Automatic, Meta and Human Evaluation for Multimodal Summarization with Multimodal Output

Haojie Zhuang[1], **Wei Emma Zhang**[1], **Leon Xie**[1],
Weitong Chen[1], **Jian Yang**[2], **Quan Z. Sheng**[2]
[1]The University of Adelaide, Adelaide, Australia
[2]Macquarie University, Sydney, Australia
{haojie.zhuang,wei.e.zhang,
leon.xie,weitong.chen}@adelaide.edu.au
{jian.yang,michael.sheng}@mq.edu.au

Abstract

Multimodal summarization with multimodal output (MSMO) has attracted increasing research interests recently as multimodal summary could provide more comprehensive information compared to text-only summary, effectively improving the user experience and satisfaction. As one of the most fundamental components for the development of MSMO, **evaluation** is an emerging yet underexplored research topic. In this paper, we fill this gap and propose a research framework that studies three research questions of MSMO evaluation: (1) **Automatic Evaluation**: We propose a novel metric mLLM-EVAL, which utilizes multimodal Large Language Model for MSMO EVALuation. (2) **Meta-Evaluation**: We create a meta-evaluation benchmark dataset by collecting human-annotated scores for multimodal summaries. With our benchmark, we conduct meta-evaluation analysis to assess the quality of different evaluation metrics and show the effectiveness of our proposed mLLM-EVAL. (3) **Human Evaluation**: To provide more objective and unbiased human annotations for meta-evaluation, we hypothesize and verify three types of cognitive biases in human evaluation. We also incorporate our findings into the human annotation process in the meta-evaluation benchmark. Overall, our research framework provides an evaluation metric, a meta-evaluation benchmark dataset annotated by humans and an analysis of cognitive biases in human evaluation, which we believe would serve as a valuable and comprehensive resource for the MSMO research community. [1]

1 Introduction

With the exponentially growing amount of multimedia data online, multimodal summarization with multimodal output (MSMO) has garnered more and more attention from researchers. Unlike unimodal output (e.g., text-only summary), MSMO aims at extracting the most salient information from different modalities and generating multimodal summaries, with a text summary and the most relevant images (Zhu et al., 2018, 2020; Jangra et al., 2023; Modani et al., 2016; Zhang et al., 2021, 2022a; Li et al., 2020). Compared to a text-only summary, a multimodal summary is more informative and could provide more comprehensive and user-friendly content to the readers, as well as effectively improve their reading experience and satisfaction (Zhu et al., 2018, 2020; Li et al., 2020), which could thus be suitable for many applications (e.g., multimedia news summarization).

Evaluation metrics for MSMO are essential to objectively measure the quality of the multimodal summary outputs. However, only a few works have focused on the evaluation for multimodal summaries (Zhu et al., 2018, 2020; Modani et al., 2016). Meanwhile, it is also important to establish a meta-evaluation system to assess the quality of the MSMO metrics to promote more exploration of effective and unbiased evaluation metrics. Despite its significance, there is no such benchmark or study in meta-evaluation. Furthermore, meta-evaluation requires human-annotated ratings for the correlation test since human annotations are considered as the gold standard, but the analysis of cognitive biases in the human annotations remains unexplored. To fill these gaps and call for more research in this area, we propose a research framework (as in Fig. 1) and investigate the following research questions:

RQ1. Automatic Evaluation: *How to properly evaluate the quality of a multimodal summary?* The existing evaluation methods proposed by Zhu et al. (2018, 2020) require the human-written references for evaluation (e.g., ROUGE score (Lin, 2004) for text quality, selection precision for image quality, supervised regression for overall quality). However, it is costly to collect human-written mul-

[1]Our code and annotated dataset are publicly available at: https://github.com/hjzhuang/MSMO-Eval.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 7768–7790
June 16-21, 2024

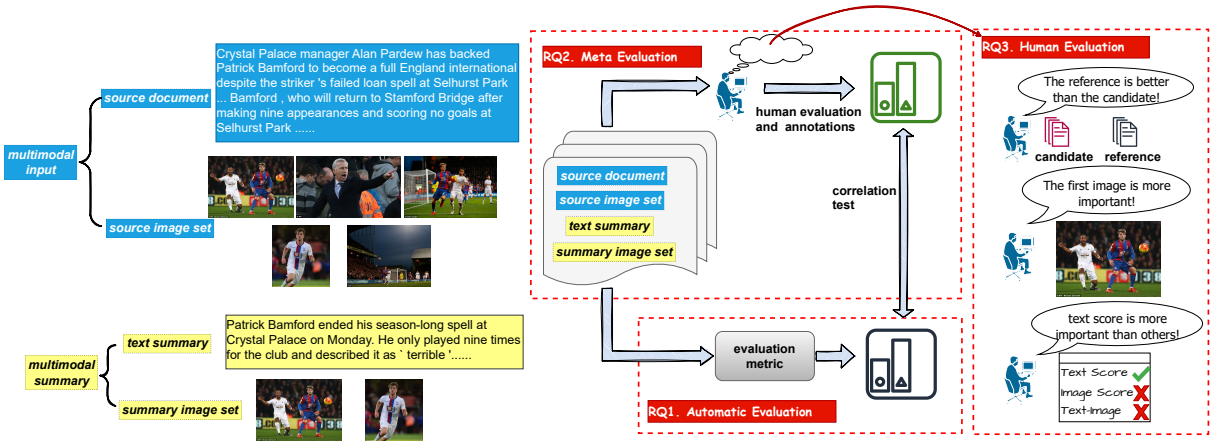


Figure 1: The illustration of our research framework. Left: MSMO extracts the most important information from both text and images and generates a multimodal summary. Right: Three research questions in our research framework: automatic evaluation (RQ1), meta-evaluation (RQ2) and human evaluation (RQ3).

timodal summaries as references for evaluation. Inspired by the success of Large Language Models (LLMs) in reference-free text summary evaluation (Fu et al., 2023; Liu et al., 2023; Chiang and Lee, 2023), as well as the development of multimodal LLMs (e.g., GPT-4 (OpenAI, 2023)), we propose mLLM-EVAL, an MSMO metric utilize the emergent abilities of multimodal LLMs that shows a high correlation with human judgments.

RQ2. Meta-Evaluation: *How to evaluate the evaluation metrics for multimodal summarization?* Furthermore, to evaluate our proposed metric and encourage further exploration of MSMO metrics, we create the first MSMO meta-evaluation benchmark dataset by asking three experts to assess the quality of multimodal summaries. Using our meta-evaluation benchmark, we evaluate and analyze the quality of the existing metrics and mLLM-EAVL by testing how well they correlate with human annotations, where our proposed method outperforms the current evaluation metrics.

RQ3. Human Evaluation: *Is human evaluation truly the gold standard?* In our meta-evaluation (RQ2), we require human evaluation to examine how well the MSMO metrics correlate with human judgments, as human evaluation is widely considered as the gold standard. However, humans could potentially bring biases into the evaluation due to cognitive biases, which would make the comparison unfair and biased. To this end, we hypothesize and verify three types of cognitive biases: anchoring bias; lead bias in image selection; text bias in overall evaluation. We also incorporate our findings to avoid such biases, thus providing more objective and unbiased human annotations for meta-evaluation (RQ2), which makes our meta-evaluation more reliable.

The main contribution of this paper is: **(1)** This is a pioneering work to study automatic evaluation, meta-evaluation and cognitive biases analysis in human evaluation for MSMO. **(2)** We propose mLLM-EVAL using multimodal LLMs for reference-free MSMO evaluation. **(3)** We create the first MSMO meta-evaluation benchmark by collecting human annotations and then conduct meta-evaluation to assess the quality of various MSMO metrics. **(4)** We study the cognitive biases in human evaluation for MSMO. Overall, we aim to establish a valuable research foundation to significantly benefit the multimodal summarization community.

2 Related Work

Multimodal Summarization. Multimodal summarization aims to generate summaries given the inputs of multiple modalities (Evangelopoulos et al., 2013; Li et al., 2017; Mademlis et al., 2016; Koupaee and Wang, 2018; Zhu et al., 2018, 2020). Zhu et al. (2018) proposed the task of multimodal summarization with multimodal output (MSMO) that requires the model to output multimodal summaries and release a large-scale dataset for MSMO. Further improvements include training with multimodal reference guidance in Zhu et al. (2020), knowledge distillation in Zhang et al. (2022a), location-aware approach in Zhang et al. (2021). Due to the issues of reference-based metrics in MSMO (being costly to collect references; different from humans' reference-free evaluation manner), we focus on the reference-free MSMO evalu-

ation metric and investigate the use of multimodal LLMs in this paper.

