

Describe Images in a *Boring* Way: Towards Cross-Modal Sarcasm Generation

Jie Ruan, Yue Wu, Xiaojun Wan, Yuesheng Zhu
Peking University

ruanjie@stu.pku.edu.cn, {zaczywy, wanxiaojun, zhuysh}@pku.edu.cn

Abstract

Sarcasm generation has been investigated in previous studies by considering it as a text-to-text generation problem, i.e., generating a sarcastic sentence for an input sentence. In this paper, we study a new problem of cross-modal sarcasm generation (CMMSG), i.e., generating a sarcastic description for a given image. CMMSG is challenging as models need to satisfy the characteristics of sarcasm, as well as the correlation between different modalities. In addition, there should be some inconsistency between the two modalities, which requires imagination. Moreover, high-quality training data is insufficient. To address these problems, we take a step toward generating sarcastic descriptions from images without paired training data and propose an Extraction-Generation-Ranking based Modular method (EGRM) for CMMSG. Specifically, EGRM first extracts diverse information from an image at different levels and uses the obtained image tags, sentimental descriptive caption, and commonsense-based consequence to generate candidate sarcastic texts. Then, a comprehensive ranking algorithm, which considers image-text relation, sarcasticness, and grammaticality, is proposed to select a final text from the candidate texts. Human evaluation at five criteria on a total of 2100 generated image-text pairs and auxiliary automatic evaluation show the superiority of our method. Code and data are publicly available¹.

1. Introduction

Sarcasm is a phenomenon in which the literal sentiment of a text differs from its implied sentiment [41]. The use of sarcasm is found to be beneficial for increasing creativity and humor in both the speakers and the addressees in conversations [4]. Researches on sarcasm have an influence on downstream application tasks such as dialogue system and content creation. Over the years, studies have investigated sarcasm detection and textual sarcasm generation. Sarcasm detection aims to detect whether the in-

put data is sarcastic, which has been explored in some research work [13, 14, 19, 33]. However, research on sarcasm generation stays in textual (text-to-text) sarcasm generation [6, 18, 32, 35, 38, 49], that is, outputting sarcastic text for the input text. Till now, there is no work attempting to generate sarcastic texts for images, while enabling machines to perceive visual information and generate sarcastic text will increase the richness and funniness of content or conversation, and serve downstream applications such as multi-modal dialogue systems, content creation, virtual worlds, entertainment, role-playing, games, and dramas to make things interesting. In this study, we for the first time formulate and investigate a new problem of cross-modal sarcasm generation (CMMSG).

CMMSG is challenging as it should not only retain the characteristics of sarcasm but also make the information generated in a different modality related to the original modality. In addition, there should be some inconsistency between the semantic information of the two modalities, which requires imagination and creativity. For example, the literal and intended meaning is reversed. The information of the two modalities should have the effect of enhancing or producing sarcasm. Sarcasm factors are defined as follows: 1) be evaluative, 2) be based on the inconsistency of the ironic utterance with the context, 3) be based on a reversal of valence between the literal and intended meaning, 4) be aimed at some target, and 5) be relevant to the communicative situation in some way [3, 4]. Moreover, there is insufficient high-quality cross-modal sarcasm training data, which makes CMMSG more difficult. Experiment on one of the baseline BLIP [23] demonstrates that the quality of the existing cross-modal sarcasm dataset [5] is too poor to be used to train supervised models to solve CMMSG problems.

To address these problems, we focus on generating sarcastic texts from images and propose an Extraction-Generation-Ranking based Modular method (EGRM) for unsupervised CMMSG (shown in Figure 2). We introduce to extract and obtain diverse image information at different levels through image tagging and sentimental descriptive captioning for generating sarcastic texts. A sarcastic texts generation module is proposed to generate a set of candidate

¹<https://github.com/EnablerRx/CMMSG-EGRM>

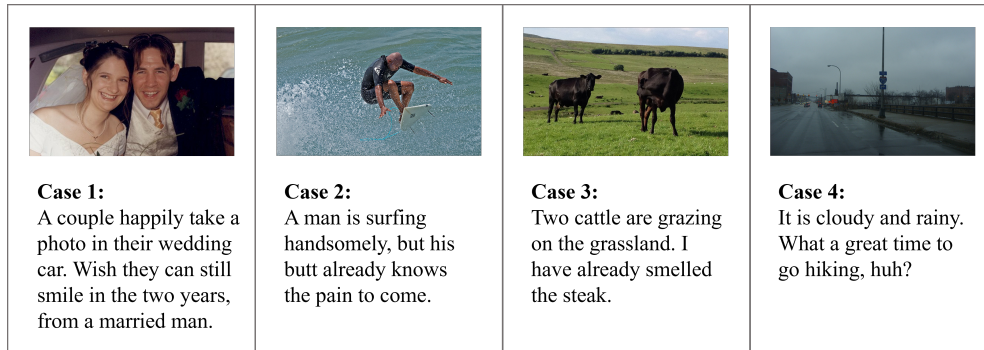


Figure 1. Sarcastic image-text pairs written by human. Case 1 satirizes the new couple may experience unhappiness in their marriage. Case 2 satirizes the man may fell. Case 3 satirizes the cattle may be killed for steak. Case 4 satirizes the bad weather when going hiking.

sarcastic texts. In the sarcastic texts generation module, we first reverse the valence (RTV) of the sentimental descriptive caption and use it as the first sentence. Then the cause relation of commonsense reasoning is adopted to deduce the consequence of the image information, and the consequence and image tags are used to generate a set of candidate sarcastic texts. As CMSG involves the evaluation from multiple perspectives, we propose a comprehensive ranking method that considers image-text relation, sarcasticness, and grammaticality to rank the candidate texts. Examples of sarcastic image-text pairs written by human are shown in Figure 1, where images are selected in MSCOCO [25].

