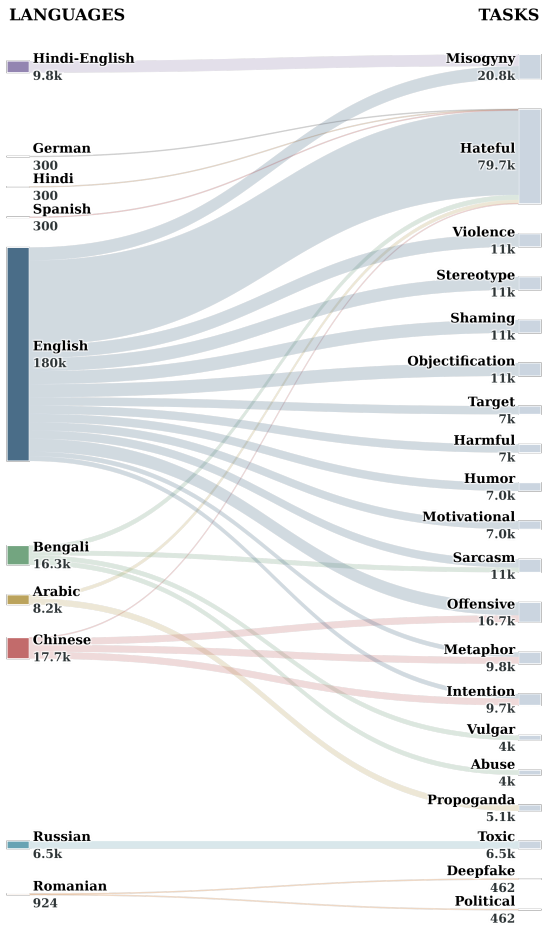

Figure 2: Language–task– coverage in MEMELENS.

non-English datasets, also in the original language, using task-specific prompts to ensure consistency across diverse phenomena (e.g., toxicity, misinformation, intent, and humor).

We generate explanations using GPT-4.1 (OpenAI, 2023). The detailed prompt is provided in section I, Listing 17. Kmainasi et al. (2025) reported that GPT-4.1 explanations can serve as high-quality references, with average human ratings above 4/5 on faithfulness, clarity, plausibility, and informativeness.

Overall, this augmentation extends *MemeLens* beyond label-only supervision and supports research on explainable multimodal learning, reasoning-centric evaluation, and explanation-conditioned training and inference. For explanation generation, we use zero-shot prompting with deterministic decoding by setting the temperature to 0 to ensure reproducibility. Explanations are constrained to approximately 114 words on average, with mean lengths of 118 words for English and 104 for native-language explanations. Detailed explanation statistics are reported in Appendix F.

Explanation quality. We evaluate explanations using four criteria: *informativeness*, *clarity*, *plausibility*, and *faithfulness*, each scored on a 1–5 Likert scale (higher is better), following the guidelines of Kmainasi et al. (2025); Hasanain et al. (2025). As human rating at MemeLens scale (38 datasets) is prohibitively costly, we use a complementary automatic assessment with two strong LLM judges, **Gemini 2.5 Pro** (Comanici et al., 2025) and **GPT-5** (Singh et al., 2026), to reduce single-judge bias.

Table 6 reports the results. Both judges assign consistently high scores across all criteria, with GPT-5 achieving a higher overall mean than Gemini (4.65 vs. 4.36). The largest gaps appear in clarity (4.80 vs. 4.57) and faithfulness (4.69 vs. 4.30). The shared ranking across judges suggests the explanations are generally coherent, informative, and grounded in the multimodal input.

Table 7 reports manual evaluation on a small sample of generated explanations from four representative datasets. Although the human scores are generally lower than the LLM-judge scores, they show a broadly consistent trend. FHM performs best, with strong and balanced scores across all criteria, while ArMeme is clearly more difficult, particularly in plausibility and faithfulness. Multi3 and MIMIC show moderate performance. Across datasets, clarity is typically higher than plausibility and faithfulness, suggesting that the model often produces explanations that are easy to follow but less reliably grounded. Overall, the human evaluation supports the broader picture from the automatic analysis while showing that explanation quality varies substantially by dataset.

4 Experiments

4.1 Instruction Dataset

To construct instruction-following datasets, we begin with a single manually designed seed instruction for each dataset. We then expand this seed using two LLMs, GPT-4.1 and Gemini-3-Pro, each generating 10 paraphrased instructions, resulting in approximately 20 English instructions per dataset.

We prepare two different output formats. The *classification-only* variant asks the model to output only the `Label`. The *classification-with-explanation* variant follows Kmainasi et al. (2025); Hasanain et al. (2025) and outputs `Label` and `Explanation` in a fixed structure. In both cases, we use a short system prompt that specifies the task and enforces the required format.

4.2 Baselines

Unimodal baselines. For text-only modeling, we fine-tune `bert-base-multilingual` (Xu et al., 2025) on OCR text. For image-only modeling, we use `ViT-B/16` (Dosovitskiy et al., 2020; Radford et al., 2021) trained solely on visual inputs.

Multimodal baselines. We fine-tune `Qwen3-VL-8B-Instruct` (Qwen3-VL-8B) (Xu et al., 2025) separately for each dataset using a dataset-specific sequence-classification (`seq_cls`) head. Unlike MEMELENS, this baseline does not share parameters across datasets nor perform unified multitask learning.

Zero-shot MLLM baselines. We additionally evaluate instruction-tuned MLLMs in a zero-shot setting, including GPT-4.1 and Qwen3-8B.

4.3 Training Setup

Unimodal models. We fine-tune text-only and image-only baselines with standard supervised training. We use batch size 32, learning rate 3×10^{-5}, and weight decay 0.01, and train for 7 epochs. Hyperparameters are tuned on the development set, and we select the checkpoint with the best development performance for test evaluation.

Multimodal models. We fine-tune all multimodal models with LoRA (Hu et al., 2021) and use a *two-stage training* following Kmainasi et al. (2025). ***Stage I*** (classification) optimises label prediction only. ***Stage II*** continues from the best Stage I checkpoint and jointly optimises label prediction and explanation generation, where the explanation loss acts as an auxiliary regulariser while keeping the inference conditioning for classification unchanged. More details on the experimental parameters are provided in Appendix J.

For the `seq_cls` baseline, we use the same setup, but train a dataset-specific head adapter for 20 epochs with learning rate 1×10^{-5}.

5 Results and Discussion

In this section we report results for *MemeLens* across multiple modeling paradigms and analyze *(i)* the impact of multimodality, *(ii)* variation across task families, *(iii)* dataset-level differences relative to prior work, and transfer under unified and single-dataset fine-tuning across languages, domains, and varying levels of task complexity and annotation noise.. We report Accuracy (Acc), Macro-F1 (M-F1) as the primary metric under class imbalance, and Weighted-F1 (W-F1).

Table 1 compares uni-modal, multimodal, zero-shot, and fine-tuned models including *MemeLens*. Uni-modal text and image models achieve comparable performance, suggesting that each modality independently provides useful signals. However, the multimodal sequence-classification baseline consistently outperforms both uni-modal variants, highlighting the importance of cross-modal interaction for meme understanding in complex, context-dependent scenarios across diverse datasets and real-world applications.

Model / Modality	Acc	M-F1	W-F1
Uni-modal (Text)	65.0	0.460	0.590
Uni-modal (Image)	63.6	0.472	0.600
Multi-modal (Seq-Classification)	71.0	0.580	0.680
Zero-Shot			
GPT-4.1	61.2	0.533	0.599
Qwen3-VL-8B-Instruct	55.1	0.482	0.539
InternVL3.5-8B (Wang et al., 2025)	55.4	0.476	0.545
Gemma-3-12B (Sellergren et al., 2025)	48.2	0.439	0.485
Qwen3-2B	45.6	0.394	0.431
Phi-3.5-Vision-4.2B (Abdin et al., 2024)	43.8	0.393	0.447
MEMELENS	**74.1**	**0.625**	**0.720**

Table 1: Competitive performance across models on the full **test set**.

Pretrained zero-shot models demonstrate competitive but inconsistent performance. GPT-4.1 performs similarly to the fine-tuned unimodal baselines despite no task adaptation, however Qwen3-VL-8B lags behind, suggesting that instruction tuning alone does not match task-specific fine-tuning for meme understanding across datasets with varying complexity, label distributions, and multimodal reasoning requirements in practice..

Fine-tuning yields the largest gains. **MEMELENS**, obtained by fine-tuning Qwen3-VL-8B on the unified multitask multimodal mixture, achieves the best overall results and produces label predictions with explanations for interpretability.

MemeLens vs. Gemma-4 (fine-tuned). To better understand whether a Qwen3-based model generalizes more effectively than another strong backbone, we also fine-tune Gemma-4. The fine-tuned model achieves an accuracy of 74.3%, which is nearly identical to the 74.1% obtained by MemeLens. The difference is not statistically significant according to the Wilcoxon signed-rank test.

5.1 Semantic Task Analysis

For task-wise analysis, we group datasets into five high-level semantic categories (Figure 1): *Safety & Moderation* (e.g., hateful, harmful, toxic, offensive, abuse), *Social & Bias Analysis* (e.g., misogyny, shaming, stereotyping, objectification, violence), *Information & Intent* (e.g., target identification, intention, metaphor, motivational content), *Misinformation* (e.g., propaganda, political manipulation, deepfakes), and *Humor & Sarcasm* (e.g., humor, sarcasm, vulgarity). This grouping reflects shared semantic properties across datasets.

