Difficulty-Based Training Strategy with MLLMs 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 different effects of samples at different levels of difficulty. In this paper, we propose a novel **M**ulti**D**imensional **S**ample **D**ifficulty (MDSD) based training strategy with the Multimodal Large Language Models (MLLMs) for MuSE. Leveraging the multidimensional sample difficulty of image-text pairs, we enable MLLMs to learn from easy to hard samples in the training stage, mitigating the impact of samples of varying difficulty and preventing local optima. We can achieve better cross-modal alignment without complicated procedures based on the alignment and innate knowledge of MLLMs. Experimental results on two open-source MLLMs on a publicly released dataset MORE demonstrate that MDSD can further enhance MLLMs and achieve 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. Although many works (Xu et al., 2020b; 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 classification results relatively superficial for further sarcasm understanding. Therefore, Multimodal Sarcasm Explanation (MuSE) aims to provide natural language explanations for given sarcastic image-text pairs and has increasingly attracted research attention. Examples of MuSE are shown in Figure 1.

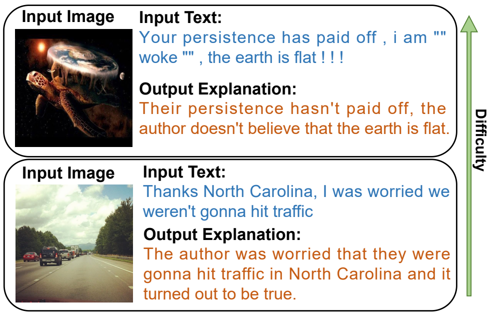

Figure 1: Examples of MuSE. The bottom example only requires identifying a traffic jam in the image. The upper harder example requires the prior knowledge that "the Earth is round" to infer the subtly hidden sarcasm.

Previous works on MuSE primarily focus on effectively injecting visual features into text-generation models. For instance, Desai et al. (2022) incorporates the image and text features through cross-modal attention in the Transformer (Vaswani et al., 2017) encoder and generates explanations by a BART (Lewis et al., 2020)-based auto-regressive decoder. Jing et al. (2023) further 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 heavily rely on complex graph constructions and the extra knowledge

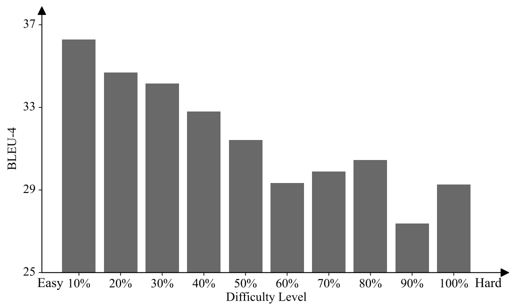

Figure 2: The average BLEU-4 score of previous methods on the MORE test set with increasing difficulty[1].

base, which limits their applicability and scalability for real-world scenarios. Moreover, they treat all samples equally during training without considering the different effects of samples at different levels of difficulty (Bengio et al., 2009; Xu et al., 2020a; Wang et al., 2022). As shown in Figure 2, the performance of MuSE models declines with the increasing difficulty of samples.

Inspired by the data-centric training of Large Language Models (LLMs) (Lin et al., 2024; McKinzie et al., 2024; Tirumala et al., 2024) and the wide applications of Multimodal Large Language Models (MLLMs) (Liu et al., 2023a; Dai et al., 2023), we propose a novel **M**ulti**D**imensional **S**ample **D**ifficulty (MDSD) based training strategy with MLLMs for MuSE. Specifically, we measure sample difficulty from three dimensions: MLLM Self-Assessment, Text-Image Consistency, and Textual Difficulty. We rank the samples according to the sample difficulty and enable MLLMs to learn from easy to hard during training, which can achieve a better sarcasm understanding. By leveraging the inherent knowledge and sufficient cross-modal alignment of MLLMs, we can achieve better alignment without cumbersome procedures. In summary, our contributions are as follows:

- We design the MDSD to measure the difficulty of image-text pairs. This helps MuSE models learn from easy to hard samples, reducing the impact of variable difficulty during training.
- We propose to use MLLMs for MuSE, which can achieve better cross-modal alignment without complex processes.
- Experimental results on a public dataset demonstrate that MDSD can enhance MLLMs and achieve state-of-the-art performance.

