MemeCULT-1K: Benchmarking South Asian Cultural Context and Humor Understanding of Multimodal Models

Tawsif Tashwar Dipto[1*] **Mehedi Ahamed**[2*] **Radib Bin Kabir**[1*]
Mueeze Al Mushabbir[1] **Mohammed Saidul Islam**[3] **Mir Rayat Imtiaz Hossain**[4]
Md Tahmid Rahman Laskar[5†] **Sabbir Ahmed**[1,6†]

[1]Islamic University of Technology [2]South East University [3]Vector Institute
[4]University of British Columbia [5]York University [6]Queen's University

{tawsiftashwar, radib}@iut-dhaka.edu mehedi.ahamed@seu.edu.bd

Abstract

Meme understanding goes beyond recognizing visual content or literal text; it requires implicit cultural knowledge and pragmatic inference that most vision-language models still lack. We introduce **MemeCULT-1K**, a multilingual benchmark of 1,000 South Asian memes in Bengali, English, and Hindi, where each meme is paired with a cultural context note and three human-written explanations, along with a supplementary set of 54 Bengali regional dialect memes. We evaluate thirteen popular Vision Language Models (VLMs) under two settings: *meme-only* and *context-aware*. Providing minimal cultural context yields consistent gains across all models and languages: mean SBERT similarity improves from 44.6 to 56.4 (+11.8), BLEURT from 37.3 to 42.3 (+5.0), and LLM-as-a-Judge scores from 2.57 to 3.43 out of 5 (+0.86). Fine-grained error analysis reveals that closed-source models fail mainly on entity and reference misidentification, while open-source models are bottlenecked by broader cultural knowledge gaps, with linguistic and phonological failures proving the most context-resistant across both. These results highlight the difficulty of culturally grounded meme understanding and motivate future work on explicit cultural knowledge integration. Our dataset and code are publicly available at TawsifDipto17/MemeCULT-1K.

1 Introduction

Memes are multimodal forms of online humor whose meaning often depends on cultural knowledge, social context, and pragmatic inference rather than on image or text recognition alone (Nguyen and Ng, 2024). A model may correctly parse both the caption and visual content, yet still fail to understand the humor when the underlying cultural reference, person, event, idiom, or social situation is not recognized (Ananthram et al., 2025). This makes meme understanding a challenging test of culturally grounded vision-language reasoning.

Existing meme benchmarks largely focus on classification tasks such as sentiment, offensiveness, or hatefulness (Sharma et al., 2020; Kiela et al., 2020), while explanation-centric resources such as MemeCap (Hwang and Shwartz, 2023) and MEMEX (Sharma et al., 2023) remain English-dominant and Western-centric (Nguyen and Ng, 2024). South Asian memes provide a particularly challenging setting because they frequently combine Bengali, Hindi, English, and code-mixed text with references to cricket, cinema and region-specific humor conventions. Since these references are often hard to infer from the image and caption alone, in this paper, we propose a context-aware evaluation setting that tests whether model failures stem from missing background knowledge.

To support this investigation, we introduce MEMECULT-1K, a benchmark of over 1,000 South Asian memes in Bengali, English, and Hindi. Each meme is paired with a short cultural context note and three independent human-written English explanations. We evaluate thirteen closed and open-source VLMs under two settings: *meme-only*, where the model sees only the meme, and *context-aware*, where it also receives the cultural context note. We additionally include 54 Bengali regional dialect memes to probe model robustness to non-standard spellings, dialectal vocabulary, and localized humor conventions. Across automatic metrics, LLM-as-a-judge scoring, and human evaluation, context consistently improves explanation quality. Our error analysis identifies weaknesses in both open-source and closed-source models for cultural contextual meme understanding.

Overall, our contributions are threefold: (i) MEMECULT-1K, a multilingual benchmark for South Asian meme explanation; (ii) a context-aware evaluation setting for measuring the cultural background in VLM reasoning; and (iii) a fine-

*Equal contribution. †Equal supervision.

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 12634–12643
October 24-29, 2026 ©2026 Association for Computational Linguistics

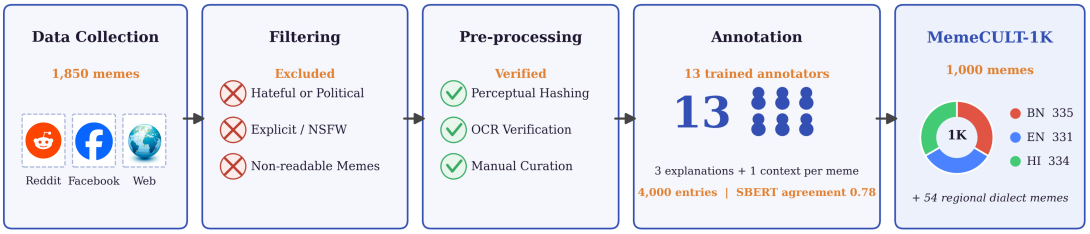

Figure 1: Dataset pipeline: sourcing, filtering, cleaning, and annotation with validation to produce the MEMECULT-1K dataset.

grained analysis of persistent VLM failures in culturally grounded meme understanding.

2 Related Work

Existing meme benchmarks mainly frame meme understanding as a *classification* task, covering sentiment, humor labels, hatefulness, misogyny, and language-specific abusive content (Sharma et al., 2020; Kiela et al., 2020; Gasparini et al., 2022; Das and Mukherjee, 2023). While useful, these classification labels do not directly assess whether a system can recover the *implicit premise* that makes a meme meaningful and funny.

Explanation-centric meme evaluation better captures interpretation, with datasets such as MemeCap (Hwang and Shwartz, 2023) and MEMEX (Sharma et al., 2023) pairing English memes with captions or explanatory evidence, but such resources remain English- and Western-centric (Nguyen and Ng, 2024). More broadly, although VLMs perform strongly on standard multimodal benchmarks (Li et al., 2022; Radford et al., 2021; Li et al., 2023), prior work shows that they degrade under real-world ambiguity (Lee et al., 2025) and in culturally specific settings, often missing culturally grounded intent despite fluent descriptions (Ananthram et al., 2025). These gaps motivate our focus on culturally grounded South Asian meme explanation with explicit context-aware evaluation.