Task Category	Text	Image	MM-Seq	MMLS	Qzs
Safety & Moderation	0.45	0.47	0.57	**0.61**	0.52
Social & Bias	0.50	0.57	0.69	**0.77**	0.61
Information & Intent	0.51	0.42	0.53	**0.60**	0.32
Misinformation	0.43	0.49	0.60	**0.67**	0.57
Humor & Sarcasm	0.41	0.42	0.46	**0.63**	0.30
Average	0.46	0.47	0.57	**0.65**	0.46

Table 2: Task-wise comparison across models on the full test set. **MMLS** refers to **MEMELENS**, and **Qzs** denotes the non-fine-tuned (zero-shot) Qwen3-VL-8B-Instruct model. Scores are reported as **Macro-F1**.

Language	Text	Image	MM-Seq	MEMELENS
English	0.430	0.438	0.517	**0.560**
Arabic	0.533	0.490	**0.666**	0.613
Chinese	0.539	0.385	0.618	**0.664**
Hindi	0.579	0.528	**0.750**	0.724
Spanish	0.358	0.661	0.661	**0.796**
Bangla	0.592	0.635	0.706	**0.723**
Romanian	0.331	0.461	0.461	**0.663**
German	0.371	0.504	0.714	**0.731**
Russian	0.493	0.495	**0.698**	0.691
Average	0.470	0.511	0.644	**0.685**

Table 3: Language-wise comparison across models. Scores are reported as **Macro-F1**.

In Table 2, we report Macro-F1 aggregated by task family. Multimodal models outperform unimodal baselines across all categories, and MEMELENS achieves the best average performance. Comparing MEMELENS with its unfine-tuned counterpart Qzs shows consistent gains on all categories.

MEMELENS improves most on *Social & Bias, Misinformation*, and *Humor & Sarcasm*, where visual context and implicit cues matter. *Safety & Moderation* also benefits from multimodality, while *Information & Intent* shows smaller but consistent gains. *Humor & Sarcasm* remains the most challenging, reflecting its reliance on pragmatic and cultural cues.

In Table 3, we report Macro-F1 by language. MEMELENS gives the strongest average performance and typically improves over unimodal baselines, while seq_cls remains competitive in several languages. Overall, unified multilingual training improves multimodal meme understanding, though transfer strength varies across languages, task categories and model families.

5.2 Dataset-Level SOTA Comparison

To assess robustness beyond aggregate metrics, we analyze performance at the dataset level and compare **MEMELENS** against previously reported state-of-the-art (SOTA) results where available. As meme datasets adopt heterogeneous evaluation metrics, we group benchmarks according to their official evaluation metric (Acc, M-F_1, or F_1 on the positive class) and report performance under the corresponding metric.

To ensure fair comparison, we exclude a small number of datasets for which prior work does not clearly specify the F_1 variant (e.g., Macro-F_1 vs. Weighted-F_1), and we additionally account for differences in dataset preprocessing. Our unified dataset removes samples without embedded text under a consistent text-over-image definition, which can affect comparability with results reported on unfiltered data.

Under these controlled comparisons, **MEMELENS** achieves performance that is broadly comparable to, and in some cases slightly exceeding, dataset-specific SOTA across benchmarks. In particular, **MEMELENS** slightly outperforms prior SOTA on accuracy-based benchmarks on average ($\Delta \approx 2\%$), while remaining close to parity under macro-F_1 ($\Delta \approx 0.01$) and F_1-POS ($\Delta \approx 0.02$). These results suggest that unified multitask training can match strong specialized models on average despite heterogeneous tasks, labels, and languages.

At the same time, performance still varies across datasets, which is expected given differences in task formulation, label granularity and domain cues. Overall, these findings highlight an inherent trade-off between unification and specialization. While dataset-specific models may achieve higher peak performance on individual benchmarks, a single unified model such as **MEMELENS** provides strong overall performance and broad coverage across tasks and languages. Detailed dataset-level results are provided in Appendix H.

5.3 Ablation: Single-Dataset Fine-Tuning

To test whether single-dataset fine-tuning is sufficient for robust meme understanding, we fine-tune the Qwen3-8B model on the Facebook Hateful Memes (FHM) dataset Kiela et al. (2020) and compare it to *MemeLens*. This diagnostic isolates the extent to which training on a single dataset transfers across related tasks. As shown in Table 4, single-dataset fine-tuning yields lower performance on *MemeLens* than unified multitask training. This suggests that models trained on a single benchmark tend to *over-specialize* to dataset-specific distributions and annotation conventions, which limits transfer across tasks and languages even when the datasets share broad semantic overlap.

Training Setup	Acc	M-F1	W-F1
MEMELENS	**74.0**	**0.620**	**0.720**
FHM-only fine-tuning	56.9	0.495	0.556

Table 4: Diagnostic comparison between unified multitask training and single-dataset fine-tuning. M-F1: Macro-F1, W-F1: Weighted-F1.

Importantly, we present this experiment as a diagnostic of when unified training across heterogeneous tasks and datasets is beneficial. The findings provide empirical support for this direction while also revealing remaining challenges in generalization. Methods that explicitly improve cross-dataset and label-set transfer remain a promising direction for future work.

6 Conclusion and Future Work

We presented MEMELENS, a unified multilingual and multitask explanation-enhanced vision–language model for meme understanding. We consolidated 38 public meme datasets and harmonized their heterogeneous annotations into a shared taxonomy of 20 tasks spanning harm, targets, figurative and pragmatic intent, and affect. Through comprehensive empirical analyses across modeling paradigms, task categories, and datasets, we showed that robust meme understanding benefits from multimodal training and exhibits substantial variation across semantic categories. Our dataset-level analysis further highlights a trade-off between unification and specialization, with unified training offering broad coverage and robustness while achieving performance that is broadly comparable to dataset-specific state-of-the-art models under controlled and fair comparison.

In future work, we aim to further improve robustness in low-resource and diverse settings while expanding to more languages, cultures, and emerging meme trends, incorporating continual learning, domain adaptation, and culturally grounded evaluation protocols to better capture evolving online communication dynamics.

Limitations

MEMELENS consolidates a large set of public resources and therefore inherits dataset-specific ar-

tifacts, including platform biases, annotation conventions, and temporal drift (e.g., evolving slang, symbols, and community norms). Consequently, performance under our unified labels may not fully reflect robustness to emerging meme formats or rapidly shifting online contexts. Our label unification maps heterogeneous annotations into a shared task taxonomy to enable cross-dataset learning and comparison; however, this mapping can compress fine-grained distinctions from the original label spaces and introduce partial ambiguity for some cases. Explanation supervision is generated and/or curated under practical constraints, so explanations may be imperfect—for instance, they can miss relevant cues, over-emphasize salient regions, or provide plausible yet incomplete rationales. Finally, although MEMELENS covers multiple languages, coverage remains uneven across languages and tasks, and low-resource settings may require targeted data collection to close these gaps.

Ethics and Broader Impact

MEMELENS is built from publicly available datasets for research on multilingual, robust, and transparent multimodal understanding. As memes can contain offensive, hateful, or otherwise sensitive content, models trained on this data may reproduce harmful language or reinforce stereotypes. We therefore encourage responsible release and use, including clear documentation, usage constraints, and guidance for safe deployment. Explanations are intended to be descriptive and evidence-grounded rather than endorsing the meme's message, but they may still surface sensitive concepts; downstream systems should apply content filtering, human oversight in high-stakes settings, and ongoing monitoring. While MEMELENS can support moderation, cross-cultural analysis of online manipulation, and improved multilingual accessibility, it could also be misused (e.g., to optimize harmful persuasion), highlighting the importance of controlled access and auditing.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *ArXiv preprint*, arXiv:2404.14219.

Siddhant Agarwal, Shivam Sharma, Preslav Nakov, and Tanmoy Chakraborty. 2024. MemeMQA: Multimodal question answering for memes via rationale-based inferencing. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5042–5078, Bangkok, Thailand. Association for Computational Linguistics.

Firoj Alam, Md Rafiul Biswas, Uzair Shah, Wajdi Zaghouani, and Georgios Mikros. 2024a. Propaganda to hate: A multimodal analysis of arabic memes with multi-agent llms. In *International Conference on Web Information Systems Engineering*, pages 380–390. Springer.

Firoj Alam, Stefano Cresci, Tanmoy Chakraborty, Fabrizio Silvestri, Dimiter Dimitrov, Giovanni Da San Martino, Shaden Shaar, Hamed Firooz, and Preslav Nakov. 2022. A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6625–6643.

Firoj Alam, Abul Hasnat, Fatema Ahmad, Md. Arid Hasan, and Maram Hasanain. 2024b. ArMeme: Propagandistic content in Arabic memes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21071–21090, Miami, Florida, USA. Association for Computational Linguistics.

Kate Barnes, Péter Juhász, Marcell Nagy, and Roland Molontay. 2024. Topicality boosts popularity: a comparative analysis of NYT articles and Reddit memes. *Social Network Analysis and Mining*, 14(1):119.

Mohamed Bayan Kmainasi, Mucahid Kutlu, Ali Ezzat Shahroor, Abul Hasnat, and Firoj Alam. 2026. Can thinking models think to detect hateful memes? *ArXiv preprint*, pages arXiv–2603.

Nicholas Brody and Sean Cullen. 2023. Meme sharing in relationships: The role of humor styles and functions. *First Monday*, 28(5).

Minh Duc Bui, Katharina Von Der Wense, and Anne Lauscher. 2025. Multi^{3}Hate: Multimodal, multilingual, and multicultural hate speech detection with vision–language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9714–9731, Albuquerque, New Mexico. Association for Computational Linguistics.

Rui Cao, Roy Ka-Wei Lee, Wen-Haw Chong, and Jing Jiang. 2022. Prompting for multimodal hateful meme classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 321–332, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dushyant Singh Chauhan, Dhanush S R, Asif Ekbal, and Pushpak Bhattacharyya. 2020. All-in-one: A deep attentive multi-task learning framework for humour, sarcasm, offensive, motivation, and sentiment on memes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 281–290, Suzhou, China. Association for Computational Linguistics.

