

A Report on the FigLang 2024 Shared Task on Multimodal Figurative Language

Shreyas Kulkarni¹, Arkadiy Saakyan¹, Tuhin Chakrabarty¹, Smaranda Muresan¹

¹Department of Computer Science, Columbia University,

shreyas.kulkarni@columbia.edu, a.saakyan@columbia.edu, tuhin.chakr@cs.columbia.edu, smara@cs.columbia.edu

Abstract

We present the outcomes of the Multimodal Figurative Language Shared Task held at the 4th Workshop on Figurative Language Processing (FigLang 2024) co-located at NAACL 2024. The task utilized the V-FLUTE dataset (Saakyan et al., 2024) which is comprised of <image, text> pairs that use figurative language and includes detailed textual explanations for the entailment or contradiction relationship of each pair. The challenge for participants was to develop models capable of accurately identifying the visual entailment relationship in these multimodal instances and generating persuasive free-text explanations. The results showed that the participants’ models significantly outperformed the initial baselines in both automated and human evaluations. We also provide an overview of the systems submitted and analyze the results of the evaluations. All participating systems outperformed the LLaVA-ZS baseline, provided by us in F1-score.

1 Introduction

Figurative language, which demands an understanding of the implied meanings behind expressions, has been extensively studied, as demonstrated in prior research (Chakrabarty et al., 2022; Saakyan et al., 2022). Similar complexities exist in visual domains, notably in visual metaphors (Chakrabarty et al., 2023; Akula et al., 2023), though most research on large multimodal models (LVMs) has primarily addressed the interpretation of literal meanings in images, as seen in benchmarks like e-ViL (Kayser et al., 2021), ScienceQA (Lu et al., 2022), and MMMU (Yue et al., 2024).

In this shared task, we aim to explore how LVMs handle figurative content in multimodal inputs. Our task, explainable figurative visual entailment, challenges a model to determine whether an image (the premise) supports or contradicts a given claim (the hypothesis) and to provide a reasoned explanation

for its decision. Examples from our dataset are shown in Table 2.

The dataset leverages extensive prior research on both figurative language and images (Chakrabarty et al., 2023; Yosef et al., 2023; Hessel et al., 2023; Hwang and Schwartz, 2023; Desai et al., 2022). It is designed specifically for the visual entailment task and is enhanced with high-quality annotations that include explanations.

This paper reports the results of the shared task that is part of the 4th Workshop on Figurative Language Processing (FigLang 2024) at NAACL 2024. Details of the task, datasets, and evaluation methods are discussed in Section 2. Summaries of each participating system are provided in Section 4, and Section 4.4 offers a comparative analysis of these systems.

2 Datasets and Task Description

Subset	Fig. Lang. Type	Fig. Part
IRFL	Metaphor, Idiom, Simile	Caption
VisMet	Metaphor, Simile	Image
MemeCap	Humor	Image
MuSE	Sarcasm	Caption
NYCC	Humor	Both
V-FLUTE	Metaphor, Idiom, Simile, Sarcasm, Humor	Image, Caption, Both

Table 1: Overview of subsets for visual entailment and multimodal figurative language understanding.

The shared task utilizes an early version of the V-FLUTE dataset, introduced by Saakyan et al. (2024). The dataset is comprised of <image, text> pairs, each annotated with labels indicating either entailment or contradiction, along with explanations for each pair (see Table 2). Originating from five previous studies (Chakrabarty et al., 2023; Yosef et al., 2023; Hessel et al., 2023; Hwang and Schwartz, 2023; Desai et al., 2022), V-FLUTE includes figurative language elements such as metaphors, idioms,






Subset	Image (Premise)	Claim (Hypothesis)	Label and Explanation
VisMet		The faculty meeting was peaceful.	Label: Contradiction <i>Explanation:</i> The image shows a faculty meeting transformed into a dramatic battlefield scene, with members dressed as knights discussing academic content on boards behind them as if they were battle tactics. This visual metaphor suggests the faculty meeting was like a war, and not peaceful.
IRFL		Their relationship is a house on fire.	Label: Entailment <i>Explanation:</i> [...] the photo suggests there is conflict or an intense emotional situation between the two individuals, which aligns with the symbolism of a house on fire representing a relationship filled with turmoil or heated arguments.
MuSE		Oh I just #love having to stare at this while I #work.	Label: Contradiction <i>Explanation:</i> the author wants to go to the disneyland and not just stare at it while working.
MemeCap	<p>DUDE DIED, BUT THEY MADE HIM GO TO WORK ANYWAY</p> 	Even death won't exempt you from going to work.	Label: Entailment <i>Explanation:</i> The image displays RoboCop [...] This entails the claim that even death won't exempt you from going to work because it humorously illustrates a character who has been reanimated as a cyborg to continue working despite having died.
NYCC		Easy for you to say, you're cured!	Label: Entailment <i>Explanation:</i> A play on the word "cured". People go to therapy to have their mental problems remedied or cured. But "cured" can also refer to a meat preparation technique — here the therapist is cured bacon, and the patient is an egg (which is not cured). The egg is saying that the therapist doesn't understand his problems because he's "cured" in both senses.

Table 2: Sample dataset instances form V-FLUTE corresponding to the source datasets.

Subset	Train		Test	
	#	%	#	%
VisMet	731	16.5	126	18.3
IRFL	1322	29.9	198	28.7
MuSE	1000	22.6	150	21.8
MemeCap	853	19.3	128	18.6
NYCC	520	11.7	87	12.6
Total	4426	100.0	689	100.0

Table 3: Summary of subset distribution statistics involved in V-FLUTE.

similes, humor, and sarcasm. It consists of 5,115 multimodal pairs of high-quality images and texts, complete with labels and explanations. For statistics on the dataset, please see Table 3.

