From Pixels to Punchlines: Investigating Figurative Meaning in Memes with VLMs

Anonymous ACL submission

Abstract

Figurative language is central to humorous and persuasive communication. Internet memes, as a popular form of multimodal online communication, often use figurative elements to convey layered meaning through the combination of text and images, but little is known on what elements vision-language models (VLMs) utilize to detect non-literal meaning in memes. To address this gap, we evaluate nine state-of-the-art generative VLMs on their ability to detect and classify six types of non-literal meaning in memes. Our results show that VLMs outperform a majority-vote baseline, and, importantly, their accuracy improves as the figurative complexity of memes increases. Model performance across figurative categories varies by modality: identifying irony relies on text, while anthropomorphism on image. Although VLMs demonstrate competitive performance on single-modality inputs, they fail to fully integrate multimodal content. We thus highlight both the capabilities and limitations of today's VLMs in figurative meme understanding.

1 Introduction

Figurative language uses words in a non-literal way to create more vivid, symbolic, or abstract meanings (Lakoff and Johnson, 2008). By enabling indirect communication and intensifying rhetorical effects, figurative expressions are a key element to persuasive and humorous discourse, and have been shown to engage the audience emotionally (Fussell and Moss, 2014; Burgers et al., 2016).

We believe generative AI needs to handle figurative language well to communicate and interact appropriately with humans. While in NLP figurative language is often studied based on the text alone (Chakrabarty et al., 2022a; Stowe et al., 2022; Lai et al., 2023), it is clearly a multimodal phenomenon (Chakrabarty et al., 2022b; Akula et al., 2023; Kulkarni et al., 2024). For example, Figure 1 shows a meme, pairing a mischievous expression together with gestures, both of which highlight the sarcastic tone to emphasize and critique selective migration preferences expressed in text.

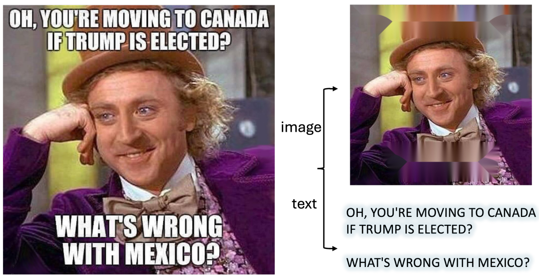
OH, YOU'RE MOVING TO CANADA IF TRUMP IS ELECTED?
WHAT'S WRONG WITH MEXICO?
image
text
OH, YOU'RE MOVING TO CANADA IF TRUMP IS ELECTED?
WHAT'S WRONG WITH MEXICO?

Figure 1: In our work we consider a meme as being composed of an image and text that together convey a message, but here we also decompose them to study their respective effects on figurativeness. This meme from FIGMEMES (Liu et al., 2022a) is human-annotated with the label *Irony/Sarcasm*.

Tasks like automatically detecting hateful memes (Zhang et al., 2024a; Pramanick et al., 2021; Hossain et al., 2024) or detecting humor, sarcasm, metaphors (Sharma et al., 2020; Tanaka et al., 2022; Nandy et al., 2024) are challenging for vision-language models (VLMs) because different modalities contribute to convey a meme's message, and processing non-literal meanings requires integrating multimodal information. Furthermore, it is less known to what degree different elements of multimodality contribute to meme interpretation, especially in today's VLMs. Therefore, evaluating how well models handle figurative meaning[1] is essential to understanding the limitations of VLMs in multimodal figurative comprehension.

To the best of our knowledge, no study has yet compared how generative VLMs perform across different figurative categories in memes. To this end, we pose the following research questions:

- **RQ1**: To what extent can VLMs *detect* the

[1] In this paper, we use the term *figurative meaning* instead of the commonly used *figurative language* to emphasize the multimodal nature of our analyses.

presence of figurative meaning in memes?

- **RQ2**: How effectively can VLMs *classify* different types of figurative meaning in memes?

We conduct controlled ablations across input types to gain a complete understanding to how each modality—text and image in isolation or combined—contributes to the construction and interpretation of figurative meaning in memes. Our results show that while VLMs outperform the baseline in general figurative meaning detection, they struggle with multimodal integration at granular figurative category classification tasks.

2 Related Work

Figurative Language in NLP Research on figurative language typically follows two main approaches. One approach studies individual rhetorical devices, such as idioms, irony, metaphors, and other expressions that go beyond literal meaning (Tay et al., 2018; Chakrabarty et al., 2021; Tong et al., 2021). Another line of work treats all non-literal expressions as a unified category, studying them collectively (Ghosh et al., 2015; Do Dinh et al., 2018; Chen et al., 2024; Lee et al., 2024).

As dialogue systems develop, research has shifted toward investigating models' ability non-literal expressions in communication (Jhamtani et al., 2021; Stowe et al., 2022; Liu et al., 2022b; Jang et al., 2023; Yerukola et al., 2024). Incorporating vision encoders into LLMs extends this focus to interpreting non-literal meaning across modalities (Hessel et al., 2023; Zhang et al., 2024b). Memes, rich in multimodal figurative content, have thus become an important object of study (Hwang and Shwartz, 2023; Nandy et al., 2024).

Multimodal Meme Understanding Internet memes are a distinctive form of multimodal communication, where visuals shape the interpretation of text through context, priming, or template-based expectations (Shifman, 2013; Nissenbaum and Shifman, 2017; Wiggins, 2019). NLP and vision-language communities have increasingly studied memes, with studies showing how images influence text via templates (Zhou et al., 2024; Bates et al., 2025). Various tasks have been explored, including sentiment analysis (Hossain et al., 2022), hateful or harmful meme detection (Gomez et al., 2020; Kiela et al., 2020; Cao et al., 2022), and emotion classification (Sharma et al., 2020).

Although memes feature a wide range of figurative categories, few studies examine how well models handle them. Liu et al. (2022a) laid groundwork with a new dataset of politically-opinionated memes (FIGMEMES), enabling cross-modal comparisons across discriminative models.

Total memes		1,542
Memes with OCR text		1,396
Category	Allusion	229
	Exaggeration	240
	Irony	315
	Anthropomorphism	113
	Metaphor	244
	Contrast	163
#labels	0 label	445
	1 label	651
	2 labels	250
	3 labels	47
	4 labels	3

Table 1: Statistics of FIGMEMES test split, label distribution and multi-label composition. #label = number of labels in the meme.