Zixin Chen, Hongzhan Lin, Kaixin Li, Ziyang Luo, Zhen Ye, Guang Chen, Zhiyong Huang, and Jing Ma. 2025. AdamMeme: Adaptively probe the reasoning capacity of multimodal large language models on harmfulness. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4234–4253, Vienna, Austria. Association for Computational Linguistics.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *ArXiv preprint*, arXiv:2507.06261.

Mithun Das and Animesh Mukherjee. 2023. BanglaAbuseMeme: A dataset for Bengali abusive meme classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15498–15512, Singapore. Association for Computational Linguistics.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2009–2026, Mexico City, Mexico. Association for Computational Linguistics.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 70–98, Online. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv preprint*, arXiv:2010.11929.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. SemEval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549, Seattle, United States. Association for Computational Linguistics.

Deepesh Gavit, Debajyoti Mazumder, Samiran Das, and Jasabanta Patro. 2025. On VLMs for diverse tasks in multimodal meme classification. *ArXiv preprint*, arXiv:2505.20937.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1470–1478.

Hexiang Gu, Qifan Yu, Saihui Hou, Zhiqin Fang, Huijia Wu, and Zhaofeng He. 2025. MemeMind: A large-scale multimodal dataset with chain-of-thought reasoning for harmful meme detection. *ArXiv preprint*, arXiv:2506.18919.

Maram Hasanain, Md Arid Hasan, Mohamed Bayan Kmainasi, Elisa Sartori, Ali Ezzat Shahroor, Giovanni Da San Martino, and Firoj Alam. 2025. PropXplain: Can LLMs enable explainable propaganda detection? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23855–23863, Suzhou, China. Association for Computational Linguistics.

Ming Shan Hee, Wen-Haw Chong, and Roy Ka-Wei Lee. 2023. Decoding the underlying meaning of multimodal hateful memes. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5995–6003. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Ming Shan Hee and Roy Ka-Wei Lee. 2025. Demystifying hateful content: Leveraging large multimodal models for hateful meme detection with explainable decisions. In *Proceedings of the Nineteenth International AAAI Conference on Web and Social Media*, ICWSM '25, pages 774–785, Copenhagen, Denmark. AAAI Press.

Ming Shan Hee, Shivam Sharma, Rui Cao, Palash Nandi, Preslav Nakov, Tanmoy Chakraborty, and Roy Ka-Wei Lee. 2024. Recent advances in online hate speech moderation: Multimodality and the role of large models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4407–4419. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MUTE: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 32–39, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv preprint*, arXiv:2106.09685.

S M Jishanul Islam, Sahid Hossain Mustakim, Sadia Ahmmed, Md. Faiyaz Abdullah Sayeedi, Swapnil Khandoker, Syed Tasdid Azam Dhrubo, and Nahid Hossain. 2024. MIMIC: Multimodal islamophobic meme identification and classification. *ArXiv preprint*, arXiv:2412.00681.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. In *Proceedings of the Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Mohamed Bayan Kmainasi, Abul Hasnat, Md Arid Hasan, Ali Ezzat Shahroor, and Firoj Alam. 2025. MemeIntel: Explainable detection of propagandistic and hateful memes. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30263–30279, Suzhou, China. Association for Computational Linguistics.

DongGeon Lee, Joonwon Jang, Jihae Jeong, and Hwanjo Yu. 2025. Are vision-language models safe in the wild? A meme-based benchmark study. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30533–30576, Suzhou, China. Association for Computational Linguistics.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. 2024. Towards explainable harmful meme detection through multimodal debate between large language models. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*, pages 2359–2370, Singapore, Singapore.

Hongzhan Lin, Ziyang Luo, Jing Ma, and Long Chen. 2023. Beneath the surface: Unveiling harmful memes with multimodal reasoning distilled from large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9114–9128, Singapore. Association for Computational Linguistics.

Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. 2025. MIND: A multi-agent framework for zero-shot harmful meme detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 923–947, Vienna, Austria. Association for Computational Linguistics.

Junyu Lu, Bo Xu, Xiaokun Zhang, Haohao Zhu, Kaichun Wang, Liang Yang, and Hongfei Lin. 2025. Is having rationales enough? Rethinking knowledge enhancement for multimodal hateful meme detection. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25)*, Padua, Italy.

Delfina S. Martinez Pandiani, Erik Tjong Kim Sang, and Davide Ceolin. 2025. 'Toxic' memes: A survey of computational perspectives on the detection and explanation of meme toxicities. *Online Social Networks and Media*, 47:100317.

Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving hateful meme detection through retrieval-guided contrastive learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5333–5347, Bangkok, Thailand. Association for Computational Linguistics.

Jingbiao Mei, Jinghong Chen, Guangyu Yang, Weizhe Lin, and Bill Byrne. 2025a. Robust adaptation of large multimodal models for retrieval augmented hateful meme detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23817–23839, Suzhou, China. Association for Computational Linguistics.

Jingbiao Mei, Mingsheng Sun, Jinghong Chen, Pengda Qin, Yuhong Li, Da Chen, and Bill Byrne. 2025b. ExPO-HM: Learning to explain-then-detect for hateful meme detection. *ArXiv preprint*, arXiv:2510.08630.

OpenAI. 2023. Gpt-4 technical report. *ArXiv preprint*, arXiv:2303.08774.

Vasile Păiş, Sara Niţă, Alexandru-Iulius Jerpelea, Luca Pană, and Eric Curea. 2024. RoMemes: A multimodal meme corpus for the Romanian language. *ArXiv preprint*, arXiv:2410.15497.

Jeongsik Park, Khoi P. N. Nguyen, Jihyung Park, Minseok Kim, Jaeheon Lee, Jae Won Choi, Kalyani Ganta, Phalgun Ashrit Kasu, Rohan Sarakinti, Sanjana Vipperla, Sai Sathanapalli, Nishan Vaghani, and Vincent Ng. 2025. MemeInterpret: Towards an all-in-one dataset for meme understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16073–16087, Suzhou, China. Association for Computational Linguistics.

Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2783–2796, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763.

Alisa Raikovskaia, Nurlan Rakhimzhanov, and Alexander Rogachev. 2023. Toxic memes detection dataset. Zenodo.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. 2025. MedGemma technical report. *ArXiv preprint*, arXiv:2507.05201.

Siddhant Bikram Shah, Shuvam Shiwakoti, Maheep Chaudhary, and Haohan Wang. 2024. MemeCLIP: Leveraging CLIP representations for multimodal meme classification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17320–17332.

Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.

Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What do you meme? Generating explanations for visual semantic role labelling in memes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(8):9763–9771.

Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and understanding harmful memes: A survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, IJCAI '22, pages 5597–5606, Vienna, Austria. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2026. OpenAI GPT-5 system card. *ArXiv preprint*, arXiv:2601.03267.

Aakash Singh, Deepawali Sharma, and Vivek Kumar Singh. 2024. MIMIC: Misogyny identification in multimodal internet content in Hindi-English code-mixed language. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *ArXiv preprint*, arXiv:2508.18265.

Bo Xu, Tingting Li, Junzhe Zheng, Mehdi Naseriparsa, Zhehuan Zhao, Hongfei Lin, and Feng Xia. 2022. MET-Meme: A multimodal meme dataset rich in metaphors. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2887–2899.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025. Qwen3-Omni technical report. *ArXiv preprint*, arXiv:2509.17765.

Jing Zhang and Yujin Wang. 2022. Srcb at semeval-2022 task 5: Pretraining based image to text late sequential fusion system for multimodal misogynous meme identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 585–596.

A Label Taxonomy

Table 5 summarises the full label harmonisation used to construct *MemeLens*, mapping each dataset-specific annotation scheme to the canonical labels adopted in our unified taxonomy. The constituent datasets adopt heterogeneous label schemes, ranging from simple binary indicators such as `0/1` or `Yes/No`, to fine-grained multi-class severity rubrics (e.g., MET-Meme offensiveness levels), structured target categories (e.g., HarMeme), and multi-label codes for misogyny subtypes (e.g., MIMIC2024).