Meta-evaluation for Summarization. In the works of Fabbri et al. (2021) and Bhandari et al. (2020), the authors re-evaluated different text summarization evaluation metrics by providing human annotations and released the meta-evaluation benchmark for further research. For multimodal summarization, (Wan and Bansal, 2022) collected human annotations on the factuality and released a benchmark to evaluate the quality of multimodal factuality metrics. To date, there is no research work on meta-evaluation for MSMO. We thus create a benchmark dataset and hope it will serve as a valuable resource for future research.

Cognitive Biases in Human Evaluation. Human evaluation is the core component of evaluation research. While human evaluation is generally considered the gold standard in many machine learning tasks, (Schoch et al., 2020) identified the cognitive biases in human evaluation for natural language generation and claimed that the lack of transparency in human evaluation will also impact the reliability of results. (Santhanam et al., 2020) studied the cognitive biases in evaluating conversational systems. However, there are no previous studies on cognitive biases in multimodal summarization evaluation. We thus investigate this topic in this paper to fill this gap.

3 Methodology

3.1 RQ1: Automatic Evaluation

Our proposed metric mLLM-EVAL evaluates the quality of multimodal summaries without the need for references. The evaluation attributes include the (1) quality of the text summary (including relevance, coherence, consistency and fluency); (2) quality of the summary images; (3) relevance of the summary text and images; (4) the overall quality of the multimodal summary. Specifically, we use a multimodal LLM as an evaluator, taking both the texts and images as inputs, and then we design attribute-specific prompts for the multimodal LLM to output the quality scores. Following (Liu et al., 2023), the prompt consists of (1) the descriptions of the evaluation task; (2) the evaluation attribute and criteria; (3) the auto chain-of-thoughts (Zhang et al., 2022b) for evaluation steps; (4) the example for evaluation; (5) scoring form for the final result. Besides the text prompt, we also add the images as the inputs to the multimodal LLM, which would consider the multimodal information. For example, for the overall quality of the multimodal summary, we use the following text prompt:

> You will be given one text summary and some summary images (ID) written for a news article with source images. The summary images is a subset of the source images (e.g., Summary Image ID: 1,3 refer to the first and third source images).
> Your task is to rate the overall quality of the text summary and the summary images.
> Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
> *Evaluation Criteria*:
> Quality (1-5) - the overall quality of the text summary and summary images as a whole, including the quality of the text summary, the quality of the summary images (how well the selected summary images represent the content of the source document), the relevance of text summary and summary images (how well the text summary and summary images match).
> *Evaluation Steps*:
> 1. Read the source document, source images, text summary and summary images carefully.
> 2. Assess the overall quality of the text summary and summary images as a whole, given the source document and source images.
> 3. Assign a score for overall quality on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.
> *Example*:
> Source Text: {{Document}}
> Summary: {{Summary}}
> Summary Images: {{Image IDs}}
> Evaluation Form (scores ONLY):
> - Quality:

For the prompts of other attribute evaluation, we refer the readers to Appendix A for more details.

3.2 RQ2: Meta-Evaluation

Human annotations are significantly important and indispensable for meta-evaluation, as we need to compare the correlations of automatic metrics and human judgments to evaluate the effectiveness of the metric. Since there are no human annotations of meta-evaluation for the MSMO task, we construct a **meta-evaluation benchmark dataset** for MSMO with 1,562 human-annotated examples, which is sufficient to evaluate the quality of automatic metrics. The construction of the benchmark consists of the following steps: (1) sampling multimodal inputs (pairs of document and source image set) and then using different summarization models and image selection (generation) algorithms to generate multimodal summaries for each multimodal input; (2) conducting human annotations for the multimodal summaries. With our proposed meta-evaluation benchmark, we conduct meta-evaluation to assess the quality of various metrics.

Topic	Number
sports	35
politics	13
culture/travel	14
crime/public safety	25
health/lifestyle	19
science/technology	7
celebrities/entertainment	24
others	5

Table 1: The topic distribution of 142 multimodal inputs.

3.2.1 Data Preparation

Multimodal Inputs. We use the dataset collected by Zhu et al. (2018), which is the most commonly used benchmark dataset for the MSMO task. The news articles are collected from the Daily Mail[2]. We randomly sample 150 examples from the test set and discard those without any summary image [3]. As a result, we end up with 142 distinct multimodal inputs for the following generation by different summarization systems. We further provide the dataset statistics on the topics of 142 documents in Table 1, as we expect the dataset could cover a vast spread of domains (e.g., sports, politics).

Multimodal Outputs. Following (Bhandari et al., 2020; Fabbri et al., 2021), we use different summarization systems to generate text summaries given the aforementioned sampled examples. Specifically, we use 9 summarization models to generate summaries, including: (1) BART (Lewis et al., 2020) (2) Distilbart (Shleifer and Rush, 2020) (3) GPT-2 (Radford et al., 2019) (4) PEGASUS (Zhang et al., 2020) (5) ProphetNet (Qi et al., 2020) (6) T5 (Raffel et al., 2020) (7) HAN (Zhu et al., 2018) (8) GPT-4 (OpenAI, 2023) (9) GPT-4 with vision (OpenAI, 2023). We refer the readers to Appendix B for details of these models. Besides, we include two additional summaries to the benchmark: (1) the reference summary; (2) a randomly sampled summary from the dataset.

Among the above models, only the HAN model could output both texts and images. Other models could only generate text-only summaries. To have multimodal outputs, we further design different algorithms to select the images from the source image set. Given a multimodal input, there is a reference summary image set (with N images). We edit this reference set with the following methods: (1) randomly delete an image; (2) randomly add an extra image from the source image set; (3) randomly sample an extra image from the source image set and replace a random image in the reference set; (4) pick the first two images from the source image set as the summary image set; (5) randomly sample N images from the source image set as the summary image set (same size); (6) being same as the reference image set.

Each time when a summarization model (except the HAN model) generates the text summary, a random image selection algorithm would be used to obtain the summary image set. Thus each multimodal input could have 11 different multimodal outputs. Finally, we have 1,562 multimodal summarization examples for further human annotation. Following the aforementioned method, we believe the examples are *diverse* enough to evaluate the metrics since we expect a good metric could assess examples of different quality levels, including *good* examples and *bad* examples. We also list a few good/bad examples of our benchmark in the Appendix C.

3.2.2 Human Annotations

After having 1,562 multimodal summarization examples, we ask three annotators to evaluate the quality of the multimodal summaries. All annotators are English native speakers and/or summarization researchers, and they all have summarization annotation experiences. We use a 1 to 5 rating scale for scoring, while 1 denotes the worst and 5 denotes the best. The annotators evaluate the following aspects:

- **Text Summary-Relevance**: how well the summary captures the most relevant and salient information while omitting the unnecessary or redundant information.
- **Text Summary-Coherence**: how well the summary is organized with sentences logically connected. It measures the flow and logical structure of the text.
- **Text Summary-Consistency**: how factual and faithful the summary is to the source document. A factually consistent summary contains only information that could be entailed (directly or indirectly) by the source document.
- **Text Summary-Fluency**: how fluent each sentence in the summary (e.g., grammar, spelling, word choice, etc).
- **Summary Images**: how well the selected

[2]http://www.dailymail.co.uk/

[3]The summary image set might be empty in the dataset if the annotators in (Zhu et al., 2018) think there are no relevant or suitable images

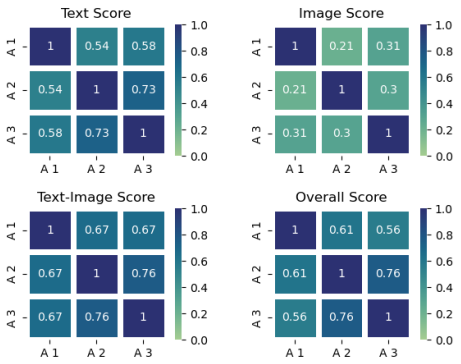


Figure 2: The cohen's kappa coefficient of three annotators (A1, A2, A3).

summary images represent the content of the source document.

- **Text Summary-Summary Images Relevance**: how well the summary text and summary images match each other.
- **Overall Quality**: the overall quality of the generated multimodal summary given the multimodal input.

The text summary evaluation attributes are adopted from Kryscinski et al. (2019) while the remaining ones are from Zhu et al. (2018). Before conducting the human annotation, we set guidelines for the annotators, which include the above evaluation attributes, as well as the instructions to avoid cognitive biases that we study in this paper (details in Section 3.3 and 4.3). To further ensure the annotators have similar annotation criteria, we sampled 5 examples and had a discussion on what scores should be given to each example as well as the reason for the scoring. We believe we use the consistent annotation criteria after the examples discussion.

We report the Cohen's kappa coefficient to indicate the inter-annotator agreement in Fig 2, which shows a moderate agreement among the three annotators. The agreement of image score is relatively low, which we think is reasonable because a source image set could have multiple reference summary image sets. We list two examples for illustration in Appendix D. We also notice the inconsistency issue in the reference text summary, which we discuss in the Appendix E.