3 Experimental Setup

3.1 Dataset

We use FIGMEMES (Liu et al., 2022a). Each meme is annotated for the presence of any of six distinct figurative categories: *allusion*, *exaggeration*, *irony*, *anthropomorphism*, *metaphor*, and *contrast* (see Appendix A for the definition of each category). The dataset also includes text extracted via OCR.

Our evaluation is conducted on the test split defined by Liu et al. (2022a). To ensure comparability across modality conditions, we restrict evaluation to the 1,396 memes with OCR-extracted text. Table 1 summarizes dataset statistics.

3.2 Image Processing

In previous work (including Liu et al. (2022a)), the whole meme – with its embedded text– is usually treated as visual input, without removing textual information that VLMs can easily extract. To investigate the contribution of different modalities comprehensively, we include an *image*-only input condition. We use OCR to detect text and mask them to exclude textual information, as shown in Figure 1. (see Appendix B.1).

3.3 Task Setup

We design two tasks to assess detection and classification of figurative meaning in memes. The first is a **binary** classification task, aimed at determining whether any non-literal meaning is generally present in a meme. The second is a more granular

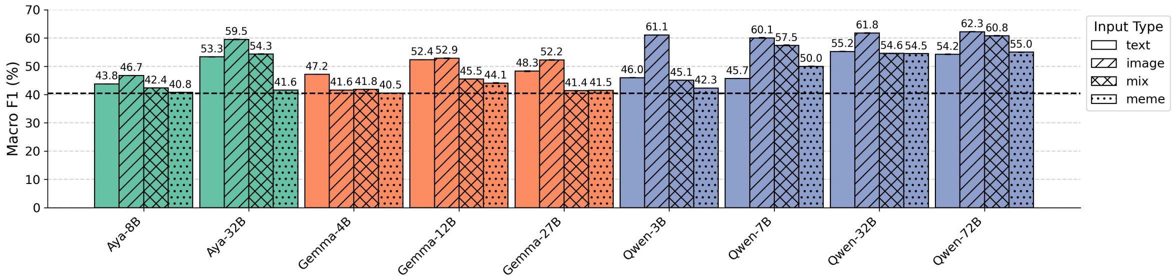

Figure 2: Macro F1 scores (%) for binary classification across input types. Each group corresponds to an input type (meme, mix, image, text), and within each group, different bars represent models grouped by family type.

multiclass classification task, aimed at predicting the presence of each specific figurative category. See Appendix B.2 for prompt templates.

In addition, to analyze how models capture and integrate features across modalities, we test four input conditions: the original *meme* and *text*-only as used in Liu et al. (2022a), as well as *image*-only, and *mix* (*image+text*).

3.4 Vision-Language Models

We evaluate three families of state-of-the-art VLMs, each including variants of different sizes: Aya-Vision (`Aya`, Üstün et al. 2024), Gemma 3 (`Gemma`, Team et al. 2025), and Qwen2.5-VL (`Qwen`, Bai et al. 2025). For `Gemma` and `Qwen`, we test their instruction-tuned versions. All models are prompted in zero-shot settings. Implementation details are provided in Appendix B.3. Due to data imbalance, we report macro F1 scores.

4 Results & Analysis

4.1 How Well Can VLMs Identify Whether a Meme Contains Figurative Meaning?

All evaluated models outperform the majority-vote baseline (macro F1 = 40.45%) under every modality configuration, as shown in Figure 2. Among the nine models, `Qwen-72B` ranks highest for the meme, image-only, and mixed input modality, and is surpassed by `Qwen-32B` in the *text* setting.

Figurative Complexity. In order to better understand the models' performance in the general detection task, we quantify the complexity of the figurative meaning in the meme by the number of figurative categories present. Based on the label distribution in Table 1, we split the memes into five subclasses with 0 (non-figurative) to 4 labels. Using `Qwen-72B` as an illustrative case, we observe a consistent trend across all modality input settings in Figure 3: as the figurative complexity of memes increases, models are more likely to give positive predictions for the presence of figurative meaning. These findings can be interpreted as: the more figurative elements are present in a meme, the more likely the model is to encounter one it can recognize—thus making the classification task easier. Other large-scale models such as `Qwen-32B` and `Aya-32B` also exhibit similar trends across input modalities (see Appendix C.1). In contrast, smaller models such as `Aya-8B` and `Gemma-4B` are biased toward positive predictions regardless of figurative complexity, explaining their weaker performance.

Input Modality. All models except for `Gemma-4B` show higher performance on image-only inputs compared to text-only inputs (see Figure 2). Surprisingly, the image-only condition—where the text is removed—often outperforms the original meme input that contains both modalities. However, further analysis in Figure 3 reveals a more nuanced pattern. Regardless of the figurative complexity of inputs, the likelihood of being classified as containing figurative meaning consistently follows the pattern: original *meme* > *mix* (image+text) > *image* or *text* alone. This suggests that instruction-tuned VLMs are more inclined to associate the meme format—where text is visually embedded in the image—with figurative intent, even beyond what the actual semantic content may indicate. This may be expected, as plain images on the web are generally less likely to be figurative than those containing text (Scott, 2021).

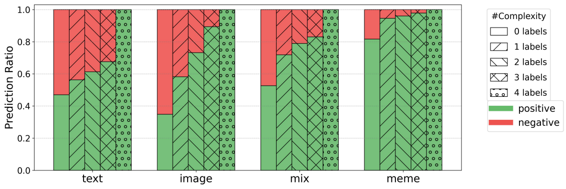

Figure 3: Predicted label distribution by Qwen-72B across groups defined by label complexity.

