FigCLIP: A Generative Multimodal Model with Bidirectional Cross-attention for Understanding Figurative Language via Visual Entailment

Qihao Yang
School of Computer Science
South China Normal University
Guangzhou, China
charlesyeung@m.scnu.edu.cn

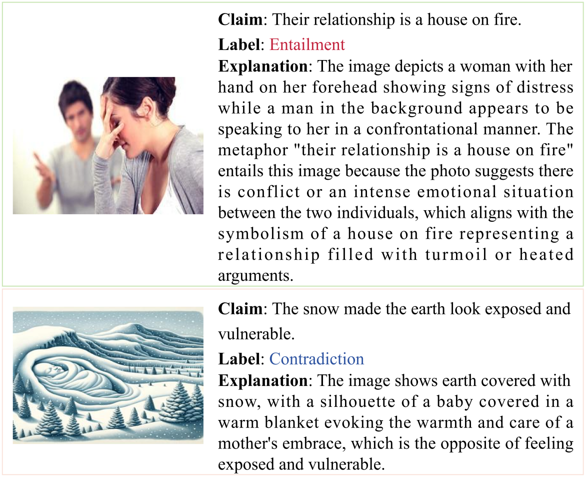
Xuelin Wang✉
College of Chinese Language and Culture
Jinan University
Guangzhou, China
wangxuelin@stu2022.jnu.edu.cn

Abstract

This is a system paper for the FigLang-2024 Multimodal Figurative Language Shared Task. Figurative language is generally represented through multiple modalities, facilitating the expression of complex and abstract ideas. With the popularity of various text-to-image tools, a large number of images containing metaphors or ironies are created. Traditional recognizing textual entailment has been extended to the task of understanding figurative language via visual entailment. However, existing pre-trained multimodal models in open domains often struggle with this task due to the intertwining of counterfactuals, human culture, and imagination. To bridge this gap, we propose FigCLIP, an end-to-end model based on CLIP and GPT-2, to identify multimodal figurative semantics and generate explanations. It employs a bidirectional fusion module with cross-attention and leverages explanations to promote the alignment of figurative image-text representations. Experimental results on the benchmark demonstrate the effectiveness of our method, achieving 70% F1-score, 67% F1@50-score and 50% F1@60-score. It outperforms GPT-4V, which has robust visual reasoning capabilities.

1 Introduction

Figurative language is typically divided into metaphor, simile, and sarcasm (Saakyan et al., 2022). It serves as an implicit way for us to convey complex and imaginative expressions. In recent years, researchers have focused on developing neural networks through mining contextual information. They also aim to construct large-scale figurative datasets to facilitate in-depth research on recognizing textual entailment (Gu et al., 2022; Bigoulaeva et al., 2022; Phan et al., 2022). Despite increasing in parameter size, pre-trained language models (Devlin et al., 2018; Liu et al., 2019) are still unable to fully comprehend cultural knowledge and the social context within figurative language.



Claim: Their relationship is a house on fire.

Label: Entailment

Explanation: The image depicts a woman with her hand on her forehead showing signs of distress while a man in the background appears to be speaking to her in a confrontational manner. The metaphor "their relationship is a house on fire" entails this image because 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.

Claim: The snow made the earth look exposed and vulnerable.

Label: Contradiction

Explanation: The image shows earth covered with snow, with a silhouette of a baby covered in a warm blanket evoking the warmth and care of a mother's embrace, which is the opposite of feeling exposed and vulnerable.

Figure 1: Illustration of the Multimodal Figurative Language Shared Task.

With the prevalence of social media, individuals sometime use images with visual metaphors (i.e., figurative images) to convey counterfactual or humorous meanings, particularly in the advertising industry (Yosef et al., 2023). Various text-to-image AI tools can also be used to create a vast number of figurative images (Chakrabarty et al., 2023). To promote the research on figurative language, the Multimodal Figurative Language Shared Task[1] (named Understanding of Figurative Language Through Visual Entailment) is first introduced by FigLang-2024[2]. Given an <image, text> pair, the goal of this task is to 1) predict whether the image entails or contradicts the text, where the text is referred to as "claims"; 2) generate an explanation for the entailment or contradiction. The illustration of this task is shown as Figure 1.