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

2 Methodology

In this section, we first present the brief task formulation of MuSE and describe the MDSD to measure the difficulty of image-text pairs, including MLLM SelfAssessment, Text-Image Consistency, and Textual Difficulty. Finally, we rank the image-text pairs based on the total difficulty and enable MLLMs to learn from easy to hard.

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. The multimodal sarcasm explanation model needs to generate the corresponding sarcasm explanation.

2.2 Difficulty Measurement

We design the multidimensional sample difficulty, which consists of MLLM SelfAssessment, Text-Image Consistency, and Textual Difficulty. We measure the samples from totally different dimensions and assume that they are independent of each other and each contributes to different extents.

2.2.1 MLLM SelfAssessment

Large language models have been found to perform a strong powerful self-decision-making capability, which has been applied in data optimization (Xu et al., 2023) and decision-making (Yang et al., 2023; Asai et al., 2023). In this paper, we aim to enable MLLMs to better understand the sarcasm in multimodal image-text pairs. Thus allowing MLLMs to self-score the difficulty of samples can distinguish samples of varying difficulty from the dimension of models.

As shown in Figure 3, MLLMs are required to assign a score from 0 to 10 to evaluate the difficulty $\mathcal{D}_{self}$ of explaining sarcasm in the given image-text pair. A higher score indicates a greater difficulty of the sample. For simple samples, the model can easily interpret the sarcasm, while more complex samples require further sarcasm understanding.

2.2.2 Text-Image Consistency

As for multimodal sarcasm, sarcasm often resides in the semantic differences between text and image pairs. For example, as shown in the bottom example of Figure 1, the text “I was worried we weren't gonna hit traffic” contrasts with the image of a traffic jam, thus creating a sarcastic expression. Additionally, current MLLMs typically use an adapter to connect the visual encoder with the

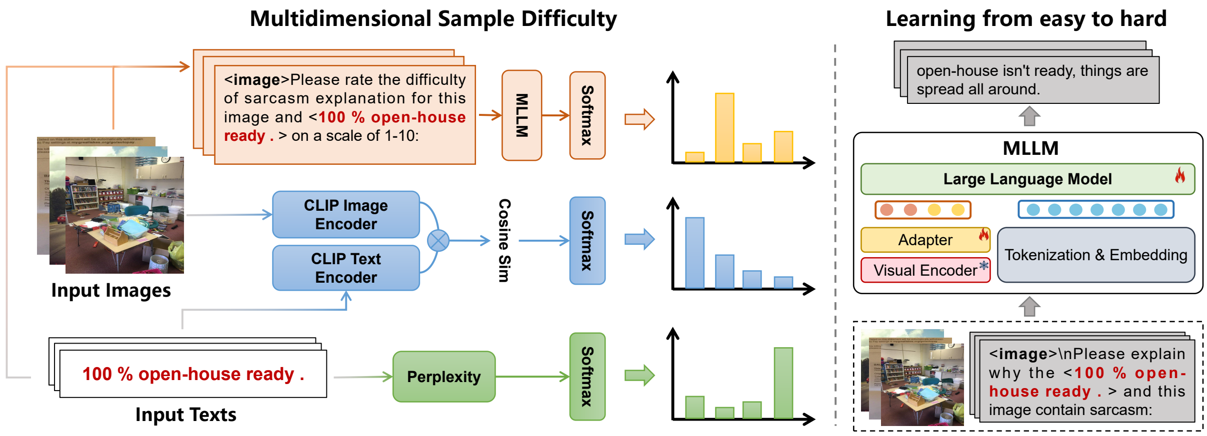

Figure 3: The overview of MDSD. First, we measure the multidimensional sample difficulty of the image-text pairs, then rank them by difficulty to enable MLLM to learn from easy to hard samples.

large language model. The degree of alignment between textual and visual modalities also influences the understanding of MLLMs of multimodal data. Therefore, we can assess the difficulty of image-text pairs from a cross-modal alignment dimension.