Dataset	Task	#	Original Labels	Canonical Labels
BanglaAbuseMeme	sarcasm	2	Yes / No	`sarcasm` / `not-sarcasm`
	vulgar	2	Vulgar / Not Vulgar	`vulgar` / `not-vulgar`
	abuse	2	Abusive / Non-abusive	`abusive` / `not-abusive`
Toxic-memes (RU)	toxic	2	1 / 0	`toxic` / `non-toxic`
RoMemes	deepfake	3	Real / Fake / DeepFake	`real` / `fake` / `deepfake`
	political	2	Yes / No	`political` / `not-political`
MIMIC2024	misogyny	2	1 / 0	`misogynous` / `not-misogynous`
	misogyny cat.†	7	multi-label codes	`obj.` / `prej.` / `hum.` / `unspec.` `obj.,prej.` / `obj.,hum.` / `prej.,hum.`
Multi3Hate (en/de/zh/es/hi)	hateful	2	1 / 0	`hateful` / `not-hateful`
MMHS†	hateful	2	any label containing "hate" / otherwise	`hateful` / `not-hateful`
Memotion	humour	4	not_funny / funny / very_funny / hilarious	`not-funny` / `funny` / `very-funny` / `hilarious`
	sarcasm	4	not_sarcastic / general / twisted_meaning / very_twisted	`not-sarcastic` / `general-sarcasm` / `twisted-meaning` / `very-twisted`
	motivational	2	motivational / not_motivational	`motivational` / `not-motivational`
	offensive	4	not_offensive / slight / very_offensive / hateful_offensive	`not-offensive` / `sl.-offensive` / `very-offensive` / `hateful-offensive`
MET-Meme (En)	intention†	4	1 / 2 / 3 / 4 / 5 → removed *other*	`expressive` / `offensive` / `interactive` / `entertaining`
	offensive	4	0 / 1 / 2 / 3	`not-offensive` / `sl.-offensive` / `mod.-offensive` / `very-offensive`
	metaphor	2	1 / 0	`metaphorical` / `literal`
MET-Meme (Zh)	intention	5	1 / 2 / 3 / 4 / 5	`expressive` / `offensive` / `interactive` / `entertaining` / `other`
	offensive	4	0 / 1 / 2 / 3	`not-offensive` / `sl.-offensive` / `mod.-offensive` / `very-offensive`
	metaphor	2	1 / 0	`metaphorical` / `literal`
HarMeme†	harmful	3	not_harmful / partially_harmful / very_harmful (removed *unspec.*)	`not-harmful` / `partially-harmful` / `very-harmful`
	target	5	none / indivisual / org. / comm. / society	`none` / `indiv.` / `org.` / `comm.` / `society`
HarMeme-Covid	harmful	3	not_harmful / partially_harmful / very_harmful	`not-harmful` / `partially-harmful` / `very-harmful`
	target	5	none / indivisual / org. / comm. / society	`none` / `indiv.` / `org.` / `comm.` / `society`
MAMI	misogynous	2	1 / 0	`misogynous` / `not-misogynous`
	shaming	2	1 / 0	`shaming` / `not-shaming`
	stereotype	2	1 / 0	`stereotype` / `not-stereotype`
	obj.	2	1 / 0	`obj.` / `not-obj.`
	violence	2	1 / 0	`violence` / `not-violence`
ArMeme†	propaganda	2	propaganda / not_propaganda (4 → 2)	`propaganda` / `not-propaganda`
FHM	hateful	2	hateful / not-hateful	`hateful` / `not-hateful`
Prop2Hate-Meme	hateful	2	hateful / not-hateful	`hateful` / `not-hateful`
MIMIC-Islamophobia	hateful	2	1 / 0	`hateful` / `not-hateful`
MUTE	hateful	2	hate / not-hate	`hateful` / `not-hateful`

Table 5: Label mapping from original annotations to canonical labels in MEMELENS. **#** = number of canonical classes. † = original label set was reduced. Abbreviations: *obj.* = objectification, *prej.* = prejudice, *hum.* = humiliation, *unspec.* = unspecified, *sl.* = slightly, *mod.* = moderately, *org.* = organization, *comm.* = community, *indiv.* = individual.

To enable consistent cross-dataset training and evaluation, we map each original label to a human-readable canonical form following a uniform `positive/not-positive` naming convention (e.g., `hateful/not-hateful`, `vulgar/not-vulgar`).

For datasets whose original label spaces are reduced, marked with † in the table, we either map rare or ambiguous subclasses into the canonical task label (e.g., the 41-way hate subtypes in MMHS are mapped to `hateful/not-hateful`) or remove under-represented categories (e.g., the `other` class in MET-Meme English and the `unspecified` class in HarMeme). Importantly, the taxonomy is *not* uniformly binary, since we preserve the full cardinality of the original label space when it is semantically meaningful and stable across benchmarks. Examples include the 3-way deepfake labels in RoMemes, the 3-way harmfulness and 5-way target schemes in HarMeme and HarMeme-COVID, the 4-way humour, sarcasm, and offensiveness scales in Memotion, and the 7-way multi-label misogyny categories in MIMIC2024, which we represent by decomposing composite codes (e.g., `prejudice+humiliation`) into shared primitives. Overall, this mapping policy minimises spurious variation introduced by annotation conventions, while preserving genuinely distinct semantic distinctions across datasets, tasks, and languages.

B Explanation Quality

We assess the quality of the generated explanations using four established criteria, namely *informativeness*, *clarity*, *plausibility*, and *faithfulness*. Each criterion is scored on a 1–5 Likert scale, following the evaluation setup reported in Kmainasi et al. (2025). Since human annotation at the full *MemeLens* scale (38 datasets, ∼40K test instances) is prohibitively expensive, we complement targeted human review with a large-scale automatic evaluation using two strong and independent LLM judges, *Gemini 2.5 Pro* (Comanici et al., 2025) and *GPT-5* (Singh et al., 2026), in order to reduce single-judge bias. Table 6 reports the LLM-judge results aggregated over the entire test set. Both judges assign consistently high scores across all four criteria, indicating that the generated explanations are generally coherent, informative, and well grounded in the multimodal input. GPT-5 yields a higher overall mean than Gemini 2.5 Pro (4.65 vs. 4.36), with the higher gaps appearing in *clarity* (4.80 vs. 4.57) and *faithfulness* (4.69 vs. 4.30). However, the ranking of criteria is consistent across the two judges, providing evidence that the explanations capture the relevant cues underlying each label.

To validate the automatic evaluation, Table 7 reports a human evaluation on a stratified subset drawn from four representative datasets spanning different languages and tasks: ArMeme (Arabic propaganda), FHM (English hate), Multi3Hate (multilingual hate), and MIMIC (Hindi–English misogyny). Although the absolute human scores are systematically lower than those of the LLM judges, the relative trend is broadly consistent. FHM receives the strongest and most balanced ratings across all criteria, while ArMeme is clearly the most difficult, particularly for *plausibility* (2.89) and *faithfulness* (2.85). Multi3Hate and MIMIC show moderate, comparable performance. Across datasets, *clarity* is typically rated higher than *plausibility* and *faithfulness*, suggesting that the model produces explanations that are easy to follow, with further room to strengthen their grounding in the multimodal evidence. Overall, the LLM-judge and human evaluations indicate that the explanation supervision in *MemeLens* is of high overall quality, while also highlighting that explanation quality varies substantially by dataset and remains most challenging in low-resource settings.

Model	Info.	Clar.	Plaus.	Faith.	Overall
Gemini-2.5	4.35	4.57	4.24	4.30	4.36
GPT-5	4.48	4.80	4.62	4.69	4.65

Table 6: LLM-judge scores for explanation quality on the whole test set (∼40K). **Info.** = informativeness, **Clar.** = clarity, **Plaus.** = plausibility, and **Faith.** = faithfulness. **Gemini-2.5** = Gemini 2.5 Pro.

Dataset	Inf.	Clar.	Plaus.	Faith.	Overall
ArMeme	3.09	3.48	2.89	2.85	3.08
FHM	4.67	4.65	4.64	4.60	4.64
Multi3	3.33	3.95	3.41	3.33	3.50
MIMIC	3.44	4.05	3.42	3.41	3.58

Table 7: Human evaluation of generated explanation on a small sample from each dataset.

C Dataset Details

BanglaAbuseMeme (Das and Mukherjee, 2023). This is a Bengali (Bangla) multimodal meme dataset curated from social platforms, where each example pairs an image with its embedded text (using OCR) for abusive-language understanding. It is designed to support abusive content detection in memes, emphasizing cases where abusiveness emerges from the interaction of image and text. We use the *abuse*, *sarcasm*, and *vulgarity* subsets of BanglaAbuseMeme, formulating each as a **binary classification task** (present vs. not-present).

RoMemes (Păiş et al., 2024). RoMemes is a Romanian-language multimodal meme benchmark consisting of image–text pairs annotated for multiple meme understanding tasks. The dataset supports evaluation beyond high-resource English settings by focusing on Romanian social media content. In our experiments, we use the *political meme detection* subset and the *fake image detection*.

HarMeme (Pramanick et al., 2021). HarMeme is an English-language multimodal meme dataset annotated for harmfulness and target identification, where each meme image is paired with embedded textual content. The dataset provides fine-grained harmfulness labels (e.g., not harmful, partially harmful, very harmful) and annotations indicating the targeted entity (e.g., individual, organization, group). HarMeme is organized into two subsets: Harm-C, consisting of COVID-19-related memes, and Harm-P, consisting of political memes from the U.S. context. In our experiments, we use Harm-C from the original HarMeme-V0 release and Harm-P from the updated HarMeme-V1 re-

lease. In addition to harmfulness, we use the target annotation; memes without an explicit target are treated as *none*, following the dataset schema.

Islamophobic memes dataset (Islam et al., 2024). The Islamophobic memes dataset is a curated collection of multimodal memes designed to study Islamophobic and anti-Muslim content in online discourse. Each instance consists of a meme image paired with its embedded textual content, with annotations supporting supervised detection of Islamophobic (hateful) versus non-hateful memes. In our experiments, we exclude samples whose images do not contain embedded text, focusing on instances where both visual and textual modalities contribute to the expression of hateful intent.

MMHS150K (Gomez et al., 2020). MMHS150K is a large-scale multimodal hate-speech dataset of social media posts that pairs images with accompanying textual content. The dataset is designed with the aim to support hate-speech detection in multimodal settings, including cases where the hateful meaning emerges from the interaction between the visual and the textual signals. In our experiments, we exclude any examples for which the images contain no embedded text, and we focus on instances where both modalities actually contribute to the multimodal context.

Multi3Hate (Hossain et al., 2022). Multi3Hate is a multilingual, parallel meme dataset created by instantiating the same meme templates across multiple languages, enabling controlled cross-lingual evaluation of hate-speech detection. By keeping visual content fixed while varying language realizations, it supports analysis of cross-lingual robustness and transfer in multimodal hate detection. In our experiments, we use the Bengali, German, English, Spanish, Hindi, and Chinese subsets of the dataset.

Prop2Hate (Alam et al., 2024a). Prop2Hate is an Arabic multimodal meme dataset constructed with the aim to study the intersection between propagandistic and hateful content in memes. The dataset extends an existing Arabic propagandistic meme collection by annotating memes for hatefulness, where each instance pairs a meme image with its embedded textual content. The annotations in this dataset support supervised hate-speech detection in multimodal settings, capturing cases where hateful meaning emerges from the interaction between visual and textual cues. In our experiments, we use the dataset in a **binary classification** setup, distinguishing between *hateful* and *non-hateful* memes.