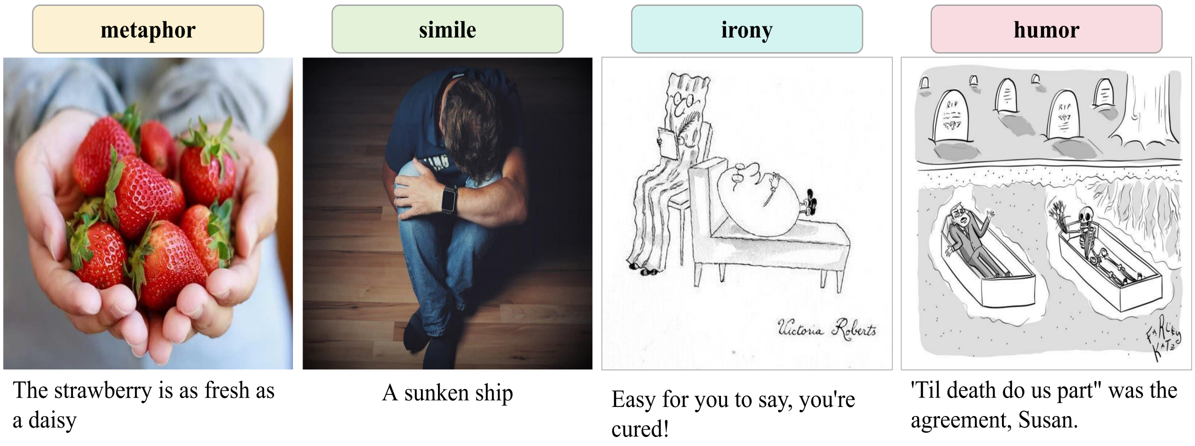
Different from previous research that focused

✉Corresponding author.

[1] https://www.codabench.org/competitions/1970

[2] https://sites.google.com/view/figlang2024/home

Proceedings of the 4th Workshop on Figurative Language Processing (FLP), pages 92–98
June 21, 2024 ©2024 Association for Computational Linguistics



The strawberry is as fresh as a daisy

A sunken ship

Easy for you to say, you're cured!

'Til death do us part" was the agreement, Susan.

Figure 2: Examples of visual entailment between images and claims.

on recognizing textual entailment (Chakrabarty et al., 2022b), the Multimodal Figurative Language Shared Task introduces an image modality to interpret figurative language. Empirically, images can carry richer contextual information than words. Awareness of abstract implications beyond literal and intuitive meanings is the most significant challenge for this task. Even CLIP (Radford et al., 2021), a state-of-the-art architecture in image-text understanding, achieves only 62% accuracy in multimodal entailment test settings, which is far less than the human accuracy of 94% (Hessel et al., 2023). Moreover, existing vision-language models (Radford et al., 2021; Li et al., 2023, 2022) and generative language models (Raffel et al., 2020; Radford et al.) are utilized separately to predict image-text labels and generate explanations. This results in a decoupling of the task, which is inconsistent with the widely accepted paradigm of end-to-end training. Although many large multimodal models (Liu et al., 2024; Jin et al., 2023) perform well on diverse downstream tasks, the availability of large-scale figurative image-text datasets and the requirement for high computational resources are prerequisites for fine-tuning them. Therefore, developing a generic, low-cost, end-to-end multimodal model for multimodal figurative language can potentially further advance the future associated research.

In this paper, we propose FigCLIP. It is built upon CLIP and GPT-2 (Radford et al.) and can jointly achieve the two requirements of label prediction and explanation generation. The main contributions of this work can be summarized as follows:

- A low-cost and end-to-end model is proposed, which is competitive in multimodal figurative language task.
- A bidirectional fusion module with cross-attention is introduced, which enhances the alignment of figurative image-text representations within the mapping space defined by CLIP and GPT-2.
- We compare the model performance for understanding multimodal figurative language at different resolutions.