Model	Size	Input	Allus.	Exagg.	Irony	Anthr.	Metaph.	Contr.
baseline	-	-	52.32	44.00	49.77	41.76	44.87	56.91
Aya	8B	text	55.79	**52.78**	51.81	51.06	36.56	**63.66**
		image	53.67	44.45	47.78	36.10	43.55	43.94
		mix	50.84	46.75	50.75	44.52	42.34	58.44
		meme	39.35	40.31	39.84	37.14	30.56	43.71
	32B	text	60.71	48.12	**54.98**	53.01	53.04	46.80
		image	**69.37**	39.62	51.46	59.77	54.04	39.07
		mix	69.17	42.57	52.01	60.00	**57.67**	43.62
		meme	63.04	31.48	41.24	**62.13**	54.55	42.11
Gemma	4B	text	57.45	41.86	41.72	49.93	25.72	49.54
		image	**68.29**	30.23	43.49	56.84	19.91	41.99
		mix	67.76	22.35	33.55	59.21	19.83	37.50
		meme	65.14	17.52	28.17	57.25	16.67	31.67
	12B	text	59.75	46.40	**57.57**	52.46	43.96	**64.80**
		image	61.90	39.45	48.38	59.64	49.01	35.29
		mix	60.49	31.10	50.82	**62.74**	49.12	33.43
		meme	54.05	30.83	42.39	60.19	48.05	33.15
	27B	text	55.23	46.54	50.34	51.12	47.69	56.92
		image	49.79	**49.37**	42.62	61.05	**52.07**	45.27
		mix	50.61	44.53	34.47	62.07	51.52	49.29
		meme	43.26	46.34	28.32	62.37	50.18	49.52
Qwen	3B	text	49.69	51.78	54.72	49.01	45.46	59.20
		image	47.61	**60.89**	45.81	60.32	45.78	66.78
		mix	54.30	58.66	53.93	59.68	45.79	65.96
		meme	52.00	58.68	57.43	59.63	47.80	**67.01**
	7B	text	51.96	51.80	54.36	49.06	48.21	54.42
		image	70.37	55.30	50.04	58.84	54.10	58.19
		mix	70.22	53.26	52.61	59.39	52.55	59.89
		meme	**72.87**	51.33	51.42	59.36	55.26	57.43
	32B	text	50.24	53.32	**62.17**	52.14	50.87	61.72
		image	55.81	60.64	49.35	55.76	57.35	55.30
		mix	46.99	55.23	60.02	56.46	56.96	55.76
		meme	43.81	52.91	58.28	56.85	58.94	54.44
	72B	text	58.82	52.15	58.90	51.16	54.16	65.40
		image	64.63	53.92	51.07	59.70	58.93	63.09
		mix	60.84	47.27	54.32	61.80	**60.12**	63.40
		meme	60.00	40.45	46.25	**61.85**	59.58	61.56

Table 2: Macro-F1 scores (in %) for each figurative category. The best score in each model family is highlighted in **bold**. Results are averaged over 5 runs.

4.2 How Well Can VLMs Distinguish Figurative Meaning Types?

We treat the best results from Liu et al. (2022a) across modalities and models as a baseline. Table 2 shows that the top-performing zero-shot results from Aya, Gemma, and Qwen all exceed this baseline. For more visualized results, see Appendix C.2.

Input Modality. Intuitively, one might expect that providing more input information would lead to better model performance—for example, using memes would outperform the mix setting, which in turn would outperform image-only or text-only inputs. However, the results deviate significantly from this expectation. For *irony*, we observe that, with the exception of the smaller models Gemma-4B and Qwen-3B, the other models achieve their best performance when given text-only input. Adding or replacing visual information appears to introduce noise rather than provide useful signals for the task. Nevertheless, for *anthropomorphism*, providing visual elements significantly enhances performance in all models except Aya-8B. In the remaining categories, patterns are less evident.

Figurative categories often align with different modalities in memes. For example, *irony* tends to rely more on textual wordplay, while *anthropomorphism* is more grounded in visual cues. These results suggest that generative VLMs capture modality-specific patterns to some extent. However, the fact that memes—which exhibit the richest combination of modalities—do not result in the best performance suggests that the models are failing to fully integrate multimodal information.

Model Type & Model Size. Based on the best scores of each model family in Table 2, Qwen achieves higher performance than Aya and Gemma in all categories except *anthropomorphism*. Unlike Aya and Gemma, Qwen achieves its best performance in most categories when visual input is provided. Specifically, Qwen's top scores in *allusion*, *anthropomorphism*, and *contrast* are attained with meme input. Instead, Aya benefits from meme input when identifying *anthropomorphism*, while Gemma never performs best with meme input for any category.

As discussed in Section 4.1, larger Qwen and Aya models perform better in general figurative meaning detection. When detecting the *metaphor* category, all nine models show consistent performance gains with increasing model size, regardless of input modality. However, for other categories, no clear trend for model size is observed. The distinct *metaphor* results likely stem from the relatively abundant training data for this category (Shutova et al., 2016; Kehat and Pustejovsky, 2020; Chakrabarty et al., 2023), and larger models leveraging data more effectively.

5 Conclusion

We evaluate Aya, Gemma, and Qwen on their ability to detect and classify figurative meaning in internet memes. All models' best setup for the respective task outperforms the baseline. For general figurative meaning detection, large-scale models like Qwen-72B are more likely to detect figurative meaning as memes' figurative complexity increases. All models show a strong bias toward predicting figurative meaning in the meme input setting. For granular figurative meaning classification, models benefit differently from modality combinations: text supports irony, while images help with anthropomorphism. Despite their generative strength, current VLMs still struggle to fully integrate multimodal signals.

Limitations

Limited Dataset Diversity Due to the scarcity of datasets annotated with multiple types of figurative meanings in memes, our experiments are conducted solely on the FIGMEMES (Liu et al., 2022a) dataset. Moreover, FigMemes collects data exclusively from the 4chan[2] platform, which may differ significantly in style and content from other popular social media platforms such as Reddit, X (Twitter), and Facebook. This limitation may affect the generalizability of our findings across different meme communities.

Simplified Masking and Inpainting Approach For masking text in memes, we adopt a relatively straightforward method. As a result, the inpainting process is sometimes imperfect and may introduce artifacts even when successful. However, in most cases, this approach is sufficient because meme creators often avoid placing text over important visual elements or follow established templates that position text outside the core image (see Section B).

Ethical considerations

We see no ethical issues related to this work. All experiments were conducted with publicly available data and open-source software, and we have made all of our code openly available for reproducibility.

Use of AI Assistants. The authors acknowledge the use of ChatGPT solely for correcting grammatical errors, enhancing the coherence of the final manuscripts, and providing assistance with coding.

References

Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, and 1 others. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23201–23211.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Luke Bates, Peter Ebert Christensen, Preslav Nakov, and Iryna Gurevych. 2025. A template is all you meme. 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 10443–10475, Albuquerque, New Mexico. Association for Computational Linguistics.

Christian Burgers, Elly A Konijn, and Gerard J Steen. 2016. Figurative framing: Shaping public discourse through metaphor, hyperbole, and irony. *Communication theory*, 26(4):410–430.

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.