3.2.3 Meta-Evaluation Method

As illustrated in Fig. 1, with our constructed benchmark, we can perform meta-evaluation to evaluate the effectiveness of an automatic evaluation metric. Specifically, we conduct the meta-evaluation for text score, image score, text-image score and overall quality score, by calculating the correlation coefficient between the scores obtained by the automatic metrics and the human scores of our benchmark. With a higher correlation, the evaluation metric is more similar to human judgments and thus is more reliable and effective. We show the meta-evaluation results in Section 4.2.

3.3 RQ3: Human Evaluation

Humans could possibly bring cognitive biases into human evaluation, which makes the results also biased and unreliable (Schoch et al., 2020). A cognitive bias refers to the systematic thought process that deviates from rationality in decision or judgment (Tversky and Kahneman, 1974; Schoch et al., 2020; Gehlbach and Barge, 2012). Humans tend to be influenced by their own cognitive biases to make decisions or interpret information, which leads to errors in judgment and decision-making. Thus, we believe it is critical to investigate and study the cognitive biases in human evaluation for multimodal summarization, which is significant for collecting more objective and unbiased human annotations. Specifically, we study three types of cognitive biases in multimodal summarization evaluation:

- **Anchoring Bias**: Humans might have an anchoring bias when evaluating the system outputs if the gold multimodal summaries are also present to the annotators;
- **Lead Bias in Image Selection**: Humans might tend to believe the early parts of the source images are likely to be more important or informative.
- **Text Bias in Overall Evaluation**: Humans are likely to rely more on the text quality (compared to the image quality) to give an overall quality score.

The methodology of our study on cognitive biases of human evaluation is described as follows,

1. hypothesizing a potential cognitive bias when humans are asked to assess the quality of a multimodal summary;
2. asking humans to evaluate under some settings that might elicit this type of bias and verifying whether humans would bring the bias into the evaluation results;
3. proposing a simple and effective method to avoid cognitive bias, which we incorporate into the meta-evaluation benchmark construction (RQ2).

4 Experiment

4.1 RQ1: Automatic Evaluation

We use GPT-4 with vision (*"gpt-4-vision-preview"*) as the multimodal LLM. To evaluate the quality of our proposed metric, we leverage our proposed meta-evaluation benchmark and describe the results in Section 4.2.

4.2 RQ2: Meta-Evaluation

After collecting human annotations for multimodal summaries as described in Section 3.2, we perform meta-evaluation for different automatic metrics, including our proposed method. The meta-evaluation is conducted for 4 different aspects: (1) the text score (i.e., text summary quality); (2) the image score (i.e., summary images quality); (3) the text-image score (i.e., the relevance of the text and images in a multimodal summary) (4) the overall score (i.e., the overall quality of a multimodal summary as a whole) respectively. We conduct three correlation tests: Pearson correlation coefficient r, Spearman's rank correlation coefficient ρ, and Kendall rank correlation τ.

4.2.1 Meta-Evaluation for Text Score

For text score, we evaluate the following baseline metrics: (1) lexical overlap metrics: ROUGE (Lin, 2004) (R1, R2, RL); BLEU (Papineni et al., 2002); (2) pretrained language models: BERTScore (Zhang* et al., 2020); BARTScore (Yuan et al., 2021); MoverScore (Zhao et al., 2019); BLANC (Vasilyev et al., 2020) (3) LLM based metrics: GPTScore (Fu et al., 2023); G-Eval (Liu et al., 2023) (4) MuSQ (Text) (Modani et al., 2016) measures the degree of coverage of input text document by text summary in a reference-free manner. We use ROUGE and BERTScore as the text similarity functions in MuSQ (Text), denoted as MUSQ-R1, MUSQ-R2, MUSQ-R3 and MUSQ-BS respectively. We also repurpose ROUGE, BLEU, BERTScore and BARTScore as reference-free metrics following Bao et al. (2022).

We list the meta-evaluation results in Table 2. The results show that our proposed metric mostly achieves the best correlation with human annotations, while the second best in some aspects. We also observe that ROUGE and BERTScore with reference-free settings have a higher correlation than a reference-based setting, which we believe is because the human scores are also annotated in a reference-free manner.

4.2.2 Meta-Evaluation for Image Score

For the image score, we evaluate the baseline Image Precision (Zhu et al., 2018) (reference-based metric) and MUSQ (Image) (Modani et al., 2016). In addition, we implement some other baseline metrics by formulating image scoring as an *image ranking* problem: (1) CLIPScore-based methods: using CLIPScore (Hessel et al., 2021) to measure the relevance between each image and the document (denoted as "CLIPScore"); (2) Caption-based methods: using the relevance between the caption of each image and the document (denoted as "CAP"). The captions are obtained by either using a pretrained model BLIP (Li et al., 2022) (denoted as "GEN") or extracting the provided captions from the MSMO dataset (Zhu et al., 2018) directly (denoted as "EXT"). We use ROUGE score and BERTScore to measure the relevance between caption and document (denoted as "R1/R2/RL/BS"). After having the relevance score of each image, we construct a pseudo-reference summary image set by two methods: (1) Avg: any image with a relevance score higher than the average is selected (denoted as "Avg"); (2) TopK: ranking the source images by the relevance score and setting the top-K images (denoted as "topK"). Finally, we calculate the precision of the summary image set with the pseudo-reference set as the image score.

The results in Table 3 indicate that our proposed method has the best performance. Surprisingly, mLLM-EVAL (reference-free) has an advantage over the reference-based metric Image Precision. We believe one of the main reasons is that the reference-based metric relies on the human-annotated images, while the same source image set could have multiple reference summary images (as discussed in the Appendix D). We also compare the caption of 'EXT" and "GEN" in Appendix F.

4.2.3 Meta-Evaluation for Text-Image Score

To demonstrate the effectiveness of our metric for the text-image score, we use MMAE (Zhu et al., 2018), MUSQ (Text-Image) (Modani et al., 2016) as our baselines. Furthermore, we utilize CLIPScore (Hessel et al., 2021) to implement four other strong baseline metrics: (1) the average or maximum CLIPScores of each summary image and the whole text summary (denoted as "CLIPScore_Whole_Avg" and "CLIPScore_Whole_Max"); (2) the average or maximum CLIPScores of all the image-sentence pairs (denoted as "CLIPScore_Sumsent_Avg" and "CLIP-

	Relevance			Coherence			Consistency			Fluency		
Metric	r	ρ	τ	r	ρ	τ	r	ρ	τ	r	ρ	τ
Reference-based Methods												
R-1	0.40	0.42	0.29	0.19	0.27	0.19	0.44	0.43	0.32	0.14	0.22	0.15
R-2	0.23	0.39	0.27	0.10	0.25	0.17	0.28	0.43	0.32	0.04	0.19	0.13
R-L	0.27	0.40	0.28	0.12	0.26	0.18	0.32	0.43	0.32	0.07	0.21	0.15
BLEU	0.01	-0.05	-0.04	0.01	-0.03	-0.02	0.02	0.03	0.02	-0.02	-0.12	-0.09
BERTScore	0.36	0.45	0.32	0.18	0.30	0.21	0.40	0.45	0.33	0.16	0.32	0.22
BARTScore	0.43	0.42	0.29	0.17	0.25	0.17	0.46	0.38	0.28	0.13	0.21	0.15
MoverScore	0.41	0.39	0.33	0.19	0.27	0.20	0.44	0.43	0.31	0.15	0.21	0.17
Reference-free Methods												
R-1	0.48	0.50	0.36	0.26	0.29	0.21	0.42	0.42	0.32	0.28	0.31	0.22
R-2	0.48	0.48	0.35	0.27	0.31	0.22	0.51	0.54	0.41	0.25	0.29	0.20
R-L	0.47	0.49	0.36	0.28	0.33	0.23	0.46	0.5	0.37	0.28	0.32	0.23
BLEU	-0.01	-0.07	-0.05	-0.01	-0.04	-0.03	0.01	0.05	0.04	-0.05	-0.17	-0.12
BERTScore	0.71	0.48	0.35	0.40	0.37	0.27	0.80	0.56	0.43	0.41	0.38	0.27
BARTScore	0.50	0.44	0.32	0.22	0.24	0.17	0.48	0.40	0.30	0.24	0.24	0.17
MUSQ-R1	0.21	0.23	0.16	0.07	0.10	0.07	0.27	0.27	0.20	0.05	0.11	0.08
MUSQ-R2	0.48	0.40	0.29	0.23	0.24	0.17	0.56	0.48	0.36	0.21	0.23	0.16
MUSQ-RL	0.23	0.25	0.18	0.08	0.12	0.08	0.31	0.31	0.22	0.07	0.13	0.09
MUSQ-BS	0.35	0.44	0.32	0.26	0.33	0.24	0.42	0.49	0.37	0.17	0.25	0.23
BlancHelp	0.57	0.51	0.37	0.30	0.33	0.24	0.57	0.49	0.37	0.31	0.32	0.23
BlancTune	0.51	0.50	0.37	0.29	0.32	0.23	0.51	0.51	0.39	0.28	0.30	0.22
GPTScore	0.77	0.52	0.36	**0.43**	0.37	0.29	0.83	0.57	0.44	0.42	0.38	0.28
G-Eval	0.79	0.55	0.35	0.41	**0.38**	0.31	0.82	0.60	**0.45**	0.46	0.38	0.32
mLLM-EVAL (ours)	**0.81**	**0.62**	**0.43**	0.41	**0.38**	**0.33**	**0.85**	**0.61**	0.44	**0.49**	**0.40**	**0.34**

Table 2: The meta-evaluation results of text score, where the score with bold text denotes the best performance.