Specifically, we use CLIPScore (Hessel et al., 2021), which is designed to evaluate text-image similarity. Given image-text pairs, we obtain 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.

The higher the text-image consistency, the more similar the text and image, allowing models to explain sarcasm just by analyzing their differences. Conversely, the lower the text-image consistency, the greater the disparity between the text and image, requiring models to perform more extensive analysis to understand sarcasm. Thus we measure the difficulty of text image consistency $\mathcal{D}_{TIS}$ by the reciprocal of cosine similarity Emb_v and Emb_t:

$$\mathcal{D}_{TIS} = 1/cos(\text{Emb}_v, \text{Emb}_t) \quad (3)$$

The higher the $\mathcal{D}_{TIS}$, the harder the sample.

2.2.3 Textual Difficulty

For MLLMs, the core component is LLMs, and the text generation capability of LLMs could influence the final generation of sarcasm explanations in natural language. Considering that the commonly used loss function of LLM's pre-training stage is perplexity, which is also often used to measure textual difficulty (Marion et al., 2023; Muennighoff et al., 2024), we employ perplexity as the metric to measure the difficulty $\mathcal{D}_{ppl}$ 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 given text and w_i is the i-th word. The higher $\mathcal{D}_{ppl}$, the more difficult it is for LLMs to generate the required explanation.

2.2.4 Total Difficulty

After obtaining the difficulties from the different dimensions mentioned above, we need to combine them to get the final total difficulty. Our goal is to rank the samples based on their difficulty. Therefore, we only need to determine the relative difficulty of each sample within the overall samples. We also treat three dimensions of difficulty with equal importance. Specifically, we normalize the three above difficulties separately by softmax, and then sum them:

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

where $\mathcal{D}_{total}^i$ is the total difficulty of the given i-th image-text pair.

2.3 Optimization Object

Finally, we rank the image-text pairs based on $\mathcal{D}_{total}$ and enable MLLMs to learn from easy to hard samples. We construct the input for MLLMs by a pre-designed template for the given image-text pair, as shown in Figure 3.

Consistent with the loss calculation in auto-regressive LLMs, we only compute the cross-entropy loss for the response of MLLMs, i.e., the corresponding sarcasm explanation of the input:

$$\mathcal{L}_{ce} = \sum_{i=1}^{n-1} - \log p_\theta(y_{i+1} | \langle \mathrm{X}_v, \mathrm{X}_{instruction} \rangle, \mathrm{Y}_i) \quad (6)$$

where X_v is the visual input, $\mathrm{X}_{instruction}$ is the textual instruction. $\mathrm{Y}_i = \langle y_1, ..., y_i \rangle$ is the under generating response, n is the response token numbers, y_i is the i-th token of generated response and θ represents the parameters of MLLMs.

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 $\langle image, text, explanation \rangle$. Statistics of the MORE dataset are shown in Table 1.

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

3.2 Experimental Settings

We choose the LLaVA-1.5-7B (Liu et al., 2023a) and ShareGPT4V-7B (Chen et al., 2023b) as the base MLLM. We use the same vision encoder of MLLMs to calculate text-image consistency, and the same LLM to calculate textual difficulty. For the image inputs, we used "BLIP2-FlanT5-XL" to extract captions as inputs for LLMs. 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, we train the model from simple to difficult samples in the hope that it will better learn the sarcastic meanings in the samples. After the model has acquired a basic capability for sarcasm explanation in the first epoch, we randomize the samples in the second epoch to enhance the training's robustness. All models are trained on 2 NVIDIA 3090Ti GPUs for several hours and tested on a single NVIDIA 3090Ti GPU.

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

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 1: 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.

3.3 Compared Methods

To valid the effectiveness of our proposed method, we compare our method with the following existing methods following previous works (Desai et al., 2022; Jing et al., 2023):

(1) **PGN** (See et al., 2017). The Pointer Generator Network is a text-based generation model that utilizes a conventional decoder and a copy mechanism to directly copy words from the input text.

(2) **Transformer** (Vaswani et al., 2017). A text-based generation baseline generates sarcasm explanations with the transformer architecture.

