Pay More Attention to Images: Numerous Images-Oriented Multimodal Summarization

Min Xiao[1,2], Junnan Zhu[1]*, Feifei Zhai[1,3], Chengqing Zong[1,2], Yu Zhou[1,3]*
[1] State Key Laboratory of Multimodal Artificial Intelligence Systems,
Institute of Automation, CAS, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Fanyu AI Laboratory, Zhongke Fanyu Technology Co., Ltd, Beijing, China
{min.xiao, junnan.zhu, yzhou, cqzong}@nlpr.ia.ac.cn, zhaifeifei@zkfy.com

Abstract

Existing multimodal summarization approaches struggle with scenarios involving numerous images as input, leading to a heavy load for readers. Summarizing both the input text and numerous images helps readers quickly grasp the key points of multimodal input. This paper introduces a novel task, **N**umerous **I**mages-Oriented **M**ulti**m**odal **S**ummarization (NIMMS). To benchmark this task, we first construct the dataset based on a public multimodal summarization dataset. Considering that most existing metrics evaluate summaries from a unimodal perspective, we propose a new **M**ultimodal **info**rmation evaluation (M-info) method, measuring the differences between the generated summary and the multimodal input. Finally, we compare various summarization methods on NIMMS and analyze associated challenges. Experimental results show that M-info correlates more closely with human judgments than five widely used metrics. Meanwhile, existing models struggle with summarizing numerous images. We hope that this research will shed light on the development of multimodal summarization. Furthermore, our code and dataset will be released to the public[1].

1 Introduction

Multimodal summarization is an emerging research area driven by advancements in multimodal learning. Images serve to either spotlight the essential content or to provide rich information. Most existing multimodal summarization studies focus on identifying the most informative visual keyframes and distilling the text content into pivotal points.

While tasks involving both multimodal input and multimodal output have been extensively explored, existing studies struggle to account for scenarios where a multitude of images are presented as input. Consequently, this leads to a lack of summarization for numerous images within the text summary. For example, as shown in Figure 1, the input contains a document and numerous images, each capturing a scene of remarkable diversity. It is impractical to expect a single image to encapsulate all the visual information from such a diverse set. However, these images can be summarized in a single sentence, *i.e.*, *A young woman enjoying yachts, luxury cars, exotic vacations, and high-end fashion*. This sentence condenses all the visual inputs and complements the original document's summary, highlighting the advantage of textual descriptions over a few images when summarizing multiple images.

Moreover, in scenarios such as disaster scenes, sports events, product appearances, hotel conditions, *etc.*, readers tend to prefer summaries that quickly encapsulate the key information. Therefore, we argue that providing a textual summary for numerous images is essential for three reasons: 1) Condensing the numerous images into a concise

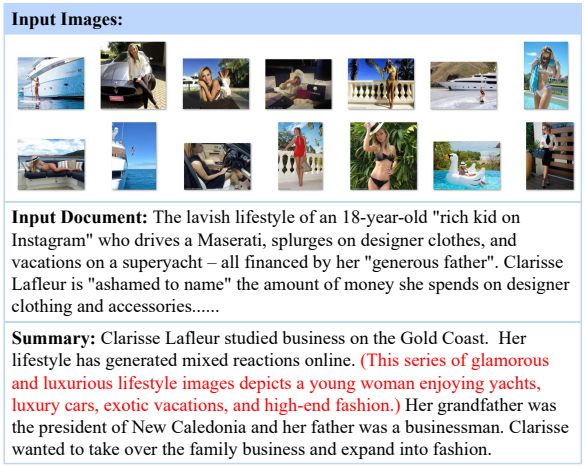


Figure 1: An example of NIIMS. In the summary, the text in black denotes the original summary, while the text in red is the additional summary of the input images that our task aims to include. Best viewed in color.

* Corresponding author.
[1]Please find code and dataset at https://github.com/xiaomin-plus/NIMMS

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9379–9392

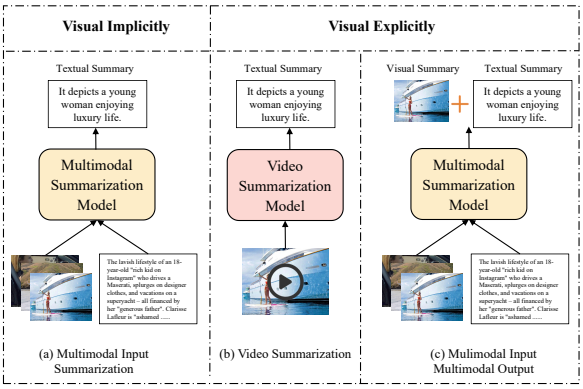


Figure 2: Existing multimodal summarization tasks.

summary significantly reduces the reading load for users. 2) While existing methods (Zhu et al., 2018; Chen and Zhuge, 2018; Zhu et al., 2020) select representative images as the image summary, a few images alone cannot fully convey the content presented across multiple images. 3) Compared to representative images, textual descriptions enable a more rapid and efficient comprehension of the input images.

Existing multimodal summarization studies can be categorized based on the role of visual input into two types: *visual implicitly* and *visual explicitly*. *Visual implicitly* means that images serve to highlight the core content of the text (Li et al., 2020a, 2018; Xiao et al., 2023), as shown in multimodal-input-text-output tasks (Figure 2 (a)). *Visual explicitly* indicates that images are either summarized as text or directly selected to form a summary, as shown in Figure 2 (b) and (c), respectively. The former is exemplified by video summarization (Li et al., 2020b; Sanabria et al., 2018; Liu et al., 2023c; Atri et al., 2021), which aims to produce a textual summary for a video; the latter is exemplified by multimodal-input-multimodal-output summarization (Zhu et al., 2018; Chen and Zhuge, 2018; Zhu et al., 2020), where the visual content is represented by the most relevant image.