2 Related Work

Understanding figurative language has been framed as a recognizing textual entailment (RTE) task (Chakrabarty et al., 2022b). Given a <premise, hypothesis> pair, a RTE model is required to determine whether the texts entail or contradict each other. Pre-trained language models like BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) are used to encode both premise and hypothesis texts. The deep representations of premise and hypothesis texts are concatenated and then input into a linear-layer classifier to output an entailment or contradiction label (Chakrabarty et al., 2021, 2022a; Hu et al., 2023). However, these methods cannot enable us to probe whether language models are right for the right reasons. Thus, researchers are committed to construct refined RTE datasets to avoid spurious correlations and annotation artifacts and provide profound figurative knowledge. Explanation-based RTE datasets such as e-SNLI (Camburu et al., 2018) and FLUTE (Chakrabarty et al., 2022b) are increasingly favored. Employing large language models (LLMs) has become

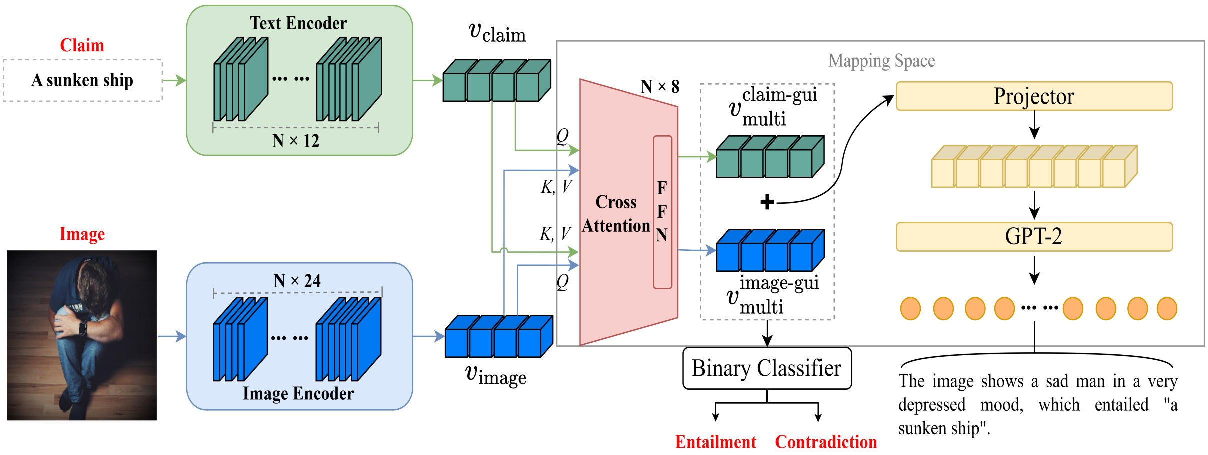


Figure 3: Overview framework of the proposed FigCLIP model.

the mainstream approach to address the RTE task (Kim et al., 2023). Premise and hypothesis texts are combined into prompts to guide the LLMs for generating answers. This implies that the RTE task is simplified into a question-answering problem, allowing the full utilization of the LLMs' capabilities in natural language inference.

Figurative language in images has recently received increasing attention (Yosef et al., 2023; Hessel et al., 2023). As shown in Figure 2, images with claim texts can present metaphors, similes, irony and humor. With the help of diffusion-based text-to-image models such as DALL-E (Ramesh et al., 2021), a number of comic-like figurative images is created based on figurative texts. A high-quality dataset is constructed by (Chakrabarty et al., 2023), containing 6,476 visual metaphors for 1,540 linguistic metaphors and their associated visual elaborations. The Image Recognition of Figurative Language (IRFL) dataset is developed by (Yosef et al., 2023), with human annotation and an automatic pipeline. Although the size of figurative multimodal datasets is increasing, it is still not enough for training a model with strong generalization ability. Thus, pre-trained multimodal models can serve as the backbone and are used to learn the fine-grained figurative image-text representations by fine-tuning on limited figurative multimodal datasets. They only perform the label prediction. For generating explanation, captions generated from images are concatenated with claim texts into pure textual questions. The questions are fed into language models such as GPT-2 and T5 (Raffel et al., 2020), then an explanations are output. To meet the two needs of prediction and explanation at the same time, several large multimodal models, such as GPT-4V (Achiam et al., 2023), MiniGPT4 (Zhu et al., 2023), Flamingo (Alayrac et al., 2022), LlaVA (Liu et al., 2024), are used to accept image and text input and then generate labels and explanations. However, they are commonly evaluated by zero-shot or few-shot due to the high training cost. Research on fine-tuning them on figurative multimodal datasets is still scarce.