3 Dataset

We introduce MEMECULT-1K, a multilingual dataset of 1,000 South Asian memes. The dataset contains **Bengali (335), English (331), and Hindi (334)** memes, with frequent code-mixing and culturally specific references. We additionally collected 54 Bengali regional dialect memes as a supplementary evaluation set. Each meme is paired with three independent English explanations and one short cultural context note. The dataset covers diverse domains, including sports, education, pop culture, irony, and social commentary. Figure 1 summarizes our data creation pipeline.

Data Collection: 1,850 memes were collected through manual curation and web scraping from publicly accessible sources, such as Reddit, Facebook, and regional meme pages. We excluded explicit, hateful, and political content.

Pre-processing: From 1,850 collected memes, we removed duplicates and low-quality images using perceptual hashing and manual review. OCR was used to verify language accuracy, and illegible memes were discarded. Three curators further checked cultural relevance, including code-mixed Banglish and Hinglish examples (a mix of English and Bengali). The final dataset contains 1,000 balanced memes.

Annotation: Thirteen trained annotators, fluent in English and at least in one target language, wrote three English explanations and one cultural context note for each meme, yielding 4,000 text entries. To ensure quality, annotations were validated through majority agreement, with curators stepping in to resolve any disagreements. All contributors were expected to write in a neutral tone, with proper grammar, and with cultural sensitivity in mind. The mean pairwise SBERT similarity (Reimers and Gurevych, 2019) among explanations was 0.78, indicating strong semantic agreement.

Difficulty Stratification: To characterize the sources of difficulty across languages, we annotate all 1,000 memes with a dominant difficulty type, four non-exclusive properties (wordplay, code-mixing, cultural prior knowledge, and entity recognition), and reference scope. As shown in Table 2, situational or relatable context is the most common difficulty overall (52.4%), particularly in Bengali (71.6%) and Hindi (58.7%), while English exhibits a more balanced distribution across situational, cultural/social, and linguistic difficulties. Cultural prior knowledge is required for 33.9% of memes and entity recognition for 17.1%, with substantial variation across languages; references are

Model	SBERT Similarity						Avg. Δ	BERTScore F1						Avg. Δ	BLEURT						Avg. Δ
	Meme-only			Context-aware				Meme-only			Context-aware				Meme-only			Context-aware			
GPT-5 Nano	48.1	36.8	49.3	54.8	55.3	52.7	+9.53	85.6	84.3	85.1	86.2	85.7	85.5	+0.80	39.1	33.1	38.7	41.2	40.7	40.6	+3.87
GPT-5 Mini	51.9	49.7	51.3	56.0	58.8	54.4	+5.43	85.7	85.1	85.3	86.1	85.9	85.5	+0.47	40.2	39.1	39.7	41.3	42.3	41.1	+1.90
Gemini 2.5 Flash Lite	57.9	51.8	56.7	59.6	59.1	58.7	+3.67	86.5	85.5	86.0	86.9	86.5	86.3	+0.57	41.0	38.7	41.5	42.9	41.1	42.9	+1.90
Gemini 2.5 Flash	55.8	54.1	53.8	58.2	56.6	54.9	+2.00	86.5	86.0	85.9	86.8	86.3	86.0	+0.23	41.8	41.1	40.9	43.2	42.6	41.9	+1.30
Gemini 2.5 Pro	57.1	56.9	54.2	59.4	58.6	54.7	+1.50	86.9	86.7	86.2	87.3	86.9	86.4	+0.27	43.7	44.0	42.1	45.0	45.2	42.3	+0.90
Gemma-3-4B	31.6	22.9	31.4	37.6	36.0	36.9	+8.20	85.6	84.9	85.4	85.8	85.4	85.5	+0.27	33.9	30.6	34.3	36.1	35.5	36.4	+3.07
Gemma-3-12B	37.0	24.8	34.4	52.9	49.5	49.9	+18.70	85.8	84.7	85.5	86.7	85.9	86.0	+0.87	34.6	28.4	33.4	40.3	38.0	39.1	+7.00
InternVL3.5-2B	49.4	25.3	44.0	66.1	62.2	62.5	+24.03	86.6	84.9	85.3	88.0	87.2	86.8	+1.73	38.5	31.0	37.2	46.7	44.3	44.4	+9.57
InternVL3.5-4B	53.7	28.6	49.4	67.7	63.4	63.3	+20.90	86.9	84.9	85.3	88.2	87.2	87.8	+2.03	40.0	31.2	38.6	46.6	44.0	44.3	+8.37
InternVL3.5-8B	52.7	30.8	51.4	61.7	58.2	57.7	+14.23	86.7	85.0	83.8	87.5	86.8	86.9	+1.90	40.2	31.1	39.7	44.4	42.3	43.0	+6.23
Qwen3VL-2B	53.6	31.7	46.8	67.7	62.8	61.6	+20.00	86.4	84.3	84.0	87.8	87.0	85.0	+1.70	39.8	32.3	37.4	46.7	43.9	44.2	+8.43
Qwen3VL-4B	47.9	28.0	45.3	56.4	55.5	53.8	+14.83	86.0	84.7	83.9	86.9	86.4	86.4	+1.70	37.9	30.3	36.8	42.3	41.4	41.3	+6.67
Qwen3VL-8B	50.2	34.1	49.5	57.0	56.8	49.7	+9.90	86.2	84.9	84.1	86.9	86.4	88.3	+2.13	39.0	32.0	38.2	42.4	41.3	40.8	+5.10

Table 1: Automatic metric results on MEMECULT-1K. Each entry reports English | Bengali | Hindi. Avg. Δ denotes the average change in performance after providing context across all three languages. Higher is better.