Memotion (Sharma et al., 2020). Memotion is a multimodal meme dataset released as part of SemEval-2020 Task 8. The dataset pairs meme images with their corresponding embedded textual content. The dataset provides annotations for multiple emotion-related categories, including *humor*, *sarcasm*, *offensiveness*, and *motivation*, thus enabling the study of affective phenomena in memes that arise from the interaction between visual and the textual cues.

MUTE (Hossain et al., 2022). MUTE is a Bengali multimodal meme dataset introduced to support hateful meme detection in low-resource language settings. The dataset pairs meme images with their textual content, including both Bengali and Bengali–English code-mixed captions, and provides annotations for supervised hate-speech detection. In our experiments, we use the Bengali subset of MUTE under a **binary classification** setup with two labels (*hateful* vs. *non-hateful*).

MET-Meme (Xu et al., 2022). MET-Meme is a metaphor-rich multimodal meme dataset that pairs images with their textual content to support multimodal meme understanding. The dataset provides annotations for multiple semantic aspects of memes, including metaphor occurrence, intention, and offensiveness, capturing cases where meaning arises from the interaction between visual and textual elements. In our experiments, we use the *metaphor occurrence detection*, *intention detection*, and *offensiveness detection* annotations for both the English and Chinese subsets of the dataset.

MAMI (Fersini et al., 2022). MAMI is a dataset released for SemEval-2022 Task 5, consisting of meme images paired with embedded textual content. The dataset provides annotations for misogyny at both a coarse level (misogynous vs. non-misogynous) and a fine-grained level, including *shaming*, *stereotype*, *objectification*, and *violence*. In our experiments, we use misogyny identification as the primary task. Fine-grained categories (*violence*, *objectification*, *shaming*, *stereotype*) are modeled as auxiliary binary prediction tasks within our unified framework, but are not treated as independent benchmarks.

MIMIC (Singh et al., 2024). MIMIC is a Hindi–English code-mixed multimodal dataset for misogyny identification in online memes and posts. Each example combines an image with short, often code-mixed textual content, and is annotated for misogyny. The dataset is provided in two variants: one formulated as a *binary misogyny detection* dataset (misogynous vs. non-misogynous), and another containing *category-level annotations* for misogyny-related subtypes. We use the category-level dataset as well, where memes are labeled with the following categories: *unspecified*, *prejudice*, *objectification*, *humiliation*, as well as their observed combinations (e.g., *objectification+humiliation*, *prejudice+humiliation*, *objectification+prejudice*).

ArMeme (Alam et al., 2024b). ArMeme is an Arabic multimodal meme dataset designed for propaganda detection. It consists of meme images paired with their overlaid text and corresponding propaganda annotations. In this work, we use the binary version of ArMeme, which includes two labels: *propaganda* and *not propaganda*.

FHM (Kiela et al., 2020). FHM (Facebook Hateful Memes) is an English-language multimodal meme dataset introduced as part of the Hateful Memes Challenge. Each example pairs a meme image with its textual content and is annotated for hate speech under a binary labeling scheme (hateful vs. non-hateful). The dataset is specifically designed to require joint reasoning over visual and textual modalities by including challenging confounders that prevent reliance on unimodal cues, making it a standard benchmark for evaluating multimodal hate-speech detection models.

Toxic Memes Detection Dataset (Raikovskaia et al., 2023). The Toxic Memes Detection Dataset is a Russian-language multimodal meme dataset released on Zenodo, consisting of images collected from popular Russian Telegram channels and annotated for toxic content according to Facebook Community Standards. The dataset supports supervised toxic content detection in memes through a binary labeling scheme. As the dataset provides image-level annotations without explicit textual transcriptions, we extract the embedded text from meme images using OCR to construct image–text pairs for multimodal modeling.

D Data Release

We make our *MemeLens* dataset publicly available to the research community with the aim to enable further research in multilingual multimodal meme analysis. Our resource extends existing public datasets through label unification and the addition of explanations. Accordingly, we release our contribution under the CC BY-SA 4.0 license, that is, the Creative Commons Attribution-ShareAlike 4.0 International License: `https://creativecommons.org/licenses/by-sa/4.0/` . As the original datasets may be distributed under different licenses, we strongly encourage users of *MemeLens* to cite the source datasets and prior work where appropriate.

E State-of-the-Art (SOTA) Results

In order to contextualize our results, we report previously published state-of-the-art (SOTA) or best-reported performance figures for each dataset and task considered in this work, as summarized in Table 9. Note that all reported numbers are taken directly from the original dataset papers or subsequent benchmark studies and are provided for reference only; we did not try to reproduce the published numbers.

F Explanation Length Statistics

In Table 10, we present the average explanation length across all 38 datasets. We measure explanation length in words for both English and native-language explanations where available.

Overall Statistics. Across all datasets, English explanations average 118 words, with lengths ranging from 109 words (MET-Meme Offensive in Chinese) to 128 words (Memotion Humor). This narrow range (19 words) indicates consistent levels of detail across tasks and languages. Native-language explanations are slightly shorter (104 words), with lengths ranging from 84 words (Toxic Memes in Russian) to 128 words (MIMIC category-level misogyny in Hindi–English code-mixed text).

Cross-Lingual Comparison. Native-language explanations are shorter than their English counterparts, with an average reduction of 14 words (12%). The gap is most pronounced in Russian (29 words shorter for Toxic Memes), Bengali (21–26 words shorter across BanglaAbuse tasks), and Arabic (24–26 words shorter for ArMeme and Prop2Hate).

Conversely, Hindi and code-mixed Hindi-English explanations are comparable to or slightly

Task	Dataset	Language	Label Type	Train	Val	Test
Abuse	BanglaAbuseMeme	Bengali	Binary	2,827	404	806
Deepfake	RoMemes	Romanian	Multi-class	322	47	93
Harmful	HarMeme (COVID-19)	English	Multi-class	3,008	174	354
Harmful	HarMeme	English	Multi-class	2,937	176	355
Hateful	FHM	English	Binary	8,500	540	2,000
Hateful	MIMIC (Islamophobia)	English	Binary	515	75	150
Hateful	MMHS	English	Binary	41,484	5,887	11,881
Hateful	Prop2Hate-Meme	Arabic	Binary	2,143	312	606
Hateful	MUTE	Bengali	Binary	3,365	375	416
Hateful	Multi3Hate	German	Binary	209	30	61
Hateful	Multi3Hate	English	Binary	209	30	61
Hateful	Multi3Hate	Spanish	Binary	209	30	61
Hateful	Multi3Hate	Hindi	Binary	209	30	61
Hateful	Multi3Hate	Chinese	Binary	209	30	61
Humor	Memotion	English	Multi-class	4,890	699	1,398
Intention	MET-Meme	English	Multi-class	2,749	396	784
Intention	MET-Meme	Chinese	Multi-class	4,067	584	1,161
Metaphor	MET-Meme	English	Binary	2,754	395	788
Metaphor	MET-Meme	Chinese	Binary	4,061	587	1,166
Misogyny	MAMI	English	Binary	9,000	1,000	1,000
Misogyny	MIMIC2024	Hindi-English	Binary	3,448	490	967
Misogyny (Cat.)	MIMIC2024	Hindi-English	Multi-label	3,429	486	988
Motivational	Memotion	English	Binary	4,890	700	1,397
Objectification	MAMI	English	Binary	9,000	1,000	1,000
Offensive	Memotion	English	Multi-class	4,888	700	1,399
Offensive	MET-Meme	English	Multi-class	2,752	396	789
Offensive	MET-Meme	Chinese	Multi-class	4,064	580	1,170
Political	RoMemes	Romanian	Binary	322	47	93
Propaganda	ArMeme	Arabic	Binary	3,604	522	1,021
Sarcasm	Memotion	English	Multi-class	4,888	700	1,399
Sarcasm	BanglaAbuseMeme	Bengali	Binary	2,827	404	806
Shaming	MAMI	English	Binary	9,000	1,000	1,000
Stereotype	MAMI	English	Binary	9,000	1,000	1,000
Target (COVID)	HarMeme	English	Multi-class	3,008	174	354
Target	HarMeme	English	Multi-class	2,938	176	355
Toxic	Toxic Memes	Russian	Binary	4,512	647	1,297
Violence	MAMI	English	Binary	9,000	1,000	1,000
Vulgar	BanglaAbuseMeme	Bengali	Binary	2,825	405	807
Total				**178,062**	**22,228**	**40,105**

Table 8: Data distribution across tasks, datasets, and languages.

longer than English versions, suggesting that explanation verbosity may be influenced by linguistic structure and expressiveness in the target language. Note that Chinese (ZH) native explanations were excluded from the native average calculation due to challenges in word-level tokenization for logographic writing systems, where character-based segmentation does not directly correspond to the word-level granularity used for other languages.

G Dataset-Level Results

In Table 11, we provide a comprehensive comparison across text-only, image-only, and multimodal sequence classification baselines, alongside our proposed *MemeLens* model, covering a wide range of tasks, languages, and label granularities. The results are presented using each dataset's official evaluation metric to ensure fair comparison. This detailed breakdown complements the aggregate analyses in the main paper and enables fine-grained inspection of model behavior across diverse multimodal reasoning scenarios.