Based on the above discussion, in this study, we introduce a novel task, **N**umerous **I**mages-oriented **M**ulti**M**odal **S**ummarization (NIMMS). Our exploration focuses on two key questions: 1) how to acquire the relevant data; 2) how to automatically evaluate the quality of the summary from a holistic multimodal perspective. For the first, based on the existing multimodal dataset, we select the samples with numerous images and reconstruct their summaries. For the second, we propose a new metric, **M**ultimodal **info**rmation evaluation (M-info), which measures the distributional differences between the summary and multimodal inputs. Specifically, M-info calculates the aggregate distribution of both input and output information, and then computes the KL divergence between these distributions. Finally, we conduct experiments using both existing summarization methods and multimodal LLM, and discuss the unique challenges presented by NIMMS.

Our main contributions are as follows:

- We introduce the NIMMS task, a pioneering effort to generate a text summary that encapsulates both the textual input and numerous image inputs. Furthermore, we construct a dataset to facilitate NIMMS research.
- We propose a novel evaluation metric, M-info, which quantifies the alignment between the generated summary and the multimodal inputs by analyzing their distributional consistency.
- We conduct a comprehensive analysis of various summarization methods and multimodal large language model applied to the NIMMS task, identifying its unique challenges and offering insights for future research to enhance the performance.

2 Related Work

Multimodal Summarization. With the rapid progress of multimedia, various multimodal summary tasks have emerged, such as multimodal input summarization (Li et al., 2018; Jangra et al., 2021; Overbay et al., 2023), multimodal summarization with multimodal output (Zhu et al., 2018; Liang et al., 2023), video summarization (Sanabria et al., 2018; Yu et al., 2021; Mahasseni et al., 2017; Liu et al., 2024), topic-aware multimodal summarization (Mukherjee et al., 2022). Although multimodal summarization receives increasing attention, current studies struggle to account for scenarios where a multitude of images are presented as input. Furthermore, a pictorial summary, consisting of a text summary and a representative image, cannot summarize all the visual information. Hence, it is vital to address summarization tasks involving numerous images.

Summarization Evaluation. Evaluation for text summarization can be divided into two categories, *unit overlapping* and *semantic embedding* (Zong et al., 2021). *Unit overlapping* metrics are based on

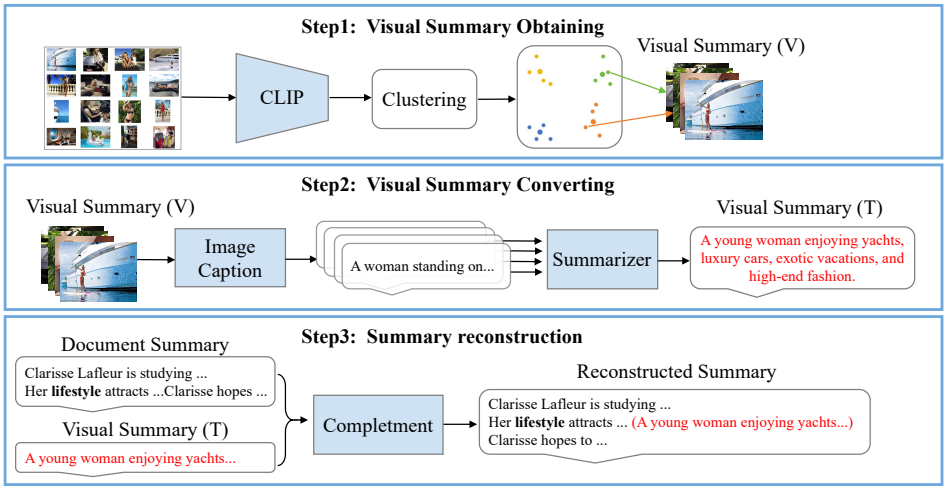


Figure 3: Dataset construction.

n-gram calculations of semantic units overlap, such as ROUGE (Lin, 2004), BLEU (Papineni et al., 2002) and Meteor (Banerjee and Lavie, 2005). *Semantic embedding* metrics recognize the semantic limitations of n-grams and employ continuous text features to measure accuracy, such as MoverScore (Zhao et al., 2019) and BERTscore (Zhang* et al., 2020). Most multimodal summarization studies consider employing text summarization metrics above to evaluate only the quality of text summaries, ignoring the quality of summaries from the other modalities. Additional studies incorporate visual modalities into multimodal summary evaluation. For instance, Zhu et al. (2018) propose image precision (IP) to measure the accuracy of image selection. Vijayan et al. (2024) introduce perplexity to measure the impact of input visual information. Takahashi et al. (2024) introduce an abstractness score to evaluate the performance of video summarization models. Wan and Bansal (2022) combines both CLIPScore (Hessel et al., 2021) and BERTScore (Zhang* et al., 2020) for scoring. Although these studies consider the role of visual modality in summarization, they can only evaluate the quality of summaries when the input or output is unimodal.

Inspired by the above studies, our objectives are two-fold: 1) to define a task that summarizes both textual and visual content; 2) to create a modality-unstricted evaluation framework for multimodal summarization.

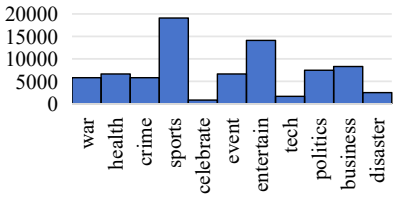


Figure 4: The distribution across different scenarios.

3 Dataset Construction

3.1 Overview

Figure 3 depicts the details of dataset construction. We construct NIMMS dataset by extending existing multimodal summarization dataset. Specifically, we select suitable samples (§3.2) from the public dataset and reconstruct their summaries (§3.3). The overall workflow of summary reconstruction includes three steps: 1) obtaining the visual summary in visual form, 2) converting the visual summary into text form, and 3) complementing the original document summary[2] with the visual summary.