Meme property	En	Bn	Hi	All
Dominant difficulty type				
Situational / Relatable	26.6	71.6	58.7	52.4
Entity / Reference	19.3	5.4	19.8	14.8
Cultural / Social knowledge	26.9	13.4	16.5	18.9
Linguistic / Code-mixing	20.5	2.4	0.9	7.9
Wordplay / Pun	6.6	7.2	4.2	6.0
Property present (% Yes)				
Pun / wordplay	7.3	8.4	4.2	6.6
Code-mixing	39.6	5.4	2.7	15.8
Cultural prior needed	41.1	31.3	29.3	33.9
Entity recognition needed	21.5	9.0	21.0	17.1
Reference scope				
International	35.0	44.5	63.5	47.7
National	62.8	54.3	35.9	51.0
Regional	2.1	1.2	0.6	1.3

Table 2: Difficulty-stratification composition of the full dataset across English (En), Bengali (Bn), and Hindi (Hi). Dominant difficulty type and reference scope each sum to approximately 100% per column due to rounding. The four "property present" rows are independent Yes-rates and therefore do not sum.

predominantly national (51.0%) or international (47.7%). Overall, these results highlight that multilingual meme understanding depends on diverse forms of contextual knowledge beyond visual and textual content, whose relative importance varies considerably across languages.

4 Experimental Setting

With real-world deployment in mind, we evaluate a diverse set of VLMs spanning both closed and open-source families. On the closed-source side, we include several models from the GPT and Gemini Flash lineages, while the open-source selection covers a range of parameter scales across multiple model families. We also include Gemini 2.5 Pro as a reference point for its strong multilingual performance (Gemini Team, Google, 2025).

Evaluation Protocol: Each model was evaluated on Bengali, English, and Hindi subsets under two conditions: a **meme-only** setting, where the model received only the image, and a **context-aware** setting, where a brief cultural context note was additionally provided. As shown in Table 8 (Appendix A.3), even minimal context can shift a generic visual description toward a more culturally grounded interpretation. A fixed prompt (Appendix A.4) instructed models to generate a single explanatory sentence, with temperature set to 0 and output capped at 100 tokens. Scores were averaged across three human references per meme. Full experimental details are provided in Appendix A.5.

Automatic Metrics: Model-generated explanations were assessed against human references using three semantic metrics: **BERTScore-F1** (Zhang et al., 2020), **SBERT similarity** (Reimers and Gurevych, 2019), and **BLEURT** (Sellam et al., 2020) chosen for their ability to capture semantic and contextual alignment beyond mere surface-level overlap.

LLM-as-a-Judge and Human Evaluation: To complement automatic metrics, we employ Gemini 2.5 Pro as an LLM judge (Zheng et al., 2023) to score nine representative models on fluency, cultural correctness, and humor relevance using a 1–5 Likert scale. A parallel human evaluation (N=3 raters, 100 memes per language) was conducted using the same rubric.

5 Results and Analysis

5.1 Automatic Metric Evaluation

Table 1 shows the performance of different VLMs on our multilingual meme dataset for SBERT Similarity, BERTScore F1, and BLEURT.

Effect of Context: Cultural context consistently improves performance across all languages and metrics, though scores remain far from saturation,

Model	LLM as a Judge						Avg. Δ	Human Evaluation						Avg. Δ
	Meme-only			Context-aware				Meme-only			Context-aware			
Gemini 2.5 Flash	3.70	3.84	4.03	4.21	4.31	4.17	+0.37	3.90	3.75	4.00	4.10	4.25	4.20	+0.30
Gemini 2.5 Flash Lite	3.32	2.49	3.53	4.02	3.74	3.91	+0.78	3.50	2.83	3.67	3.91	3.95	3.98	+0.61
GPT-5 Nano	3.04	1.98	3.37	3.95	3.64	3.77	+0.99	3.45	2.40	3.67	4.05	3.78	3.82	+0.71
GPT-5 Mini	3.82	3.54	3.99	4.36	4.33	4.26	+0.53	3.60	3.25	3.40	4.20	4.00	3.80	+0.58
Gemma-3-12B	2.19	1.32	2.16	2.87	2.74	2.82	+0.92	2.15	1.63	1.80	3.00	2.63	2.25	+0.77
InternVL3.5-2B	1.91	1.02	1.61	3.04	2.15	2.64	+1.10	2.11	1.41	1.95	3.23	2.75	2.96	+1.16
InternVL3.5-8B	2.39	1.38	2.41	3.33	3.03	3.31	+1.16	2.51	1.94	2.74	3.64	3.45	3.56	+1.15
Qwen3VL-2B	1.95	1.33	1.50	3.05	2.31	2.49	+1.02	2.34	1.97	2.04	3.24	3.05	2.50	+0.81
Qwen3VL-8B	2.71	2.33	2.41	3.60	3.25	3.30	+0.90	2.91	2.91	2.36	3.96	3.78	3.65	+1.07

Table 3: Results using LLM-as-a-Judge and Human Evaluation. Each entry reports English | Bengali | Hindi. Avg. Δ denotes the average improvement after providing context across all three languages. Higher is better.

	Gemini 2.5 Pro (closed)						Gemma 12B (open)					
	English		Bengali		Hindi		English		Bengali		Hindi	
Error Type	E%	R%	E%	R%	E%	R%	E%	R%	E%	R%	E%	R%
Cultural Knowledge Gap	16.3	13.3	26.3	33.3	16.7	21.4	34.0	9.4	63.0	22.2	49.0	18.7
Entity Misidentification	41.3	26.3	23.1	52.4	39.3	12.1	12.0	27.3	7.0	17.4	8.0	6.7
Linguistic Failure	13.0	16.7	19.8	38.9	15.5	0.0	21.0	5.3	12.0	15.0	9.0	11.8
Literal Interpretation	9.8	22.2	8.8	50.0	7.1	0.0	24.0	22.7	12.0	22.0	22.0	25.0
Partial Humor Explanation	9.8	22.2	18.7	52.9	8.3	42.9	3.0	33.3	1.0	66.7	5.0	20.0
Hallucination	7.6	28.6	2.2	0.0	11.9	30.0	5.0	20.0	4.0	35.7	5.0	33.3
Visual Grounding Failure	2.2	0.0	1.1	100.0	1.2	100.0	1.0	100.0	1.0	0.0	2.0	0.0

Table 4: Error analysis of meme-humor explanations from Gemini 2.5 Pro and Gemma 12B across English, Bengali, and Hindi. **E%** shows error distribution; **R%** shows context-based recovery per category.

reflecting the inherent difficulty of meme understanding.