3 Method

3.1 Task formulation

The Multimodal Figurative Language (MFL) Shared Task can be treated as a classification and generation problem. Given an <image, claim> pair, a MFL model is required to align image-claim representations, learn a binary classification function F_c to predict entailment or contradiction labels by following Eq. 1, and learn a generation function F_g to generate explanations by following Eq. 2.

$$label = \arg\max F_c\left(image, claim\right) \quad (1)$$

$$explanation = \arg\max F_g\left(image, claim\right) \quad (2)$$

3.2 The FigCLIP model

Architecture. The proposed FigCLIP model employs 12-layer transformers as the text encoder and 24-layer vision transformers as the image encoder. The text encoder and the image encoder are both initialized by CLIP. A GPT-2 model is utilized to generate explanations. The framework of the FigCLIP model is shown in Figure 3.

Specifically, a given claim is input to the text encoder and a claim vector v_{claim} is output. A given image is fed into the image encoder and an image vector v_{image} is output. For label prediction, the FigCLIP model needs to consider whether the claim is semantically entailed by the image. To fuse the deep representations of the claim and the image, a bidirectional fusion module with 8-layer cross-attention is designed. The fusion process is divided into two steps. The claim vector v_{claim} serves as Q, and the image vector v_{image} serves as K and V. They are fed into the fusion module and then a claim-guided multimodal vector $v_{\text{multi}}^{\text{claim-gui}}$ is calculated by $\operatorname{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}V\right)$, where d_k denotes the dimension of 768. This claim-guided multimodal vector achieves an effective interaction of observing details in images based on text. Similarly, the image vector v_{image} serves as Q, and the claim vector v_{claim} serves as K and V. They are fed into the fusion module and then a image-guided multimodal vector $v_{\text{multi}}^{\text{image-gui}}$ is calculated by the same cross-attention calculation process. This image-guided multimodal vector achieves an effective interaction of observing details in text based on images. These two mentioned-above steps share parameters, enhancing the alignment of figurative image-text representations. After that, the $v_{\text{multi}}^{\text{claim-gui}}$ and the $v_{\text{multi}}^{\text{image-gui}}$ are concatenated and input to a binary linear-layer classifier to predict a label of entailment or contradiction.

The original representation space of CLIP is inconsistent with that of GPT-2. GPT-2 relies on a 50257-dimensional vocabulary to generate text, while the CLIP multimodal space is 768-dimensional. For generating explanation, the FigCLIP model needs to match the low-dimensional multimodal representations to 50257 dimensions in a mapping space. Inspired by ClipCap (Mokady et al., 2021), we stack multiple linear layers of different dimensions as a projector. This projector is composed of three sets of linear layers of (768*2→2048), (2048→4096), (4096→50257). In order to further compress the size of parameters to reduce training costs, the parameters of this (4096→50257) linear layer are frozen and treated as a fixed matrix. This is the reason why FigCLIP is more lightweight than ClipCap, despite their similar model architectures. The multimodal representations after projector mapping is fed into GPT-2 to generate an explanation about why the image and claim are semantically entailed or contradicted.

Loss. Two cross-entropy losses are defined to optimize the FigCLIP model jointly, comprising a classification loss ($\mathcal{L}_{\text{cls}}$) and a generation loss ($\mathcal{L}_{\text{gen}}$). The predicted labels and the ground-truth labels are used to calculate the classification loss, which can promote semantic alignment between images and claims to learn more fine-grained details of entailment or contradiction. The generated explanations and the ground-truth explanations are used to calculate the generation loss, which also can facilitate the mapping of multimodal deep representations to establish a reliable mapping space. Finally, the sum of the $\mathcal{L}_{\text{cls}}$ and the $\mathcal{L}_{\text{gen}}$ is regarded as the complete training objective.