H Dataset-Level Comparison with SOTA

As meme datasets vary substantially in task formulation, label spaces, and evaluation protocols, we group them by their official evaluation metrics, including Accuracy, Macro-F_1, and positive-class F_1 (F_1-POS), ensuring fair and consistent comparison across diverse experimental settings and benchmarks. This analysis assesses how a single unified multilingual, multitask model performs across heterogeneous meme benchmarks, rather than aiming to establish new dataset-specific state of the art. For

Dataset	Task	Ref.	Score
BanglaAbuse.	Vulg. (Ma-F1)	2023	71.66
	Sarc. (Ma-F1)		68.28
	Abuse (Ma-F1)		70.51
RoMemes	Deepfake (Acc.)	2024	97.1
	Political (Acc.)		62.0
HarMeme	Harm-C (Ma-F1)	2021	53.85
	Harm-P (Ma-F1)		64.70
Multi3Hate	ENG (Acc.)	2025	75.8
	DE (Acc.)		72.2
	ES (Acc.)		69.2
	HI (Acc.)		63.1
	ZH (Acc.)		68.7
Prop2Hate	Hate (Ma-F1)	2024a	70.9
MAMI	Misog. Classif. (F1)	2022	83.4
MUTE	Classif. (W-F1)	2022	67.2
MIMIC	Islam. (Ma-F1)	2024	69.5
MMHS150K	Hate (Acc.)	2020	68.4
FHM	Hate (Acc.)	2025	79.2
MET-Meme	EN (F1-pos)	2022	82.39
	ZH (F1-pos)		77.23
Russ. Toxic	–	N/A	–

Table 9: SOTA reference results. Ma-F1: Macro-F1; W-F1: Weighted F1; Acc.: Accuracy.

each dataset, we report the performance of **MEME-LENS** under the official metric and compare it with the corresponding SOTA results from prior work.

To ensure fair comparison, we exclude datasets for which prior work does not clearly specify the F_1 variant used (e.g., Macro-F_1 vs. Weighted-F_1). Moreover, our unified training and evaluation pipeline filters out samples that do not contain textual content within the image, following a consistent text-over-image meme definition applied uniformly across all datasets. As a result, some dataset-level comparisons may not be strictly equivalent to prior results reported on unfiltered data or under differing preprocessing and experimental setups. We therefore treat these comparisons as indicative rather than definitive, emphasizing relative trends and cross-dataset consistency rather than absolute performance differences, and interpret results with caution given these methodological differences.

H.1 Accuracy-Based Evaluation

Table 12 reports dataset-level results for benchmarks evaluated using Accuracy. Under this metric, **MEMELENS** slightly outperforms dataset-specific SOTA on average ($\Delta \approx 2.0\%$) under controlled and fair comparison settings. While performance

Dataset	Task	L	EN Avg	Nat Avg
BanglaAbuse	Abuse	BN	116	93
RoMemes	Deepfake	RO	122	110
HarMeme (Co.)	Harmful	EN	122	-
HarMeme	Harmful	EN	125	-
FHM	Hateful	EN	110	-
MIMIC_Isl	Hateful	EN	114	-
MMHS	Hateful	EN	122	112
Prop2Hate	Hateful	AR	123	94
MUTE	Hateful	BN	121	95
Multi3Hate	Hateful	DE	113	102
Multi3Hate	Hateful	EN	112	-
Multi3Hate	Hateful	ES	113	114
Multi3Hate	Hateful	HI	115	118
Multi3Hate	Hateful	ZH	112	-
Memotion	Humor	EN	128	-
MET-Meme	Intention	EN	122	-
MET-Meme	Intention	ZH	114	-
MET-Meme	Metaphor	EN	120	-
MET-Meme	Metaphor	ZH	113	-
MAMI	Misogyny	EN	116	-
MIMIC2024	Misogyny	HI-EN	119	127
MIMIC2024 (Cat.)	Misogyny	HI-EN	123	128
Memotion	Motiv.	EN	114	-
MAMI	Objectif.	EN	122	-
Memotion	Offensive	EN	124	-
MET-Meme	Offensive	EN	119	-
MET-Meme	Offensive	ZH	109	-
RoMemes	Political	RO	117	104
ArMeme	Propaganda	AR	125	99
Memotion	Sarcasm	EN	125	-
BanglaAbuse	Sarcasm	BN	121	92
MAMI	Shaming	EN	118	-
MAMI	Stereotype	EN	120	-
HarMeme (CO.)	Target	EN	126	-
HarMeme	Target	EN	126	-
Toxic Memes	Toxic	RU	113	84
MAMI	Violence	EN	116	-
BanglaAbuse	Vulgar	BN	113	87
Average	-	-	**118**	**104**

Table 10: Average explanation length (in words) per dataset, task, and language. English explanations are available for all datasets, while native-language (Nat. = Native) explanations are reported where applicable. L. = Language.

varies across individual datasets, **MEMELENS** matches or exceeds prior SOTA on several benchmarks, reflecting the effectiveness of unified multitask training across heterogeneous datasets, languages, and label spaces, and demonstrating strong generalization despite differences in annotation schemes, domains, and underlying data distributions across benchmarks, as well as improved robustness to cross-dataset variation and noise, particularly in low-resource and cross-lingual scenarios.

H.2 F_1-POS Evaluation

Table 13 reports results for datasets evaluated using F1 on the positive class. Across these benchmarks,

Dataset	Task	Lang.	Text-Only			Image-Only			MM-Seq			MEMELENS		
			Acc	Ma	W	Acc	Ma	W	Acc	Ma	W	Acc	Ma	W
BanglaAbuse	Abuse	BN	66.0	.564	.615	68.0	.628	.663	73.1	.698	.723	**78.7**	**.759**	**.782**
RoMemes	Deepfake	RO	63.4	.259	.493	64.5	.399	.630	57.5	.338	.551	**77.0**	**.491**	**.753**
HarMeme (Co)	Harmful	EN	71.2	.499	.706	70.3	.443	.677	**81.1**	**.546**	**.797**	74.8	.523	.740
HarMeme	Harmful	EN	49.9	.338	.489	53.5	.362	.527	59.0	.400	.590	**62.2**	**.467**	**.617**
Prop2Hate	Propaganda	AR	74.6	.427	.637	74.3	.426	.636	**80.0**	**.650**	**.760**	77.2	.546	.703
MUTE	Propaganda	BN	64.2	.556	.602	68.8	.659	.682	**73.0**	**.710**	**.730**	71.9	.700	.718
Multi3Hate	Hateful	DE	59.0	.371	.438	55.7	.504	.533	72.0	.710	.720	**75.4**	**.731**	**.745**
MIMIC_Isl	Hateful	EN	64.7	.633	.635	58.0	.576	.577	51.0	.340	.350	**70.7**	**.707**	**.707**
MMHS	Hateful	EN	63.1	.387	.488	62.1	.495	.561	**63.0**	.500	**.570**	61.4	**.516**	.568
Multi3Hate	Hateful	EN	57.4	.573	.573	50.8	.507	.507	**77.0**	**.770**	**.770**	74.1	.735	.734
FHM	Hateful	EN	63.3	.541	.592	62.3	.507	.567	76.0	.740	.760	**79.8**	**.782**	**.798**
Multi3Hate	Hateful	ES	55.7	.358	.399	67.2	.661	.668	62.0	.600	.610	**80.0**	.796	.799
Multi3Hate	Hateful	HI	65.6	.579	.618	57.4	.528	.559	**75.0**	**.750**	**.760**	75.4	.724	.744
Multi3Hate	Hateful	ZH	63.9	.390	.499	57.4	.470	.535	64.0	.610	.640	**77.0**	**.740**	**.765**
Memotion	Humor	EN	**35.3**	.204	.276	32.5	.235	.297	35.0	.250	.310	35.2	**.248**	**.316**
MET-Meme	Intention	EN	46.4	.353	.444	38.3	.295	.363	32.0	.220	.340	**52.4**	**.442**	**.514**
MET-Meme	Intention	ZH	62.1	.443	.611	44.3	.212	.376	67.0	.470	.660	**71.0**	**.521**	**.701**
MET-Meme	Metaphor	EN	81.0	.725	.796	81.4	.724	.797	**87.0**	**.820**	**.870**	86.7	.821	.863
MET-Meme	Metaphor	ZH	84.7	.838	.846	67.8	.636	.662	**90.0**	**.890**	**.890**	86.6	.859	.865
MAMI	Misogyny	EN	62.3	.620	.620	62.8	.611	.611	75.0	.740	.740	**84.9**	**.849**	**.849**
MIMIC2024	Misogyny (Cat.)	HI-EN	47.0	.146	.412	47.0	.146	.412	66.0	.290	.430	**76.6**	**.592**	**.659**
MIMIC2024	Misogyny	HI-EN	67.3	.671	.671	63.0	.246	.570	85.0	.850	.850	**89.9**	**.899**	**.899**
Memotion	Motivational	EN	**64.7**	.399	.512	60.8	**.451**	.537	64.0	.450	.540	63.7	.450	**.545**
MAMI	Objectification	EN	67.0	.503	.590	73.2	.671	.714	81.0	.780	.810	**83.5**	**.797**	**.826**
Memotion	Offensive	EN	38.8	.140	.217	37.1	**.227**	**.334**	**39.0**	.220	.330	38.6	.215	.325
MET-Meme	Offensive	EN	74.8	.242	.667	74.2	.233	.658	74.0	**.310**	**.710**	**74.8**	.309	.708
MET-Meme	Offensive	ZH	80.3	.485	.787	74.2	.223	.684	81.0	.500	.790	**83.0**	**.535**	**.819**
RoMemes	Political	RO	67.7	.404	.547	65.6	.524	.613	83.0	.780	.820	**86.7**	**.834**	**.858**
ArMeme	Propaganda	AR	75.5	.639	.734	73.5	.554	.685	**79.0**	**.690**	**.770**	78.9	.679	.765
BanglaAbuse	Sarcasm	BN	63.9	.568	.599	65.6	.636	.651	**68.0**	.660	.670	67.4	**.661**	**.672**
Memotion	Sarcasm	EN	50.2	.167	.335	46.8	**.195**	**.352**	**51.0**	.170	.340	50.1	.167	.337
MAMI	Shaming	EN	85.4	.461	.787	83.4	.610	.819	87.0	.710	.870	**89.8**	**.719**	**.883**
MAMI	Stereotype	EN	66.1	.561	.624	72.9	.336	.707	74.0	.700	.730	**78.4**	**.739**	**.772**
HarMeme (Co)	Target	EN	77.7	.449	.788	72.9	.336	.707	**87.0**	.550	.870	82.3	.420	.840
HarMeme	Target	EN	48.5	.314	.479	45.1	.204	.404	**59.0**	.350	**.580**	56.2	**.493**	.565
Toxic	Toxic	RU	82.6	.493	.771	83.9	.495	.777	86.0	.700	.850	**86.6**	**.691**	**.853**
MAMI	Violence	EN	85.3	.504	.793	72.2	.618	.702	91.0	.770	.890	**92.3**	**.809**	**.914**
BanglaAbuse	Vulgar	BN	74.3	.680	.740	74.3	.680	.740	80.0	.750	.800	**82.7**	**.772**	**.821**
	Average		65.0	.460	.590	63.6	.472	.600	70.6	.579	.678	**74.1**	**.625**	**.720**