Score_Sumsent_Max"), as the text summary usually contains multiple sentences. As shown in Table 4, our proposed metric outperforms all the strong baselines.

4.2.4 Meta-Evaluation for Overall Score

For the overall score, we evaluate our proposed metric and MMAE (Zhu et al., 2018), MUSQ (Overall) (Modani et al., 2016). As in Table 5, our overall score correlates with human assessments the best, suggesting its high quality and effectiveness.

4.3 RQ3: Human Evaluation

For RQ3, we study the cognitive biases in human evaluation for MSMO. Specifically, we hypothesize three types of potential cognitive biases and verify them through experiments. We also incorporate our findings to avoid such biases in human annotation when constructing the meta-evaluation benchmark (details in Section 3.2), thus making our meta-evaluation more reliable.

4.3.1 Anchoring Bias

Anchoring bias is a type of cognitive bias where a particular “anchor" influences humans’ decision-making (Tversky and Kahneman, 1974; Gehlbach and Barge, 2012). In human evaluation for multimodal summarization, we hypothesize that humans tend to give lower scores for the candidate summaries that are generated by a multimodal summarization system if they are shown both the candidate summaries and references, where the references, as “anchors", make the annotators tend to flavor the gold summaries more than the system-generated summaries. To elicit the anchoring bias, we ask the annotators to evaluate 100 candidate

Metric	r	ρ	τ
Image Precision *	0.31	0.27	0.22
MUSQ-Image	-0.12	-0.16	-0.11
CAP_GEN_R1_Avg	0.17	0.16	0.12
CAP_GEN_R2_Avg	0.14	0.15	0.11
CAP_GEN_RL_Avg	0.12	0.12	0.09
CAP_GEN_BS_Avg	0.13	0.13	0.10
CAP_EXT_R1_Avg	0.14	0.12	0.10
CAP_EXT_R2_Avg	0.17	0.17	0.13
CAP_EXT_RL_Avg	0.17	0.16	0.12
CAP_EXT_BS_Avg	0.07	0.06	0.04
CAP_GEN_R1_topK	0.20	0.19	0.15
CAP_GEN_R2_topK	0.20	0.17	0.13
CAP_GEN_RL_topK	0.21	0.19	0.15
CAP_GEN_BS_topK	0.20	0.17	0.14
CAP_EXT_R1_topK	0.24	0.22	0.17
CAP_EXT_R2_topK	0.20	0.18	0.14
CAP_EXT_RL_topK	0.24	0.22	0.17
CAP_EXT_BS_topK	0.21	0.19	0.15
CLIPScore_Avg	0.09	0.06	0.05
CLIPScore_topK	0.22	0.20	0.16
mLLM-EVAL (ours)	**0.33**	**0.32**	**0.25**

Table 3: The meta-evaluation results of image score, where the score with bold text denotes the best performance. Metric with * is a reference-based method while others are reference-free ones.

Metric	r	ρ	τ
MMAE	0.41	0.37	0.26
MUSQ	0.19	0.17	0.12
CLIPScore_Whole_Avg	0.67	0.46	0.34
CLIPScore_Whole_Max	0.65	0.45	0.33
CLIPScore_Sumsent_Avg	0.65	0.44	0.32
CLIPScore_Sumsent_Max	0.69	0.48	0.35
mLLM-EVAL (ours)	**0.73**	**0.52**	**0.42**

Table 4: The meta-evaluation results of the text-image relevance score.

Metric	r	ρ	τ
MMAE	0.55	0.47	0.32
MUSQ (Overall)	0.33	0.28	0.17
mLLM-EVAL (ours)	**0.63**	**0.50**	**0.43**

Table 5: The meta-evaluation results of overall quality score.

multimodal summaries, where each summary is evaluated twice. In the first stage of evaluation, the summaries are present with the corresponding references, while without the references in the second stage. To eliminate the influences of first-stage evaluation results on the second stage (e.g., humans might remember what scores they give out in the first stage), the annotators conduct the second-stage evaluation 1 week after the first stage. Furthermore, we conduct another experiment same as above but present each summary with a randomly sampled system-generated summary as the "*false reference*" while telling the annotators that it is the true reference. We show the evaluation results in Fig 3. As shown in the results, annotators are more likely to give a lower score for a summary if it is shown along with the references, even with the false references.

To this end, we believe that showing the candidate summary along with the corresponding reference to the annotators would bring anchoring bias to the evaluation results. We thus present the candidate multimodal summary alone to the annotators when we perform the human evaluation to construct the meta-evaluation benchmark.

4.3.2 Lead Bias in Image Selection

In text summarization, the lead bias (Xing et al., 2021; Zhu et al., 2021) is a common phenomenon in news articles. We hypothesize that humans tend to select the first few images as the summary image sets in MSMO evaluation. To verify that, we sample 100 examples and ask the annotators to select the most important source images. The annotators conduct the selection twice for each example,

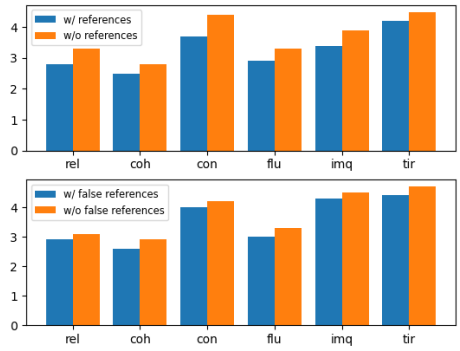


Figure 3: The evaluation results of anchoring bias. The"rel","coh", "con","flu","imq","tir" means "relevance", "coherence", "consistency", "fluency", "image quality", "text image relevance" respecitively.

	1st Image	2nd Image	First N Image
Original Order	82.7%	56.7 %	38.0%
Shuffled Order	76.0%	52.7%	36.7%

Table 6: The experiment results of lead bias in image selection. First/Second is how frequently the annotators would pick the first/second image. First N means that the N images chosen by the annotators are exactly the first N images.

where the source images are present in an original order and random order respectively. The results are shown in Table 6, where we could observe that the annotators prefer to choose the *lead* images from the source image set. Even when presenting with the shuffled order, the lead bias still exists in the image selection process. Therefore, we believe that there is a lead bias phenomenon when humans are selecting the most important and representative images.

For the human ratings collecting in meta-evaluation, we accordingly present the source images and summary images in a shuffled order and explicitly tell the annotators (1) the images are present in a random order (2) focus on the importance and representativeness of the images.

4.3.3 Text Bias in Overall Evaluation

When humans are rating an overall score for a multimodal summary, we hypothesize that the annotators would heavily rely on the text evaluation results. To this end, we sample 100 examples and the annotators are asked to evaluate each example twice. In the first stage, the annotators are shown with only the source documents and the text summaries, while being shown the full example

(including the images) in the second stage. The mean overall score of the first and second stages is 3.38 and 3.35, while the standard deviation is 1.51 and 1.56. In the second stage, the mean image score and text-image score are 2.14 and 2.43. Although the image and text-image scores are much lower than the text score, the mean overall score is still very similar to the first stage. The empirical results indicate that the evaluators tend to rely on the text evaluation results, which is able to verify our hypothesis. For the text bias, we explicitly tell the annotators to focus on the overall quality of the multimodal summaries without any prior preferences, and the annotators have to penalize the overall quality if the image score or text-image score is low, with a few examples for demonstration in the evaluation criteria discussion before we conduct the annotation.

5 Calls for Future Research

We hope this paper will be a valuable resource for future research on multimodal summarization evaluation and models. The study in this work demonstrates the need for future research on (1) better automatic evaluation metrics that could properly and objectively evaluate the quality of multimodal summaries; (2) more meta-evaluation in multimodal summarization to update with the current advanced systems and datasets (3) more study on cognitive biases in human evaluation for multimodal summarization, to have a more fair and accurate comparison of different systems or evaluation metrics. We hope this work could demonstrate the importance of these issues and call for future research (evaluation, meta-evaluation, human evaluation), which would significantly benefit the multimodal summarization research community.