(3) **MFFG-RNN** and **MFFG-Trans**. Two variations of MFFG (Liu et al., 2020), which is a multimodal-based generation model 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. M-Transf adopts the concatenation of text and image features for query and text representation for key and value in the cross-modal attention.

(5) **ExMore** (Desai et al., 2022). This method is designed for multimodal sarcasm explanation, which adopts BART (Lewis et al., 2020) as the model backbone and employs cross-modal attention in the encoder to inject the visual information into BART. Different from M-Transf, ExMore uses text representation for query and image representa-

Model	BLEU				ROUGE			METEOR	BERTScore			SentBERT
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
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
ChatGPT-zero-shot	12.64	6.83	4.40	3.01	18.56	19.18	6.51	25.39	83.62	86.77	85.15	60.85
ChatGPT-one-shot	26.20	15.34	9.91	5.99	28.98	30.22	11.46	28.61	86.95	87.84	87.38	63.19
ChatGLM2-6B	53.51	44.28	37.98	33.26	52.98	55.46	38.71	46.82	91.96	90.94	91.42	75.46
Llama2-7B	57.54	47.37	40.61	35.57	53.41	56.76	39.55	49.65	91.85	91.51	91.66	78.31
LLaVA1.5-7B	57.92	47.83	41.21	36.18	54.63	56.95	39.72	50.54	92.01	91.74	91.85	78.43
ShareGPT4V-7B	**59.07**	48.67	41.84	36.62	54.64	57.76	40.17	51.61	92.07	91.95	91.99	78.65
MDSD (LLaVA)	<u>58.82</u>†	**49.43**†	<u>43.16</u>†	<u>38.38</u>†	<u>56.65</u>†	<u>59.63</u>†	<u>42.77</u>†	<u>52.26</u>†	<u>92.49</u>†	**92.03**†	<u>92.24</u>†	<u>79.32</u>†
MDSD (ShareGPT4V)	58.33	<u>49.27</u>†	**43.16**†	**38.48**†	**57.05**†	**59.73**†	**43.19**†	**52.37**†	**92.57**†	<u>91.98</u>†	**92.25**†	**79.32**†

Table 2: Experimental results on MORE. † means our method outperforms the base MLLM (LLaVA, ShareGPT4V) significantly with $p < 0.05$. The best results are highlighted in bold, and the second-best results are underlined.

tion for key and value projections.

(6) **TEAM-w/o-Know** and **TEAM** (Jing et al., 2023). This is the previous SOTA method, which is a graph-based method utilizing the object-level meta-data and external knowledge base like ConceptNet (Speer et al., 2017) for multimodal sarcasm explanation. TEAM conducts the multi-source semantic graph construction process through the graph convolutional network in the BART encoder, and generates explanations in the BART decoder. TEAM-w/o-Know means TEAM that does not use external knowledge like ConceptNet.

We also compare our method with recent LLMs and MLLMs for a comprehensive comparison:

(7) **ChatGPT-zero-shot** and **ChatGPT-one-shot** [5]. A closed-source LLM for chat, as known as GPT-3.5-turbo. For the one-shot setting, we randomly choose an example of the training set of MORE as the demonstration.

(8) **ChatGLM2-6B** (Du et al., 2022). An open bilingual language model based on the general language model, with 6.2 billion parameters.

(9) **Llama-2-7B** (Touvron et al., 2023b). The foundation LLM pre-trained on 2 trillion tokens, with 7 billion parameters.

(10) **LLaVA-1.5-7B** (Liu et al., 2023a). An open-source MLLM adopts a multi-layer perceptron as an adapter to connect the vision encoder and LLM, which has 7 billion parameters.

(11) **ShareGPT4V-7B** (Chen et al., 2023b). An open-source MLLM with high-quality data annotated by GPT4V.

For a fair comparison, we apply MDSD on the two MLLMs, LLaVA and ShareGPT4V, to validate the effectiveness of our method. We utilize image captions as the visual inputs for the LLMs: ChatGPT, ChatGLM, and Llama.

3.4 Main Results

As shown in Table 2, our method achieves improvements on the majority of metrics across two MLLMs, demonstrating the effectiveness of MDSD. Additionally, MDSD (ShareGPT4V) outperforms MDSD (LLaVA), indicating the importance of the choice of base models.