The main contributions are as follows: 1) For the first time, we formulate the problem of cross-modal sarcasm generation and analyze its challenges. 2) We propose a non-trivial extraction-generation-ranking based modular method (EGRM) to address the challenging CMSG task. EGRM uses commonsense-based consequence and image tags to generate imaginative sarcastic texts, which makes the two modalities relevant and inconsistent to produce sarcasm. Moreover, we consider the performance of candidate sarcastic texts from multiple perspectives, including image-text relation, semantic inconsistency, and grammar, and propose a comprehensive ranking method that simultaneously considers the performance of candidate texts from multiple perspectives to select the best-generated text. EGRM doesn't rely on cross-modal sarcasm training data. 3) Human evaluation results show the superiority of EGRM in terms of sarcasticness, humor, and overall performance. Code and data are released.²

2. Related Work

2.1. Textual Sarcasm Generation

Research on Textual Sarcasm Generation is relatively preliminary. The limited amount of research on textual sarcasm generation is mainly divided into two categories, one is to generate a *sarcasm response* based on the in-

put utterance [18, 35], and the other is to generate a *sarcasm paraphrase* based on the input utterance [6, 32, 38]. Joshi et al. [18] introduced a rule-based sarcasm generation module named SarcasmBot. SarcasmBot implements eight rule-based sarcasm generators, each of which generates a kind of sarcasm expression. Peled and Reichart [38] proposed a novel task of sarcasm interpretation which generate a non-sarcastic utterance conveying the same message as the original sarcastic utterance. They also proposed a supervised sarcasm interpretation algorithm based on machine translation. However, it is impractical to train supervised generative models with deep neural networks due to the lack of large amounts of high-quality cross-modal sarcasm data. Therefore, we turn to unsupervised approaches. Mishra et al. [32] introduced a retrieval-based framework that is trained only using unlabeled non-sarcastic and sarcastic opinions. Chakrabarty et al. [6] presented a retrieve-and-edit-based framework to make reversal of valence and semantic incongruity with the context. Oprea et al. [35] proposed Chandler that generates sarcastic responses and explanations. However, these works mainly generate sarcastic text based on input utterance, and there is no existing research on cross-modal sarcasm generation. Enabling machines to perceive visual information and generate sarcasm information for communication will increase the richness and humor of communication and serve downstream tasks such as multi-modal dialogue system and content creation. Therefore, we focus on cross-modal sarcasm generation.

2.2. Image Captioning

Image Captioning is the task of describing the content of an image in words. Recent works on image captioning have concentrated on using the deep neural network to solve the MS-COCO Image Captioning Challenge³. CNN family is often used as the image encoder and the RNN family is used as the decoder to generate sentences [20, 45, 47]. Many methods have been proposed to improve the performance of

²Previous version of this paper: <https://arxiv.org/abs/2211.10992>

³<http://mscoco.org/dataset/#captions-challenge2015>

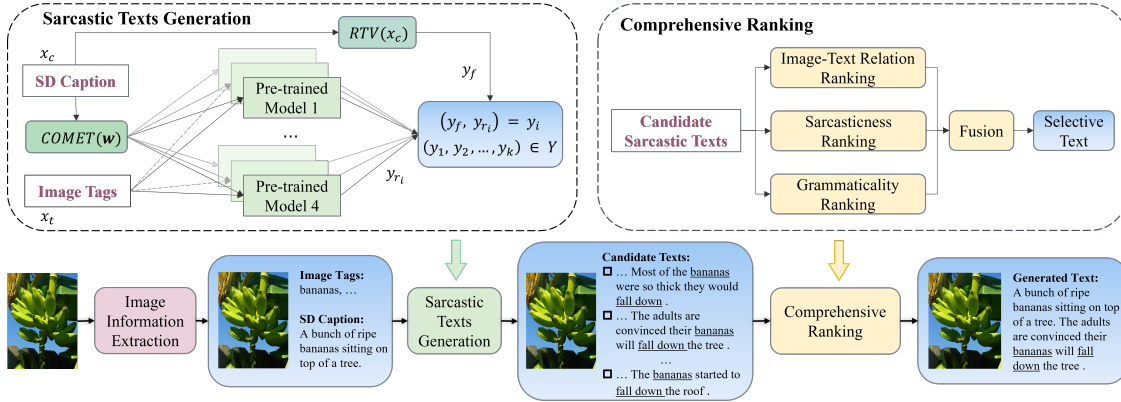


Figure 2. The overall framework of EGRM. EGRM consists of three modules: image information extraction, sarcastic texts generation, and comprehensive ranking. In the sarcastic texts generation module, RTV reverses the valence of the SD Caption. COMET is a commonsense reasoning method used to infer the consequence of the SD Caption.

image captioning. Previous work used reinforcement learning methods [26, 40], visual attention mechanism [1, 36, 46], contrastive or adversarial learning [9, 10], and transformer [8, 22, 28, 29, 44]. A slightly related branch of our research in image captioning is sentimental image captioning which generates captions with emotions. Mathews et al. [30] proposed SentiCap, a switching architecture with factual and sentimental caption paths, to generate sentimental descriptive captions. You et al. [48] introduced Direct Injection and Sentiment Flow to better solve the sentimental image captioning problem. Nezami et al. [34] proposed an attention-based model namely SENTI-ATTEND to better add sentiments to image captions. Li et al. [24] introduce an Inherent Sentiment Image Captioning method via an attention mechanism. However, cross-modal sarcasm generation involves creativity as well as correlations and inconsistencies among different modalities, existing image captioning methods cannot meet the requirement.