6 Conclusion

In this paper, we propose a research framework, where we investigate the automatic evaluation, meta-evaluation and cognitive bias analysis in human evaluation for multimodal summarization. For automatic evaluation, we propose a reference-free metric based on multimodal LLMs that correlates well with human judgments. Furthermore, we collect human annotations for multimodal summaries and release a meta-evaluation benchmark to evaluate various evaluation metrics. For the human evaluation, we study and verify three types of cognitive biases in human evaluation. We believe our work would be a valuable resource for multimodal summarization research, and hope this work could demonstrate the importance of these topics as well as encourage further research in this area.

Limitations

One limitation of our work is that we only consider the image as the vision modality. VMSMO (Li et al., 2020) introduces the video-text-image summarization. Besides, multimodal summarization could also contain audio or other modalities. We leave these as our future work. In addition, our work only focuses on the English language, which could also be extended to multilingual settings for a more comprehensive evaluation.

Cognitive biases in human evaluation could also include other biases that are not investigated in this paper, such as framing effects (Schoch et al., 2020). We would also like to have more exploration of this research topic in the future.

Ethical Considerations

We use the publicly available dataset MSMO (Zhu et al., 2018) to build our meta-evaluation benchmark dataset. Also, we did not collect any personal information or free-form text, therefore we consider the risk of releasing our data very low. We intend to use the dataset only for research purposes.

In addition, we hire three annotators to label our benchmark dataset and the compensation ($35/h) for the annotators is higher than the local minimum wage.

Acknowledgments

This work is supported by the Early Career Industry Fellowship IE230100119 (Australian Research Council), Discovery Project DP240103070 (Australian Research Council), and Sustainability FAME Strategy Internal Grant (The University of Adelaide). The authors sincerely thank Hu Wang for the insightful discussion, as well as all the anonymous reviewers for their valuable comments and feedback.

References

Forrest Sheng Bao, Ruixuan Tu, and Ge Luo. 2022. Docasref: A pilot empirical study on repurposing reference-based summary quality metrics reference-freely. *arXiv preprint arXiv:2212.10013*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359.

Cheng-Han Chiang and Hung-yi Lee. 2023. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, pages 1553–1568.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, pages 391–409.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Hunter Gehlbach and Scott Barge. 2012. Anchoring and adjusting in questionnaire responses. *Basic and Applied Social Psychology*, pages 417–433.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 2015 International Conference on Neural Information Processing Systems*, page 1693–1701.

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.

Anubhav Jangra, Sourajit Mukherjee, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2023. A survey on multi-modal summarization. *ACM Comput. Surv.*

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset.

Wojciech Kryscinski, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Neural text summarization: A critical evaluation. 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 540–551.

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.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1092–1102.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020. VMSMO: Learning to generate multimodal summary for video-based news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522. Association for Computational Linguistics.

Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis, and Ioannis Pitas. 2016. Multimodal stereoscopic movie summarization conforming to narrative characteristics. *IEEE Transactions on Image Processing*, pages 5828–5840.

Natwar Modani, Pranav Maneriker, Gaurush Hiranandani, Atanu R. Sinha, Utpal, Vaishnavi Subramanian, and Shivani Gupta. 2016. Summarizing multimedia content. In *Proceedings of the 17th International Conference on Web Information Systems Engineering - Volume 10042*, page 340–348.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 2016 SIGNLL Conference on Computational Natural Language Learning*, pages 280–290.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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 on Association for Computational Linguistics*, page 311–318.

Weizhen Qi, Yu Yan, Yeyun Gong, Dayiheng Liu, Nan Duan, Jiusheng Chen, Ruofei Zhang, and Ming Zhou. 2020. Prophetnet: Predicting future n-gram for sequence-to-sequence pre-training. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2401–2410.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Sashank Santhanam, Alireza Karduni, and Samira Shaikh. 2020. Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, page 1–13.

Stephanie Schoch, Diyi Yang, and Yangfeng Ji. 2020. "this is a problem, don't you agree?" framing and bias in human evaluation for natural language generation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 10–16.

Sam Shleifer and Alexander M Rush. 2020. Pre-trained summarization distillation. *arXiv preprint arXiv:2010.13002*.

Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, pages 1124–1131.

Oleg Vasilyev, Vedant Dharnidharka, and John Bohannon. 2020. Fill in the BLANC: Human-free quality estimation of document summaries. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 11–20.

David Wan and Mohit Bansal. 2022. Evaluating and improving factuality in multimodal abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9632–9648.

Linzi Xing, Wen Xiao, and Giuseppe Carenini. 2021. Demoting the lead bias in news summarization via alternating adversarial learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 948–954.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 2020 International Conference on Machine Learning*, pages 11328–11339.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2022a. Unims: A unified framework for multimodal summarization with knowledge distillation. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11757–11764.

Zhengkun Zhang, Jun Wang, Zhe Sun, and Zhenglu Yang. 2021. Lams: A location-aware approach for multimodal summarization (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence*, (18):15949–15950.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022b. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2021. Leveraging lead bias for zero-shot abstractive news summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1462–1471.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal summarization with multimodal output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4154–4164.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9749–9756.

A Prompts for Evaluation

A.1 Prompts for the text score

For the text summary evaluation, we mainly follow the prompts in Liu et al. (2023) and show the attribute-specific prompts (relevance, coherence, consistency, fluency) as follows,

relevance of text summary
You will be given one summary written for a news article.
Your task is to rate the summary on one metric.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Relevance (1-5) - selection of important content from the source document and images. The summary should include only important information from the source document and images. Annotators were instructed to penalize summaries which contained redundancies and excess information.
Evaluation Steps:
1. Read the summary and the source document carefully.
2. Compare the summary to the source document and identify the main points of the article.
3. Assess how well the summary covers the main points of the article, and how much irrelevant or redundant information it contains.
4. Assign a score for relevance on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.
Example:
Source Text: {{Document}}
Summary: {{Summary}}
Evaluation Form (scores ONLY):
- Relevance:

For the evaluation of *coherence*, *consistency* and *fluency*, the descriptions of the evaluation task, the example for evaluation as well as the scoring form are the same as the above prompt for *relevance*. We thus list the evaluation criteria and steps for these evaluation attributes as follows,

coherence of text summary
Evaluation Criteria:
Coherence (1-5) - the collective quality of all sentences. We align this dimension with the DUC quality question of structure and coherence whereby "the summary should be well-structured and well-organized. The summary should not just be a heap of related information, but should build from sentence to a coherent body of information about a topic."
Evaluation Steps:
1. Read the news article carefully and identify the main topic and key points.
2. Read the summary and compare it to the news article. Check if the summary covers the main topic and key points of the news article, and if it presents them in a clear and logical order.
3. Assign a score for coherence on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

consistency of text summary
Evaluation Criteria:
Consistency (1-5) - the factual alignment between the summary and the summarized source. A factually consistent summary contains only statements that are entailed by the source document. Annotators were also asked to penalize summaries that contained hallucinated facts.
Evaluation Steps:
1. Read the news article carefully and identify the main facts and details it presents.
2. Read the summary and compare it to the article. Check if the summary contains any factual errors that are not supported by the article.
3. Assign a score for consistency on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

fluency of text summary
Evaluation Criteria:
Fluency (1-5) - the quality of individual sentences in terms of grammar, spelling, punctuation, word choice, and sentence structure.
Evaluation Steps:
1. Read the given summary carefully.
2. Assign a score for fluency on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.

A.2 Prompts for the other scores

For the other scores (image score, text-image relevance score and overall quality score), the multimodal LLM is required to evaluate the quality of the images in the multimodal summary. Thus, we design the following prompts.

quality of the summary images
You will be given one text summary and some summary images (ID) written for a news article with source images. The summary images is a subset of the source images (e.g., Summary Image ID: 1,3 refer to the first and third source images).
Your task is to rate the quality of summary images.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Quality (1-5) - how well the summary images could represent the content of the news article.
Evaluation Steps:
1. Read the source document and source images carefully.
2. According to the source document, identify which source images are most important and representative.
3. Assess how well the summary images could represent the content of the news article.
4. Assign a score for quality on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.
Example:
Source Text: {{Document}}
Summary: {{Summary}}
Summary Images: {{Image IDs}}
Evaluation Form (scores ONLY):
- Quality:

relevance of the text summary and summary images
You will be given one text summary and some summary images written for a news article with source images.
Your task is to rate the relevance of the text summary and summary images.
Please make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
Evaluation Criteria:
Relevance (1-5) - how well the text summary and summary images match. A well-matched example means the text summary and summary images could represent each other.
Evaluation Steps:
1. Read the text summary and images carefully.
2. Assess how well the summary images could represent the content of the text summary, and how well the text summary could represent the summary images.
4. Assign a score for relevance on a scale of 1 to 5, where 1 is the lowest and 5 is the highest based on the Evaluation Criteria.
Example:
Summary: {{Summary}}
Summary Images: {{Images}}
Evaluation Form (scores ONLY):
- Relevance:

For the overall quality evaluation, we show the text prompt in Section 3.1.