3.2 Sample Selected

We employ the MSMO dataset (Zhu et al., 2018) as our data source and select samples based on two criteria. First, summarizing visual input is necessary if there is a magnitude of image input, thus we select samples with more than $N(N = 8)$ images. Second, referring to the news categories on the CNN website [3], we divide the news category of NIMMS into 11 scenarios: *war, health, crime, sports, celebration, event, entertainment,technology, politics,*

[2]In this paper, "document" refers to the input text, "document summary" refers to the summary of the input text.

[3]https://edition.cnn.com/

Tasks	Datasets	Input					Output	
		Visual	Textual	Asyn	NI	V-Words	V-Implicit	V-Explicit
video	TVSum (Song et al., 2015)	✓	✗	✗	-	-	✓	✓
	SumMe (Gygli et al., 2014)	✓	✗	✗	-	-	✓	✗
	VSUMM (de Avila et al., 2011)	✓	✗	✗	-	-	✓	✗
textual	X-Sum (Narayan et al., 2018)	✗	✓	✗	-	-	✗	✗
	Pubmed (Sen et al., 2008)	✗	✓	✗	-	-	✗	✗
multi-modal input	AVIATE (Atri et al., 2021)	✓	✓	✗	-	-	✗	✓
	VMSMO (Li et al., 2020b)	✓	✓	✗	-	-	✗	✓
	MM-AVS (Fu et al., 2021)	✓	✓	✗	-	-	✗	✓
	mRedditSum (Overbay et al., 2023)	✓	✓	✓	1.00	3.10	✗	✓
	MMSS (Li et al., 2018)	✓	✓	✓	1.00	2.80	✓	✗
	EC-product (Li et al., 2020a)	✓	✓	✓	1.00	2.56	✓	✗
	M3LS (Verma et al., 2023)	✓	✓	✓	1.93	-	✓	✗
	MCLS (Xiaorui, 2023)	✓	✓	✓	3.15	-	✓	✗
	MSMO (Zhu et al., 2018)	✓	✓	✓	6.57	14.48	✗	✓
	E-DailyMail (Chen and Zhuge, 2018)	✓	✓	✓	5.42	-	✗	✓
	NIMMS (ours)	✓	✓	✓	**15.52**	**46.40**	✗	✓

Table 1: Comparison of different summarzation tasks and datasets. NI and V-words represents the number of images, the number of scenario words contained in input images, respectively.

	train	valid	test
#Documents	78,214	3,060	2,492
#AvgImages	15.52	13.45	14.12
#AvgTokens(Doc)	722.08	766.98	732.85
#AvgTokens(DocSum)	72.02	73.14	77.03
#AvgTokens(AllSum)	106.54	109.68	112.37

Table 2: Corpus statistics. #AvgImages is the average number of images in each sample. #AvgTokens(Doc), #AvgTokens(DocSum) and #AvgTokens(AllSum) denote the average number of tokens in the document, document, and overall summary respectively.

Hyp	Annotator$_A$	Annotator$_B$	Overall
summ-d	3.21/3.34	3.32/3.24	3.27/3.29
summ-d+v*	3.63/3.74	3.72/3.69	3.68/3.71
summ-r	4.07/3.98	4.15/3.87	4.11/3.93

Table 3: User satisfaction test results. 1 stands for the worst, and 5 stands for the best for all three metrics, two blocks represent the summary cover all input content/quality of the summary. “summ-d",“summ-d+v*" and “summ-r" represent document, pictorial and overall summary, respectively.

business, disaster. We randomly select 500 samples from these scenarios for manual annotation, with each sample assigned to two annotators. Samples are retained if both annotators agree that the input images contain information not present in the input text. It results in 88.9% samples retention and a kappa coefficient of 0.75 for the annotation. Finally, GPT-4 is exploited to automatically classify the scenario for each sample with details provided in the Appendix A.

3.3 Summary Reconstruction

A good reconstructed summary could both summarize the visual input and complement the original document summary. Given that current large models have demonstrated strong performance in image captioning and text summarization tasks, we can leverage a pipeline technology to summarize multiple images, as illustrated in Figure 3.

Step1: Visual Summary Obtaining. To remove outliers and redundant information from the visual input, the images are clustered into k groups, where $k = \frac{1}{2}n$, n is the number of input images. Specifically, we first adopt a pre-trained model CLIP (Radford et al., 2021) to obtain the feature representation f_i of each image v_i. Then, we perform clustering on the image representation f_i exploiting KMeans. Finally, the images that are closest to their respective cluster centers are selected as the visual summary results $(v_1, v_2, ..., v_k)$.

By obtaining the visual summary in visual form, those redundant and outlier images are eliminated.

Step2: Visual Summary Converting. We exploit the LLava (Liu et al., 2023a) and GPT-3.5 [4] to obtain the visual summary in textual form. First, LLava generates the caption for each image v_j in the visual summary.

$$g_j = \text{LLava}(v_j) \quad (1)$$

[4]We use the version of gpt-3.5-turbo-1106.

Second, GPT-3.5 condenses all image captions to a visual summary in textual form. Each image caption consists of two parts: a generated caption g_j and a self-provided caption c_j. Compared to the generated captions, the self-provided captions are more closely related to the input document but are less accurate in reflecting the image content. Thus, the visual summary in textual form s^v is:

$$s_q^v = \mathrm{GPT}(\bigcup_{j=1}^{k}(g_j, c_j))) \quad (2)$$

where Equation 2 is executed Q times to obtain multiple summary candidates $\{s_1^v, ..., s_q^v, ..., s_Q^v\}$. By combining the generated captions with the self-provided ones, we ensure that the visual summary accurately represents the visual input while maintaining a close connection with the document. We provide details for Step2 in the appendix E.