Tuhin Chakrabarty, Yejin Choi, and Vered Shwartz. 2022a. It‘s not rocket science: Interpreting figurative language in narratives. *Transactions of the Association for Computational Linguistics*, 10:589–606.

Tuhin Chakrabarty, Debanjan Ghosh, Adam Poliak, and Smaranda Muresan. 2021. Figurative language in recognizing textual entailment. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3354–3361, Online. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7370–7388, Toronto, Canada. Association for Computational Linguistics.

Tong Chen, Akari Asai, Niloofar Mireshghallah, Sewon Min, James Grimmelmann, Yejin Choi, Hannaneh Hajishirzi, Luke Zettlemoyer, and Pang Wei Koh. 2024. CopyBench: Measuring literal and non-literal reproduction of copyright-protected text in language model generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15134–15158, Miami, Florida, USA. Association for Computational Linguistics.

Erik-Lân Do Dinh, Steffen Eger, and Iryna Gurevych. 2018. Killing four birds with two stones: Multi-task learning for non-literal language detection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1558–1569, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[2] https://4chan.org/

Susan R Fussell and Mallie M Moss. 2014. Figurative language in emotional communication. In *Social and cognitive approaches to interpersonal communication*, pages 113–141. Psychology Press.

Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. 2015. SemEval-2015 task 11: Sentiment analysis of figurative language in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, Denver, Colorado. Association for Computational Linguistics.

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.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do androids laugh at electric sheep? humor "understanding" benchmarks from the new yorker caption contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. MemoSen: A multimodal dataset for sentiment analysis of memes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1542–1554, Marseille, France. European Language Resources Association.

Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah Masud Preum. 2024. Deciphering hate: Identifying hateful memes and their targets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8347–8359, Bangkok, Thailand. Association for Computational Linguistics.

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.

Hyewon Jang, Qi Yu, and Diego Frassinelli. 2023. Figurative language processing: A linguistically informed feature analysis of the behavior of language models and humans. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9816–9832, Toronto, Canada. Association for Computational Linguistics.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Taylor Berg-Kirkpatrick. 2021. Investigating robustness of dialog models to popular figurative language constructs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7476–7485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gitit Kehat and James Pustejovsky. 2020. Improving neural metaphor detection with visual datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5928–5933, Marseille, France. European Language Resources Association.

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. *Advances in neural information processing systems*, 33:2611–2624.

Shreyas Kulkarni, Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2024. A report on the FigLang 2024 shared task on multimodal figurative language. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 115–119, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2023. Multilingual multi-figurative language detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9254–9267, Toronto, Canada. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 2008. *Metaphors we live by*. University of Chicago press.

Gyeongeun Lee, Christina Wong, Meghan Guo, and Natalie Parde. 2024. Pouring your heart out: Investigating the role of figurative language in online expressions of empathy. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 519–529, Bangkok, Thailand. Association for Computational Linguistics.

Chen Liu, Gregor Geigle, Robin Krebs, and Iryna Gurevych. 2022a. FigMemes: A dataset for figurative language identification in politically-opinionated memes. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7069–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Emmy Liu, Chenxuan Cui, Kenneth Zheng, and Graham Neubig. 2022b. Testing the ability of language models to interpret figurative language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics:*

Human Language Technologies, pages 4437–4452, Seattle, United States. Association for Computational Linguistics.

Abhilash Nandy, Yash Agarwal, Ashish Patwa, Millon Madhur Das, Aman Bansal, Ankit Raj, Pawan Goyal, and Niloy Ganguly. 2024. ***YesBut***: A high-quality annotated multimodal dataset for evaluating satire comprehension capability of vision-language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16878–16895, Miami, Florida, USA. Association for Computational Linguistics.

Asaf Nissenbaum and Limor Shifman. 2017. Internet memes as contested cultural capital: The case of 4chan's /b/ board. *New Media & Society*, 19(4):483–501.

Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. MOMENTA: A multimodal framework for detecting harmful memes and their targets. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kate Scott. 2021. Memes as multimodal metaphors: A relevance theory analysis. *Pragmatics & Cognition*, 28(2):277–298.

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.

Limor Shifman. 2013. *Memes in digital culture*. MIT press.

Ekaterina Shutova, Douwe Kiela, and Jean Maillard. 2016. Black holes and white rabbits: Metaphor identification with visual features. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 160–170, San Diego, California. Association for Computational Linguistics.

Kevin Stowe, Prasetya Utama, and Iryna Gurevych. 2022. IMPLI: Investigating NLI models' performance on figurative language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5375–5388, Dublin, Ireland. Association for Computational Linguistics.

Kohtaro Tanaka, Hiroaki Yamane, Yusuke Mori, Yusuke Mukuta, and Tatsuya Harada. 2022. Learning to evaluate humor in memes based on the incongruity theory. In *Proceedings of the Second Workshop on When Creative AI Meets Conversational AI*, pages 81–93, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Yi Tay, Anh Tuan Luu, Siu Cheung Hui, and Jian Su. 2018. Reasoning with sarcasm by reading in-between. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1010–1020, Melbourne, Australia. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Xiaoyu Tong, Ekaterina Shutova, and Martha Lewis. 2021. Recent advances in neural metaphor processing: A linguistic, cognitive and social perspective. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4673–4686, Online. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Bradley E Wiggins. 2019. *The discursive power of memes in digital culture: Ideology, semiotics, and intertextuality*. Routledge.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275, Bangkok, Thailand. Association for Computational Linguistics.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024a. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Linhao Zhang, Jintao Liu, Li Jin, Hao Wang, Kaiwen Wei, and Guangluan Xu. 2024b. GOME: Grounding-based metaphor binding with conceptual elaboration for figurative language illustration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18500–18510, Miami, Florida, USA. Association for Computational Linguistics.

Naitian Zhou, David Jurgens, and David Bamman. 2024. Social meme-ing: Measuring linguistic variation in memes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3005–3024, Mexico City, Mexico. Association for Computational Linguistics.

A Definitions for the categories of figurative language

The definitions for the categories of figurative language below are adapted from Liu et al. (2022a):

- **Allusion**: Referencing historical events, figures, symbols, art, literature, or pop culture.
- **Exaggeration/Hyperbole**: Use of exaggerated terms for emphasis, including exaggerated visuals (including unrealistic features portraying minorities).
- **Irony/Sarcasm**: Use of words that convey a meaning opposite to their usual meaning/-mock someone or something through caustic or bitter expression.
- **Anthropomorphism/Zoomorphism**: Attributing human qualities to animals, objects, natural phenomena, or abstract concepts, or applying animal characteristics to humans in a way that conveys additional meaning.
- **Metaphor/Simile**: Implicit or explicit comparisons between two items or groups, attributing the properties of one thing to another. This category includes dehumanizing metaphors.
- **Contrast**: Comparison between two positions/people/objects (usually side-by-side).