Language-wise Trends: English consistently achieves the highest scores across all metrics and models, reflecting stronger cultural coverage during pretraining. Bengali consistently scores lowest overall but gains the most from added context, suggesting that meme text alone carries limited semantic signal and models rely heavily on explicit background cues.

Model Comparison: Among closed-source models, Gemini 2.5 Pro and Gemini 2.5 Flash Lite achieve the strongest results. Within open-source models, InternVL3.5 and Qwen3-VL families benefit most from context, with InternVL3.5-4B and Qwen3-VL-2B reaching up to 67.7 SBERT on English, surpassing several closed-source models. Notably, smaller variants often outperform larger ones, as InternVL3.5-4B surpasses InternVL3.5-8B across most context-aware metrics, suggesting parameter scale does not reliably improve context utilization. The Gemma family consistently underperforms across all metrics and languages. Bootstrap confidence intervals and the significance test results can be found in Appendix A.7.

5.2 Judge-Based Evaluation

Table 3 reports Likert scores (1–5) from both the LLM judge and human raters. Notably, we find that the LLM-judge scores correlated strongly with humans (Spearman $\rho = 0.92$).

Context gap: Judge-based evaluation reveals larger gaps between meme-only and context-aware settings than automatic metrics indicate. Gemma-3-12B in Bengali improves from (1.32→2.74) under LLM-as-a-Judge and (1.63→2.63) under human evaluation after adding context. GPT-5 Nano in Bengali rises from (1.98→3.64) and (2.40→3.78) respectively. These gains suggest contextual grounding yields qualitative improvements only partially reflected by surface-level metrics such as BERTScore or BLEURT.

Open vs. closed source gap: Judge-based evaluation highlights the open/closed-source disparity that automatic metrics often compress. In the meme-only setting, open-source models perform substantially worse, particularly in Bengali (Gemma-3-12B: 1.32; InternVL3.5-8B: 1.38 under LLM-as-a-Judge). Both improve notably with context (1.32→2.74; 1.38→3.03), suggesting these models can leverage external context effectively but struggle to infer cultural cues independently. To verify that these findings are not an artifact of judge–model family overlap, we further re-scored all models with **Qwen3-32B** in Appendix A.6.

5.3 Error Analysis

To complement aggregate metrics, we analyze failures of the best closed model, Gemini 2.5 Pro, and the largest open model, Gemma 12B, using seven error categories (full definitions in Appendix A.8) and measuring whether added context resolves each error (Table 4). The models fail in distinct ways. Gemini is primarily dominated by *Entity Misidentification*, often understanding the meme structure but failing to identify the correct cultural entity or reference. In contrast, Gemma is dominated by *Cultural Knowledge Gap* errors, particularly for Bengali and Hindi, indicating weaker regional cultural priors in open-source models.

Model	SBERT Similarity		BERTScore F1		BLEURT	
	Meme	CA	Meme	CA	Meme	CA
Gemini 2.5 Pro	59.6	59.5	86.7	86.8	42.1	42.9
Gemini 2.5 Flash	52.8	55.9	85.7	86.3	36.0	39.1
Gemini 2.5 Flash Lite	42.4	51.6	84.8	85.5	33.6	37.0

Table 5: Regional Bengali dialect meme results. Meme and CA denote the meme-only and context-aware settings, respectively.

Across both models, *Linguistic Failure* remains among the least recoverable categories, suggesting that short context notes are insufficient for resolving phonological, cross-script, or dialect-specific wordplay. *Literal Interpretation* and *Partial Humor Explanation* also remain common, especially in Bengali memes, where humor often depends on implicit cultural assumptions. Gemini generally benefits more consistently from context, whereas Gemma shows relatively weak recovery for *Cultural Knowledge Gap* errors. Finally, *Visual Grounding Failure* remains rare ($\leq$2.2%), confirming that the primary bottleneck is cultural-linguistic reasoning rather than visual perception.

5.4 Regional Meme Analysis

Table 5 reports results on memes written in our additional Bengali regional dialectal variant subset. Compared to standard Bengali memes, these memes are substantially more challenging due to non-standard spellings, dialect-specific vocabulary, and localized humor conventions. Among the evaluated models, Gemini 2.5 Pro achieves the strongest overall performance in the meme-only setting, with only marginal gains from contextual information, suggesting robust regional-cultural priors even without explicit context. Context improves performance consistently across all models, with the largest gains in Gemini 2.5 Flash Lite.

6 Conclusion and Future Work

We presented MEMECULT-1K, a multilingual benchmark for culturally grounded meme understanding in South Asian languages. Across thirteen VLMs, minimal cultural context consistently improves explanation quality across automatic metrics, LLM judging, and human evaluation, though overall performance remains far from ceiling. Error analysis shows distinct failure modes: closed-source models often misidentify cultural references, while open-source models lack regional background knowledge. Linguistic and phonological failures remain largely resistant to contextual cues, motivating future work on culturally grounded and cross-script multimodal reasoning.

Limitations

Our dataset focuses on South Asian culture, specifically Bengali, Hindi, and English memes, and findings may not generalize to other regions or linguistic traditions. The benchmark is balanced across languages but modest in per-language scale, so results computed within a single language-metric cell or within a single error category should be read as indicative rather than definitive. Cultural humor is inherently subjective and temporally fluid; interpretations grounded in current cultural references may shift as memes and social conventions evolve. Automatic metrics such as BERTScore, SBERT, and BLEURT approximate semantic similarity but do not fully capture humor understanding, irony, or pragmatic nuance, which is why we complement them with LLM-as-a-judge and human evaluation. The LLM judge evaluation covers a representative subset of models selected to span both open and closed-source families. Our cultural context notes are concise by design, which may limit their ability to fully resolve deeply entrenched cultural or phonological wordplay, as also observed in our error analysis. Future work will expand to additional South Asian languages such as Tamil, Urdu, and Marathi, explore retrieval-augmented context grounding, and conduct larger-scale human evaluation across diverse annotator populations.