Step3: Summary Complementing. The visual summary s_q^v complements the unit of the document summary s^d. We divide the document summary into P units and each unit s_p^d is a sentence. To organically integrate the document and visual summaries, we consider: 1) the redundancy between the visual and document summaries $\mathrm{Red}(s^d, s_q^v)$, where $\mathrm{Red}(\cdot)$ is ROUGE-1 score (Lin, 2004), and 2) the relevance of document summary units to the visual summary $\mathrm{Rel}(s_p^d, s_q^v)$, where $\mathrm{Rel}(\cdot)$ is BERTScore (Zhang* et al., 2020). The formula is:

$$f(s_p^d, s_q^v) = -\beta \sum_{p=1}^{P} \frac{1}{P} \mathrm{Red}(s_p^d, s_q^v) + \mathrm{Rel}(s_p^d, s_q^v) \quad (3)$$

where β is set to 10. This ensures that the visual summary s_q^v is prioritized first, followed by the corresponding document summary unit s_q^v. Subsequently, the complementing form is:

$$p^*, q^* = \arg\max_{p,q} f(s_p^d, s_q^v) \quad (4)$$

Through p^*, q^*, visual summary $s_{q^*}^v$ can be used to complement document summary unit $s_{p^*}^d$. Finally, the overall summary s^r is $(s_1^d, s_2^d, ..., s_{p*}^d, s_{q*}^v, s_{p*+1}^d ...)$. In order to provide a higher quality test set, we invite three annotators to manually remove samples that have the following errors: (1) the image summary is inaccurate or weakly related to the text content; (2) the image summary that is in incorrect position. This process filters out 9.71% of the samples.

3.4 User Satisfaction Test

We conduct an experiment to investigate whether the overall summary can improve user satisfaction. 150 samples are randomly selected from the training and test set (1:1). We invite two annotators to score: (1) whether the summary covers all the input content; (2) the comprehensive quality of the summary includes informativeness, fluency and non-redundancy. All annotators provide a score from 1 to 5. Table 3 shows the results of the user satisfaction test. Ratings of the overall summaries are 11.7%/5.9% higher than pictorial summaries and 25.7%/19.5% higher than document summaries. It shows that users prefer this way of summarizing, thus confirming our motivation for NIMMS.

3.5 Comparison with Existing Datasets

Figure 4 and Table 2 present a comprehensive analysis of the NIMMS datasets statistics. And we give two examples of NIMMS in the appendix C. In addition, as shown in Table 1, the key distinguishing features of NIMMS are summarized as follows: (1) Each sample contains a large number of image inputs, and they cannot be adequately summarized by a single image. (2) To quantify the richness of the image content, a multimodal model (GPT-4) is adopted to generate scenario words for all input images. The total number of scenario words reflects the richness of the image. NIMMS has more scenario words for images compared to other datasets. (3) The output includes explicit summaries of the visual input, and both visual and document summaries are unified into the text modality.

4 Task and Experiment Setup

4.1 Task Definition

Given a document d with n images depicting various scenes as input I, *i.e.*, $I = \{d, (v_1, v_2, ..., v_n)\}$, NIMMS task is to generate an overall summary s^r that includes both the document summary s^d and visual summary s^v as output O.

4.2 Comparative Methods

We select text-only model, small model (SM) that can be modified for multi-image input, and multimodal large language model (MLLM) that accepts multi-image input for comparative experiments:
PGN (See et al., 2017): It generates the current summary word by copying words from the source text or producing new words from the generator.
MAtt (Li et al., 2018): It adopts modality attention

and image filtering for multimodal summarization.
MSMO (Zhu et al., 2018): It proposes multimodal input and multimodal output method.
BertAbs (Liu and Lapata, 2019): It exploits a text encoder BERT, an image encoder VGG and a transformer decoder.
LLava (Liu et al., 2023b): an MLLM based on LLama (Touvron et al., 2023).
idefics2 (Laurençon et al., 2024): an MLLM that accepts any text and image sequence as input.
InternLM2 (Cai et al., 2024): an MLLM based on InternLM.

Notably, the small models are specifically trained for our task, while the large models are used off-the-shelf without any additional training.

4.3 M-info: Evaluation

We employ five widely used automatic metrics to evaluate the above methods, including ROUGE-1, ROUGE-2, ROUGE-L (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang* et al., 2020). More details are given in the appendix B. Considering these metrics only evaluate the summary quality from a unimodal perspective, we expect an evaluation method for with the following two capabilities: 1) measuring the impact of visual input $v_1, v_2, ..., v_n$; 2) applying to both textual and multimodal outputs. Specifically, the metric should evaluate the textual summary $O = (s^d, s^v)$ and the multimodal summary $O = (s^d, v_k)$, where v_k is the representative image.

Inspired by Peyrard (2019) that summarization aims to minimize differences between input and output distribution, we focus on two questions:

How to obtain the multimodal distribution ? A unified distribution format is necessary for representing the input I and output O. Colombo et al. (2021) suggest using the pre-trained language model and word masking to obtain the aggregate distribution, which can then be compared for textual similarity. Inspired by this study, we use a pre-trained multimodal model Vilt (Kim et al., 2021) to obtain the multimodal distribution of input I or output O as shown in Figure 5. A sequence of text tokens and image units x with a mask at position j is denoted as $[x]^j$, the pre-trained multimodal model predicts a distribution $p_\Omega(\cdot \mid [x]^k)$ over the vocabulary Ω given the masked context. For example, for a masked sequence $[x]^1$ =[mask] near sea $v_{1,1}, v_{1,2}, v_{1,3}$, the model might assign high probabilities to the token "woman", because the unit $v_{1,1}$ is an image region showing a woman. Therefore, the above process is formulated as:

$$p_\Omega(\cdot \mid x) \cong \sum_{k=1}^{M} p_\Omega(\cdot \mid [x]^k) + \sum_{k=M+1}^{M+N} (\overline{p}_\Omega(\cdot \mid [x]^k)) \quad (5)$$

where M, N indicates the length of text tokens and the number of image units, respectively. It is worth noting that mapping the distribution of image units to the vocabulary remains a gap between modalities. This gap results in the distributions of image units being very similar and lacking good distinguishability. To obtain a higher quality distribution for image units, we subtract an average distribution score p_{avg} for each image unit distribution. Here, p_{avg} is the average of the distribution for all image units in the dataset $\mathcal{D}$. As shown in the right of Figure 5, the image unit distribution $\overline{p}_\Omega(\cdot \mid [x]^k)$ is computed as:

$$p_{avg} = \frac{1}{|\mathcal{D}|}\frac{1}{N}\sum_{\mathcal{D}}\sum_{k=M+1}^{M+N} p_\Omega(\cdot \mid [x]^k) \quad (6)$$

$$\overline{p}_\Omega(\cdot \mid [x]^k) = p_\Omega(\cdot \mid [x]^k) \div p_{avg} \quad (7)$$

Finally, $p_\Omega(\cdot|O)$ and $p_\Omega(\cdot|I)$ can be derived from formula 5.

How to measure the distribution differences? Considering redundancy and relevance, the quality of a summary can be formalized by calculating the KL divergence between the input and output texts (Peyrard, 2019). Similarly, for multimodal summarization, the target is to minimize the KL divergence between the input and output $\mathrm{KL}(O||I)$.

Therefore, the multimodal summarization evaluation metric M-info is computed as:

$$\text{M-info}(I, O) = \mathrm{KL}(p_\Omega(\cdot|O)||p_\Omega(\cdot|I)) \quad (8)$$

5 Experiment

5.1 Automatic Evaluation Results

Table 4 presents the automatic evaluation results of different models. In general, overall summaries tend to perform better than document summaries, which aligns with the objective of the NIMMS task to summarize both document and visual input. **BertAbs** performs the best on NIMMS. Specifically, **MSMO** performs worse compared to **BertAbs**. This is because **MSMO** adopts visual coverage to select matching image during summarization, which suppresses visual summary generation. **PGN** is a pure text summarization model

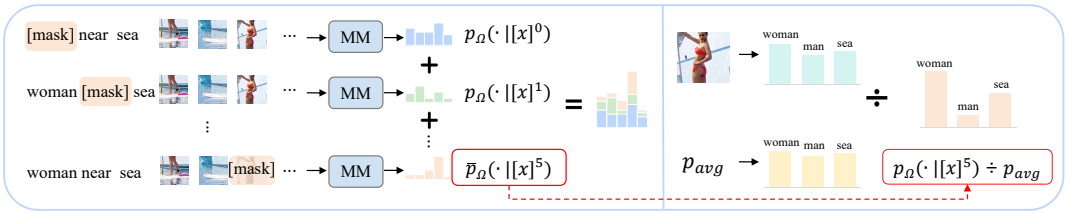


Figure 5: To obtain the distribution in M-info evaluation. "MM" is a pre-trained multimodal model.

	Method	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑	BLEU ↑	BERTScore ↑	M-info ↓
Text	PGN	29.60/30.25	10.26/10.31	27.91/28.11	35.67/36.09	83.89/83.98	- /1.86
SM	Matt	31.60/37.25	11.86/12.31	28.91/32.61	37.67/42.89	84.44/84.63	- /1.40
	MSMO	34.06/39.28	13.50/13.59	31.19/34.34	39.01/44.65	85.17/85.19	2.32/1.44
	BertAbs	**35.62/40.57**	**14.25/14.21**	**32.77/35.59**	**40.27/45.41**	**85.66/85.59**	- **/1.29**
MLLM	LLava	28.76/26.70	10.11/ 8.49	23.53/21.39	15.27/22.42	81.09/81.54	2.11/1.40
	idefics2	34.12/34.40	11.13/10.83	27.73/27.24	39.00/31.08	84.69/84.92	3.41/3.28
	InternLM2	26.05/26.54	6.25/ 5.89	21.08/20.50	36.34/29.94	81.25/81.26	**1.98**/1.32

Table 4: Automatic metric results for different summaries. For ROUGE, BLEU, and BERTScore, two blocks represent the document summary/overall summary. For M-info, two blocks represent the pictorial/overall summary. The ↑ indicates that a higher value of the indicator is better, while ↓ indicates the opposite.

Hyp	Informativeness	Fluency	Non-Redundancy
summ-d	3.58	3.60	3.50
summ-v	3.20	3.75	3.77
summ-r	3.46	3.52	3.63

Table 5: Human evaluations on test dataset. 1 stands for the worst, and 5 stands for the best for three metrics. summ-d,summ-v and summ-r represents document summary, visual and overall summary, respectively.

and cannot handle the visual summary. **LLava**, **idfics2**, and **InternLM2** underperform compared to **BertAbs** due to a lack of pre-training tasks related to summarizing multiple images. Among MLLMs, **idefics2** excels in text metrics but scores the lowest on the M-info metric, as it tends to ignore images in the NIMMS task, focusing solely on text. This is further supported by **idefics2**'s weaker performance on overall summaries compared to document summaries. Regarding the M-info metrics, all methods perform worse on pictorial summaries than on overall summaries. Because M-info emphasizes comprehensive visual information, which a single image cannot fully capture and may even introduce distractions. Additionally, compared to **MSMO**, **LLava**, and **InternLM2** generate more redundant summaries, resulting in lower text metric scores, though their superior image modeling boosts M-info values.