As for ChatGPT, we believe the low performance is due to a gap between the content output of the alignment standard by GPT after Reinforcement Learning with Human Feedback (RLHF) and the human-annotated reference standard of MORE. As a result, the metrics calculated based on the GPT output and reference results are not high. The one-shot performs better than the zero-shot, indicating the effectiveness of in-context learning.

For LLMs such as ChatGLM and Llama, even when the input images are converted into textual captions, LLMs can still perform well on the multimodal task MuSE. Llama2-7B even surpasses the previous SOTA method TEAM. The performance differences between Llama and ChatGLM are attributed to the differences in pre-training data, which result in inherent performance differences between the base models.

[5] https://chatgpt.com/

Model	BLEU B1	BLEU B2	BLEU B3	BLEU B4	ROUGE RL	ROUGE R1	ROUGE R2	METEOR	BERTScore Pre	BERTScore Rec	BERTScore F1	SentBERT (Cosine)
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
ChatGPT-zero-shot	12.69	7.04	4.56	3.12	18.90	19.32	6.83	26.54	83.94	87.15	85.35	60.40
ChatGPT-one-shot	25.74	15.45	9.24	5.38	27.24	28.25	10.19	26.15	86.83	87.24	87.02	63.39
ChatGLM2-6B	54.60	45.21	38.38	33.16	55.08	57.36	40.57	49.02	92.20	91.36	91.75	74.79
Llama2-7B	59.34	49.22	42.15	36.73	54.93	58.04	40.84	51.88	92.13	92.01	92.05	73.35
LLaVA1.5-7B	60.05	50.11	43.12	37.65	57.70	59.29	42.58	53.67	92.41	92.24	92.30	78.41
ShareGPT4V-7B	59.64	49.11	41.83	36.05	56.59	58.89	41.12	53.37	92.26	92.38	92.30	79.68
MDSD (LLaVA)	**60.72**†	**51.29**†	**44.70**†	**39.49**†	59.44†	61.83†	44.94†	54.91†	92.83†	92.51†	92.65†	79.92†
MDSD (ShareGPT4V)	60.65†	51.15†	44.56†	39.29†	**59.87**†	**62.15**†	**45.24**†	**55.40**†	**92.93**†	**92.59**†	**92.74**†	**80.17**†
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
ChatGPT-zero-shot	12.56	6.78	4.35	2.90	18.54	18.90	6.63	24.68	83.61	86.49	85.01	61.55
ChatGPT-one-shot	25.58	14.85	9.24	5.38	27.24	28.25	10.19	26.15	86.83	87.24	87.02	63.39
ChatGLM2-6B	52.70	43.65	37.74	33.36	51.62	54.11	37.43	44.90	91.81	90.60	91.18	75.32
Llama2-7B	56.19	46.08	39.53	34.72	52.36	55.84	38.69	47.92	91.65	91.06	91.33	77.98
LLaVA1.5-7B	56.46	46.24	39.89	35.17	52.21	54.96	37.40	47.70	91.67	91.27	91.44	77.86
ShareGPT4V-7B	**58.72**	**48.48**	42.01	37.14	53.14	56.57	39.41	48.93	91.87	91.54	91.68	77.30
MDSD (LLaVA)	57.63†	48.24†	42.13†	37.58†	54.53†	**57.73**†	40.88†	**49.95**†	92.18†	**91.59**†	**91.86**†	78.51†
MDSD (ShareGPT4V)	56.75	48.03	**42.24**†	**37.96**†	**54.92**†	57.67†	**41.46**†	49.71†	**92.24**†	91.43	91.81†	**78.14**†

Table 3: Experimental results on MORE. † means our method outperforms the base MLLM (LLaVA, ShareGPT4V) significantly with $p < 0.05$. The best results are highlighted in bold, and the second-best results are underlined.