3. Methodology

Due to the low quality and insufficient quantity of existing cross-modal sarcasm training data, which is confirmed in the experimental results of the pre-trained supervised baseline BLIP, we focus on unsupervised cross-modal sarcasm generation. However, retrieval-based methods for generating sarcasm sentences are limited by the quality of the retrieval corpus and the ability of multi-keyword retrieval. The texts directly generated by rule-based methods are easily limited by the proposed rules and have worse performance on tasks requiring creativity and imagination like sarcastic texts generation. Therefore, we propose a modular cross-modal sarcasm generation method, which has a key component of constrained text generation and is able to generate more imaginative and creative sarcastic texts.

The overall framework of our proposed Extraction-Generation-Ranking based Modular method (EGRM) is

shown in Figure 2. And the pseudo-code for EGRM is shown in Algorithm 1. Given an image, EGRM generates a sarcastic text related to the input image. EGRM consists of three modules: image information extraction (lines 1-2 in Algorithm 1), sarcastic texts generation (lines 3-4 in Algorithm 1), and comprehensive ranking (lines 5-10 in Algorithm 1), as shown in Figure 2. The image information extraction module extracts and obtains diverse image information at different levels, including image tags and sentimental descriptive caption (SD Caption). In the sarcastic texts generation module, we first reverse the valence (RTV) of the sentimental descriptive caption and use it as the first sentence. Then the cause relation of commonsense reasoning is adopted to deduce the consequence of the image information, and the consequence and image tags are used to generate a set of rest texts via constrained text generation. The first sentence and each rest text are concatenated to form a candidate sarcastic text set. At last, we propose a comprehensive ranking module with multiple metrics (shown in Figure 2) to measure various aspects of the generated candidate texts and the highly ranked one is selected.

3.1. Image Information Extraction

As a cross-modal sarcasm generation task, it is crucial to extract and obtain important and diverse information from the input image that is useful for generating sarcastic texts. We obtain image tags x_t and sentimental descriptive caption x_c from the image. Particularly, a popular object detection method YOLOv5 [17] is adopted to detect objects in the image and record image tags. SentiCap [30], a switching recurrent neural network with word-level regularization, is used to generate sentimental descriptive image caption.

3.2. Sarcastic Texts Generation

As shown in the upper-left part of Figure 2, there are two branches in the sarcastic texts generation module. The top

Algorithm 1 EGRM

Input: image x **Output:** generated text $y = (y_f, y_r)$, where y_f is the first sentence and y_r is the rest text

- 1: Extract image tags $x_t \leftarrow \text{YOLOv5}(x)$
 - 2: Extract sentimental descriptive caption $x_c \leftarrow \text{SentiCap}(x)$
 - 3: Generate the first sentence $y_f \leftarrow \text{RTV}(x_c)$
 - 4: Generate a set of rest texts $(y_{r_1}, y_{r_2}, \dots, y_{r_k}) \leftarrow \text{SarcasticTextsGeneration}(x_t, x_c)$
 - 5: **for** $i = 1, 2, \dots, k$ **do**
 - 6: Calculate the Image-Text Relation score $p(x | [y_i, s_t]) \leftarrow \text{CLIPScore}(x, y_i)$
 - 7: Calculate the Sarcasm score $p(s_t | y_i) \leftarrow \text{FinetunedRoBERTa}(y_f, y_i)$
 - 8: Calculate the Grammaticality score $p(y_i) \leftarrow \text{PPL}(y_i)$
 - 9: Calculate the comprehensive ranking score $p_{crank}(y_i | x, s_t) \leftarrow p(x | [y_i, s_t]) p(s_t | y_i) p(y_i)$
 - 10: **end for**
 - 11: Select the rest texts with the highest comprehensive ranking score as rest text y_r of the generated text y
-

Algorithm 2 SarcasticTextsGeneration

Input: image tags x_t , sentimental descriptive caption x_c **Output:** a set of rest texts $Y = (y_{r_1}, y_{r_2}, \dots, y_{r_k})$

- 1: Extract verbs, nouns, adverbs and adjectives w from x_c
 - 2: Obtain commonsense-based consequence $c \leftarrow \text{COMET}(w)$
 - 3: Initialize $Y \leftarrow []$
 - 4: pretrainedModels $\leftarrow [\text{base} - \text{One} - \text{Billion} - \text{Word}, \text{base} - \text{Yelp}, \text{large} - \text{One} - \text{Billion} - \text{Word}, \text{large} - \text{Yelp}]$
 - 5: **for** pre in pretrainedModels **do**
 - 6: **for** $numKeywords$ in range(len(c), len(c) + len(x_t)) **do**
 - 7: $maskSentences.append(c)$
 - 8: $maskSentences.append(x_t[0 : numKeywords - len(c)])$
 - 9: $Y.append(\text{CBART}(maskSentences, pre))$
 - 10: **end for**
 - 11: **end for**
-

branch generates the first sentence y_f from the sentimental descriptive caption (SD Caption) x_c . The bottom branch generates a set of rest texts $(y_{r_1}, y_{r_2}, \dots, y_{r_k})$ from the given sentimental descriptive caption x_c and image tags x_t . k is the total number of generated texts. Concretely, we generate multiple rest texts by using different pre-trained models with different image tags and consequence collocations as input. The first sentence is then concatenated with each generated rest text to produce a set of candidate sarcastic texts Y , where each candidate text $y_i \in Y$. The pseudo-code of this module is shown in Algorithm 2.

