A DETAILED ABLATION STUDY

A.1 Detailed Ablation Methods

The 6 detailed ablation methods of the proposed EGRM method are as follows:

- **EGRM-woRT**: EGRM method without generating rest text y_r. The goal of EGRM-woRT is to analyze the effect of the first sentence y_f of the generated text.
- **EGRM-woFS**: EGRM method without generating first sentence y_f. The goal of EGRM-woFS is to analyze the effect of the rest text y_r of the generated text.
- **EGRM-woG**: EGRM method without using grammaticality ranking during comprehensive ranking. The goal of EGRM-woG is to analyze the effect of grammaticality ranking.
- **EGRM-woI**: EGRM method without using image-text relation ranking during comprehensive ranking. The goal of EGRM-woI is to analyze the effect of image-text relation ranking.
- **EGRM-woR**: EGRM method replacing comprehensive ranking with randomly selecting final text from candidate sarcastic texts. The goal of EGRM-woR is to explore the effect of the Comprehensive Ranking module.
- **EGRM-woRTV**: EGRM method without reversing the valence of the sentimental descriptive caption. The goal of EGRM-woRTV is to analyze the effect of the RTV component.

A.2 Experimental Results

Experimental results of the other 6 detailed ablation methods are shown in the bottom part of Table 1. For the convenience of comparison, we also list the experimental results of EGRM together with other comparison methods and main ablation methods in the upper and middle parts of Table 1. Since the CMSG task focuses on the Sarcasticness of image-text pairs, we mainly analyze the Sarcasticness and the overall performance and briefly analyze other aspects.

For the Sarcasticness of the image-text pairs, we can learn from the experimental results of EGRM-woRT that the rest text y_r is most significant to Sarcasticness and the overall performance of the generated text. This shows the effectiveness of the Sarcastic Texts Generation module and the Comprehensive Ranking module. Results of EGRM-woRT, EGRM-woFS and EGRM-woRTV show that the first sentence y_f of the generated text together with the RTV play a small role in Sarcasticness and the overall performance. Experimental results of EGRM-woG show that removing the Grammaticality ranking component slightly reduces the Sarcasticness score and the overall performance score. Removing the Image-Text Relation ranking component (EGRM-woI) reduces the Sarcasticness score and the overall performance score by a larger margin than removing the Grammaticality ranking component. The reason is that Image-Text Relation is the premise of Sarcasticness. If the text and the image are not related, the image-text pair can be very confusing. The Sarcasticness score of replacing comprehensive ranking with randomly selecting final text from candidate sarcastic texts is 0.52 lower than that of EGRM, which demonstrates the importance of the Comprehensive Ranking module.

For Image-Text Relation of the image-text pairs, we can learn from the experimental results of EGRM-woFS and EGRM-woI that the first sentence y_f of the generated text and the Image-Text Relation ranking is the most important to Image-Text Relation score. The sentimental descriptive captions ensure the image-text relevance of the first sentence, while the Image-Text Relation ranking ensures the relevance of the rest text to the image. Interestingly, EGRM-woRT gets the highest Image-Text Relation score. EGRM gets a lower Image-Text Relation score because the rest text has a certain imagination which may lead to Image-Text inconsistency, which is the key to producing Sarcasticness.

For the Humor score of the image-text pairs, experimental results of EGRM-woRT demonstrates that the rest text plays a key role in producing humor. As shown in Table 1, EGRM-woRT gets the highest Grammaticality score. This is because the first sentence y_f is generated by a pretrained supervised sentimental descriptive image captioning method, which may result in fewer grammar mistakes. Forcing CBART to generate sentences with image tags and commonsense-based consequences may result in some grammatical mistakes. However, EGRM-woRT gets the lowest Grammaticality score, which shows the necessity of using CBART to generate sarcastic candidate rest texts. In conclusion, the Sarcastic Texts Generation module and Comprehensive Ranking module play an important role in generating image-text-related sarcastic texts from images.

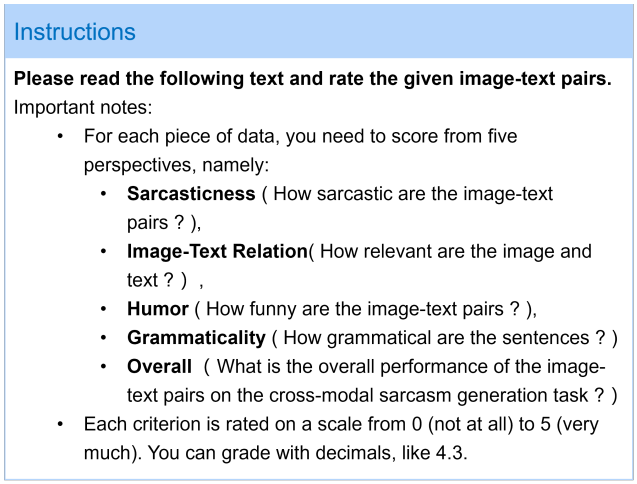

Figure 1: Instructions for human evaluation.

B HUMAN EVALUATION DETAILS

We design an MTurk CMSG task where each Turker was asked to score the image-text pairs from all methods. Each Turker was given the image together with a set of sarcastic texts generated by all systems. Each criterion is rated on a scale from 0 (not at all) to 5 (very much). The Turker can grade with decimals like 4.3. As

Table 1: Evaluation results of all methods. The scores in columns 4~8 are human evaluation results, and the scale ranges from 0 (not at all) to 5 (very). The upper part of the table shows the comparison of our method and three baseline methods, and the middle part shows the results of our main ablation study. The bottom part shows the results of our detailed ablation study.

Method	TL	CLIPScore	Sarcasticness	Image-Text Relation	Humor	Grammaticality	Overall
SC-MTS	9.43	19.70	0.65	0.98	0.71	0.88	0.73
BLIP	9.87	**27.23**	1.31	**3.29**	1.91	3.31*	1.95
SC-R^3	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
EGRM-woRT	10.47	25.7	0.90	4.00	0.95	3.96	1.57
EGRM-woFS	13.35	22.3	2.29	2.85	2.25	3.22	2.36
EGRM-woG	25.73	25	2.46	3.16	2.42	3.24	2.56
EGRM-woI	25.54	24.3	2.21	2.90	2.16	3.16	2.28
EGRM-woR	26.91	24.4	2.33	2.94	2.26	3.16	2.40
EGRM-woRTV	25.41	24.8	2.25	2.93	2.21	3.21	2.33

CMSG is a difficult task requiring imagination, each image-text pair was scored by 3 individual Turkers. We select annotators with good English reading and writing skills, judging by their English qualification certificates. Each Turker is paid $490 for the whole evaluation of 2,100 image-text pairs, which is roughly $0.23 per image-text pair. Figure 1 shows the instructions released to the Turkers. The final score for each method is an average of scores from all Turkers.