4 Analysis

4.1 Multidimensional Sample Difficulty Benefit

Taking LLaVA as the base MLLM, we also calculate the BLEU-4 scores based on different difficulties to further validate the effectiveness of MDSD. As shown in Figure 4, both MDSD (LLaVA) and LLaVA significantly outperform TEAM, indicating the promising performance of simply adopting MLLMs. Furthermore, MDSD can especially enhance MLLMs to learn difficult samples, e.g., the samples of 50%-80% difficulty level, demonstrating the effectiveness of MDSD. It is worth noting that the test set of the MORE dataset has a relatively small sample size. When samples are divided according to difficulty level, the number of samples for each difficulty level is different, which will lead to fluctuations in the calculated BLEU score curve, as shown in 4. However, the trend of the curve still allows us to draw the conclusions.

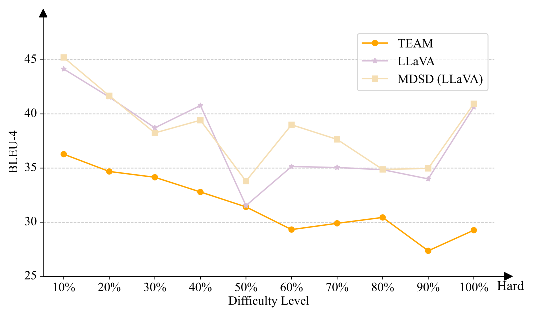

Figure 4: The average BLEU-4 score of our method, LLaVA (Liu et al., 2023a) and TEAM (Jing et al., 2023) at different difficulties on the MORE test set.

4.2 Non-OCR and OCR Settings

Following previous works (Desai et al., 2022; Jing et al., 2023), we also conduct the performance comparison of different methods across three dataset settings: all samples (as shown in Table 2), Non-OCR samples, and OCR samples. OCR samples denote the samples whose images contain embedded texts, while Non-OCR samples do not. As shown

Model	BLEU				ROUGE			METEOR	BERTScore			SentBERT
	B1	B2	B3	B4	RL	R1	R2		Pre	Rec	F1	(Cosine)
All samples												
MDSD (LLaVA)	**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
Non-OCR samples												
MDSD (LLaVA)	**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
OCR samples												
MDSD (LLaVA)	**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 4: Ablation study of MDSD (LLaVA) on MORE dataset. $\mathcal{D}_{self}$, $\mathcal{D}_{TIS}$, $\mathcal{D}_{ppl}$ are the difficulties in Sec 2.2.

in Table 3, most methods exhibit a performance decline on the OCR setting, indicating that the embedded text in the image poses a greater challenge for MuSE models to understand the image inputs, thereby increasing the difficulty of MuSE. Nevertheless, our proposed MDSD still achieves improvements on the majority of metrics in both Non-OCR and OCR settings, with the enhancements being particularly significant in the non-OCR setting.

4.3 Ablation Study

We also choose MDSD (LLaVA) to conduct the ablation study, as shown in Table 4. Without our multidimensional sample difficulty, i.e. pure LLaVA, 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 although adapters are introduced to align image and text representations, enabling MLLMs to understand multimodal data. For samples with low text-image consistency, MLLMs require more knowledge and analysis to understand the image and text to figure out the correct result. Therefore, allowing MLLMs to learn from easy samples with high image-text consistency can better facilitate the understanding of sarcasm in image-text pairs. Furthermore, allowing MLLMs to perform the self-assessment of sample difficulty $\mathcal{D}_{self}$ and requiring MLLMs to learn gradually can also boost MLLMs performance. As for $\mathcal{D}_{ppl}$, given that the widely used loss of LLMs pre-training is already perplexity and that LLMs have been trained on a large number of unsupervised samples, focusing solely on perplexity in MuSE may have minimal impact.

4.4 Case Study

To further demonstrate that ordering samples by difficulty level helps in training, we also conduct a case study, as shown in Figure 5. We select ShareGPT4V (Liu et al., 2023a) as the base MLLM and select cases from different difficulty levels.

From the perspective of MDSD difficulty levels, the low-level case just requires recognizing that "the baby" in the image is unhappy to explain the sarcasm. The medium-level case needs to infer that the image depicts a babysitting scenario and that the job is "tiring" rather than "great". The High-level case further requires analyzing that the windshield is covered with ice, which causes inconvenience for the author, and better integrating the textual description to infer that the author "hates winter". The difficulty of explaining the sarcasm of the sample is consistent with our difficulty level, demonstrating the effectiveness of MDSD.