Ethics Statement

All memes were publicly available and screened to remove NSFW or offensive content. While efforts were made to minimize bias, humor interpretation remains subjective and culturally situated, and explanations may reflect annotator perspectives. Users should exercise caution in downstream applications. All participating annotators were paid well above their country-specific minimum wages.

Acknowledgements

We thank Islamic University of Technology for funding the annotation effort, CUPE 3903 Research Grant Fund at York University for providing API credits, Google for Gemini API credits, and the Digital Research Alliance of Canada for computing resources. We are also grateful to the anonymous reviewers for their constructive feedback, which substantially improved this work.

References

Amith Ananthram, Elias Stengel-Eskin, Mohit Bansal, and Kathleen McKeown. 2025. See it from my perspective: How language affects cultural bias in image understanding. *Preprint*, arXiv:2406.11665.

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.

Francesca Gasparini, Giulia Rizzi, Aurora Saibene, and Elisabetta Fersini. 2022. Benchmark dataset of memes with text transcriptions for automatic detection of multi-modal misogynistic content. *Data in Brief*, 44:108526.

Gemini Team, Google. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261. Technical report.

EunJeong Hwang and Vered Shwartz. 2023. MemeCap: A dataset for captioning and interpreting memes. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.

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 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

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 30545–30588, Suzhou, China. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Preprint*, arXiv:2201.12086.

Khoi P. N. Nguyen and Vincent Ng. 2024. Computational meme understanding: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. 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. *Preprint*, arXiv:2103.00020.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

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, Ramaneswaran S, Udit Arora, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. MEMEX: Detecting explanatory evidence for memes via knowledge-enriched contextualization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5272–5290, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

A Appendix

A.1 Annotation Guidelines

13 Annotators were hired among university students based on their urban culture and meme knowledge. They were hired in contract and were paid in full upon the completion of the annotation work. To ensure consistent and culturally sensitive meme

Filename	English_0001.jpg
Language	English
Context	The meme is about Mushfiqur Rahim and Nurul Hasan Shohan making similar cricketing mistakes.
Explanation_1	The meme contextualizes how both Mushfiqur Rahim and Nurul Hasan Shohan commit similar cricketing errors by breaking the stump before catching the ball.
Explanation_2	This meme humorously points out that both Mushfiqur Rahim and Nurul Hasan Shohan make the same mistake, breaking the stump before completing the catch.
Explanation_3	The image exaggerates the comedic similarity in their blunders, turning a cricket mishap into a light-hearted joke.
Annotator IDs	A03, A05, A09

Table 6: Representative example entry from the MEMECULT-1K dataset (CSV format). Values are illustrative.

explanations, all annotators followed a standardized set of guidelines. Each meme was annotated independently by three annotators, fluent in English and familiar with at least one cultural region represented in the dataset (Bengali, Hindi, or English).

General Instructions. Annotators were instructed to provide short, objective explanations that clearly described each meme's meaning or humor while maintaining a neutral tone. They were asked to avoid reproducing or quoting any potentially offensive text verbatim, ensure grammatical correctness, and remain culturally balanced in their phrasing. Personal commentary, subjective humor, or political interpretation was explicitly discouraged to maintain consistency and fairness across all samples.

Explanation Writing Protocol. Each meme was annotated with three distinct English explanations (`exp1-3`), written independently by different annotators. While the three versions preserved the same core interpretation, they varied in style and phrasing to capture natural linguistic diversity. Annotators were instructed to focus on accurately describing the visual–textual relationship within the meme and to include only as much cultural context as necessary for a reader unfamiliar with the source culture to grasp the intended humor or message.

Cultural Context Notes. Each meme was accompanied by a short (2-3 sentences) context note summarizing cultural or situational background. Annotators were encouraged to identify cultural markers (e.g., cricket, film tropes, or local idioms) influencing humor.

Annotation Format. Each record in MEMECULT-1K corresponds to a single meme image stored as a CSV row. Columns include the meme filename, cultural context note, three English explanations written by independent annotators, and their anonymized IDs. Table 6 shows one representative example from the dataset, formatted vertically for readability.

Inter-Annotator Agreement. Since explanations are open-ended natural language texts rather than categorical labels, inter-annotator consistency was assessed using pairwise SBERT cosine similarity between the three explanations per meme. The mean similarity score across all samples was 0.78, indicating strong semantic agreement in annotators' interpretations.

A.2 Example Annotations and Contexts

Table 7 presents representative samples from the Bengali, Hindi, and English subsets. All images are resized to fit within column width and anonymized for ethical compliance.

A.3 Example explanations

Example explanations for memes generated by VLMs are provided in Table 8.

A.4 VLM Evaluation Prompt

For consistent automated evaluation, all vision–language models were prompted with the following standardized instruction:

```
Write one clear and self-contained sentence that explains the meaning or humor expressed in the content. Focus only on the underlying idea or irony implied. Avoid using words such as "meme", "image", "picture", "joke", or "pun". Do not describe visual details or add commentary. Return only the explanatory sentence.
```

A.5 Experimental Setup

For each meme, we use a fixed instruction prompt. We send the image as a base64-encoded input, prepend the context note only in the context-aware condition, and use deterministic decoding (temperature = 0; outputs limited to roughly 100 tokens). For each item, model outputs are compared against

Context: Depicts unsold cows at the Qurbani market after the festival ends.	**Context:** Depicts the repeated interaction between a failing student and an exam guard.	**Context:** Contrasts children's reactions to their mother's versus father's return from the market.
Explanation: Shows the sorrow of cows that return unsold, using emotional irony to anthropomorphize livestock.	**Explanation:** The guard mocks the student's persistence, symbolizing repeated academic failure humorously.	**Explanation:** Children cheer when their mother arrives expecting snacks but act studious when their father returns home.

Table 7: Example memes with corresponding context notes and human explanations.