The sarcastic texts generation method needs to satisfy the correlation between image and text and also the inconsistency of the two modalities. This means that the content of the generated text should be related to the image. At the same time, there is some inconsistency in the semantic information of the generated text with regard to the image, such as forming inversion or obtaining some contrast content, which is related to the image but not directly reflected by the image, through certain imagination and reasoning. Firstly, we obtain the first sentence y_f based on the SD cap-

tion generated from the input image to achieve image-text relevance. We reverse the valence (RTV) of the caption to make the text and image inconsistent. Considering that sarcasm usually occurs in positive sentiment towards a negative situation (i.e., sarcastic criticism) [6,21], we invert the negative sentiment expressed by the caption, so that the first sentence contains context with positive sentiment. Specifically, we obtain the negative score of the evaluative word from SentiWordNet [12] and use WordNet [31] to replace the evaluative words with its antonyms similar to the R^3 method [6]. We do nothing if there is no negative sentiment in the caption. To sum up, the first sentence is obtained as $y_f = \text{RTV}(x_c)$. For example, for a raining image, we may reverse the first sentence “a **bad** rainy day” to “a **good** rainy day”, which produces sarcasm and humor and may enhance sarcasm by the rest generated text.

The key to producing sarcasm is the reversal of valence between the literal and intended meaning as well as the relevance of the communicative situation. In the CMSG task, we should make some semantic inconsistency between the connotation expressed by the text and the real information shown by the image in the specific situation of the image. To achieve this goal, we propose to use the commonsense-based consequence inferred by information from the image modality and the image tags to generate the rest texts, which will be concatenated after the first sentence. The reason we use the image information to deduce the consequence c is that commonsense reasoning can infer the cause relation and the possible consequence in the scene shown in the image, making the intention of the sarcasm clearer and the effect of the sarcasm more intense. Taking the image of Case 2 in Figure 1 as an instance, commonsense reasoning result shows that information in the image may cause a **crash**. EGRM generates text “a man on a surfboard riding a wave in the ocean. He was cited as a reckless person in the surfboard incident due to an avoidable crash in 2006.” We may not feel sarcastic when we read the first sentence. However, we feel sarcastic and funny when we imagine a man riding

a wave and suddenly falls down from the surfboard which causes a crash. By using the commonsense-based consequence, the model is able to capture the deeper information contained in the image and imagine possible situations based on the commonsense-based consequence to generate more realistic sarcastic texts. For inferring commonsense-based consequence, we extract verbs, nouns, adverbs, and adjectives, which denote as \mathbf{w} , from the sentimental descriptive caption x_c and feed them to COMET to infer the consequence. Detailed information can be seen in these papers [2, 6, 43]. Therefore, the commonsense-based consequence c is obtained by $c = \text{COMET}(\mathbf{w})$. The process of inferring commonsense-based consequence is shown in lines 1-2 of Algorithm 2.

Using image tags makes the image and text more relevant and makes it clearer who caused the consequence. In this way, we can generate sarcastic texts related to the image and inconsistent with the real semantic content. For instance, both SC- R^3 and our method infer the consequence ‘‘crash’’ of the image of Case 2 in Figure 1. SC- R^3 retrieves sentences from the corpus according to the commonsense-based consequence and gets a sentence ‘‘The ceiling came down with a terrific crash.’’, which is irrelevant to the image. The result is not only non-ironic but also confusing. Our method considers image tags and the commonsense-based consequence, and the generated text ‘‘He was cited as a reckless person in the surfboard incident due to an avoidable crash in 2006.’’ has image-text correlation and inconsistency, which produce sarcasm.

To implement the cross-modal sarcastic texts generation module, we generate the rest texts based on a recently proposed constrained text generation method CBART [15]. For instance, given image tag ‘‘bananas’’ and consequence ‘‘fall down’’ as input, the model may generate ‘‘The adults are convinced their bananas will fall down the tree’’, which can be seen in Figure 2. As shown in the upper-left part of Figure 2 and Algorithm 2 lines 3-11, by using different numbers of tags, changing different pre-trained models, and using commonsense-based consequence inferred by information from the image modality, the sarcastic texts generation module can generate a variety of different sarcastic texts for selection. We use four pre-trained models to generate texts which are the base model initialized with BART-base model training on One-Billion-Word [7] dataset (base-One-Billion-Word), the base model initialized with BART-base model training on Yelp⁴ dataset (base-Yelp), the large model initialized with BART-large model training on One-Billion-Word dataset (large-One-Billion-Word), and the large model initialized with BART-large model training on Yelp dataset (large-Yelp). Different pre-trained models can generate diverse rest texts, making the candidate sarcastic texts more abundant. For more details about

⁴<https://www.yelp.com/dataset>

CBART, please read the original paper of CBART [15].

3.3. Comprehensive Ranking

Since EGRM needs to consider the performance of image-text pairs in terms of image-text relation, sarcasm, and grammaticality, it is necessary to comprehensively rank candidate texts to select the text with the best comprehensive performance. In the CMSG task, we need to convert the image to the text of the target sarcasm style s_t . Given an input image x , the conditional likelihood of the generated sarcastic text y is divided into three terms:

$$\begin{aligned} p(y | x, s_t) &= \frac{p(y, x, s_t)}{p(x, s_t)} \propto p(x, [y, s_t]) \\ &= p(x | [y, s_t]) p([y, s_t]) \quad (1) \\ &= \underbrace{p(x | [y, s_t])}_{\text{Image-Text Relation}} \underbrace{p(s_t | y)}_{\text{Sarcasticness}} \underbrace{p(y)}_{\text{Grammaticality}}, \end{aligned}$$

where $[\cdot]$ groups related terms (e.g., $[y, s_t]$) together. In the CMSG task, the first term of Equation 1, $p(x | [y, s_t])$ measures the *Image-Text Relation* between the input image x and the output target text y . It calculates the correlation between the image and the generated text. The second term, $p(s_t | y)$, is a measure of *Sarcasticness*. The third term, $p(y)$, measures the overall *Grammaticality* of the output text y , which also shows the fluency of the generated text.