For the analysis of output results, our method is closer to the target explanation compared to the base MLLM. At the low level, our result accurately identifies "the baby" to explain the sarcasm. At the mid level, our result correctly recognizes that babysitting is "exhausting". At the high level, our result also accurately identifies "hates the winter". This indicates that enabling MLLMs to learn from

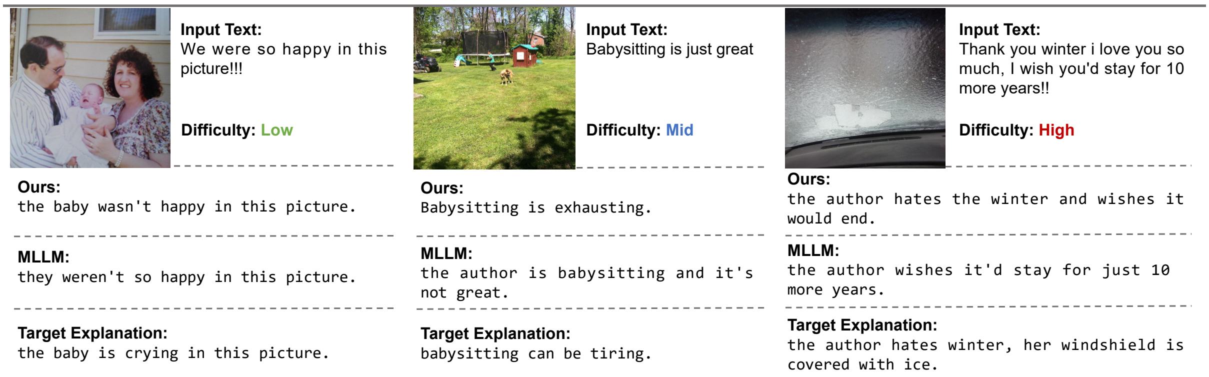

Figure 5: Case study on the test set of MORE, $\mathcal{D}_{total}$, $\mathcal{D}_{self}$, $\mathcal{D}_{TIS}$, $\mathcal{D}_{ppl}$ are at the same level.

easy to hard based on our proposed MDSD during training can significantly enhance their understanding of sarcasm, leading to better performance.

5 Related Work

5.1 Multimodal Sarcasm Detection and Explanation

Traditional sarcasm detection aims to identify user sentiments and detect sarcasm in textual data (Zhang et al., 2016; Tay et al., 2018; Babanejad et al., 2020). With the rise of multimodal data on social media, the focus has shifted to multimodal sarcasm detection (Schifanella et al., 2016; Cai et al., 2019). Further research on multimodal sarcasm detection has explored the integration of visual and textual data through various methods, such as decomposition and relation networks (Xu et al., 2020b), BERT-based models with modified attention mechanisms (Pan et al., 2020; Wang et al., 2020), graph neural networks (Liang et al., 2021, 2022), optimal transport (Pramanick et al., 2022), hierarchical framework with external knowledge (Liu et al., 2022), dynamic routing (Tian et al., 2023) and utilization of CLIP (Radford et al., 2021) from multi views (Qin et al., 2023).

However, the lack of corresponding natural language explanations for those sarcasm samples makes further understanding of sarcasm and its applications difficult. Thus Desai et al. (2022) further proposes the multimodal sarcasm explanation with a cross-modal BART-based model. Jing et al. (2023) adopts the graph neural network with extra meta-data and knowledge bases to enhance the performance of the multimodal sarcasm explanation model. Compared with those methods, our proposed methods can utilize MLLMs without extra data resources and enable MLLMs to learn from easy to hard for a better understanding of multimodal sarcasm samples.