3 Evaluation Setup

To evaluate the participant models, we developed a test set by randomly selecting 689 instances, each comprising an $\langle \text{image}, \text{text} \rangle$ pair with corresponding explanations, from our dataset. We describe below the automatic metrics used to evaluate the models’ capability in interpreting figurative language.

Automatic Metrics We used BERTScore (using `microsoft/deberta-xlarge-mnli`), termed here as the *explanation score*, which ranges from 0 to 100, to evaluate the quality of the explanations. Rather than just reporting label accuracy, we report the label F1 score at three explanation score thresholds: 0, 50, and 60. An $F1@0$ score corresponds to basic label F1, while an $F1@50$ score includes only those correct label predictions with an explanation score above 50.

4 Participants and Results

4.1 Training Phase

The competition began on January 25, 2024, with the release of training data and auxiliary scripts to all registered participants. Participants had the option to further divide the training data into a validation set for tuning hyperparameters or to use the data for cross-validation.

4.2 Evaluation Phase

The test instances were made available on February 15, 2024, for evaluation. The deadline for submissions was March 25, 2024. From the submissions, two system papers were accepted for presentation at the Workshop. Submissions were made through

the Codalab site and evaluated against the test instances’ gold labels. We utilized Codabench (Xu et al., 2022) for the competition due to its user-friendly interface, its ability to facilitate communication (such as mass emailing) with participants, and its real-time leader-board updates. Additionally, we established our own GPU-based evaluation system using custom Docker architecture. The leader-board showcased the $F1@60$ scores in descending order.

4.3 Participants

Overall, five teams participated in the competition, excluding the organizing team. The following section details the two systems that were accepted.

Baselines We report a fine-tuned baseline and a zero-shot baseline for the task. These baselines utilize a Zero-shot LLaVA-v1.6-mistral-7B model and a fine tuned LLaVA-v1.6-mistral-7B model on the V-FLUTE dataset.

MAPPER (map, 2024) is a modal-supplement framework, consisting of a describer and a thinker. The describer uses a frozen large vision model (LLaVA-7B-v1.5) to detail images capturing essential semantic information. The thinker, enhanced with LoRA (Hu et al., 2021) on a fine-tuned large multi-modal model (LLaVA-7B-v1.5), leverages these descriptions along with claims and images to form predictions and explanations. MAPPER’s vision component uses CLIP (Radford et al., 2021) for image understanding.

FigCLIP (fig, 2024) merges CLIP and GPT-2 to identify and elucidate multimodal figurative semantics. It features separate text and image encoders initialized by CLIP (CLIP-ViT-L/14), connected via a bidirectional fusion module with cross-attention mechanisms. A GPT-2 model generates explanations, and a special projector aligns multimodal embeddings with explanation representations, enhancing the model’s efficiency in handling figurative image-text alignment. The projector involved makes FigCLIP lightweight.

4.4 Analysis

The best performing method according to (Table 4) is MAPPER. The system outperforms others on both $F1@0$ and $F1@60$ metrics. We note that the system improvement is quite high compared to the zero-shot system. Interestingly, the FigCLIP system performs very well and only slightly lower

#	Participant	F1@0	F1@50	F1@60
1	MAPPER	0.90	0.89	0.75
2	LLaVA-FT	0.73	0.72	0.59
3	FigCLIP	0.70	0.67	0.50
4	GPT-4V	0.70	0.64	0.49
5	mrshu	0.63	0.62	0.43
6	yangst	0.51	0.48	0.31
7	LLaVA-ZS	0.45	0.38	0.21

Table 4: Automatic evaluation results by team with rank. FT refers to fine-tuned model and ZS represents the Zero-Shot model. The GPT-4V model submitted is not our baseline but a participants submission.

than the fine-tuned LLaVA model that utilizes a much stronger language model backbone.

5 Conclusion

This paper presents the outcomes of the shared task on multimodal figurative language, conducted at the 4th Workshop on Figurative Language Processing at NAACL 2024 (FigLang 2024). The goal of this shared task was to accurately classify figurative <image, text> instances and provide a persuasive explanation for the classification. We included a brief overview of each system that participants submitted to the shared task. All systems submitted by participants surpassed the LLaVA-ZS baseline in terms of F1-score. In conclusion, we anticipate that this shared task will encourage continued research into the understanding of figurative language.

References

2024. Figclip: A generative multimodal model with bidirectional cross-attention for understanding figurative language via visual entailment.
2024. A textual modal supplement framework for understanding multi-modal figurative language.
- Arjun R Akula, Brendan Driscoll, Pradyumna Narayana, Soravit Changpinyo, Zhiwei Jia, Suyash Damle, Garima Pruthi, Sugato Basu, Leonidas Guibas, William T Freeman, et al. 2023. Metaclue: Towards comprehensive visual metaphors research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23201–23211.
- Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. [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.
- Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. [Nice perfume. how long did you marinate in it? multimodal sarcasm explanation](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):10563–10571.
- 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.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- 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.
- Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1244–1254.
- Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Taffjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Arkadiy Saakyan, Tuhin Chakrabarty, Debanjan Ghosh, and Smaranda Muresan. 2022. [A report on the FigLang 2022 shared task on understanding figurative language](#). In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 178–183, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

- Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. V-flute: Visual figurative language understanding with textual explanations.
- Zhen Xu, Sergio Escalera, Adrien Pavão, Magali Richard, Wei-Wei Tu, Quanming Yao, Huan Zhao, and Isabelle Guyon. 2022. [Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform](#). *Patterns*, 3(7):100543.
- Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. [IRFL: Image recognition of figurative language](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhua Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of CVPR*.