Finally, we rank our k candidate sarcastic texts generated in the cross-modal sarcastic texts generation module according to the decomposition in Equation 1. For the i -th candidate text y_i , the ranking score is computed as:

$$p_{crank}(y_i | x, s_t) \propto p(x | [y_i, s_t]) p(s_t | y_i) p(y_i), \quad (2)$$

where p_{crank} represents the comprehensive ranking probability for y_i . We choose the size of candidate sarcastic texts k by conducting experiments on the validation data and we find that CMSG has good performance when k is 36.

All that remains is how to calculate each term in Equation 2. To calculate the first term, image-text relation, we adopt a reference-free metric CLIPScore [16] which measures the cosine similarity between the visual CLIP [39] embedding v of the image x and the textual CLIP embedding e of a candidate text y_i . We presume $p(x | [y_i, s_t]) = \text{CLIPScore}(x, y_i) = w \cdot \max(\cos(e, v), 0)$ and w is 2.5 following the settings of CLIPScore. For calculating the second term, sarcasm, we use semantic incongruity ranking [6] which fine-tunes RoBERTa-large [27] on the MultiNLI [42] dataset to calculate the contradictory score between the first sentence of the image description after reversing the valence and the rest text. For the third term, we use perplexity (PPL) to calculate the existing probability of the texts, and we use BERT [11] to calculate the probability.

4. Experimental Setup

4.1. Dataset

As we do not need parallel cross-modal sarcasm data for training, we conduct the experiment on a testing subset of 503 images in the SentiCap [30] dataset, which uses images from the MSCOCO [25] validation partition and adds sentiment captions to those images. Automatic metrics (see Section 4.3) for each method are calculated on these 503 images. Considering the time and economic cost of human evaluation, we randomly selected 150 images as the test set for human evaluation. Since there are fourteen systems, the human evaluation is conducted on a total of 2100 image-text pairs. Datasets for training pre-trained models for the sarcastic texts generation module are the One-Billion-Word [7] and Yelp⁵. One-Billion-Word is a public dataset for language modeling produced from the WMT 2011 News Crawl data. The Yelp dataset contains business reviews on Yelp.

4.2. Compared Methods

As CMSG is a new task, we design three comparison methods, and the first two methods do not need parallel cross-modal sarcasm training data while the third one relies on such data for training. The comparison methods are as follows. **SC- R^3** : We use the R^3 released by Chakrabarty et al. [6] as it is the state-of-the-art textual sarcasm generation system to transform input texts into sarcastic paraphrases. We input the captions generated by SentiCap [30] to R^3 to generate sarcastic texts. **SC-MTS**: We input the captions generated by SentiCap to MTS [32] to generate sarcastic texts. **BLIP**: This is a pre-trained image captioning model [23], and we fine-tune it on the parallel cross-modal sarcasm dataset proposed by Cai et al. [5]. It is considered a representative of the supervised methods.

To explore the effectiveness of main parts of EGRM, we ablate some important components of EGRM and evaluate their performance. These are termed as: **EGRM-woCS**: EGRM without the commonsense-based consequence to generate sarcastic texts. **EGRM-woTag**: EGRM without using image tags to generate sarcastic texts. **EGRM-woS**: EGRM without using sarcasm ranking during comprehensive ranking. **EGRM-woGI**: EGRM without using grammaticality ranking and image-text relation ranking during comprehensive ranking. **EGRM**: the complete method. Moreover, we ablate other components of EGRM in more detail, and specifically evaluate their performance and analyze the effects. The other 6 detailed ablation methods containing **EGRM-woRTV** (EGRM method without reversing the valence of the sentimental descriptive caption), **EGRM-woG** (EGRM method without using grammaticality ranking during comprehensive ranking), **EGRM-woI**

(EGRM method without using image-text relation ranking during comprehensive ranking), **EGRM-woR** (EGRM method replacing comprehensive ranking with randomly selecting final text from candidate sarcastic texts), **EGRM-woRT** (EGRM method without generating rest text y_r) and **EGRM-woFS** (EGRM method without generating first sentence y_f). Due to space limitations, we briefly show the experimental conclusions in Section 5.2. And we show detailed experimental results and corresponding analysis in the Appendix in the supplementary material.

4.3. Evaluation Criteria

Evaluation for CMSG is challenge as it is a creative and imaginative task, and there is no standard sarcastic text for reference. In addition, the difference in the average text length generated by different methods may cause problems in traditional generation evaluation metrics. These reasons make traditional generation evaluation metrics like BLEU [37], one of the most popular evaluation metrics in text generation tasks, unsuitable in CMSG involving creativity and imagination. This problem also exists in textual sarcasm generation task [6, 32]. Therefore, human evaluation is mainly used for evaluation, and we use ClipScore [16], a popular reference-free image captioning metric, to evaluate the image-text relevance. Referring to the textual sarcasm generation metric WL [32] for calculating the percentage of length increment, the notion behind which is that sarcasm typically requires more context than its literal version and requires to have more words present at the target side, we calculate the length of the generated text to assist in evaluating the model, and we name this metric total length (TL). For human evaluation, we evaluate 2100 generated image-text pairs on 14 systems with 150 image-text pairs each.

Inspired by previous work [6], we propose five criteria to evaluate the performance of the CMSG methods: 1) **Sarcasm-ticness** (How sarcastic is the image-text pair?), 2) **Image-Text Relation** (How relevant are the image and text?), 3) **Humor** (How funny is the image-text pair?) [42], 4) **Grammaticality** (How grammatical are the texts?) [6], 5) **Overall** (What is the overall quality of the image-text pair on the cross-modal sarcasm generation task?). The human evaluation details are shown in the Appendix.

