Multidimensional Difficulty Metric for Multimodal Sarcasm Explanation

Anonymous ACL submission

Abstract

Multimodal Sarcasm Explanation (MuSE) is a new yet challenging task, which aims at generating natural language explanations for sarcasm in social media image-text pairs. MuSE can further enhance sarcasm understanding and has attracted increasing research interest. Previous works design cross-modal attention or multi-source semantic graphs and achieve promising performance. However, these works either ignore the semantic gap between visual features and textual decoder or introduce complex graph constructions, which limits their practical applicability and scalability for real-world scenarios. Furthermore, they treat each sample equally during training, overlooking the impact of the inherent differences in sample difficulty. In this paper, we propose a novel multidimensional sample difficulty metric approach with the Multimodal Large Language Model (MLLM) for MuSE. Leveraging the feature alignment and innate knowledge of MLLM, we can achieve better cross-modal alignment without complicated processes. Using the multidimensional difficulty metric to measure the difficulty of image-text pairs, we enable MLLM to learn from simple to difficult samples, mitigating the impact of samples of varying difficulty on training, and preventing local optima. Experimental results on a publicly released dataset MORE demonstrate that our method is effective and achieves state-of-the-art performance.

1 Introduction

Sarcasm is a linguistic phenomenon where the literal meaning is contradictory to the actual intent of speakers. Sarcasm detection aims to identify the actual sentiments of users and can be widely applied in various scenarios such as opinion mining (Pang et al., 2008; Riloff et al., 2013) and social media analysis (Tsur et al., 2010). In the multimodal domain, multimodal sarcasm detection (Cai et al., 2019) focuses on analyzing the incongruity in image-text pairs to detect underlying sarcasm.

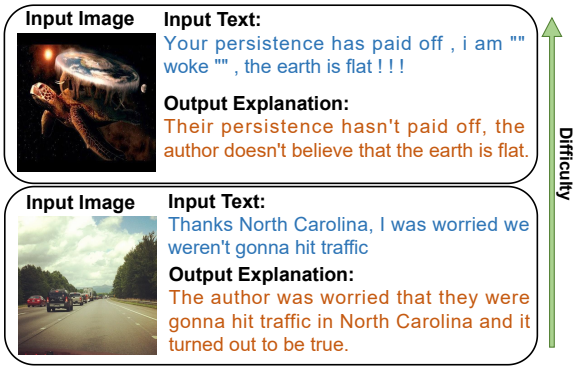

Figure 1: Examples of MuSE. The bottom example only requires identifying a traffic jam in the image. The upper harder example requires the prior knowledge that 'the Earth is round' to infer the subtly hidden sarcasm.

Although many works (Xu et al., 2020; Liang et al., 2021, 2022; Liu et al., 2022; Qin et al., 2023) provide accurate sarcasm classification results, the lack of corresponding explanations for why they are sarcastic makes the results relatively superficial for further sarcasm understanding. Thus Multimodal Sarcasm Explanation (MuSE) (Desai et al., 2022) aims to provide natural language explanations for given sarcastic image-text pairs to reveal the intended sarcasm, which has increasingly attracted research attention (Desai et al., 2022; Jing et al., 2023). Examples of MuSE are shown in Figure 1.

Previous works effectively inject the visual features into MuSE models. For instance, Desai et al. (2022) incorporates the image and text features through cross-modal attention in the Transformer encoder and generates explanations by a BART (Lewis et al., 2020)-based auto-regressive decoder. Jing et al. (2023) uses image object-level metadata, an external knowledge base, and a multi-source semantic graph for sarcasm reasoning. Despite their effectiveness, they either overlook the semantic gap between visual features and the textual decoder or rely on complex graph constructions