Table 11: Performance comparison across modalities and models on the *MemeLens* benchmark. Results are reported for text-only, image-only, multimodal sequence classification (MM-Seq), and *MemeLens* (ours). For each dataset, the best score is highlighted under the official evaluation metric, where Acc denotes accuracy, Ma denotes macro-F_1, W denotes weighted-F_1, and pos in the SOTA column denotes class-specific F_1. The SOTA column reports the best previously published result and its metric. Bold indicates the strongest performance among the compared methods under the selected metric. *MemeLens* is based on Qwen3-VL-8B-Instruct and fine-tuned with a classify-then-explain strategy. Lang. denotes the dataset language.

Dataset	MEMELENS	SOTA	Δ
Hateful_de_Multi3Hate	75.4	72.0	-3.0
Deepfake_ro_RoMemes	77.0	80.0	-3.0
Hateful_en_MMHS	61.4	68.0	+7.0
Hateful_en_Multi3Hate	74.1	76.0	-2.0
Hateful_en_FHM	79.8	79.0	-1.0
Hateful_es_Multi3Hate	80.0	69.0	+11.0
Hateful_hi_Multi3Hate	75.4	63.0	-12.0
Hateful_zh_Multi3Hate	77.0	69.0	+8.0
Political_ro_RoMemes	86.7	62.0	-25.0
Average			**-2.0**

Table 12: Dataset-level comparison on Accuracy-based benchmarks. Accuracy is reported as a percentage (%). Δ = SOTA − **MEMELENS** (percentage points).

MEMELENS achieves competitive performance relative to previously reported results, with a small average absolute difference ($\mathbf{\Delta} \approx 0.02$).

Dataset	MEMELENS	SOTA	Δ
Metaphor_en_MET	0.729	0.82	0.09
Metaphor_zh_MET	0.828	0.77	-0.06
Average			**0.02**

Table 13: Dataset-level comparison on F_1-POS benchmarks. Δ = SOTA − **MEMELENS**

H.3 Macro-F_1 Evaluation

Table 14 reports results for datasets evaluated using Macro-F_1. Under this metric, **MEMELENS** achieves performance on par with prior SOTA, suggesting that unified multitask training can match dataset-specific models even under class-imbalance-sensitive evaluation.

Dataset	MEMELENS	SOTA	Δ
Abuse_bn_Bangla	0.759	0.71	-0.05
Harmful_Covid_en	0.523	0.54	+0.02
Harmful_en_HarMeme	0.467	0.65	+0.18
Hateful_ar_Prop2Hate	0.546	0.71	+0.16
Hateful_en_MIMIC	0.707	0.70	-0.01
Intention_en_MET	0.442	0.42	-0.03
Intention_zh_MET	0.521	0.55	+0.03
Misogynous_en_MAMI	0.849	0.83	-0.02
MisogynyCat_hi_MIMIC	0.592	0.53	-0.07
Misogyny_hi_MIMIC	0.899	0.73	-0.17
Sarcasm_bn_Bangla	0.661	0.68	+0.02
Vulgar_bn_Bangla	0.772	0.72	-0.06
Hateful_bn_MUTE	0.700	0.67	-0.03

Table 14: Dataset-level comparison on Macro-F_1 benchmarks. Δ = SOTA − MEMELENS

I Prompts

Listing 17 presents the system prompt used to elicit *gold-standard* explanations for memes given an assigned label. The prompt explicitly frames the task as *justification* rather than prediction. It enforces explanation quality by requiring 4–6 sentences with grounding in at least two visual cues and one textual cue, a causal link to the label (and rubric, when available), and faithful, non-speculative language. Finally, it incorporates safety and consistency constraints, including neutral handling of sensitive content and meaning-preserving bilingual explanations when a native language is requested.

J Experimental Details

We use AdamW with cosine scheduling and a 5% warm-up. Stage I trains for 3 epochs with a learning rate of 1×10^{-4}, batch size 4 per device on 4 GPUs, and max gradient norm 1.0. We fix the seed to 42, use data-parallel training in `bfloat16`, and apply LoRA to all linear layers with r=16, α=32, and dropout 0.05, keeping the vision encoder and alignment modules frozen. We evaluate after each epoch and select the best model by validation loss. Stage II trains for 6 epochs with a learning rate of 1×10^{-5}. For the `seq_cls` baseline, we use the same setup, but train a dataset-specific head adapter for 20 epochs with learning rate 1×10^{-5}.

K Qualitative Error Analysis

To complement the aggregate results in Section 5, we present three representative cases highlighting: *(i)* how explanation-augmented training (*Stage II*) corrects errors made by classification-only training (*Stage I*), and *(ii)* why several MEMELENS tasks require genuinely multimodal reasoning beyond unimodal baselines, particularly in cases involving implicit meaning, context, and cross-modal dependencies.

```
user_prompt = f"""
TASK
- Task name: {task}
- {task_definition}

LABEL SET / RUBRIC (use this to justify the
    assigned label)
{labels}

MEME
- Assigned label: {label}
- Meme text (verbatim): "{text}"

OUTPUT LANGUAGE
{output_language_instruction}

Write a gold-standard explanation that
    justifies why this meme matches the
    assigned label using evidence from BOTH the
    image and the meme text. Return only the
    JSON object.
"""
```

Table 15: User prompt template for English only datasets, providing task context, meme text, and output language constraints for explanation generation.

```
user_prompt = f"""
TASK
- Task name: {task}
- {task_definition}

LABEL SET / RUBRIC (use this to justify the
    assigned label)
Labels in English: {labels}
Labels in Native Language: {native_labels}

MEME
- Assigned label: {label}
- Meme text (verbatim): "{text}"

OUTPUT LANGUAGE
{output_language_instruction}

Write a gold-standard explanation that
    justifies why this meme matches the
    assigned label using evidence from BOTH the
    image and the meme text. Return only the
    JSON object.
"""
```

Table 16: User prompt template for non-English datasets, providing task context, bilingual label rubric, meme text, and output language constraints for explanation generation.

```
system_prompt = """
You are an expert annotator for multimodal meme analysis. Your job is to write gold-standard
    explanations that justify a GIVEN (already assigned) label for a meme using evidence from the
    meme's image and its text.

You are NOT predicting the label. You must justify the provided label.

You will receive for each meme:
1) Task name
2) Task definition (what the task about) and explanation specifics (what to consider for this task)
3) The meme image
4) The meme text (verbatim, as provided)
5) The assigned label (and possibly a label definition/rubric)
6) The requested output language(s)

CRITICAL OUTPUT RULES
- Output MUST be a single valid JSON object and nothing else (no markdown, no extra keys, no
    commentary).
- Do NOT include line breaks inside JSON string values. Use normal spaces between sentences.
- Do not output any leading or trailing text outside the JSON object.
- If English only is requested, output exactly:
{"en_explanation":"..."}
- If English + another language is requested, output exactly:
{"en_explanation":"...","native_explanation":"..."}
- Escape any double quotes inside explanations using a backslash (\\") so the JSON remains valid.

GOLD EXPLANATION QUALITY REQUIREMENTS (apply to each requested language)
- Length: 4 to 6 sentences (prefer approximately 100 words; acceptable range 80 to 120 words).
- Evidence: explicitly reference (a) at least TWO concrete visual elements AND (b) at least ONE
    concrete textual element (quote a short phrase from the meme text when helpful).
- Reasoning: connect those visual + textual cues directly to why the assigned label fits (use the
    label definition/rubric if provided).
- Interaction: explain how image and text work together (reinforce, contrast, irony/sarcasm,
    punchline, etc.).
- Be precise and faithful: do not invent details that are not visible or not in the provided text.
    If unclear, describe generally but accurately.
- Be objective and analytical; do not endorse the meme's message.
- Sensitive content: describe neutrally. If the meme contains slurs/profanity, do not repeat them
    verbatim; replace them with placeholders like [SLUR] or [PROFANITY].
- If a label definition is provided, explicitly align at least one sentence with a key
    phrase/concept from that definition.
- Avoid generic statements like "This is funny" or "This is offensive" unless you explain exactly
    which visual/textual cue makes it so.

TRANSLATION CONSISTENCY (when requested)
- The native_explanation must preserve the same meaning and evidence as en_explanation (no
    added/removed claims).

INTERNAL SELF-CHECK (silent)
- >=2 visual cues? >=1 textual cue? 4 to 6 sentences? Clear link to label? Exact JSON keys only?
    Valid JSON?
"""
```

Table 17: System prompt for gold-standard multimodal meme explanation writing.