5. Experimental Results

5.1. Quantitative Results

Table 1 shows the scores on automatic metrics and human evaluation metrics of different methods. As shown in the upper part of the table, our proposed EGRM has the best performance among all comparison methods on all metrics except CLIPScore, on which EGRM ranks second. The ablation study in Table 1 demonstrates that our full model EGRM is superior to ablation methods in all criteria ex-

⁵<https://www.yelp.com/dataset>

Table 1. Evaluation results of all methods. The scores in columns 4~8 are human evaluation results. The upper part of the table shows the comparison of our method and three baseline methods, and the lower part shows the results of the ablation study. EGRM outperforms other baseline methods on all metrics except CLIPScore, on which EGRM is ranked 2nd (denoted by *).

Method	TL	CLIPScore	Sarcasticness	Image-Text Relation	Humor	Grammaticality	Overall
SC-MTS [32]	9.43	19.70	0.65	0.98	0.71	0.88	0.73
BLIP [23]	9.87	27.23	1.31	3.29	1.91	3.31*	1.95
SC-R ³ [6]	19.11*	25.15	2.22*	2.86	2.21*	3.30	2.29*
EGRM (Ours)	25.65	25.31*	2.85	3.29	2.78	3.41	2.90
EGRM-woCS	24.99	25.14	2.24	2.97	2.27	3.37	2.38
EGRM-woTag	25.99	24.78	2.26	2.91	2.28	3.32	2.37
EGRM-woS	30.99	24.12	2.39	2.91	2.33	3.16	2.42
EGRM-woGI	26.24	25.25	2.34	2.90	2.28	3.18	2.39




Image	Method	Generated Text
	SC-MTS	the ora person is an amazing kite in the region and the anus is a sunny beautiful day
	BLIP	it's not fair to see a kite shaped like a
	SC-R ³	A person flying a kite on the beach on a sunny day. Retreated hastily back to my to stop car.
	EGRM (Ours)	A person flying a kite on the beach on a sunny day. They will reduce sunburn and headaches .
	EGRM-woS	A person flying a kite on the beach on a sunny day. So , after doing a sunburn in another shop . after spending over \$ 100 to get treated we had a bill for the cost that was almost more than quoted on their
	EGRM-woGI	A person flying a kite on the beach on a sunny day. Great service from umbrella movers . I chose them instead of giving kite a shot , but just knowing that they to stop carried my to stop car is out was super relaxing .
	EGRM-woCS	A person flying a kite on the beach on a sunny day. Its umbrella agency creates regulatory rules on mortgages all over the world and garners \$ 1 billion in annual revenues.
	EGRM-woTag	A person flying a kite on the beach on a sunny day. This was my first to stop car wash , so I had a lot of questions .
	SC-MTS	dumbledore is a group of people walking in love with the irony
	BLIP	a black and white photo of people holding umbrellas in the
	SC-R ³	A group of people walking down the street in the rain. I got caught in the traffic jam.
	EGRM (Ours)	A group of people walking down the street in the rain. Dozens of persons were wearing umbrellas , as you might imagine in a traffic jam at peak times .
	EGRM-woS	A group of people walking down the street in the rain. Few persons got to cross into atlanta and the highways as he said there were high seas , with cars in a traffic jam on main street .
	EGRM-woGI	A group of people walking down the street in the rain. The pair of two persons standing under umbrellas , caught in a traffic jam .
	EGRM-woCS	A group of people walking down the street in the rain. This is for each persons first experience with this review is an understatement .
	EGRM-woTag	A group of people walking down the street in the rain. Officers said it had caused a traffic jam at the time .
	SC-MTS	good to see the illuminati close admin to inexperienced food good
	BLIP	yum yum beef lol lol sarcasm
	SC-R ³	A close up of a plate of food on a table. Look out the milk doesn't spill.
	EGRM (Ours)	A close up of a plate of food on a table. The veggies are perfect , carrots and celery so fresh I could spill the soda over a cup of ranch .
	EGRM-woS	A close up of a plate of food on a table. Took my car in with a spill of pool film on the freeway , took an unlimited time just to have it left out .
	EGRM-woGI	A close up of a plate of food on a table. I had a spill of a bag with food on the subway .
	EGRM-woCS	A close up of a plate of food on a table. My husband and I buy broccoli , carrots along with some other items at this health food store .
	EGRM-woTag	A close up of a plate of food on a table. I came into spill by accident a couple weeks ago at our house .

Figure 3. Examples of generated outputs from different systems.

cept the total length. In terms of sarcasticness, our full model attains the highest average score, which shows our model meets the most important requirement of the CMSG task. According to the scores, EGRM gets the highest score on the humor criteria, which shows the potential contribution of our method for improving the interestingness and humor in content creation and communication. Moreover, the grammaticality of EGRM is good and the overall score of EGRM is the highest among all the methods. The total length of the generated paragraph of EGRM is longer than SC-MTS, BLIP, and SC-R³. This can be seen as an auxiliary basis for sarcasm as sarcasm typically requires more context than its literal version and requires to have more words present on the target side.

On the CLIPScore, we observe that EGRM does not have better performance than the pre-trained image captioning

method BLIP, which is designed for generating textual descriptions of images. However, the CMSG task requires imagination and the method should imagine and generate text that is inconsistent with the image as well as relevant to the image, which leads to the CLIPScore of our method designed for the CMSG task being no better than the pre-trained image captioning method BLIP. Moreover, EGRM and BLIP have the best performance among all the four methods on the image-text relation criteria in human evaluation. This is because when human consider whether the text is related to the image, they may allow reasonable imagination. Although BLIP has a higher CLIPScore, it cannot solve the CMSG problem due to the poor performance on sarcasticness. This also shows the existing parallel cross-modal sarcasm data is unable to train a good supervised model for CMSG, due to the limitations in scale and quality.