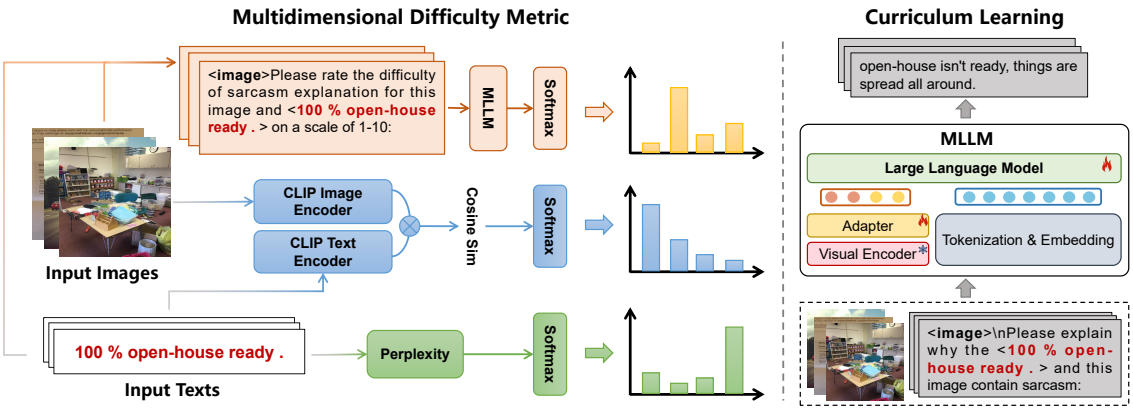

Figure 2: The overview of multimodal curriculum learning. First, we apply the multidimensional difficulty metric to the image-text pairs, then rank them by difficulty to enable MLLM to learn from easy to hard samples.

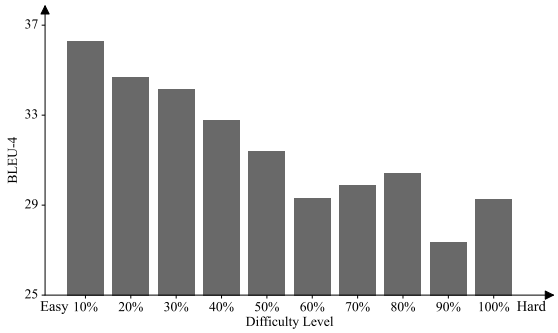

Figure 3: The average BLEU-4 score of previous methods on the MORE test set with increasing difficulty[1].

and the extra knowledge base, which limits their applicability and scalability for real-world scenarios. Moreover, they treat all samples equally during training and ignore the essential impact of different samples with different difficulties. As shown in Figure 3, the performance of MuSE models declines with the increasing difficulty of samples.

Inspired by human learning behavior, i.e., curriculum learning (Bengio et al., 2009), we propose a novel approach integrating multidimensional sample difficulty metric with the Multimodal Large Language Model (MLLM) for MuSE. By leveraging the inherent knowledge of MLLM and sufficient cross-modal alignment, we can achieve better alignment without cumbersome processes. The multidimensional difficulty metric includes requiring the MLLM itself to assess the difficulty of samples, measuring image-text consistency, and evaluating text complexity. Based on the total difficulty, we enable the MLLM to learn from easy to hard samples for better sarcasm understanding and achieving global optima. In summary, our contributions are as follows:

- We design the multidimensional difficulty metric to measure the difficulty of image-text pairs. This helps the MuSE model learn from easy to hard samples, reducing the impact of variable difficulty during training.
- We propose using the MLLM for MuSE, achieving better cross-modal alignment without complex operations.
- Experimental results on a public dataset MORE demonstrate that our method is effective and achieves state-of-the-art performance.

2 Methodology

In this section, we first present the brief task formulation and describe the multi-dimensional metric to measure the difficulty of image-text pairs.

2.1 Task Formulation

Given image-text pairs $\langle v_i, t_i \rangle$, where v_i is the i-th image input and t_i is the i-th text input caption. The multimodal sarcasm explanation model needs to generate the corresponding sarcasm explanation.

2.2 Difficulty Measurement

We design the multidimensional difficulty metric, which consists of MLLM SelfAssessment, Text-Image Consistency, and Textual Difficulty.

MLLM SelfAssessment Considering the powerful capabilities of MLLM, we first require the MLLM to self-assess the difficulty of the given samples. As shown in Figure 2, MLLM assigns a score from 0 to 10 to evaluate the difficulty of explaining the sarcasm in the current image-text pair.

[1]The difficulty is $\mathcal{D}_{total}$, which is obtained in Section 2.2.

Text-Image Consistency Sarcasm often exists in slight semantic mismatches, thus the difference in semantic consistency between the image and text can be used to measure the difficulty of a sample. Specifically, we use CLIPScore (Hessel et al., 2021) for this measurement. Given the image-text pairs, we obtain the embeddings from the visual and textual encoder of CLIP (Radford et al., 2021):

$$\text{Emb}_v = \text{CLIP}_{\text{vis}}(V) \quad (1)$$

$$\text{Emb}_t = \text{CLIP}_{\text{text}}(T) \quad (2)$$

where Emb_v and Emb_t are the visual and textual embeddings of images V and texts T.