5.2 Human Evaluation

We randomly select 100 samples from the generation of test set and invite two postgraduates to score document, visual and overall summaries from 1 to 5. 1 stands for the worst, and 5 stands for the best. The evaluation metrics include informativeness, fluency, and non-redundancy. (1) Informativeness: Does the system summary contain comprehensive reference content? (2) Fluency: Is the system summary grammatically correct and readable? (3) Non-Redundancy: Does the system summary not have redundant or incorrect information relative to the input? Table 5 shows the human evaluation results. Visual summary has the lowest informativeness, indicating that models are weak in summarizing multiple images. Besides, the overall summary is less fluent than the other two, suggesting difficulty in combining document and visual summaries. Finally, all three summaries achieve comparable redundancy scores, showing visual summaries can effectively complement document ones.

In conclusion, NIMMS faces two main challenges: 1) Summarizing the salience of numerous images. Existing multimodal models are pre-trained with single-image and textual tasks while our task requires summarizing a sequence of images. 2) Integrating the visual and document summaries. The model must identify how visual summary can complement document summary to create a more fluent and reader-friendly overall summary.

5.3 Correlation Test

To demonstrate the effectiveness of M-info, we conduct an experiment on correlations between these metrics and human judgment scores. Human anno-

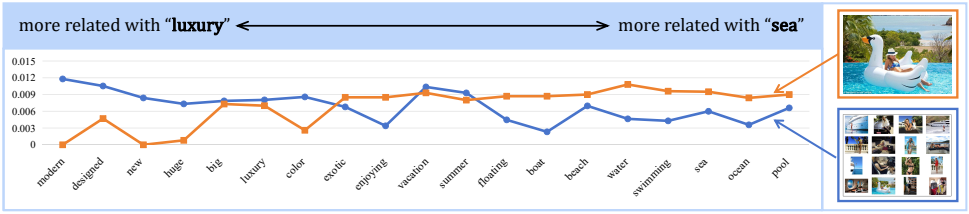


Figure 6: Visualize of M-info metric.

Hyp	Metrics	γ	ρ	τ
summ-d	ROUGE-1	.4458	.4665	.3589
	ROUGE-2	.1332	.2079	.1717
	ROUGE-L	.4280	.4453	.3434
	BLEU	**.6013**	**.6818**	**.5191**
	BertScore	.2061	.1952	.1557
summ-d+v*	MMAE	.3446	.3344	.3165
	M-info	**.3988**	**.4931**	**.3740**
summ-r	ROUGE-1	.3361	.2225	.2001
	ROUGE-2	.1084	.1000	.1059
	ROUGE-L	.2485	.1206	.1239
	BLEU	.2095	.1853	.1647
	BertScore	.0759	-.0178	-.0082
	CLIPBert	.3674	.4860	.3531
	M-info	**.6236**	**.6334**	**.4708**

Table 6: Correlation with human evaluation, measured with Pearson γ, Spearman ρ, and Kendall τ coefficients. "summ-d","summ-d+v*" and "summ-r" represents document, pictorial and overall summary, respectively.

tators score document, pictorial and overall summaries from 1 to 5, where 1 is the worst and 5 is the best, based on how well they capture the essence of multimodal input. We randomly select 100 samples from the outputs of each model on the test set, with each sample being scored by two annotators. We take the average score as the final score. The correlation is assessed using three metrics: 1) pearson correlation coefficient (γ), 2) spearman rank coefficient (ρ), and 3) kendall rank coefficient (τ). The correlation results are presented in Table 6.

We assess the human correlation of from three perspectives: document, pictorial, and overall summaries. In pictorial and the overall summaries, **M-info** correlates best with human assessments across all three correlation coefficients. This confirms that when numerous images are provided, humans tend to focus more on the visual content. For pictorial summaries, although **MMAE** evaluation shows good human correlation, human evaluators find a single image is insufficient to capture the essence of all images. Conversely, **M-info** integrates the salience of all visual content, achieving higher relevance. For overall summaries, **M-info** outperforms other metrics. Because when a summary effectively

	Hyp Summ	M-info
#1	summ-r	1.41
#2	summ-d	1.70
#3	summ-d + sent	1.55
#4	summ-d + v*	2.13

Table 7: The ablation studies for different summary measured by M-info.

captures the key content of the input images, humans tend to give higher scores. However, other textual metrics focus on textual overlap, failing to reflect the visual accuracy. **CLIPBert** independently evaluates different modalities which hurts the evaluation of the overall summary.

5.4 Analysis

Effectiveness of Overall Summary. Table 7 compares various grounded summaries using M-info on the test set: 1) document summary; 2) document summary with a random sentence from the document; 3) pictorial summary. The analysis leads to the following conclusions: 1) summary #1 outperforms #2, showing that the overall summaries provide more comprehensive information than the document summary alone; 2) Comparison between Summaries #1 and #3 indicates that the overall summaries include additional information beyond what the document provides; 3) Comparison between Summaries #2 and #4 suggests that a single image is insufficient to convey all the information from multiple images and may negatively impact the summary, resulting in a lower M-info score than the document summary.

Visualize of M-info. Since M-info measures the information distribution difference by aligning multimodal to text-modal, we expect analyzing the distribution $\overline{p}_\Omega(\cdot|[x]^k)$ of all image units x^k. Figure 6 visualizes $\overline{p}_\Omega(\cdot|[x]^k)$ for different visual inputs, displaying the top 10 words for each. The orange represents the distribution of a single image, while the blue represents the distribution of multiple images. These top 10 words accurately describe the

visual content. Moreover, we manually sort these words, with those on the right being more related to "sea" and those on the left related to "luxury". The distributions differ significantly: multiple images input aligns more with "luxury," while the single image aligns more with "sea". This phenomenon illustrates the interpretability of M-info and explains why a single image cannot capture the full content of multiple images.