Meme Image	Explanation
	Ground Truth: This meme humorously depicts the common struggle of trying to put on a second long-sleeved shirt over a first one to combat cold weather. The sleeve of the inner shirt often gets bunched up and pulled back towards the elbow, causing frustration.
	Meme-only: It shows a cartoon person wearing a long-sleeved sweater, with the sleeve pulled up, revealing sweat stains around the armpit area. Basically, it's a relatable joke about sweating in sweaters when the season shifts.
	Context-aware: This meme humorously highlights the everyday struggle of layering clothes during cold weather. When someone tries to put on a second long-sleeved shirt or sweater over the first one, the inner sleeve often gets dragged upward and bunched around the elbow, making it both uncomfortable and frustrating.
	Ground Truth: The meme is about how an employee tries to fit in his/her office even though they have differences with the rest of the team. The image hilariously portrays the contrast with a slice of orange trying to fit into a garlic.
	Meme-only: Garlic is known for its strong, pungent flavor and smell, and its tough, papery outer skin. In this metaphor, the garlic represents the tough, often demanding, or perhaps dull external environment of the office.
	Context-aware: The meme signifies that the employee is masking their authentic self and forcing themselves to fit a rigid professional mold, showing only a tough, neutral shell to their colleagues while keeping their true nature hidden inside to avoid conflict or stand out.

Table 8: Example explanations for memes generated by VLMs under meme-only vs. context-aware settings.

three independent reference explanations, and automatic scores are computed by averaging scores over references. We run this pipeline separately for each language split and save per-meme outputs and scores for analysis.

A.6 Cross-Judge Robustness of Evaluation

Because our LLM judge (Gemini 2.5 Pro) shares a model family with several evaluated systems, one may worry that judge scores are inflated by self-preference. To test this, we re-scored all models with **Qwen3-32B**, an open-weight judge from a different vendor, using the identical prompt and rubric as Section 4. We note that the Gemini 2.5 Pro judge never scored its own outputs. Table 9 reports human evaluation alongside both LLM judges, with all scores averaged over English, Bengali, and Hindi.

Human evaluation corroborates the Gemini judge. If the judge favored its own family, human ratings should diverge from it. Instead, they agree within 0.06 points on the Gemini systems in the context-aware setting (Flash: 4.18 human vs. 4.23 judge; Flash Lite: 3.95 vs. 3.89), indicating that the high Gemini scores reflect independent human judgment rather than self-preference.

Family overlap does not inflate scores. Qwen3-32B likewise judges its own vendor's models yet shows no favoritism, scoring Qwen3-VL-8B at 3.27 versus 3.80 (human) and 3.38 (Gemini judge). Its uniformly lower scores apply broadly across the

Model	Human		Gemini 2.5 Pro		Qwen3-32B	
	M	C	M	C	M	C
Gemini 2.5 Flash	3.88	4.18	3.86	4.23	3.42	3.68
GPT-5 Mini	3.42	4.00	3.78	4.32	3.45	3.85
Gemini 2.5 Flash Lite	3.33	3.95	3.11	3.89	3.05	3.58
GPT-5 Nano	3.17	3.88	2.80	3.79	2.83	3.53
Qwen3-VL-8B	2.73	3.80	2.48	3.38	2.30	3.27
InternVL3.5-8B	2.40	3.55	2.06	3.22	2.16	3.16
InternVL3.5-2B	1.82	2.98	1.51	2.61	1.88	2.99
Qwen3-VL-2B	2.12	2.93	1.59	2.62	1.90	2.93
Gemma-3-12B	1.86	2.63	1.89	2.81	2.13	2.84

Table 9: Cross-judge robustness. Scores (1–5) averaged over English/Bengali/Hindi. **M/C** denote meme-only and context-aware settings. Human evaluation uses 100 memes per language (as in the main paper); LLM judges score the full subsets.

Gemini, GPT-5, and Qwen families alike, reflecting stricter grading rather than family-directed bias.

Conclusions are evaluator-invariant. Context gains are positive for every model under all three evaluators, and model rankings agree strongly (Spearman ρ over the nine common models: human–Qwen3-32B 0.98; human–Gemini judge 0.92; between judges 0.93). No conclusion in the paper depends on absolute judge scores; accordingly, LLM judge scores should be read as relative, not absolute, quality measures.

A.7 Statistical Significance and Confidence Intervals

To quantify variance and significance, we report 95% percentile bootstrap confidence intervals (10,000 resamples over memes) for every cell across all three metrics, and test each context gain with a paired bootstrap test under Holm–Bonferroni correction in Table 10.

A.8 Error Category Definitions

For the error analysis in Section 5.3, two annotators independently labeled each incorrect or incomplete model output into one of seven mutually exclusive categories. When multiple error types co-occurred, annotators selected the most salient one- the error type most directly responsible for the explanation failing to match the ground truth. Disagreements were resolved through discussion. Definitions and representative examples are provided below.

Cultural / Social Knowledge Gap. The model fails to recognize the broader cultural, social, or situational background required to interpret the meme. The model correctly perceives the visual and textual content but cannot connect it to the relevant cultural reference (e.g., a regional festival, a local custom, or a social trope). Example: failing to recognize that "Qurbani" refers to the Eid al-Adha sacrificial festival.

Entity / Reference Misidentification. The model identifies that a specific entity is being referenced but attributes it to the wrong person, place, event, or meme template. Distinguished from Cultural Knowledge Gap in that the model recognizes the *kind* of reference but gets the specific anchor wrong. Example: identifying a cricketer in the meme but naming the wrong player.

Linguistic Failure. The model fails to correctly parse the meme's text due to code-mixing, non-standard orthography, cross-script wordplay, or phonological/transliteration patterns. Example: misreading romanized Bengali ("Banglish") as English, or failing to recognize a Hindi pun that depends on Devanagari–Roman script switching.

Surface-Level / Literal Interpretation. The model describes only the visible content of the meme without inferring its intended humorous or ironic meaning. Output is factually correct but pragmatically empty (e.g., describing the meme as "a man holding a phone" without noting the irony).

Incomplete Humor Explanation. The model partially captures the humor but misses a critical component, such as the punchline, the cultural twist, or the subject of the joke. Distinguished from Surface-Level in that some humorous intent is recognized but the explanation is incomplete.

Hallucination / Fabricated Content. The model introduces information not present in the meme or its context, such as inventing names, events, or backstories that do not exist in the source. Includes confidently asserting incorrect cultural facts.

Multimodal Grounding Failure. The model's explanation contradicts or ignores clear visual evidence in the meme (e.g., describing a person as smiling when they are crying, or referring to objects not present in the image). Distinguished from Hallucination in that the failure is specifically a vision–text misalignment rather than fabricated world knowledge.