Then we measure the text-image consistency from the cosine similarity of Emb_v and Emb_t:

$$\mathcal{D}_{TIS} = cos(\text{Emb}_v, \text{Emb}_t) \quad (3)$$

Textual Difficulty Given that MuSE requires to generate natural language explanations, we employ perplexity as the metric to measure the difficulty of the input text:

$$\mathcal{D}_{ppl} = \left(\prod_{i=1}^{N} \frac{1}{P(w_i | w_1, \ldots, w_{i-1})} \right)^{\frac{1}{N}} \quad (4)$$

where N is the length of the text and w_i is the i-th word.

Finally, we normalize the three above difficulties separately using softmax, and then aggregate them.

$$\mathcal{D}^i_{total} = \frac{e^{\mathcal{D}^i_{self}}}{\sum_j e^{\mathcal{D}^j_{self}}} + \frac{e^{\mathcal{D}^i_{TIS}}}{\sum_j e^{\mathcal{D}^j_{TIS}}} + \frac{e^{\mathcal{D}^i_{ppl}}}{\sum_j e^{\mathcal{D}^j_{ppl}}} \quad (5)$$

where $\mathcal{D}^i_{total}$ is the total difficulty of i-th image-text pair. Then We rank the image-text pairs based on $\mathcal{D}_{total}$ and enable MLLM to learn from easy to hard samples.

3 Experiments

3.1 Dataset and Metrics

We evaluate our method on the only public multi-modal sarcasm explanation dataset **MORE** (Desai et al., 2022), which contains sarcastic image-text pairs from various social media sites (Twitter[2], Instaram[3] and Tumblr[4]) and the corresponding sarcasm explanation for each pair is manually annotated, including $2,983$ for training, 175 for validation, and 352 for testing. Each sample of MORE is a triplet of $\langle image, caption, explanation \rangle$.

[2] https://twitter.com/
[3] https://www.instagram.com/
[4] https://www.tumblr.com/

Following previous works (Desai et al., 2022; Jing et al., 2023), we adopt BLEU-{1,2,3,4} (Papineni et al., 2002), ROUGE-{1,2,L} (Lin, 2005), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2019) and Sentence-BERT (Reimers and Gurevych, 2019) to assess the performance of our model.

3.2 Experimental Settings

Based on LLaVA-1.5-7B (Liu et al., 2023), we use " CLIP-ViT-L-336px" as the vision encoder and "Vicuna-v1.5-7B" as the LLM. We use the same vision encoder to calculate text-image consistency, and the same LLM to calculate textual difficulty. We adopt Parameter-Efficient-Fine-Tuning (PEFT) for the training stage,i.e., LoRA (Hu et al., 2021), and inject the low-rank matrices as adapters into MLLM. The rank of the update matrices is 128 and the scaling factor of LoRA is 256. We freeze the vision encoder and fine-tune the vision-language adapter and LLM. The learning rate for the adapter is 2e-5 and the learning rate for LLM is 2e-4. The batch size is 12 and the training epoch is 2. In the first epoch, training proceeds from easy to difficult samples based on difficulty levels. In the second epoch, all the samples are randomly shuffled for training. All models are trained on 2 NVIDIA 3090Ti GPUs and tested on a single NVIDIA 3090Ti GPU.

3.3 Compared Methods

To valid the effectiveness of our method, we compare our method with the following existing methods: (1) **PGN** (See et al., 2017). A Pointer Generator Network. (2) **Transformer** (Vaswani et al., 2017) A text-based generation transformer model generates sarcasm explanations. (3) **MFFG-RNN** and **MFFG-Trans**. Two variations of MFFG (Liu et al., 2020) for video summarization. MFFG-RNN and MFFG-Trans use RNN and transformer as the decoder respectively. (4) **M-Transf** (Yao and Wan, 2020). The multimodal Transformer model for machine translation. (5) **ExMore** (Desai et al., 2022). A BART-based multimodal sarcasm explanation model with cross-modal attention. (6) **TEAM-w/o-Know** and **TEAM** (Jing et al., 2023). The previous SOTA method, which is a BART-based method utilizing the object-level meta-data and external knowledge base for multi-source semantic graph construction. TEAM-w/o-Know means TEAM that does not use the knowledge base.