B Implementation Details

B.1 Examples of Image Preprocessing

We use EasyOCR[3] to detect overlaid text in the meme, and then apply the OpenCV[4] inpainting algorithm to mask the text. Since meme creators often avoid placing text over important visual elements—or follow templates that place text outside the core image region—the OpenCV inpainting process generally produces satisfactory results (see Figure 4). However, in cases where the text is slanted or handwritten, the inpainting performance can degrade due to the irregularity of the text structure (see Figure 5).

B.2 Prompt Templates

We design two prompt templates corresponding to our main tasks. The first prompt (see Figure 6) is used for the *figurative meaning detection* task, where the model determines whether figurative meaning is present in a given multimodal input and provides a structured explanation. The second prompt (see Figure 7) is designed for the *figurative meaning classification* task, where the model categorizes the types of figurative meaning expressed in the input.

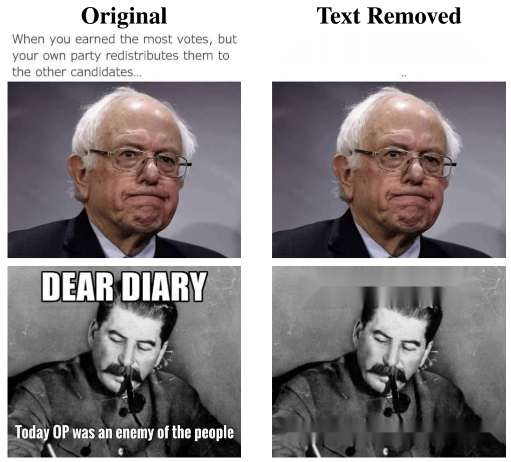
Original
Text Removed
When you earned the most votes, but your own party redistributes them to the other candidates...
DEAR DIARY
Today OP was an enemy of the people

Figure 4: Examples of successful inpainting where meme text is placed on uniform or non-critical regions.

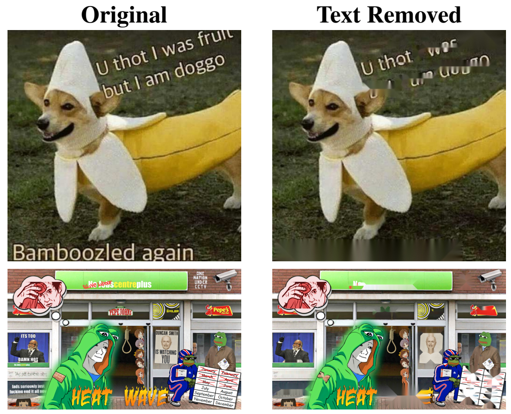

Figure 5: Examples where inpainting is less effective due to slanted or handwritten text.

We approach category detection as a multi-label classification task. Instead of performing binary classification for each label directly, we prompt the model to produce probabilities, and assign a positive label when the probability exceeds 0.5.

B.3 Model Setup

We evaluate the following vision-language models: Aya-Vision[5] (Üstün et al., 2024)), Gemma 3[6] (Team et al., 2025), and Qwen2.5-VL[7] (Bai et al.,

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

[4] https://github.com/opencv/opencv

[5] https://huggingface.co/collections/CohereLabs/cohere-labs-aya-vision-67c4ccd395ca064308ee1484

[6] https://huggingface.co/collections/google/gemma-3-release-67c6c6f89c4f76621268bb6d

[7] https://huggingface.co/collections/Qwen/qwen25-vl-6795ffac22b334a837c0f9a5

Figurative Meaning Detection Prompt

### Task Description:
You will be shown an input that may include an image, text, or a combination of both.

<VISUAL IMPUT>
The following text is written inside the image: <OCR TEXT>

Your task is to evaluate the input for the presence and usage of figurative meaning.
Please proceed in three steps:

- **Step 1:** State whether any figurative meaning is present in the input. Respond only with "Yes" or "No".
- **Step 2:** If you answered "Yes" in Step 1, estimate the probability (a value between 0 and 1) that figurative meaning is present.
- **Step 3:** Explain the reasoning behind the probability assessment. Be specific and refer to both the visual and textual elements of the input where relevant.

Definitions of figurative meaning:
sFigurative meaning refers to a non-literal interpretation of language or imagery, where expressions convey meanings that differ from their standard, literal sense–often to illustrate abstract ideas, create emphasis, or evoke imagery.

### Output Format:
Your final response should follow this JSON format:

```
{
  ''Is any figurative meaning present in the
      input?'': ''Yes'' or ''No'',
  ''Probability'': float between 0 and 1,
  ''Explanation'': ''Explanation for detection
      of figurative meaning''
}
```

Figure 6: The prompt used for the zero-shot figurative meaning detection tasks when instructing VLMs. Text enclosed in sharp brackets <. . .> is replaced by the actual examples. In *meme* mode, the original meme is added as <VISUAL INPUT>; in *image* and *mix* mode, the text-removed version is added as <VISUAL INPUT>; in *text* and *mix* mode, "The following text is written inside the image: <OCR TEXT>" line is activated.

2025).

All models are conducted locally on NVIDIA A100 and H200 GPUs with the vLLM[8] framework (Kwon et al., 2023) for efficient and consistent inference. Table 3 summarizes the decoding hyperparameters used during inference across all models. Each experiment is repeated 5 times using fixed random seeds: 2024, 3024, 4024, 5024, and 6024.

Parameter	Value
Temperature	0.7
Top-p	0.1
Repetition Penalty	1.05
Max Tokens	512

Table 3: Sampling parameters used during inference.

C Supplementary Results on Figurative Meaning in Memes

C.1 Detection Results Supplement

While the detection results for Qwen-72B are already presented in the main text, here we provide a supplementary figure showing the predicted label distributions for the remaining eight models. Figure 9 illustrates the performance of each model on meme samples categorized by figurative complexity and content type (text, image, mix, meme). For each model, the proportion of samples predicted as positive or negative is shown, with hatch patterns indicating different levels of figurative complexity (from 0 to 4).

In addition, we analyze category co-occurrence patterns across the 1,396 test memes, as illustrated in Figure 8.