Algorithm 1: Pseudocode of Training FigCLIP

```
data: a claim $c$, an image $i$;
      a ground-truth label $l_{\text{gt}}$, a ground-truth explanation $e_{\text{gt}}$;
while $c$, $i$, $l_{gt}$, $e_{gt}$ do
    # the claim vector
    $v_{\text{claim}} \leftarrow$ Text-Encoder($c$);
    # the image vector
    $v_{\text{image}} \leftarrow$ Image-Encoder($i$);
    # the claim-guided multimodal vector
    # the parameter order is $Q$, $K$, $V$
    $v_{\text{multi}}^{\text{claim-gui}} \leftarrow$ Fusion($v_{\text{claim}}, v_{\text{image}}, v_{\text{image}}$);
    # the image-guided multimodal vector
    # the parameter order is $Q$, $K$, $V$
    $v_{\text{multi}}^{\text{image-gui}} \leftarrow$ Fusion($v_{\text{image}}, v_{\text{claim}}, v_{\text{claim}}$);
    # the concatenated multimodal vector
    $v_{\text{multi}} \leftarrow v_{\text{multi}}^{\text{claim-gui}} + v_{\text{multi}}^{\text{image-gui}}$
    ------------------------------------------------------------
    # the classification loss
    $label \leftarrow$ Classifier($v_{\text{multi}}$);
    $\mathcal{L}_{cls} \leftarrow$ CrossEntropyLoss($label$, $l_{\text{gt}}$);
    ------------------------------------------------------------
    # the generation loss
    $v_{\text{multi}}^{\text{mapping}} \leftarrow$ Projector($v_{\text{multi}}$);
    $explanation \leftarrow$ GPT-2($v_{\text{multi}}^{\text{mapping}}$);
    $\mathcal{L}_{gen} \leftarrow$ CrossEntropyLoss($explanation$, $e_{\text{gt}}$);
    ------------------------------------------------------------
    # the complete training objective
    $\mathcal{L} \leftarrow \mathcal{L}_{cls} + \mathcal{L}_{gen}$;
end
```

The FigCLIP model enables end-to-end training because it can jointly address the problems of label prediction and explanation generation. The whole training procedure of the PigCLIP model can be abstracted in Algorithm 1.

4 Experiments and Results

4.1 Datasets

According to the official data description, the training data is compiled from the following five datasets about visual metaphors and multimodal understanding:

(1) a subset of 731 Visual Metaphors dataset (Chakrabarty et al., 2023);

(2) a subset of 1,322 textual metaphors with images (Yosef et al., 2023);

(3) a susbet of 853 memes with annotated claims and explanations (Hwang and Shwartz, 2023);

Data Source	Train/Valid		Test	
	absolute	proportion	absolute	proportion
nycartoons (Hessel et al., 2023)	520	11.7%	87	12.6%
irfl (Yosef et al., 2023)	1322	29.9%	198	28.7%
muse (Desai et al., 2022)	1000	22.6%	150	21.8%
memecap (Hwang and Shwartz, 2023)	853	19.3%	128	18.6%
vismet (Chakrabarty et al., 2023)	731	16.5%	126	18.3%
total	4426	100%	689	100%

Table 1: The statistical details of the datasets for the MFL task.

Model	V-FLUTE test set (%)		
	F1	F1@50	F1@60
jalor	90	89	75
FigCLIP$_{336\times336}$	70	67	50
FigCLIP$_{224\times224}$	68 (-2)	65 (-2)	49 (-1)
GPT-4V (zero-shot)	70	64	49
mrshu	63	62	43
yangst	51	48	31
LlaVA (baseline)	45	38	21

Table 2: Evaluation results on the V-FLUTE test set.

(4) a subset of 1,000 sarcastic captions with images (Desai et al., 2022);

(5) a subset of 520 unique images with captions accompanied with textual explanations (Hessel et al., 2023).

The test data is available at huggingface[3]. Table 1 displays the statistical details of the datasets (named V-FLUTE (Saakyan et al., 2024)) for the MFL task.

4.2 Settings

Our model is implemented on Pytorch 2.0.1 and only one RTX 4090 GPU. Both the text encoder and image encoder are initialized by CLIP-ViT-L/14 or CLIP-ViT-L/14@336px (Radford et al., 2021). All parameters of the text encoder and GPT-2 are optimized, while the image encoder is completely frozen for reducing the training costs. The batch size is set to 32, and the epoch is set to 20. AdamW is applied to optimize model parameters with a learning rate of 1e-04 and weight decay of 0.05. The image resolution is specified as 224×224 or 336×336, and the maximum text length is set to 77. Following previous work (Saakyan et al., 2022), three metrics are used to evaluate the model performance, including F1@0 (pure F1 score), F1@50 (F1 score computed where only instances which had their explanation match the reference with BERTscore (Zhang et al., 2019) above 50 are counted as correct), and similarly F1@60.