Following previous works (Desai et al., 2022;

Model	BLEU				ROUGE			METEOR	BERTScore			SentBERT
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
(a) All samples												
PGN	17.54	6.31	2.33	1.67	16.00	17.35	6.90	15.06	84.80	85.10	84.90	49.42
Transformer	11.44	4.79	1.68	0.73	15.90	17.78	5.83	9.74	83.40	84.90	84.10	52.55
MFFG-RNN	14.16	6.10	2.31	1.12	16.21	17.47	5.53	12.31	81.50	84.00	82.70	44.65
MFFG-Transf	13.55	4.95	2.00	0.76	15.14	16.84	4.30	10.97	81.10	83.80	82.40	41.58
M-Transf	14.37	6.48	2.94	1.57	18.77	20.99	6.98	12.84	86.30	86.20	86.20	53.85
ExMore	19.26	11.21	6.56	4.26	25.23	27.55	12.49	19.16	88.30	87.50	87.90	59.12
TEAM-w/o-Know	52.63	42.42	35.80	30.91	48.67	49.28	33.18	48.53	90.90	91.40	91.10	71.58
TEAM	55.32	45.12	38.27	33.16	50.58	51.72	34.96	50.95	91.80	91.60	91.70	72.92
Ours	**58.82**†	**49.43**†	**43.16**†	**38.38**†	**56.65**†	**59.63**†	**42.77**†	**52.26**†	**92.49**†	**92.03**†	**92.24**†	**79.32**†
(b) Non-OCR samples												
PGN	17.87	6.37	1.92	1.26	16.43	17.80	6.92	15.62	84.70	85.20	84.90	48.77
Transformer	11.65	5.65	1.73	0.69	16.16	17.41	6.26	10.13	83.60	85.10	84.30	48.40
MFFG-RNN	15.43	6.82	2.46	1.33	17.40	18.61	5.71	12.98	81.60	84.30	82.90	42.72
MFFG-Transf	13.28	5.35	1.49	0.26	14.90	16.80	4.35	11.19	81.30	84.00	82.60	41.68
M-Transf	14.91	6.90	2.66	0.83	19.34	21.05	7.08	13.91	86.50	86.30	86.40	51.77
ExMore	19.47	11.69	6.82	4.27	24.92	27.12	12.12	19.20	88.30	87.60	88.00	56.95
TEAM-w/o-Know	53.43	43.41	36.77	31.78	49.72	51.12	34.78	49.24	91.50	91.90	91.80	71.62
TEAM	56.45	46.34	39.58	34.34	52.79	53.81	36.78	51.62	92.40	92.90	92.30	73.35
Ours	**60.72**†	**51.29**†	**44.70**†	**39.49**†	**59.44**†	**61.83**†	**44.94**†	**54.91**†	**92.83**†	**92.51**†	**92.65**†	**79.92**†
(c) OCR samples												
PGN	17.19	6.08	2.49	1.79	15.55	16.92	6.76	14,64	84.90	84.90	84.90	49.53
Transformer	10.68	4.01	1.49	0.71	15.04	17.25	5.32	8.99	83.20	84.70	83.90	53.94
MFFG-RNN	12.18	4.92	1.73	0.88	14.01	15.18	4.56	10.64	81.20	83.70	82.40	45.91
MFFG-Transf	12.87	4.12	1.69	0.62	14.20	15.54	3.53	9.70	81.00	83.60	82.30	41.13
M-Transf	14.06	6.25	3.22	2.28	18.42	21.04	7.01	12.06	86.20	86.10	86.10	55.66
ExMore	19.40	11.31	6.83	4.76	25.66	28.02	12.10	19.15	88.20	87.50	87.90	60.82
TEAM-w/o-Know	51.91	41.51	34.85	29.85	47.53	49.00	32.77	47.94	90.50	91.00	90.70	71.43
TEAM	52.88	43.08	36.81	32.34	48.46	49.68	33.83	49.25	90.90	90.00	90.80	71.93
Ours	**57.63**†	**48.24**†	**42.13**†	**37.58**†	**54.53**†	**57.73**†	**40.88**†	**49.95**†	**92.18**†	**91.59**†	**91.86**†	**78.51**†

Table 1: Experimental results on MORE. † means our method outperforms TEAM significantly with $p < 0.001$.