A.9 Ethical Considerations

All memes were sourced from publicly available social media platforms (Reddit, Facebook, and regional meme pages) for non-commercial research purposes. NSFW or sensitive material was filtered

		SBERT			BERTScore F1			BLEURT		
Model	**Lang**	Meme-only	Context	Δ	Meme-only	Context	Δ	Meme-only	Context	Δ
GPT-5 Mini	EN	51.9 [50.7, 53.2]	56.0 [55.0, 57.0]	+4.1	85.7 [85.5, 85.9]	86.1 [85.9, 86.3]	+0.4	40.2 [39.6, 40.9]	41.3 [40.7, 41.9]	+1.1
	BN	49.7 [48.2, 51.1]	58.8 [57.7, 59.8]	+9.1	85.1 [84.9, 85.3]	85.9 [85.7, 86.1]	+0.8	39.1 [38.4, 39.8]	42.3 [41.7, 42.9]	+3.2
	HI	51.3 [50.0, 52.6]	54.4 [53.2, 55.6]	+3.1	85.3 [85.1, 85.5]	85.5 [85.3, 85.7]	+0.2	39.7 [39.0, 40.3]	41.1 [40.5, 41.7]	+1.4
GPT-5 Nano	EN	48.1 [46.7, 49.4]	54.8 [53.6, 55.9]	+6.7	85.6 [85.4, 85.8]	86.2 [86.0, 86.4]	+0.6	39.1 [38.5, 39.7]	41.2 [40.5, 41.8]	+2.1
	BN	36.8 [35.3, 38.4]	55.3 [54.1, 56.5]	+18.5	84.3 [84.1, 84.5]	85.7 [85.5, 85.9]	+1.4	33.1 [32.5, 33.7]	40.7 [40.1, 41.3]	+7.6
	HI	49.3 [48.0, 50.6]	52.7 [51.5, 53.9]	+3.4	85.1 [84.9, 85.3]	85.5 [85.3, 85.7]	+0.4 *(n.s.)*	38.7 [38.1, 39.3]	40.6 [40.0, 41.2]	+1.9
Gemini 2.5 Pro	EN	57.1 [56.0, 58.2]	59.4 [58.3, 60.5]	+2.3	86.9 [86.7, 87.1]	87.3 [87.1, 87.5]	+0.4	43.7 [43.0, 44.4]	45.0 [44.4, 45.6]	+1.3
	BN	56.9 [55.8, 58.0]	58.6 [57.5, 59.7]	+1.7	86.7 [86.5, 86.9]	86.9 [86.7, 87.1]	+0.2	44.0 [43.3, 44.7]	45.2 [44.5, 45.9]	+1.2
	HI	54.2 [52.8, 55.6]	54.7 [53.5, 55.9]	+0.5 *(n.s.)*	86.2 [86.0, 86.4]	86.4 [86.2, 86.6]	+0.2 *(n.s.)*	42.1 [41.5, 42.7]	42.3 [41.6, 43.0]	+0.2 *(n.s.)*
Gemini 2.5 Flash	EN	55.8 [54.5, 57.1]	58.2 [57.0, 59.4]	+2.4 *(n.s.)*	86.5 [86.3, 86.7]	86.8 [86.6, 87.0]	+0.3	41.8 [41.1, 42.5]	43.2 [42.5, 43.9]	+1.4
	BN	54.1 [52.9, 55.3]	56.6 [55.5, 57.7]	+2.5	86.0 [85.8, 86.2]	86.3 [86.1, 86.5]	+0.3	41.1 [40.4, 41.8]	42.6 [41.9, 43.2]	+1.5
	HI	53.8 [52.5, 55.1]	54.9 [53.6, 56.1]	+1.1 *(n.s.)*	85.9 [85.7, 86.1]	86.0 [85.8, 86.2]	+0.1 *(n.s.)*	40.9 [40.3, 41.5]	41.9 [41.3, 42.5]	+1.0
Gemini 2.5 Flash-Lite	EN	57.9 [56.7, 59.1]	59.6 [58.4, 60.8]	+1.7	86.5 [86.3, 86.7]	86.9 [86.7, 87.1]	+0.4	41.0 [40.3, 41.7]	42.9 [42.3, 43.6]	+1.9
	BN	51.8 [50.5, 53.1]	59.1 [58.0, 60.1]	+7.3	85.5 [85.3, 85.7]	86.5 [86.3, 86.7]	+1.0	38.7 [38.0, 39.3]	41.1 [40.5, 41.7]	+2.4
	HI	56.7 [55.4, 58.0]	58.7 [57.6, 59.8]	+2.0	86.0 [85.8, 86.2]	86.3 [86.1, 86.5]	+0.3 *(n.s.)*	41.5 [40.9, 42.1]	42.9 [42.3, 43.5]	+1.4
Gemma 3 4B	EN	31.6 [30.4, 32.8]	37.6 [36.5, 38.7]	+6.0	85.6 [85.5, 85.7]	85.8 [85.7, 85.9]	+0.2	33.9 [33.3, 34.5]	36.1 [35.5, 36.7]	+2.2
	BN	22.9 [21.9, 23.9]	36.0 [34.9, 37.1]	+13.1	84.9 [84.8, 85.0]	85.4 [85.3, 85.5]	+0.5	30.6 [30.1, 31.1]	35.5 [34.9, 36.1]	+4.9
	HI	31.4 [30.3, 32.5]	36.9 [35.8, 38.0]	+5.5	85.4 [85.3, 85.5]	85.5 [85.3, 85.7]	+0.1 *(n.s.)*	34.3 [33.7, 34.9]	36.4 [35.8, 37.0]	+2.1
Gemma 3 12B	EN	37.0 [35.9, 38.2]	52.9 [51.8, 54.1]	+15.9	85.8 [85.6, 86.0]	86.7 [86.5, 86.9]	+0.9	34.6 [34.0, 35.2]	40.3 [39.7, 40.9]	+5.7
	BN	24.8 [23.8, 25.8]	49.5 [48.3, 50.7]	+24.7	84.7 [84.5, 84.9]	85.9 [85.7, 86.1]	+1.2	28.4 [27.9, 28.9]	38.0 [37.4, 38.6]	+9.6
	HI	34.4 [33.3, 35.5]	49.9 [48.8, 51.1]	+15.5	85.5 [85.3, 85.7]	86.0 [85.8, 86.2]	+0.5	33.4 [32.8, 34.0]	39.1 [38.5, 39.7]	+5.7