B Text Summarization Models

For our meta-evaluation benchmark construction, we use multiple summarization models to generate text summaries given the sampled multimodal inputs (details in Section 3.2). Specifically, we use the following text summarization models:

- **BART** (Lewis et al., 2020): BART is a Transformer-based denoising autoencoder that is pretrained by reconstructing the original text given the corrupted document, which shows a competitive results in summarization. We use the BART model that is fine-tuned on CNN/Daily Mail summarization dataset (Hermann et al., 2015; Nallapati et al., 2016).

- **Distilbart** (Shleifer and Rush, 2020): Distilbart is a student model distilling knowledge from the teacher model BART. Similar to BART, we also use the model fine-tuned on CNN/Daily Mail dataset (Hermann et al., 2015; Nallapati et al., 2016).

- **GPT2** (Radford et al., 2019) GPT2 is a large transformer-based language model trained on massive data by predicting the next token in a sequence. As GPT2 is not specifically designed for summarization, we add "TL;DR:" after the source document and set it as the input to GPT2 to generate the summary.

- **PEGASUS** (Zhang et al., 2020): PEGASUS is a large Transformer-based encoder-decoder model. The pretraining for PEGASUS is specifically designed for abstractive summarization, where the model is trained to reconstruct the masked sentences given the unmasked context in a self-supervised learning manner.

- **ProphetNet** (Qi et al., 2020): ProphetNet is pretrained to predict the next n tokens given the previous context tokens at each time step, which drives the model to plan and strategize for future tokens generation. The model shows its effectiveness in abstractive summarization.

- **T5** (Raffel et al., 2020): T5 is an encoder-decoder model pre-trained on multiple tasks, where each task is transformed into a text-to-text format. To guide the model to generate text summary for the given document, we add a "summarize:" prefix to the source document as input to the model.

- **HAN** (Zhu et al., 2018): HAN model is specifically designed for multimodal summarization with multimodal outputs, including a text encoder, an image encoder, a multimodal attention layer and a summary decoder. HAN utilizes hierarchical visual attention for generating text summaries and selecting the most relevant images.

- **GPT4** (OpenAI, 2023): GPT-4 is a Transformer-based large language model that could perform various natural language processing tasks, demonstrating its understanding and generation ability. To perform summarization, we use "*Summarize content you are provided with and please do not exceed 100 words.*" as the prompt to the GPT-4.

- **GPT4 with vision** (OpenAI, 2023): GPT-4 with vision is a multimodal large language model, which is able to take both texts and images as inputs. In this work, we provide the multimodal input (source document and source image set) to GPT-4 with vision and obtain the summary.

C Examples in Meta-Evaluation Benchmark

We construct the meta-evaluation benchmark for evaluating the quality of automatic evaluation metrics (details in Section 3.2). A good metric should be able to assess the quality of both *good* examples and *bad* examples. Thus, the benchmark should also contain both good examples and bad examples for assessing the quality of an evaluation metric. Here we list two good examples in Fig. 4, 5 and bad examples in Fig. 6, 7 of our benchmark for illustration.

D Multiple Summary Image Set

We report the inter-annotator agreement in the paper (details in Section 3.2), where the agreement of image score is not as high as other aspects (such as text score). We believe it is reasonable because for a given multimodal input (including a source document and a source image set), there might be more than one summary image set that is acceptable. Here we provide two examples for demonstration in Fig. 8 and 9, where the summary image sets that are selected by three annotators are different from each other.

E Inconsistency in the References

During the human annotation in our work, we notice that the reference multimodal summaries could not always be perfect. For the text part, the reference text sometimes contains some information that is not supported by the source document. We believe this is because of the process of collecting the dataset in (Zhu et al., 2018), where the body of news articles is considered as the source document while the highlight is treated as the summary. The highlight of the news articles could possibly include some information that is not present in the body. We show two examples here to demonstrate the inconsistency issue in the reference text summary, as in Fig. 10 and 11.

F Caption Extraction vs. Generation

We observe that in the caption-based method ("CAP"), "EXT" (captions extracted from the dataset) slightly outperforms "GEN" (captions generated by a pretrained model) mostly. We believe the main reason is the caption's quality and informativeness. In the MSMO dataset (Zhu et al., 2018), each image is paired with an expert-annotated caption to describe the content of the image. Thus the extracted caption would be more informative or accurate than the generated caption. In Fig. 12, we list four examples of the images as well as their captions.

Source Document

SK Gaming won # 189,000 after fending off tough competition to be crowned the Epicenter 2017 CS : GO champions . They were taken to the limit as Virtus Pro forced the series to a final fifth game , testing the endurance of the players who are used to playing best of threes . SK Gaming held their nerve to win the final game but their celebrations have cut short as the airport security in Russia has sieged their trophy as it has bullets in the design . Gabriel ` FalleN ' Toledo posted a picture on Twitter of the trophy stuck at the airport and said the team were trying their best to get it back . It would be a shame if they did lose the trophy considering how hard they had to work for it . Virtus Pro came flying out of the gates and took the first game of the series 16-10 . SK Gaming then responded with a thumping 16-6 win . Virtus Pro halted their momentum and looked set to regain the lead with a thrashing of their own with a score of 12-3 . However , SK Gaming mounted an incredible comeback and for the first time , took the lead in the series , winning 19-16 . SK only needed one more win yet Virtus Pro held their nerve . They were 9-6 down but mounted a comeback of their own and won 16-12 to force a final game . Both teams traded rounds in an intense finish , SK were able to force overtime and win 22-18 to claim the Epicenter trophy , their sixth championship of the year .

Source Image Set


Summary

SK Gaming have won the Epicenter 2017 CS : GO champions. They were taken to the limit as Virtus Pro forced the series to a final fifth game. They held their nerve to win the final game 22-18 to claim the trophy. But their celebrations have cut short as the airport security in Russia has sieged their trophy as it has bullets in the design.

Summary Image Set


Human Annoatation

	rel	coh	con	flu	imq	tir	ova
A1	5	4	5	4	5	5	4
A2	5	5	5	5	5	5	5
A3	5	5	5	5	5	5	5

Figure 4: A good example (with the annotation scores of three annotators) in our meta-evaluation benchmark. The “rel",“coh", “con",“flu",“imq",“tir", “ova" means “relevance", “coherence", “consistency", “fluency", “image quality", “text-image relevance", “overall quality" respectively. (also same in Fig. 5, 6, 7)

Source Document

Angry villagers have torched a huge mock-up of a BT van on a bonfire in protest at slow rural internet speeds . The annual bonfire night in the small hamlet of Templeton , Devon , is celebrated each year with a different theme . This year fed-up locals made an effigy of a BT Openreach transit van - with the word ` open ' replaced by ` wo n't ' - during a Bonfire Night protest . The two-dimensional model , which was 1.5 times the size of a real Openreach van , went up in smoke at the Templeton Bonfire and Fireworks Night . The rural hamlet was not part of the commercial roll-out of fibre broadband by BT , or the first phase of the Connecting Devon and Somerset partnership . This means many villagers are struggling with speeds of less than 1 megabit . Villager Adam Short , who moved to Templeton last year , said he helped create the van effigy on the floor of his barn . He said : ` We knew it was terrible before we moved , but we hoped there would be a solution . ` Trying to run my business from home is nigh on impossible at times , and I 'm one of the lucky ones because I have a 4G signal on the roof with some specialist kit . ` It also has an impact on the children in the village as it 's restricting their homework . ' He added : ` Upload speeds are almost zero . There really are very few places in our village where a 2G phone signal can be reliably found , let alone 4G . ' Roger Linden said villagers were told the problem would be looked at three years ago , but nothing has happened . He said : ` They managed to get a cable to the nearby hamlet of Nomansland , but just eight kilometres further and there 's nothing . ` It 's incompetence of the first order ... but we all had a great evening watching the bonfire . ' He said he can not stream anything online and is only able to look at emails and occasionally browse the internet . A spokesman from BT said : ` Templeton is an extremely rural community which makes rolling out fibre broadband much more challenging . ` Templeton was not included in Openreach 's commercial roll-out of fibre broadband or the first phase of the Connecting Devon and Somerset partnership but we 're working hard to find alternative ways of bringing faster broadband to residents . ' .....

Source Image Set

Summary

A model of a BT van was an effigy burnt to protest at broadband in Templeton Locals are furious it was not part of the commercial roll-out of fibre broadband It was also not in the first phase of Connecting Devon and Somerset partnership This means many villagers are struggling with speeds of less than 1 megabit

Summary Image Set

Human Annoatation

	rel	coh	con	flu	imq	tir	ova
A1	4	4	5	5	5	5	4
A2	5	5	5	5	4	5	5
A3	5	5	5	5	3	5	5

Figure 5: A good example (with the annotation scores of three annotators A1, A2, A3) in our meta-evaluation benchmark.