Jing et al., 2023), we conduct the performance comparison of different methods across three dataset configurations: a) on all samples, b) only on Non-OCR samples, and c) only on OCR samples. OCR samples denote the samples whose images contain embedded texts, while Non-OCR samples do not.

3.4 Main Results

As shown in Table 1, our method significantly outperforms the previous state-of-the-art methods across different dataset configurations on all metrics, demonstrating the superiority of our approach. Furthermore, our method performs better on Non-OCR samples than on OCR samples, which might be attributed to the inherent ability of MLLM to understand images. The introduction of OCR information could have interfered with the model's understanding of textual sarcasm.

3.5 Ablation Study

The results of the ablation study are shown in Table 3. Without our multidimensional metric, the model performs the worst, which demonstrates the effectiveness of our method. Among the three different dimensions of difficulty, the impact of $\mathcal{D}_{TIS}$ is the greatest, while the impact of $\mathcal{D}_{ppl}$ is the smallest. This indicates that despite the introduction of adapters to align image-text modal information, there is still room for further optimization in image-text understanding. Therefore, allowing the model to learn from simple sample pairs with high image-text consistency of $\mathcal{D}_{TIS}$ can better facilitate the learning of sarcasm in image-text pairs. While in the MLLM, the LLM component itself undergoes extensive unsupervised text training based on perplexity, hence the effect of $\mathcal{D}_{ppl}$ is not significant.

4 Conclusion

In this paper, we design the multidimensional difficulty metric to measure the difficulty of image-text pairs and apply the MLLM to MuSE. Specifically, we develop MLLM self-assessment, image-text consistency, and textual difficulty as the multidimensional difficulty metric. Based on the overall difficulty, we rank the samples and enable the MLLM to learn from easy to hard samples. Experimental results on a public dataset demonstrate that our method is effective and outperforms previous SOTA methods by a large margin.

Limitations

Our method is constrained by the foundational performance of LLM, the visual encoder, and the adapter themselves.

Ethics Statement

We affirm that our work here does not exacerbate the biases already inherent in the large language models and does not have ethics problems.

References

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Yitao Cai, Huiyu Cai, and Xiaojun Wan. 2019. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2506–2515, Florence, Italy. 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.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. 2023. Multi-source semantic graph-based multimodal sarcasm explanation generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11349–11361, Toronto, Canada. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Bin Liang, Chenwei Lou, Xiang Li, Lin Gui, Min Yang, and Ruifeng Xu. 2021. Multi-modal sarcasm detection with interactive in-modal and cross-modal graphs. In *Proceedings of the 29th ACM international conference on multimedia*, pages 4707–4715.

Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. 2022. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1767–1777. Association for Computational Linguistics.

C Lin. 2005. Recall-oriented understudy for gisting evaluation (rouge). *Retrieved August*, 20:2005.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Hui Liu, Wenya Wang, and Haoliang Li. 2022. Towards multi-modal sarcasm detection via hierarchical congruity modeling with knowledge enhancement. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1845, Online. Association for Computational Linguistics.

Bo Pang, Lillian Lee, et al. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Libo Qin, Shijue Huang, Qiguang Chen, Chenran Cai, Yudi Zhang, Bin Liang, Wanxiang Che, and Ruifeng Xu. 2023. MMSD2.0: Towards a reliable multi-modal sarcasm detection system. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10834–10845, Toronto, Canada. 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.

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

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. Icwsm—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 4, pages 162–169.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020. Reasoning with multimodal sarcastic tweets via modeling cross-modality contrast and semantic association. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3777–3786, Online. Association for Computational Linguistics.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. 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*.

A Statistics of MORE

The statistics of the MORE dataset are shown in Table 2. This dataset contains $3,510$ triplets in the form of $\langle image, caption, explanation \rangle$, including $2,983$ for training, 175 for validation, and 352 for testing.

MORE	Samples	Input Avg.L	Explanation Avg.L
Train	2,983	19.75	15.47
Val	175	18.85	15.39
Test	352	19.43	15.08
Total	3,510	19.68	15.43

Table 2: The statistics of MORE datasets. Input Avg.L denotes the average length of input text. Explanation Avg.L denotes the average length of output explanation.

B Multidimensional Difficulty Metric Benefit

As shown in Figure 4, our method outperforms TEAM across all difficulty levels, especially in the latter half under hard conditions. This demonstrates the effectiveness of our proposed multidimensional difficulty metric.