5.2. Ablation Study

We concentrate our ablation study on the criteria of sarcasm and overall performance, as we consider these metrics as the main criteria for the success of CMSG. As shown in Table 1, the full model (EGRM) outperforms the other four main ablation methods.

EGRM-woCS has the worst performance in terms of sarcasm among the ablation methods. This indicates that the commonsense-based consequence used in the sarcastic texts generation module, which is the inferring result of the image information based on commonsense reasoning, is important for sarcasm. This is because the inconsistency between the commonsense reasoning consequence and the information of the image modality is the key to generating sarcasm. EGRM-woTag has the worst overall performance among the ablation methods. Because the combination of image information and inferring consequence can generate sarcastic image-text pairs where the two modalities are relevant, a text unrelated to the image may be regarded as incomprehensible in the generated text. The experimental results of EGRM-woCS and EGRM-woTag show that the use of image tags and commonsense-based consequences in the generation module is crucial to generating image-text related and imaginative sarcastic texts.

EGRM-woS ranks first among the four main ablation methods in terms of sarcasm and overall performance while EGRM-woGI is slightly worse than EGRM-woS. However, both EGRM-woS and EGRM-woGI are worse than EGRM with a large margin, which demonstrates the importance of the three ranking criteria. Moreover, image-text relation are significant for sarcasm as sarcasm is based on the correlation between text and image. If the text is not related to the image, the sarcasm is more likely to be poor, and sometimes it will be incomprehensible.

Experimental results of the other 6 detailed ablation methods (detailed results are shown in the Appendix) show the importance of the Comprehensive Ranking module, RTV, and the first sentence as well as the rest text of the generated text. Specifically, removing the grammaticality ranking (EGRM-woG) slightly reduces the sarcasm score, and removing the image-text relation ranking (EGRM-woI) greatly reduces the image-text relation score, and the sarcasm score also decreases. Replacing comprehensive ranking with random selection (EGRM-woR) resulted in lower scores for sarcasm, image-text relation, and grammaticality, indicating the importance of comprehensive ranking. EGRM without RTV (EGRM-woRTV) performs worse than EGRM and better than SC- R^3 , suggesting that RTV slightly increases sarcasm. Generated sarcastic text without the rest text y_r (EGRM-woRT) has the worst performance in sarcasm, demonstrating the effectiveness of the Sarcastic Texts Generation module and Comprehensive Ranking module. More detailed exper-

imental results and analysis are demonstrated in Appendix.

5.3. Qualitative Analysis

Figure 3 demonstrates several examples generated from different methods. Taking the text generated by EGRM from the first image in Figure 3 as an example, the image shows a kite flying in the sun. The person flying the kite is more likely to be full of joy. However, they may suffer from sunburn from overexposure to the sun and headaches from heat stroke. The pleasure of the image modality and the pain of the sunburn and the headache in the text modality are inconsistent, which produces sarcasm. Moreover, the kite-flyers may think that the kite can help them block the sun and reduce sunburn and headaches, which is sarcastic about the stupidity of the kite-flyers. However, the results of SC-MTS and BLIP seem not to be sarcastic and the result of SC- R^3 seems to be confusing. The second example shows that our approach is imaginative and humorous. EGRM imagines many people wearing umbrellas as traffic jams, and it satirizes road congestion caused by many umbrellas. The third image shows a plate of food that does not look delicious. However, EGRM says that the veggies are perfect and the carrots are fresh, which makes the deliciousness displayed in the text and the bad taste displayed in the image reversed and inconsistent, making the image-text pair sarcastic. The text is not sarcastic itself but produces sarcasm when combined with the image, which is different from textual sarcasm generation. The other three comparison methods do not seem to produce sarcasm.

6. Conclusion and Future Work

We are the first to formulate the problem of cross-modal sarcasm generation and analyze the challenges of this task. We focus on generating sarcastic texts from images and proposed an extraction-generation-ranking based modular method with three modules to solve the problem without relying on any cross-modal sarcasm training data. Quantitative results and qualitative analysis reveal the superiority of our method. In future work, we will explore generating sarcasm of different styles or categories. We will also try to build a large-scale high-quality parallel cross-modal sarcasm dataset for future researches in this field.

Acknowledgement

This work was supported by National Key R&D Program of China (2021 YFF0901502), National Science Foundation of China (No. 62161160339), State Key Laboratory of Media Convergence Production Technology and Systems and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We appreciate all the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