C.2 Classification Results Supplement

Table 4 reports macro-F1 scores for each figurative category. The *text* rows show the baseline scores, while the *image*, *mix*, and *meme* rows indicate performance differences compared to these baselines. Darker cell colors highlight larger differences.

Figures 10 and 11 visualize the results of multi-class figurative meaning classification, grouped by input modality and by model, respectively.

[8] https://github.com/vllm-project/vllm

Model	Size	Input	Allusion	Exaggeration	Irony	Anthrop	Metaphor	Contrast
Aya	8B	text	56.76	52.34	48.8	51.48	37.1	63.84
		image	-3.12	-7.13	-0.08	-14.24	5.76	-19.61
		mix	-6.07	-5.18	2.19	-5.08	5.01	-3.29
		meme	-18.53	-11.68	-10.37	-13.12	-8.12	-19.83
	32B	text	62.13	50.38	52.66	51.21	53.46	50.66
		image	7.28	-9.97	0	11.12	0.43	-10.91
		mix	7.04	-6.87	-0.42	10.83	4.12	-5.63
		meme	0.08	-18.92	-13.08	13.44	0.84	-7.37
Gemma	4B	text	57.45	41.86	41.72	49.93	25.72	49.54
		image	10.92	-10.6	3.88	8.25	-6.62	-7.06
		mix	10.71	-19.2	-7.86	11.11	-6.82	-12.14
		meme	7.21	-24.59	-13.91	8.53	-10.14	-18.71
	12B	text	59.75	46.4	57.57	52.46	43.96	64.8
		image	2.01	-6.25	-7.6	8.58	4.2	-28.63
		mix	0.47	-15.39	-4.95	12.33	4.41	-30.9
		meme	-6.66	-15.73	-14.95	9.36	3.47	-31.22
	27B	text	55.44	48.62	47.1	51.61	48.35	62.55
		image	-5.48	1.71	-2.22	11	2.68	-16.76
		mix	-4.47	-3.36	-11.65	12	2.17	-12.16
		meme	-12.46	-1.69	-18.71	12.34	0.89	-12.02
Qwen	3B	text	50.54	52.26	53.75	49.3	45.91	59.36
		image	-2.35	8.2	-8.53	12.51	0.34	7.71
		mix	4.85	5.84	-0.6	11.14	0.38	6.88
		meme	2.48	5.71	2.27	10.8	2.69	7.94
	7B	text	53.1	52.01	53.73	49.36	49.04	54.46
		image	17.3	4.26	-2.72	10.81	4.77	4.48
		mix	17.28	1.71	-0.22	10.52	3.05	6.68
		meme	20.15	-0.18	-2.11	11.16	6.37	3.06
	32B	text	49.8	53.21	61.09	52.66	51.09	61.45
		image	6.34	7.22	-11.96	4.23	4.56	-4.89
		mix	-2.99	1.24	-1.23	4.98	4.74	-4.77
		meme	-6.55	-1.28	-4.14	5.76	7.22	-6.1
	72B	text	59.86	51.72	57.28	51.68	55.13	65.57
		image	5.03	2.79	-5.64	9.71	3.63	-2.03
		mix	1.27	-4.41	-2.78	12.43	4.61	-1.51
		meme	0.03	-11.66	-12.53	12.33	4.67	-3.71

Table 4: Macro-F1 scores (in %) for each figurative categories. The *text* rows show the macro-F1 scores. The *image*, *mix*, and *meme* rows represent the difference in performance compared to the corresponding *text* configuration. Positive values indicate better performance than the *text*, while negative values indicate worse performance. Cell background colors become darker as the absolute value of the difference increases, highlighting greater variation.

Figurative Meaning Classification Prompt

### Task Description:

You will be shown an input that may include an image, text, or a combination of both.

Your task is to evaluate the input for the presence and usage of figurative meaning.
Please proceed in three steps:

- **Step 1:** State whether any figurative meaning is present in the input. Respond only with "Yes" or "No".
- **Step 2:** If you answered "Yes" in Step 1, estimate the probability (a value between 0 and 1) that figurative meaning is present.
- **Step 3:** Explain the reasoning behind the probability assessment. Be specific and refer to both the visual and textual elements of the input where relevant.

Definitions of figurative meaning:
Allusion: Referencing historical events, figures, symbols, art, literature or pop culture.
Exaggeration/Hyperbole: Use of exaggerated terms for emphasis, including exaggerated visuals (including unrealistic features portraying minorities).
Irony/Sarcasm: Use of words that convey a meaning that is the opposite of its usual meaning/mock someone or something with caustic or bitter use of words.
Anthropomorphism/Zoomorphism: Attributing human qualities to animals, objects, natural phenomena or abstract concepts or applying animal characteristics to humans in a way that conveys additional meaning.
Metaphor/Simile: Implicit or explicit comparisons between two items or groups, attributing the properties of one thing to another. This category includes dehumanizing metaphors.
Contrast: Comparison between two positions/people/objects (usually side-by-side).

### Output Format:

Your final response should follow this JSON format:

```
{
  "Is any figurative meaning present in the input?": "Yes" or "No",
  "figurative meaning": {
    "Allusion": probability (float between 0 and 1),
    "Exaggeration": probability (float between 0 and 1),
    "Irony": probability (float between 0 and 1),
    "Anthropomorphism": probability (float between 0 and 1),
    "Metaphor": probability (float between 0 and 1),
    "Contrast": probability (float between 0 and 1)
  },
  "Explanations": {
    "Allusion": "Explanation for Allusion",
    "Exaggeration": "Explanation for Exaggeration",
    "Irony": "Explanation for Irony",
    "Anthropomorphism": "Explanation for Anthropomorphism",
    "Metaphor": "Explanation for Metaphor",
    "Contrast": "Explanation for Contrast"
  }
}
```

Figure 7: The prompt used for the zero-shot figurative meaning classification tasks when instructing VLMs. Text enclosed in sharp brackets <. . .> is replaced by the actual examples. In *meme* mode, the original meme is added as <VISUAL INPUT>; in *image* and *mix* mode, the text-removed version is added as <VISUAL INPUT>; in *text* and *mix* mode, “The following text is written inside the image: <OCR TEXT>” line is activated.