6 Conclusion

In this paper, we introduce the NIMMS task, aiming at generating summaries by integrating both textual and numerous image inputs. Besides, we propose M-info, a new evaluation method proven effective for NIMMS. Through comprehensive analysis and comparison of various summarization methods applied to NIMMS, we identify significant challenges, particularly the limitations of current multimodal models in deeply understanding long-sequence asynchronous images and generating cohesive summaries that unify both textual and visual information. We hope NIMMS will serve as a valuable benchmark to enhance multimodal large language models.

Limitations

Since many existing multimodal summarization datasets do not meet our needs, we only select MSMO (Zhu et al., 2018) as the data source. Therefore, NIMMS is currently focused only on the news domain. In the future, we need to expand to the commerce or accommodation domains.

Acknowledgements

The research work has been supported by the Natural Science Foundation of China under Grant No. 62106263.

References

Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. 2021. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *Preprint*, arXiv:2105.09601.

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

Zheng Cai, Maosong Cao, Haojiong Chen, et al. 2024. Internlm2 technical report. *Preprint*, arXiv:2403.17297.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4046–4056, Brussels, Belgium. Association for Computational Linguistics.

Pierre Colombo, Chloe Clave, and Pablo Piantanida. 2021. Infolm: A new metric to evaluate summarization & data2text generation. In *AAAI Conference on Artificial Intelligence*.

Sandra Eliza Fontes de Avila, Ana Paula Brandão Lopes, Antonio da Luz, and Arnaldo de Albuquerque Araújo. 2011. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recogn. Lett.*, 32(1):56–68.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. MM-AVS: A full-scale dataset for multi-modal summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926, Online. Association for Computational Linguistics.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *European Conference on Computer Vision*.

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

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammed Hasanuzzaman. 2021. Multi-modal supplementary-complementary summarization using multi-objective optimization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 818–828, New York, NY, USA. Association for Computing Machinery.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *Preprint*, arXiv:2102.03334.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8188–8195.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020b. 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.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2023. Summary-oriented vision modeling for multimodal abstractive summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2934–2951, Toronto, Canada. Association for Computational Linguistics.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Nayu Liu, Xian Sun, Hongfeng Yu, Fanglong Yao, Guangluan Xu, and Kun Fu. 2023c. Abstractive summarization for video: A revisit in multistage fusion network with forget gate. *IEEE Transactions on Multimedia*, 25:3296–3310.

Nayu Liu, Kaiwen Wei, Yong Yang, Jianhua Tao, Xian Sun, Fanglong Yao, Hongfeng Yu, Li Jin, Zhao Lv, and Cunhang Fan. 2024. Multimodal cross-lingual summarization for videos: A revisit in knowledge distillation induced triple-stage training method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10697–10714.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. 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 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. 2017. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211.

Sourajit Mukherjee, Anubhav Jangra, Sriparna Saha, and Adam Jatowt. 2022. Topic-aware multimodal summarization. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 387–398, Online only. Association for Computational Linguistics.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Keighley Overbay, Jaewoo Ahn, Fatemeh Pesaran zadeh, Joonsuk Park, and Gunhee Kim. 2023. mRedditSum: A multimodal abstractive summarization dataset of Reddit threads with images. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4117–4132, Singapore. Association for Computational Linguistics.

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

Maxime Peyrard. 2019. A simple theoretical model of importance for summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1059–1073, Florence, Italy. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. *Preprint*, arXiv:1811.00347.

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

Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. 2008. Collective classification in network data. *AI Mag.*, 29(3):93–106.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5179–5187.

Rikito Takahashi, Hirokazu Kiyomaru, Chenhui Chu, and Sadao Kurohashi. 2024. Abstractive multi-video captioning: Benchmark dataset construction and extensive evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 57–69, Torino, Italia. ELRA and ICCL.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Yash Verma, Anubhav Jangra, Raghvendra Verma, and Sriparna Saha. 2023. Large scale multi-lingual multimodal summarization dataset. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3620–3632, Dubrovnik, Croatia. Association for Computational Linguistics.

Vipin Vijayan, Braeden Bowen, Scott Grigsby, Timothy Anderson, and Jeremy Gwinnup. 2024. The case for evaluating multimodal translation models on text datasets. *Preprint*, arXiv:2403.03014.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Min Xiao, Junnan Zhu, Haitao Lin, Yu Zhou, and Chengqing Zong. 2023. CFSum coarse-to-fine contribution network for multimodal summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8538–8553, Toronto, Canada. Association for Computational Linguistics.

Shi Xiaorui. 2023. MCLS: A large-scale multimodal cross-lingual summarization dataset. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 862–874, Harbin, China. Chinese Information Processing Society of China.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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*.

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.

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, Brussels, Belgium. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9749–9756.

Chengqing Zong, Rui Xia, and Jiajun Zhang. 2021. *Text data mining*, volume 711. Springer.

A Sample Scenario Selecting

We exmploit GPT-3.5 to obtain the scenario of the example. Only the example fitting the following scenario are selected: war, health, crime, sports, celebration, event, entertainment,technology, politics, business and disaster. As illustrated in Figure 7, GPT-3.5 API helps us identify the scenario for each sample.

B Evaluation details

We also provide evaluation scripts for reproduction. For ROUGE score, we use file2rouge[5] with default settings. For BERTScore[6], we use the official API. It exploits the pre-trained contextual embeddings from BERT to calculate the similarity between the hypothesis sentences and the reference sentences.