5.2 Multimodal Large Language Models

In multimodal research, applying powerful LLMs (Touvron et al., 2023a,b) to multimodal tasks has garnered increasing attention. Early work, such as Frozen (Tsimpoukelli et al., 2021), achieved impressive performance by training a visual encoder to encode image inputs as a prefix in a frozen pre-trained language model. BLIP (Li et al., 2022) pre-trained a multimodal mixture of encoder-decoder model to enhance vision-language tasks further, while BLIP2 (Li et al., 2023) introduced a Q-former to efficiently align visual features to LLMs. Other studies, such as MiniGPT4 (Zhu et al., 2023; Chen et al., 2023a), LLaVA (Liu et al., 2023a,b), and Qwen-VL (Bai et al., 2023), utilized adapters like linear layers or multi-layer perceptrons to align image features extracted from visual encoders like ViT (Dosovitskiy et al., 2020). ShareGPT4V (Chen et al., 2023b) adopts GPT4V-distilled data to construct a stronger MLLM based on LLaVA.

6 Conclusion

In this paper, we propose the MultiDimensional Sample Difficulty (MDSD) based training strategy with MLLMs for MuSE. Specifically, we develop MLLM self-assessment, image-text consistency, and textual difficulty as the multidimensional difficulty. We rank the samples based on the total difficulty and enable MLLMs to learn from easy to hard. Experimental results on two open-source MLLMs on a public dataset demonstrate that MDSD can boost MLLMs for MuSE and outperform previous SOTA methods by a large margin.

Limitations

Our method is constrained by the foundational performance of MLLMs themselves, such as the components of LLM, the visual encoder, and the adapter. Due to limited resources, we do not evaluate more recent larger MLLMs.

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

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

Nastaran Babanejad, Heidar Davoudi, Aijun An, and Manos Papagelis. 2020. Affective and contextual embedding for sarcasm detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 225–243, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023b. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

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.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

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.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

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.

Xinyu Lin, Wenjie Wang, Yongqi Li, Shuo Yang, Fuli Feng, Yinwei Wei, and Tat-Seng Chua. 2024. Data-efficient fine-tuning for llm-based recommendation. *arXiv preprint arXiv:2401.17197*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

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.

Max Marion, Ahmet Üstün, Luiza Pozzobon, Alex Wang, Marzieh Fadaee, and Sara Hooker. 2023. When less is more: Investigating data pruning for pretraining llms at scale. *arXiv preprint arXiv:2309.04564*.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. 2024. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv preprint arXiv:2403.09611*.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2024. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36.

Hongliang Pan, Zheng Lin, Peng Fu, Yatao Qi, and Weiping Wang. 2020. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1383–1392, 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.

Shraman Pramanick, Aniket Roy, and Vishal M Patel. 2022. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3930–3940.

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.

Rossano Schifanella, Paloma De Juan, Joel Tetreault, and Liangliang Cao. 2016. Detecting sarcasm in multimodal social platforms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 1136–1145.

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.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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.

Yuan Tian, Nan Xu, Ruike Zhang, and Wenji Mao. 2023. Dynamic routing transformer network for multimodal sarcasm detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2468–2480, Toronto, Canada. Association for Computational Linguistics.

Kushal Tirumala, Daniel Simig, Armen Aghajanyan, and Ari Morcos. 2024. D4: Improving llm pretraining via document de-duplication and diversification. *Advances in Neural Information Processing Systems*, 36.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212.

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.

Peiyi Wang, Liang Chen, Tianyu Liu, Damai Dai, Yunbo Cao, Baobao Chang, and Zhifang Sui. 2022. Hierarchical curriculum learning for AMR parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 333–339, Dublin, Ireland. Association for Computational Linguistics.

Xinyu Wang, Xiaowen Sun, Tan Yang, and Hongbo Wang. 2020. Building a bridge: A method for image-text sarcasm detection without pretraining on image-text data. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 19–29, Online. Association for Computational Linguistics.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020a. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, Online. Association for Computational Linguistics.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

Nan Xu, Zhixiong Zeng, and Wenji Mao. 2020b. 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.

Hui Yang, Sifu Yue, and Yunzhong He. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*.

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.

Meishan Zhang, Yue Zhang, and Guohong Fu. 2016. Tweet sarcasm detection using deep neural network.

In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2449–2460, Osaka, Japan. The COLING 2016 Organizing Committee.

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. *arXiv preprint arXiv:2304.10592*.