- [1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 3
- [2] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. Comet: Commonsense transformers for automatic knowledge graph construction. In *ACL*, pages 4762–4779, 2019. 5
- [3] Christian Burgers, Margot Van Mulken, and Peter Jan Schellens. Finding irony: An introduction of the verbal irony procedure (vip). *Metaphor and Symbol*, 26(3):186–205, 2011. 1
- [4] Christian Burgers, Margot Van Mulken, and Peter Jan Schellens. Verbal irony: Differences in usage across written genres. *Journal of Language and Social Psychology*, 31(3):290–310, 2012. 1
- [5] Yitao Cai, Huiyu Cai, and Xiaojun Wan. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In *ACL*, pages 2506–2515, 2019. 1, 6
- [6] Tuhin Chakrabarty, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng. R³: Reverse, retrieve, and rank for sarcasm generation with commonsense knowledge. In *ACL*, pages 7976–7986, 2020. 1, 2, 4, 5, 6, 7
- [7] Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Philipp Koehn. One billion word benchmark for measuring progress in statistical language modeling. *CoRR*, pages 1–6, 2013. 5, 6
- [8] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, pages 10578–10587, 2020. 3
- [9] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, pages 2970–2979, 2017. 3
- [10] Bo Dai and Dahua Lin. Contrastive learning for image captioning. *NeurIPS*, 30, 2017. 3
- [11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, 2019. 5
- [12] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *LREC*, 2006. 4
- [13] Aniruddha Ghosh and Tony Veale. Magnets for sarcasm: Making sarcasm detection timely, contextual and very personal. In *EMNLP*, pages 482–491, 2017. 1
- [14] Debanjan Ghosh, Alexander Richard Fabbri, and Smaranda Muresan. The role of conversation context for sarcasm detection in online interactions. *arXiv preprint arXiv:1707.06226*, 2017. 1
- [15] Xingwei He. Parallel refinements for lexically constrained text generation with bart. In *EMNLP*, pages 8653–8666, 2021. 5
- [16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, pages 7514–7528, 2021. 5, 6
- [17] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jébastien Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh. ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference, Feb. 2022. 3
- [18] Aditya Joshi, Anoop Kunchukuttan, Pushpak Bhattacharyya, and Mark James Carman. Sarcasmbot: An open-source sarcasm-generation module for chatbots. In *WISDOM Workshop at KDD*, 2015. 1, 2
- [19] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. Harnessing context incongruity for sarcasm detection. In *ACL*, pages 757–762, 2015. 1
- [20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015. 2
- [21] Roger J Kreuz and Kristen E Link. Asymmetries in the use of verbal irony. *Journal of Language and Social Psychology*, 21(2):127–143, 2002. 4
- [22] Deepika Kumar, Varun Srivastava, Daniela Elena Popescu, and Jude D Hemanth. Dual-modal transformer with enhanced inter-and intra-modality interactions for image captioning. *Applied Sciences*, 12(13):6733, 2022. 3
- [23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 6, 7
- [24] Tong Li, Yunhui Hu, and Xinxiao Wu. Image captioning with inherent sentiment. In *ICME*, pages 1–6. IEEE, 2021. 3
- [25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014. 2, 6
- [26] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. Context-aware visual policy network for sequence-level image captioning. In *ACMMM*, pages 1416–1424, 2018. 3
- [27] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 5
- [28] Yunpeng Luo, Jiayi Ji, Xiaoshuai Sun, Liujuan Cao, Yongjian Wu, Feiyue Huang, Chia-Wen Lin, and Rongrong Ji. Dual-level collaborative transformer for image captioning. In *AAAI*, volume 35, pages 2286–2293, 2021. 3
- [29] Yangjun Mao, Long Chen, Zhihong Jiang, Dong Zhang, Zhimeng Zhang, Jian Shao, and Jun Xiao. Rethinking the reference-based distinctive image captioning. *arXiv preprint arXiv:2207.11118*, 2022. 3
- [30] Alexander Mathews, Lexing Xie, and Xuming He. Senticap: Generating image descriptions with sentiments. In *AAAI*, volume 30, 2016. 3, 6

- [31] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 4
- [32] Abhijit Mishra, Tarun Tater, and Karthik Sankaranarayanan. A modular architecture for unsupervised sarcasm generation. In *EMNLP*, pages 6144–6154, 2019. 1, 2, 6, 7
- [33] Smaranda Muresan, Roberto Gonzalez-Ibanez, Debanjan Ghosh, and Nina Wacholder. Identification of nonliteral language in social media: A case study on sarcasm. *Journal of the Association for Information Science and Technology*, 67(11):2725–2737, 2016. 1
- [34] Omid Mohamad Nezami, Mark Dras, Stephen Wan, and Cecile Paris. Senti-attend: image captioning using sentiment and attention. *arXiv preprint arXiv:1811.09789*, 2018. 3
- [35] Silviu Oprea, Steven Wilson, and Walid Magdy. Chandler: An explainable sarcastic response generator. In *EMNLP*, pages 339–349, 2021. 1, 2
- [36] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, pages 10971–10980, 2020. 3
- [37] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 6
- [38] Lotem Peled and Roi Reichart. Sarcasm sign: Interpreting sarcasm with sentiment based monolingual machine translation. In *ACL*, pages 1690–1700, 2017. 1, 2
- [39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 5
- [40] Marc’Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015. 3
- [41] John Schwoebel, Shelly Dews, Ellen Winner, and Kavitha Srinivas. Obligatory processing of the literal meaning of ironic utterances: Further evidence. *Metaphor and Symbol*, 15(1-2):47–61, 2000. 1
- [42] Stephen Skalicky and Scott Crossley. Linguistic features of sarcasm and metaphor production quality. In *Proceedings of the Workshop on Figurative Language Processing*, pages 7–16, 2018. 5, 6
- [43] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, 2017. 5
- [44] Chi Wang, Yulin Shen, and Luping Ji. Geometry attention transformer with position-aware lstms for image captioning. *Expert Systems with Applications*, 201:117174, 2022. 3
- [45] Yanhui Wang, Ning Xu, An-An Liu, Wenhui Li, and Yongdong Zhang. High-order interaction learning for image captioning. *T-CSTV*, 2021. 2
- [46] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057. PMLR, 2015. 3
- [47] Zhilin Yang, Ye Yuan, Yuexin Wu, William W Cohen, and Russ R Salakhutdinov. Review networks for caption generation. *NeurIPS*, 29, 2016. 2
- [48] Quanzeng You, Hailin Jin, and Jiebo Luo. Image captioning at will: A versatile scheme for effectively injecting sentiments into image descriptions. *arXiv preprint arXiv:1801.10121*, 2018. 3
- [49] Mengdi Zhu, Zhiwei Yu, and Xiaojun Wan. A neural approach to irony generation. *arXiv e-prints*, pages arXiv–1909, 2019. 1