Explanation-augmented training corrects Stage I errors across label schemas. Figure 3 shows a COVID-related meme evaluated under two HarMeme schemas: *harm severity* (Task A) and *target* (Task B). The Stage I model predicts `partially-harmful` and `individual`, respectively—errors also made by the text-only (BERT), image-only (ViT), and zero-shot MLLM baselines. After Stage II training, the model correctly outputs `not-harmful` and `none`. The generated rationale attributes this correction to a metaphorical reading ("deport COVID-19") and correctly identifies the target as the virus rather than a human entity, providing an interpretable and contextual justification grounded in multimodal evidence and task-specific annotation criteria.

Unimodal baselines under-detect harm in highly harmful cases. Figure 4 presents a meme combining unverifiable case-count claims with the accusation "*hugest mass murderer ever*." Both unimodal baselines concentrate probability mass on `partially-harmful` (BERT: 83.6%, ViT: 60.0%), failing to capture the intensified and highly accusatory framing conveyed by the joint visual and textual context. In contrast, all multimodal systems we epxerimented with (Stage I, Stage II, and zero-shot MLLM) correctly predict `very-harmful`. The Stage II explanation explicitly links this escalation to the co-occurrence of a political figure and stigmatizing claims, illustrating a failure mode that is not recoverable from either modality alone. More broadly, this example highlights how unimodal models underestimate harm when severity emerges from cross-modal reinforcement rather than isolated cues.

Unimodal baselines over-detect harm in benign cases. Figure 5 shows the complementary failure mode. A pop-culture character appears alongside advocacy text promoting whistle-blower protection and government accountability. Unimodal baselines again default to `partially-harmful` (BERT: 76.6%, ViT: 60.7%), over-relying on surface lexical cues such as "*government accountable*." All multimodal variants correctly predict `not-harmful`, with Stage II explanations grounding the decision in the supportive tone, broader contextual framing, and absence of a targeted individual or group. This demonstrates how unimodal systems are prone to false positives when salient keywords appear without sufficient contextual grounding.

Takeaway. Overall, we saw that across many cases, Stage I models struggle with metaphor, satire, and implicit references that require integrating information across modalities. We further saw that Stage II training not only improves label accuracy, but also yields more faithful, interpretable rationales grounded in multimodal evidence. The consistent unimodal failures further demonstrate that harm detection in MEMELENS is inherently multimodal and cannot be decomposed into independent text and image predictions. Instead, effective modeling requires capturing cross-modal dependencies and contextual nuance, underscoring the importance of unified multimodal learning for robust meme understanding. This observation can inform future research.

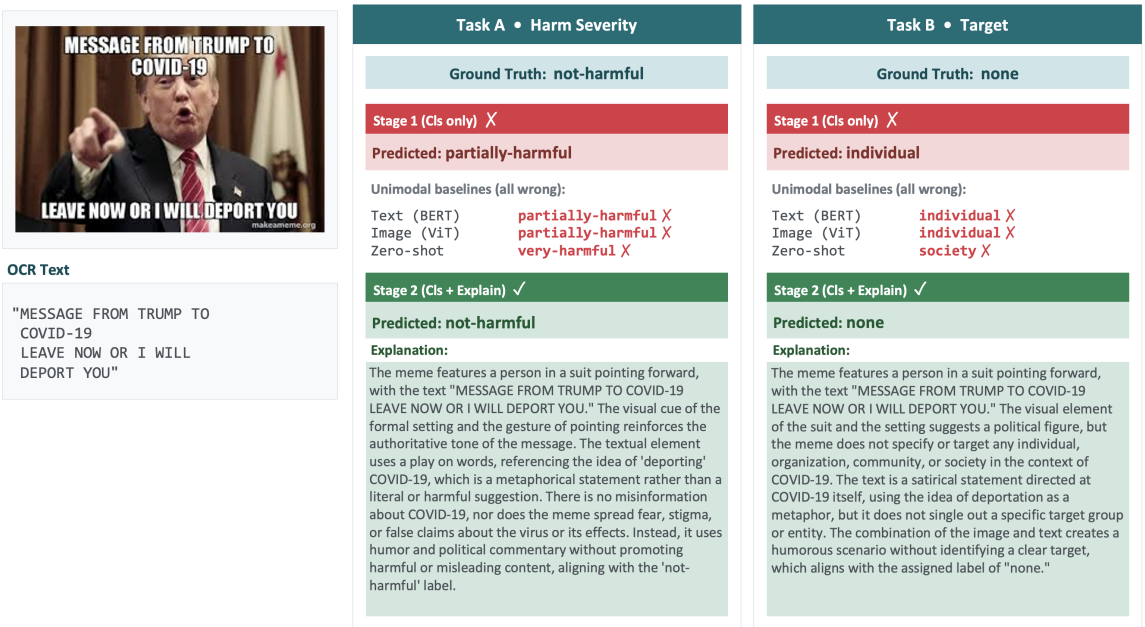

Figure 3: **Explanation corrects classification errors across schemas.** A single HarMeme instance evaluated under harm severity and target labeling. Stage I (classification-only) fails on both tasks, while Stage II (classification + explanation) recovers correct predictions and provides a grounded rationale.

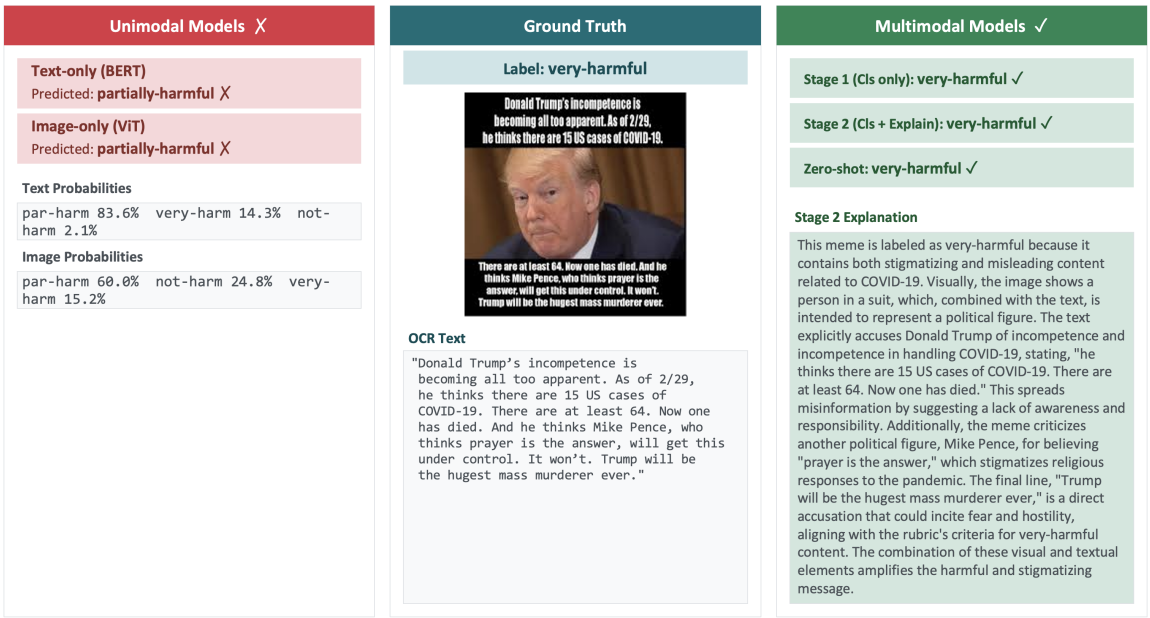

Figure 4: **Unimodal under-detection of harm.** A highly harmful meme misclassified as `partially-harmful` by unimodal baselines, but correctly identified as `very-harmful` by multimodal models. The Stage II explanation attributes the decision to combined visual and textual cues.

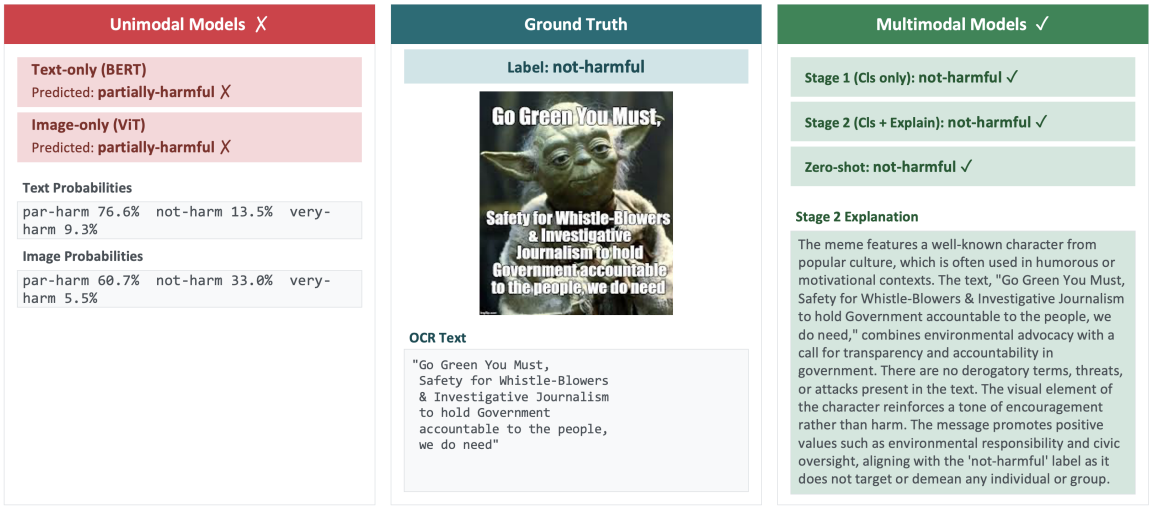

Figure 5: **Unimodal over-detection of harm.** A benign meme misclassified as harmful by unimodal baselines due to lexical bias, while multimodal models correctly predict `not-harmful`. The *Stage II* rationale highlights supportive intent and lack of a target.