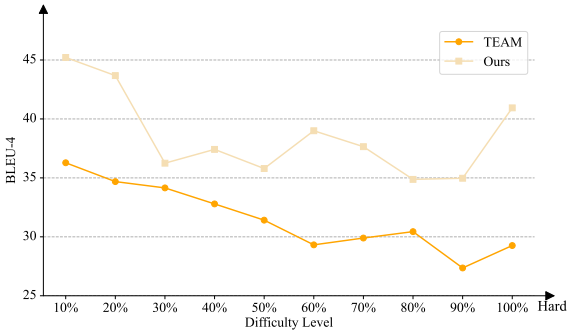

Figure 4: The average BLEU-4 score of our method and TEAM (Jing et al., 2023) at different difficulties on the MORE test set.

C Case Study

As shown in Figure 5, in the upper case, our method achieves accurate sarcasm explanation "anxiety isn't cured by someone saying 'don't be anxious'.", which is consistent with the target explanation. While TEAM (Jing et al., 2023) fails to understand the sarcasm and outputs "the author's got a miracle worker.". In the bottom case, our output explanation precisely describes the author's drawing ability as 'not good,' whereas TEAM directly states that the author cannot draw a picture.

Model	BLEU				ROUGE			METEOR	BERTScore			SentBERT
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
(a) All samples												
Ours	**58.82**	**49.43**	**43.16**	**38.38**	**56.65**	**59.63**	**42.77**	**52.26**	**92.49**	92.03	**92.24**	79.32
w/o $\mathcal{D}_{self}$	57.51	48.18	41.91	37.17	56.09	59.19	42.12	51.93	92.01	92.62	91.98	79.85
w/o $\mathcal{D}_{TIS}$	57.74	47.19	41.75	36.84	55.78	58.11	41.33	51.81	92.44	91.98	92.19	78.89
w/o $\mathcal{D}_{ppl}$	58.13	49.22	43.03	38.12	56.45	59.44	42.48	52.15	92.44	92.02	92.21	79.23
w/o $\mathcal{D}_{ppl}, \mathcal{D}_{TIS}, \mathcal{D}_{self}$	57.92	47.83	41.21	36.18	54.63	56.95	39.72	50.54	92.01	91.74	91.85	78.43
(b) Non-OCR samples												
Ours	**60.72**	**51.29**	**44.70**	**39.49**	**59.44**	**61.83**	**44.94**	**54.91**	92.83	92.51	92.65	79.92
w/o $\mathcal{D}_{self}$	59.14	49.59	42.90	37.71	58.60	61.35	44.32	54.84	92.95	92.55	92.73	80.68
w/o $\mathcal{D}_{TIS}$	59.58	49.67	42.78	37.40	57.91	60.56	42.55	54.33	92.60	92.41	92.48	79.13
w/o $\mathcal{D}_{ppl}$	59.96	51.13	44.58	39.16	59.19	61.22	44.19	54.78	92.73	92.58	92.64	80.14
w/o $\mathcal{D}_{ppl}, \mathcal{D}_{TIS}, \mathcal{D}_{self}$	60.05	50.11	43.12	37.65	57.70	59.29	42.58	53.67	92.41	92.24	92.30	78.41
(c) OCR samples												
Ours	**57.63**	**48.24**	42.13	37.58	**54.53**	57.73	40.88	**49.95**	92.18	**91.59**	91.86	78.51
w/o $\mathcal{D}_{self}$	56.76	47.66	41.69	37.25	54.19	57.67	40.78	49.85	92.38	91.52	91.93	78.79
w/o $\mathcal{D}_{TIS}$	56.75	47.44	41.28	36.68	54.31	57.41	40.53	49.81	92.30	91.58	91.92	78.26
w/o $\mathcal{D}_{ppl}$	56.99	47.78	42.74	38.27	54.39	57.90	41.02	50.15	92.16	91.53	91.82	78.30
w/o $\mathcal{D}_{ppl}, \mathcal{D}_{TIS}, \mathcal{D}_{self}$	56.46	46.24	39.89	35.17	52.21	54.96	37.40	47.70	91.67	91.27	91.44	77.86

Table 3: Ablation study on MORE dataset.

Figure 5: Two specific cases from the test set of MORE. Our results accurately explain the sarcasm, whereas TEAM (Jing et al., 2023) does not.

D Ablation Study Results

The experimental results of the ablation study are shown in Table 3.