Source Document

One of Germany 's richest women is taking her former best friend to court - claiming her new novel actually reveals parts of her private life . Babette Albrecht , the widow of Aldi heir Berthold Albrecht , is demanding several passages in Dorothee Achenbach 's best-selling Now Everybody Knows my Laundry are removed . Despite never being mentioned by name , the heiress says a character known only as ` the widow ' - who avoids a multi-million euro inheritance bill - is actually a thinly disguised version of herself . This is the latest in a long-running saga between the two families , who became firm friends after running into each other in a restaurant in 2007 , according to German tabloid Bild . In particular , the women 's husbands , the late Berthold - who died in 2012 - and art dealer Helge Achenbach , hit it off , travelling the world together to art exhibitions . But the relationship has soured in recent years , culminating in Babette taking Helge to court , accusing him of selling her husband art and cars at over inflated costs . Helge was ordered to pay back 20million euros to Babette , 55 , and later jailed for six years . Dorothee 's new novel , which has sold 40,000 copies so far , describes their fall from grace . However , the 52-year-old has changed their names to provide ` distance ' between the events in the book , and her family . But it has apparently incensed her former friend , who feels the character she believes to be her is an ` insult ' . In total , she is asking for eight passages to be removed , with Bild claiming she is seeking damages to the sum of six figures .

Source Image Set

Summary

Here's hoping for just five. But please note - none of Dorothee Albrecht"s previous novels were mentioned by name.

Summary Image Set

Human Annoatation

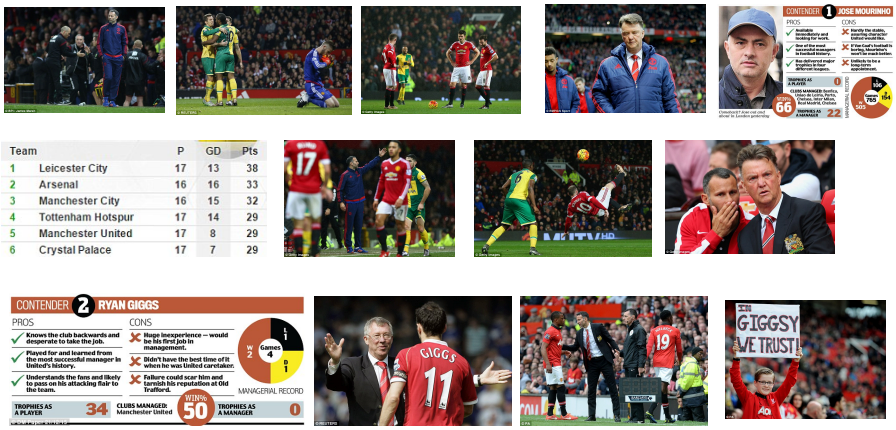
	rel	coh	con	flu	imq	tir	ova
A1	1	2	2	5	5	2	2
A2	1	3	1	4	4	1	1
A3	1	1	1	1	2	1	1

Figure 6: A bad example (with the annotation scores of three annotators A1, A2, A3) in our meta-evaluation benchmark. The text summary is in poor quality without any important information of the source document.

Source Document

It was perhaps not deliberate but the distance between Ryan Giggs and Louis van Gaal on Saturday afternoon was symbolic nevertheless . During a second half that saw Manchester United briefly entertain hopes of an old-school comeback , Giggs was resident throughout in the technical area , that small piece of Old Trafford real estate that Van Gaal seems to treat with such suspicion . At full-time meanwhile , Giggs was off down the touchline and down the tunnel , not waiting to witness the dissent he knew was about to roll down the Stretford End in the direction of the United manager . Giggs , 42 , has been a significant figure at United for a quarter of a century now and knows what he is witnessing this season -- what he is part of -- does not fit . When he briefly took over as interim manager after David Moyes was sacked in April 2014 , Giggs ' first act was to gather the United players together and tell them that a return to traditional principles of attacking play was called for . One can only imagine , therefore , what runs through his mind these days . United 's football , in terms of the team 's strategies and philosophies , have got worse since Moyes ' time , not better . There are those who know Giggs who wonder if the Welshman would be tempted to walk away , look to begin his management career somewhere else . There has been gossip to that effect but nothing concrete . What we do know is that Giggs does feel some loyalty towards Van Gaal , despite the fact he would prefer his team to play rather more on the front foot . He remains grateful to the former Barcelona and Bayern Munich manager for the opportunity to work under him and already feels he has learned a great deal . Undoubtedly , though , there is a groundswell of opinion among the United support that would like to see him given his opportunity to take charge of the first team again , at least until the end of the season .....

Source Image Set



	Team	P	GD	Pts
1	Leicester City	17	13	38
2	Arsenal	16	16	33
3	Manchester City	16	15	32
4	Tottenham Hotspur	17	14	29
5	Manchester United	17	8	29
6	Crystal Palace	17	7	29

Summary

What we know will come true, of those who have watched last night's match after Giggs' exit -- as well as the number of football players who will join the U20 team and, if done right, even more...

Summary Image Set

Human Annoatation

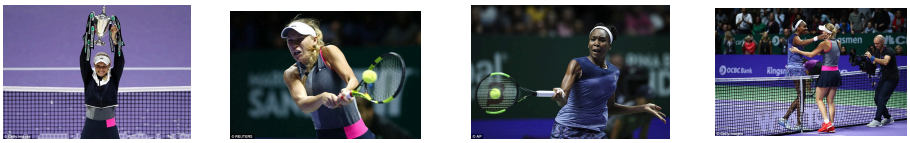
	rel	coh	con	flu	imq	tir	ova
A1	1	2	1	4	3	1	2
A2	1	2	1	4	4	1	2
A3	1	2	1	2	2	2	2

Figure 7: A bad example (with the annotation scores of three annotators A1, A2, A3) in our meta-evaluation benchmark. The text summary does not capture any the key content while being incoherent and not easy to understand for humans.

Source Document

The six biggest prizes in women 's tennis going to six different players tells the story of this season on the WTA Tour . Caroline Wozniacki further emphasised the lack of a defining player in 2017 when she defeated Venus Williams 6-4 , 6-4 to win the $ 7million WTA Finals , open to the top eight performers of 2017 . It means that the four Grand Slam titles , the year-end No 1 spot and winner of the year-end championships are all in different hands . Serena Williams (Australian Open), Jelena Ostapenko(French Open), Garbine Muguruza(Wimbledon), Sloane Stephens (US Open), Caroline Wozniacki (WTA Finals), Simona Halep(World No 1). Neither Wozniacki (right)nor the season 's No 1 , Simona Halep , won a major . With Serena Williams absent since taking the Australian Open , it has been an unpredictable scramble for trophies . This was Wozniacki 's first win over Venus Williams in their eighth meeting . ` Eight is my lucky number so I was hoping if I was going to beat her it would be today , ' she said . Meanwhile , Roger Federer beat Juan Martin del Potro 6-7 , 6-4 , 6-3 on Sunday to win the Swiss Indoors event in Basle . After winning his seventh title of the season , Federer withdrew from this week 's Paris Masters , the last event before the ATP Finals at the O2 Arena in London . Rafael Nadal now needs to win just one more match to clinch the year-end world No 1 spot .

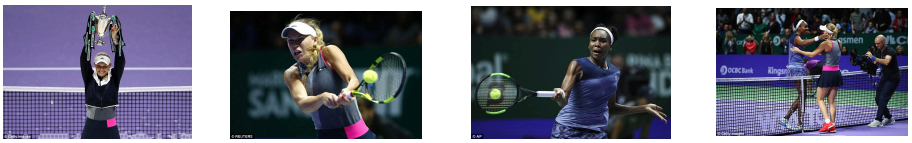
Source Image Set



Summary

Caroline Wozniacki beat Venus Williams 6-4 , 6-4 to win the WTA Finals The Dane broke her opponent 's serve to secure her second title of the year Williams was the oldest woman to reach the final at the age of 37

Summary Image Set (A1)



Summary Image Set (A2)

Summary Image Set (A3)

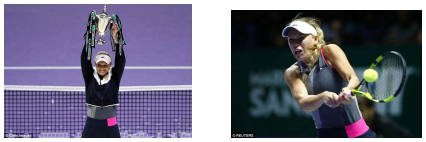
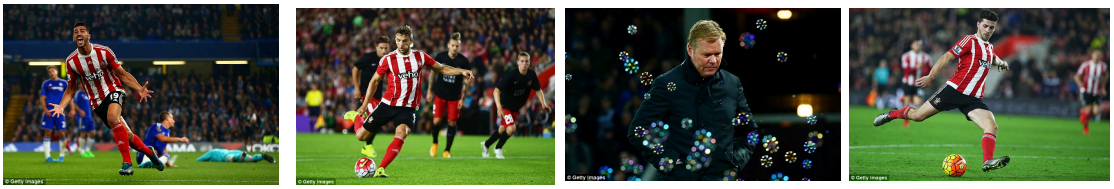


Figure 8: An example with multiple summary image sets annotated by three annotators. A1 selects all images from the source image set, while A2 and A3 only select one and two images respectively.