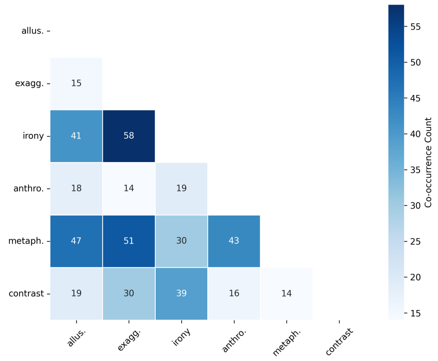

Figure 8: Category co-occurrence statistics among 1,396 test memes.

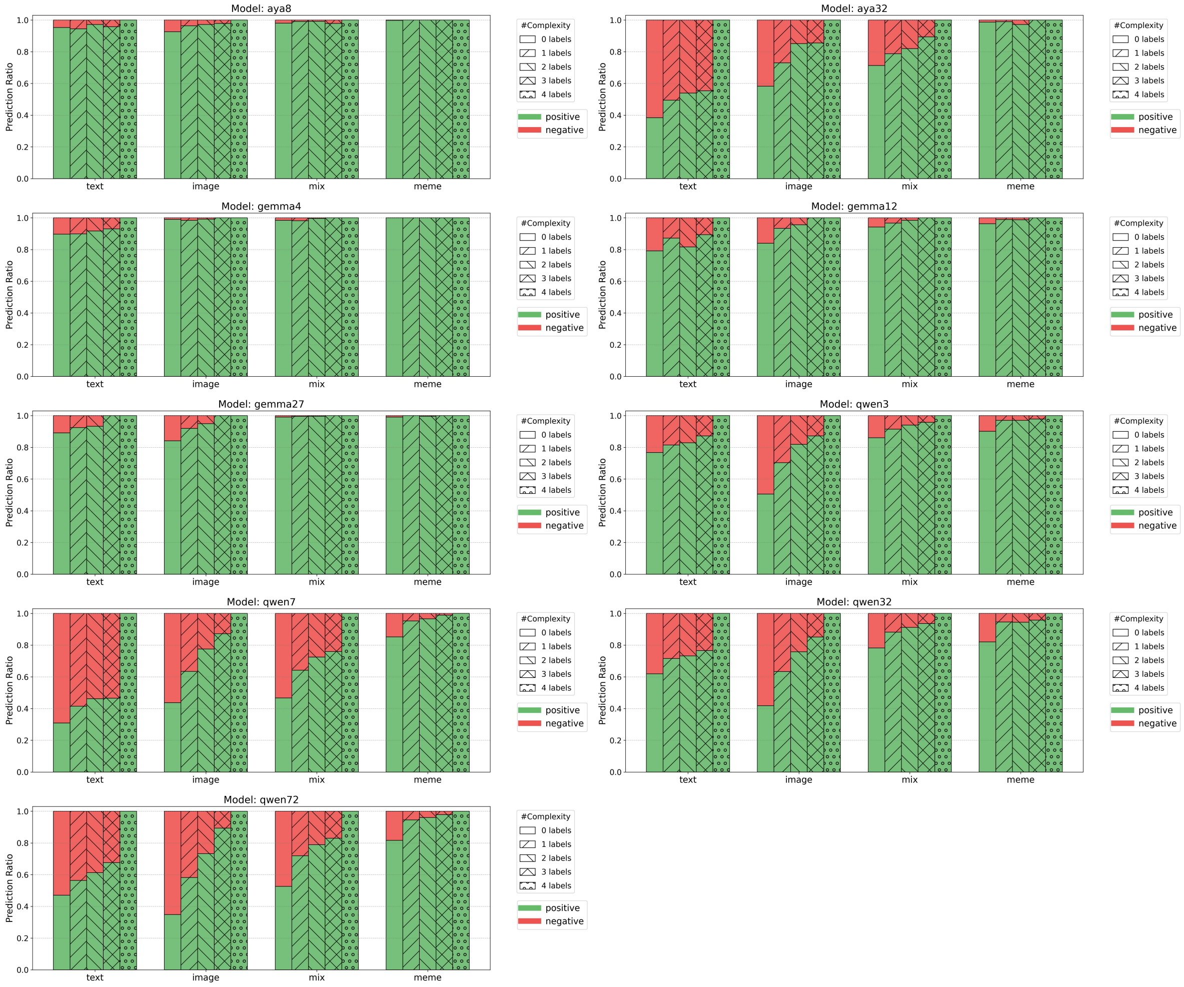

Figure 9: Prediction ratio of positive and negative labels for each model across different figurative complexity levels and content types. Hatch patterns denote complexity levels.