InternVL 2B	EN	49.4 [47.9, 50.9]	66.1 [64.8, 67.4]	+16.7	86.6 [86.4, 86.8]	88.0 [87.8, 88.2]	+1.4	38.5 [37.8, 39.2]	46.7 [45.9, 47.5]	+8.2
	BN	25.3 [23.9, 26.7]	62.2 [61.0, 63.4]	+36.9	84.9 [84.7, 85.1]	87.2 [87.0, 87.4]	+2.3	31.0 [30.4, 31.6]	44.3 [43.5, 45.1]	+13.3
	HI	44.0 [42.5, 45.5]	62.5 [61.2, 63.8]	+18.5	85.3 [85.1, 85.5]	86.8 [86.6, 87.0]	+1.5	37.2 [36.5, 37.9]	44.4 [43.7, 45.1]	+7.2
InternVL 4B	EN	53.7 [52.3, 55.1]	67.7 [66.6, 68.8]	+14.0	86.9 [86.7, 87.1]	88.2 [88.0, 88.4]	+1.3	40.0 [39.3, 40.7]	46.6 [45.8, 47.4]	+6.6
	BN	28.6 [27.1, 30.1]	63.4 [62.2, 64.6]	+34.8	84.9 [84.8, 85.0]	87.2 [87.0, 87.4]	+2.3	31.2 [30.6, 31.8]	44.0 [43.3, 44.7]	+12.8
	HI	49.4 [48.0, 50.8]	63.3 [62.1, 64.5]	+13.9	85.3 [85.1, 85.5]	87.8 [87.5, 88.1]	+2.5	38.6 [37.9, 39.3]	44.3 [43.6, 45.0]	+5.7
InternVL 8B	EN	52.7 [51.4, 54.0]	61.7 [60.6, 62.9]	+9.0	86.7 [86.5, 86.8]	87.5 [87.4, 87.7]	+0.8	40.2 [39.5, 40.9]	44.4 [43.7, 45.1]	+4.2
	BN	30.8 [29.6, 32.1]	58.2 [57.0, 59.4]	+27.4	85.0 [84.9, 85.1]	86.8 [86.6, 86.9]	+1.8	31.1 [30.6, 31.7]	42.3 [41.6, 43.0]	+11.2
	HI	51.4 [50.0, 52.8]	57.7 [56.4, 58.9]	+6.3	83.8 [83.6, 84.0]	86.9 [86.7, 87.1]	+3.1	39.7 [39.0, 40.3]	43.0 [42.3, 43.6]	+3.3
Qwen3-VL 2B	EN	53.6 [52.3, 54.9]	67.7 [66.6, 68.8]	+14.1	86.4 [86.3, 86.5]	87.8 [87.7, 87.9]	+1.4	39.8 [39.2, 40.4]	46.7 [46.0, 47.4]	+6.9
	BN	31.7 [30.3, 33.1]	62.8 [61.7, 63.9]	+31.1	84.3 [84.2, 84.4]	87.0 [86.9, 87.1]	+2.7	32.3 [31.7, 32.9]	43.9 [43.2, 44.6]	+11.6
	HI	46.8 [45.3, 48.3]	61.6 [60.3, 62.9]	+14.8	84.0 [83.8, 84.2]	85.0 [84.8, 85.2]	+1.0	37.4 [36.8, 38.0]	44.2 [43.6, 44.8]	+6.8
Qwen3-VL 4B	EN	47.9 [46.5, 49.3]	56.4 [55.2, 57.6]	+8.5	86.0 [85.8, 86.2]	86.9 [86.7, 87.1]	+0.9	37.9 [37.3, 38.5]	42.3 [41.6, 43.0]	+4.4
	BN	28.0 [26.7, 29.3]	55.5 [54.2, 56.8]	+27.5	84.7 [84.6, 84.8]	86.4 [86.2, 86.6]	+1.7	30.3 [29.7, 30.9]	41.4 [40.7, 42.1]	+11.1
	HI	45.3 [43.9, 46.7]	53.8 [52.4, 55.2]	+8.5	83.9 [83.8, 84.0]	86.4 [86.2, 86.6]	+2.5	36.8 [36.1, 37.5]	41.3 [40.7, 41.9]	+4.5
Qwen3-VL 8B	EN	50.2 [48.7, 51.6]	57.0 [55.8, 58.3]	+6.8	86.2 [86.0, 86.4]	86.9 [86.7, 87.1]	+0.7	39.0 [38.3, 39.7]	42.4 [41.7, 43.1]	+3.4
	BN	34.1 [32.7, 35.5]	56.8 [55.5, 58.0]	+22.7	84.9 [84.7, 85.1]	86.4 [86.2, 86.6]	+1.5	32.0 [31.4, 32.6]	41.3 [40.7, 42.0]	+9.3
	HI	49.5 [48.2, 50.8]	49.7 [48.4, 51.0]	+0.2	84.1 [83.9, 84.3]	88.3 [88.1, 88.5]	+4.2 *(n.s.)*	38.2 [37.5, 38.9]	40.8 [40.2, 41.5]	+2.6

Table 10: Per-language results for all 13 models under meme-only and context-aware settings, on SBERT similarity, BERTScore F1, and BLEURT (all $\times 100$). Each cell reports the mean with its 95% percentile-bootstrap confidence interval (10,000 resamples over memes). Δ is the mean paired per-meme context gain (context-aware $-$ meme-only); gains are significant at the 95% level under a paired bootstrap test with Holm–Bonferroni correction within each metric family unless marked *(n.s.)*.

in multiple stages during dataset creation and annotation. Hateful and potentially harmful memes were also removed. The dataset will be released for academic and non-commercial research use only; original meme creators retain copyright over the underlying images.