Source Document

Southampton boss Ronald Koeman admits he may have to dip into the transfer market for a striker if his current frontmen remain unfit . Koeman is currently without Graziano Pelle and Jay Rodriguez through injury , although both are expected to be back in action in the near future . Pelle has been missing since Christmas with a knee injury and will also sit out this weekend 's trip to Norwich , while Rodriguez is due back in late January after foot surgery . But Koeman is weighing up his attacking options with the January transfer window due to open on Saturday . When asked whether he was about to dip into the transfer market , Koeman said at his pre-match press conference : ` At the moment , no . ` Like I mentioned before , if we need to sign a player , it needs to be a striker . ` But that also depends on the situation of Jay and what will happen with Graziano . It 's too early to make conclusions about that . ` We still have to wait to see how they recover and how long it takes , and then we 'll make a decision about what we have to do . ' Shane Long is set to lead the line again for Saints at Carrow Road .

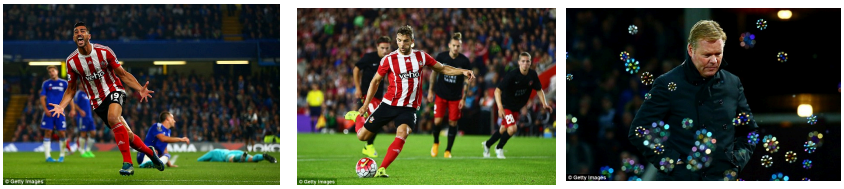
Source Image Set



Summary

Southampton forwards Graziano Pelle and Jay Rodriguez are both injured Both players are expected back next month after knee and foot problems Boss Ronald Koeman says he could move to sign a new striker in January Saints take on Norwich City at Carrow Road on Saturday , kick-off at 3pm

Summary Image Set (A1)



Summary Image Set (A2)

Summary Image Set (A3)

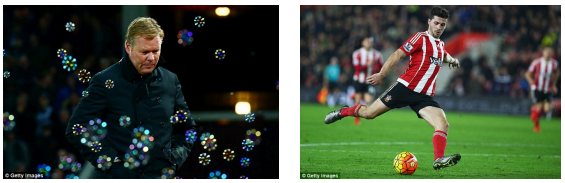
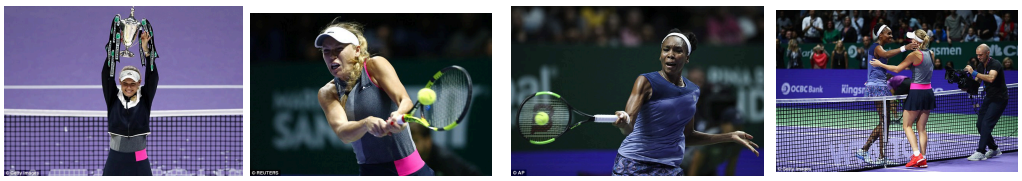


Figure 9: An example with multiple summary image sets annotated by three annotators. The results of A1, A2 and A3 are different from each other.

Source Document

The six biggest prizes in women 's tennis going to six different players tells the story of this season on the WTA Tour . Caroline Wozniacki further emphasised the lack of a defining player in 2017 when she defeated Venus Williams 6-4 , 6-4 to win the $ 7million WTA Finals , open to the top eight performers of 2017 . It means that the four Grand Slam titles , the year-end No 1 spot and winner of the year-end championships are all in different hands . Serena Williams (Australian Open), Jelena Ostapenko(French Open), Garbine Muguruza(Wimbledon), Sloane Stephens(US Open), Caroline Wozniacki(WTA Finals), Simona Halep(World No 1). Neither Wozniacki (right) nor the season 's No 1 , Simona Halep , won a major . With Serena Williams absent since taking the Australian Open , it has been an unpredictable scramble for trophies . This was Wozniacki 's first win over Venus Williams in their eighth meeting . ` Eight is my lucky number so I was hoping if I was going to beat her it would be today , ' she said . Meanwhile , Roger Federer beat Juan Martin del Potro 6-7 , 6-4 , 6-3 on Sunday to win the Swiss Indoors event in Basle . After winning his seventh title of the season , Federer withdrew from this week 's Paris Masters , the last event before the ATP Finals at the O2 Arena in London . Rafael Nadal now needs to win just one more match to clinch the year-end world No 1 spot .

Source Image Set



Summary

Caroline Wozniacki beat Venus Williams 6-4 , 6-4 to win the WTA Finals The Dane broke her opponent 's serve to secure her second title of the year. Williams was the oldest woman to reach the final at the age of 37.

Summary Image Set

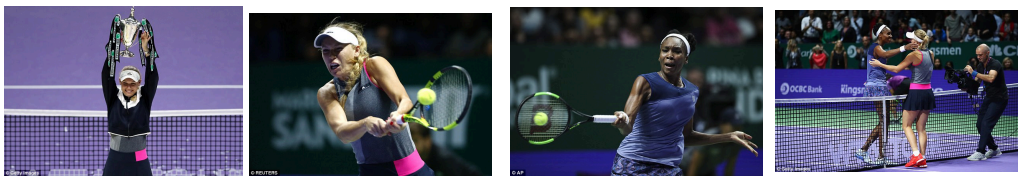


Figure 10: An example of reference summary with inconsistent content (the blue text: we are not able to infer that Williams was 37 from the source document.)

Source Document

When police knocked on Kelly McPherson 's door to ask if she 'd witnessed a stabbing nearby , her heart sank . Her stepson was yet to return home and she instinctively knew he was the victim . She said : ` My first thought was , `` Oh no I hope it 's not Michael . '' ` I rang his mobile three times and there was no answer . I knew it was him who had been stabbed , I just knew it . ' Mrs McPherson ran to a nearby park on Thursday evening after finding out about the attack . She discovered air ambulance paramedics trying to resuscitate her stepson Michael Jonas , 17 . He was pronounced dead nearly an hour later . The attack in Betts Park , near Crystal Palace in South London , comes as police battle a knife crime epidemic in the capital . A total of 22 teenagers have been murdered in London so far this year -- 16 of whom were stabbed . The number of teenagers murdered in the capital is now at its highest in nine years .....There has been a 47 per cent increase in stabbings in London this year . London Mayor Sadiq Khan held a knife crime summit with teachers earlier this week and Croydon Central MP Sarah Jones has called on Home Secretary Amber Rudd to tackle the problem . Miss Jones said : ` Time and time again we 've said `` enough is enough '' as knife crime has doubled in a year . Well over 1,000 young people were stabbed in London last year . ` Now we demand action , not words . I 've been pushing the Home Secretary to prioritise this epidemic among our young . It ca n't be fixed with short-term thinking . ` We need a ten-year , cross-government strategy . ' Police said there have been no arrests .

Source Image Set

Summary

Police and paramedics called to south London park last night after stabbing The victim , 17 , was treated at the scene but tragically died from his wounds. A knife crime takes place in England and Wales every 14 minutes on average.

Summary Image Set

Figure 11: An example of reference summary with inconsistent content (the blue text: there is no information about "the frequency of knife crime" in the source document.)

Caption Generation:
a man and woman holding up trophies in front of a crowd.

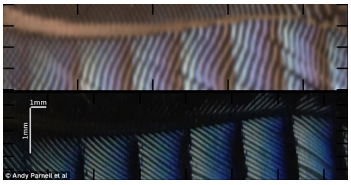
Caption Extraction:
Hingis celebrates with Jamie Murray after winning the Wimbledon mixed doubles title in 2017

Caption Generation:
there are two men sitting next to each other in a stadium

Caption Extraction:
Rafa Benitez will be watching the Under-17 World Cup final between England and Spain

Caption Generation:
a large fire is seen in the distance as it burns

Caption Extraction:
Residents of Lorne, which is less than 15 kilometres from the festival, were issued with an emergency evacuation warning early on Saturday morning. They were told they were in danger and it was 'too late to leave'



Caption Generation:
a close up of a camera lens with a picture of a person on a cell phone

Caption Extraction:
The researchers found as the structure of the feather barbs could be altered along their length, they would form complex and multicoloured patterns (pictured)

Figure 12: Four examples with captions by generation or extraction. The extracted captions are more informative or accurate. In the first two examples, the extracted caption has the **names**(e.g., Hingis, Jamie Murray, Rafa Benitez), **time**(e.g., 2017), and **events**(e.g., Wimbledon mixed doubles, Under-17 World Cup). In the third example, the extracted caption provides more **details** than the generated one. In the last example, the generated caption has an erroneous description while the extracted caption accurately provides the content of the image.