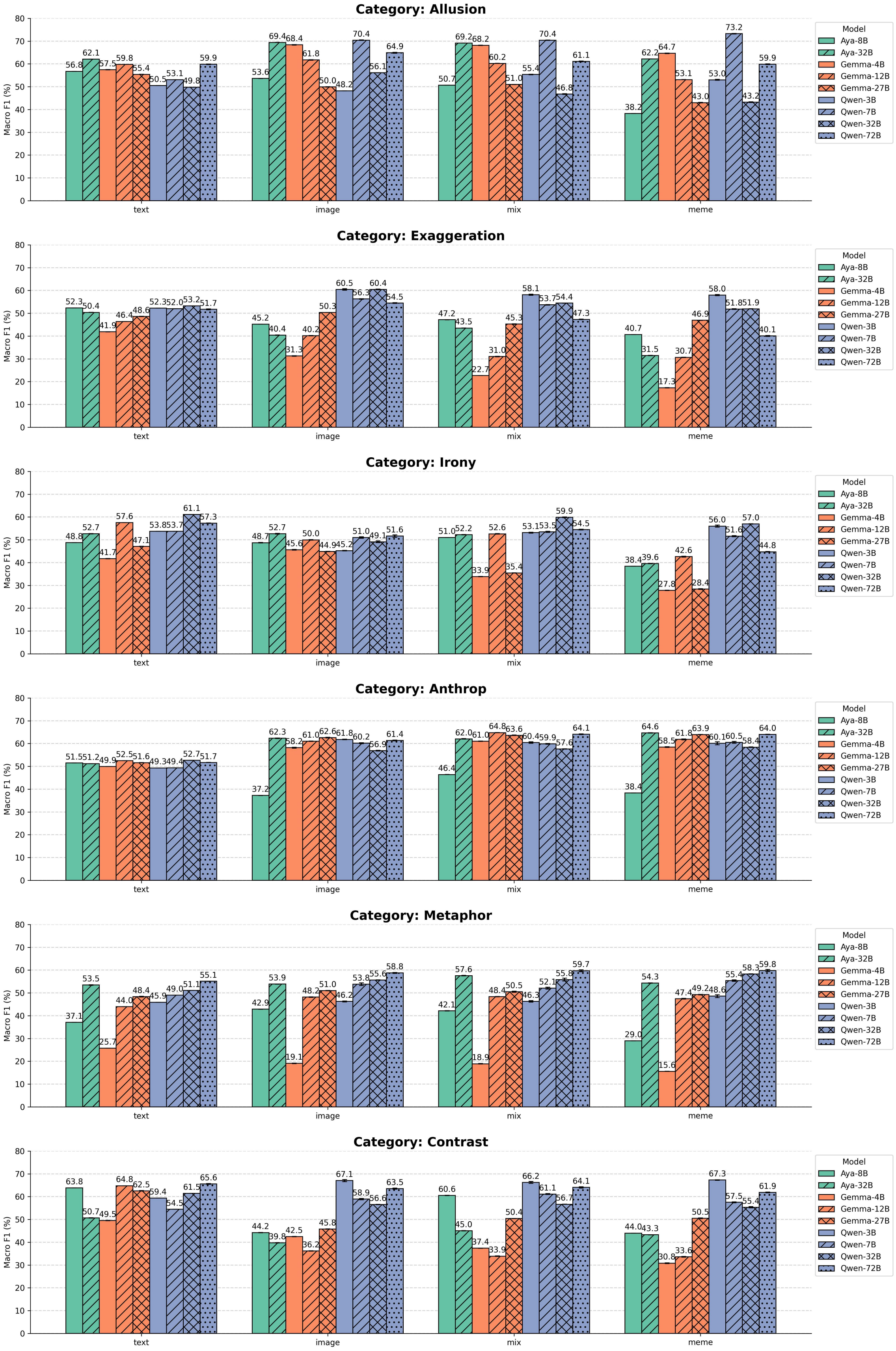

Figure 10: F1 scores (%) for different models across input types. Each group corresponds to an input type (meme, mix, image, text), and within each group, different bars represent models. Aya, Gemma, and Qwen model families are shown using the same color but distinguished with different hatch patterns. The exact F1 score is annotated above each bar, with error bars indicating standard deviation across seeds.

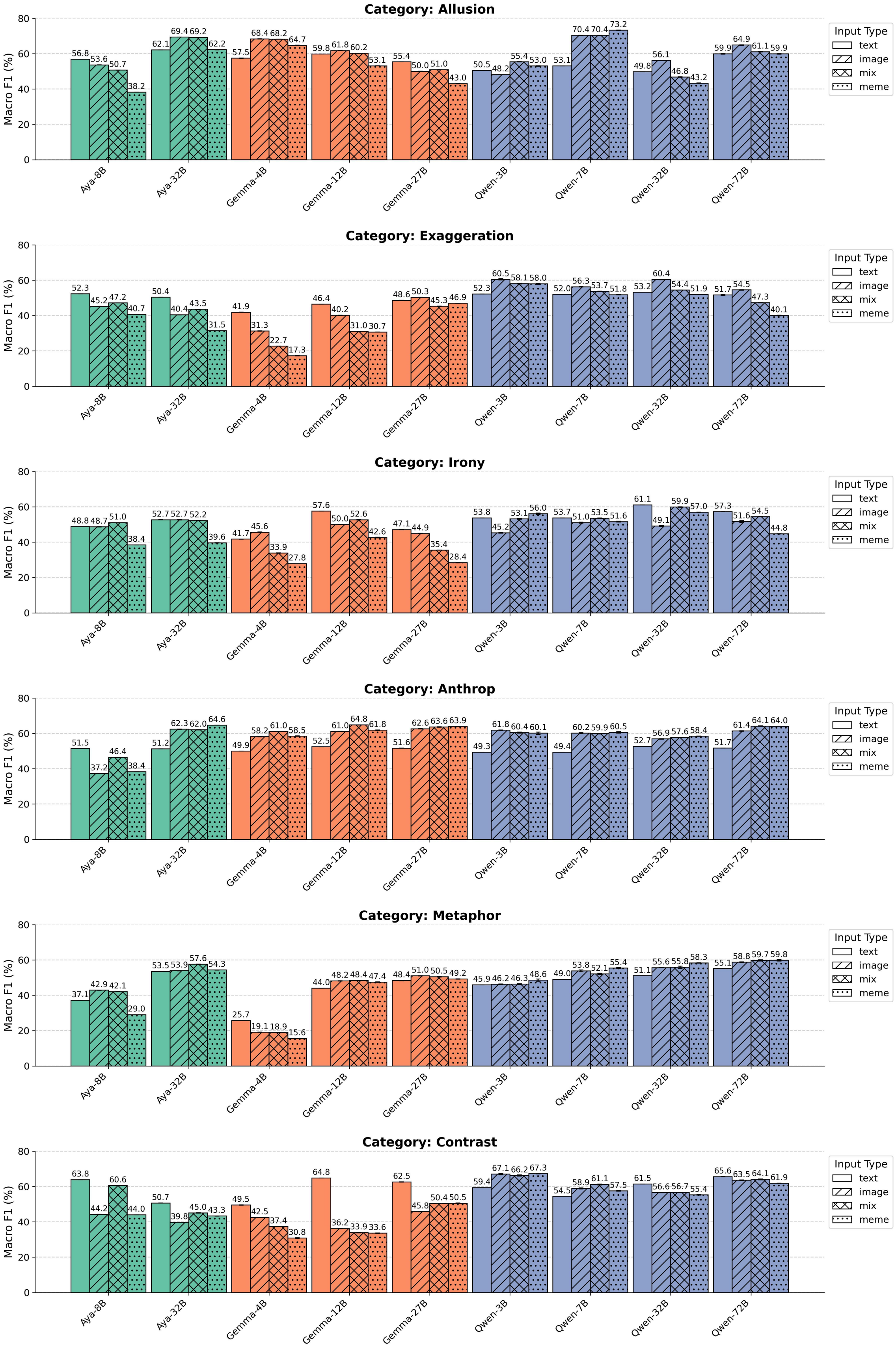

Figure 11: F1 scores (%) for different models across input types. Each group corresponds to an input type (meme, mix, image, text), and within each group, different bars represent models. Aya, Gemma, and Qwen model families are shown using the same color but distinguished with different hatch patterns. The exact F1 score is annotated above each bar, with error bars indicating standard deviation across seeds.