4.3 Results

The official evaluation results are reported in Table 2. Our submission ranked second on the leaderboard, where the FigCLIP model was initialized by CLIP-ViT-L/14@336px. The FigCLIP$_{336\times336}$ model achieved 70% F1-score, 67% F1@50-score and 50% F1@60-score on the benchmark test set. LlaVA, the official baseline, only obtained 45% F1-score, 38% F1@50-score and 21% F1@60-score by zero-shot. This means that LlaVA can be applied to this task but it is not proficient in multimodal figurative language understanding. Nevertheless, the FigCLIP$_{336\times336}$ model outperformed LlaVA by 25% F1-score, 29% F1@50-score and 29% F1@60-score respectively. Compared with GPT-4V (a state-of-the-art model in image-text understanding), the FigCLIP$_{336\times336}$ model leaded by 3% and 1% in F1@50-score and F1@60-score respectively, even though their F1-scores ware the same. It is worth noting that calling GPT-4V's API for zero-shot on the test set took approximately $19 and 2 hours, while training an epoch of the FigCLIP model only took less than 1 minute on one 24GB GPU. This demonstrates the low cost and effectiveness of our method. Moreover, we initialized FigCLIP using CLIP-ViT-L/14 to explore the impact of low resolution (224×224). We found that all three metrics dropped slightly when understanding images at low resolution. This shows that the FigCLIP$_{336\times336}$ model can capture more subtle image semantics and facilitate the identification of fine-grained implication relationships with claims.

5 Conclusion

This paper propose an end-to-end model FigCLIP for the FigLang-2024 Multimodal Figurative Language shared task. We introduce a shared bidirectional fusion module with cross-attention to advance the alignment of figurative image-text pairs. In the mapping space defined by CLIP and GPT-2, we utilize a projector to bridge multimodal representations and explanation representations and make FigCLIP lightweight. Experimental results on the benchmark test set demonstrates the effectiveness of our method, which achieves competitive performance and outperforms GPT-4V. Moreover,

[3] https://huggingface.co/datasets/ColumbiaNLP/V-FLUTE-test

understanding images at high resolution has been proven to be beneficial for capturing more fine-grained details of figurative language.

Limitations

To alleviate the training burden and reduce training costs, the image encoder was completely frozen. This may prevent the model from learning richer and more accurate knowledge of multimodal figurative language. Limited by the short duration of this task, we did not explore the impact of different generative models on model performance. In future work, we will optimize the different layers of the image encoder to find the optimal trade-off between performance and cost. Furthermore, we will replace the current generative model with several large language models such as Llama and Vicuna to enhance FigCLIP's generalization ability in understanding and explaining multimodal figurative language.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: A visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Irina Bigoulaeva, Rachneet Singh Sachdeva, Harish Tayyar Madabushi, Aline Villavicencio, and Iryna Gurevych. 2022. Effective cross-task transfer learning for explainable natural language inference with t5. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 54–60, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

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.

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.

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yuling Gu, Yao Fu, Valentina Pyatkin, Ian Magnusson, Bhavana Dalvi Mishra, and Peter Clark. 2022. Just-DREAM-about-it: Figurative language understanding with DREAM-FLUTE. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 84–93, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. 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.

Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, CHEN Bin, Chengru Song, Di ZHANG, Wenwu Ou, et al. 2023. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. In *The Twelfth International Conference on Learning Representations*.

Jeongyeon Kim, Sangho Suh, Lydia B Chilton, and Haijun Xia. 2023. Metaphorian: Leveraging large language models to support extended metaphor creation for science writing. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, pages 115–135.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Khoa Thi-Kim Phan, Duc-Vu Nguyen, and Ngan Luu-Thuy Nguyen. 2022. NLP@UIT at FigLang-EMNLP 2022: A divide-and-conquer system for shared task on understanding figurative language. In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 150–153, Abu Dhabi, United Arab Emirates (Hybrid). 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, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. 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 Dataset. https://huggingface.co/datasets/ColumbiaNLP/V-FLUTE. Dataset associated with the paper "V-FLUTE: Visual Figurative Language Understanding with Textual Explanations".

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.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.