C Samples of NIMMS

We give two samples of NIMMS as shown in Figure 8 and 9. In the summary, the text in black

[5]https://github.com/pltrdy/files2rouge

[6]https://pypi.org/project/bert-score/0.2.1

Model GPT-3.5

REVEALED : Luxury lifestyle of 18-year-old ` Rich Kid of Instagram ' who drives a Maserati , splurges on designer clothes and holidays on a super yacht - all funded by her ` generous father ' Clarisse Lafleur is ` ashamed to tell ' the amount she spends on designer clothes and accessories . The 18-year-old flies first class or by private jet , holidays aboard a charter super yacht and credits her favourite designers as Christian Louboutin , Karl Lagerfeld and Valentino Garavani . But while her lifestyle is funded by her ` very generous father ' , Miss Lafleur told Daily Mail Australia she ` can not wait ' to start work so she can make her own money . Scroll down for video The New Caledonian teenager , who is currently studying business on the Gold Coast , is one of the faces regularly featured on the Rich Kids of Instagram page . The page is an insight in to a lifestyle most people can only dream of : private jets , wads of cash and flashy sports cars . ' I completely understand that people would generalise and perceive our lifestyle very negatively when they come across content such as teenagers filling hot tubs with $ 600 Champagne , ' Miss Lafleur said .
...

Please choose a suitable topic for the above text. The candidate topics are: (a) war (b) health (c) crime (d) sports (e) celebration (f) event (g) entertainment (h) technology (i) politics (j) business (k) disaster

The suitable topic for the text provided would be **(g) entertainment.** It focuses on the luxurious lifestyle and activities of a young person, which is often associated with entertainment and lifestyle reporting.

Figure 7: GPT-3.5 identify the scenario for each sample.

denotes the original summary, while the text in red denotes the additional summary of the input images that our task aims to include.

D Efficiency of M-info

M-info calculates each masked word or image patch independently, allowing for parallel computation. The parallel size P depends on the GPU resources. On one 3090 Ti, we set the parallel parameter P=50, which means M-info computes the distribution of 50 masked tokens simultaneously. Correspondingly, we provide the M-info time cost on 200 samples in the table below:

run time	5.42s
number of input words	712.08
number of input input image patches	350.93
number of summary words	103.45

Table 8: The efficiency of M-info.

E Visual Summary Converting

The details of visual summary converting (in step2) is shown in Figure 10.

F NIMMS Dataset Distribution

The details of NIMMS dataset distribution is shown in Figure 11.

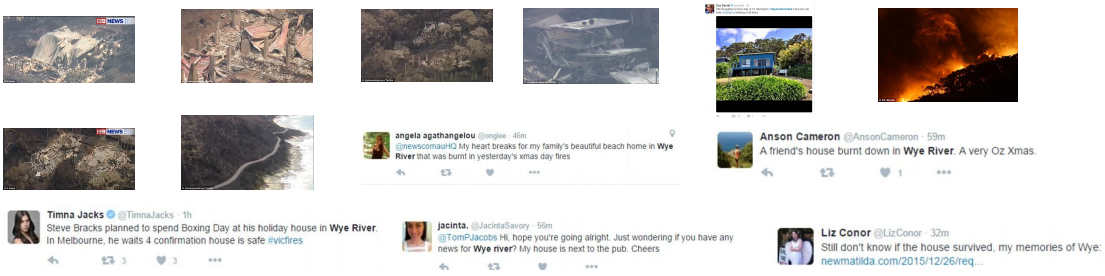
Input Images:

Document: Manchester City 's rivals may find ruling Europe easier than catching Pep Guardiola 's side in the Premier League. So , worst case scenario , Manchester City ca n't be caught . The lead , five points as it stands now after just 10 games , continues to grow . Over 38 matches , they exert their superiority . Try as they might , Manchester United and Tottenham fall away . What then ?
Well , it may not be as bad as you think . Watching Tottenham run amok against Real Madrid at Wembley , seeing Manchester United cruise through qualifying , all is not lost .
Even if City do take the Premier League , that does not mean they can not be beaten in Europe . These are set-piece events in a cup competition . The best team does n't always win .

Overall Summary: Cup competitions have historically served Jose Mourinho 's pragmatism well. Manchester United and Tottenham 's paths in Champions League could open up. Spurs may not have to overcome feelings of inferiority in random format. (A series of diverse sports images capture moments of celebration, disappointment, and anticipation among cup competition players. The referees and coaches on the sidelines are highly engaged, and there is a large audience present.) With City dominating domestically , knockout format in cup could suit rivals

Figure 8: The first example of NIMMS.

Input Images:



Document: Counting the losses of the Black Christmas bushfires : At least 116 homes destroyed , more than 2000 hectares burned and almost $ 40m in damage after Victoria infernos ... as shocking photos show the carnage. Images of the destruction wreaked upon the popular holiday towns caught in massive bushfires on Christmas day have emerged . Photographs of collapsed , ruined homes , blackened earth and scorched trees were the norm for Wye River and Separation Creek , on Victoria 's surf coast , where the fire ripped through houses and forest alike , burning right to the seaside in places on what has been dubbed Black Christmas . Victoria 's Country Fire Authority said late on Saturday increased the number of homes confirmed to have been destroyed to 116 . The massive fire burned more than 2,000 hectares around the coastal towns of Wye River and Separation Creek on Victoria 's surf coast . While the threat from the fire had eased , it was still ....

Overall Summary: 116 homes confirmed destroyed by the blaze near the Great Ocean Road. Almost 2,200 hectares of land burned in the fire , expected to continued. Insurance industry estimated so far , damages to cost almost $ 40million. No-one reported injured so far but the danger is not over yet , warnings say. People have been trying to discover the fate of their properties in the area. (The images depict the devastation caused by a large fire, including destruction of buildings and landscapes, efforts of emergency crews and firefighters to battle the blaze, and individuals seeking help and information on social media about their properties and homes.)

Figure 9: The second example of NIMMS.

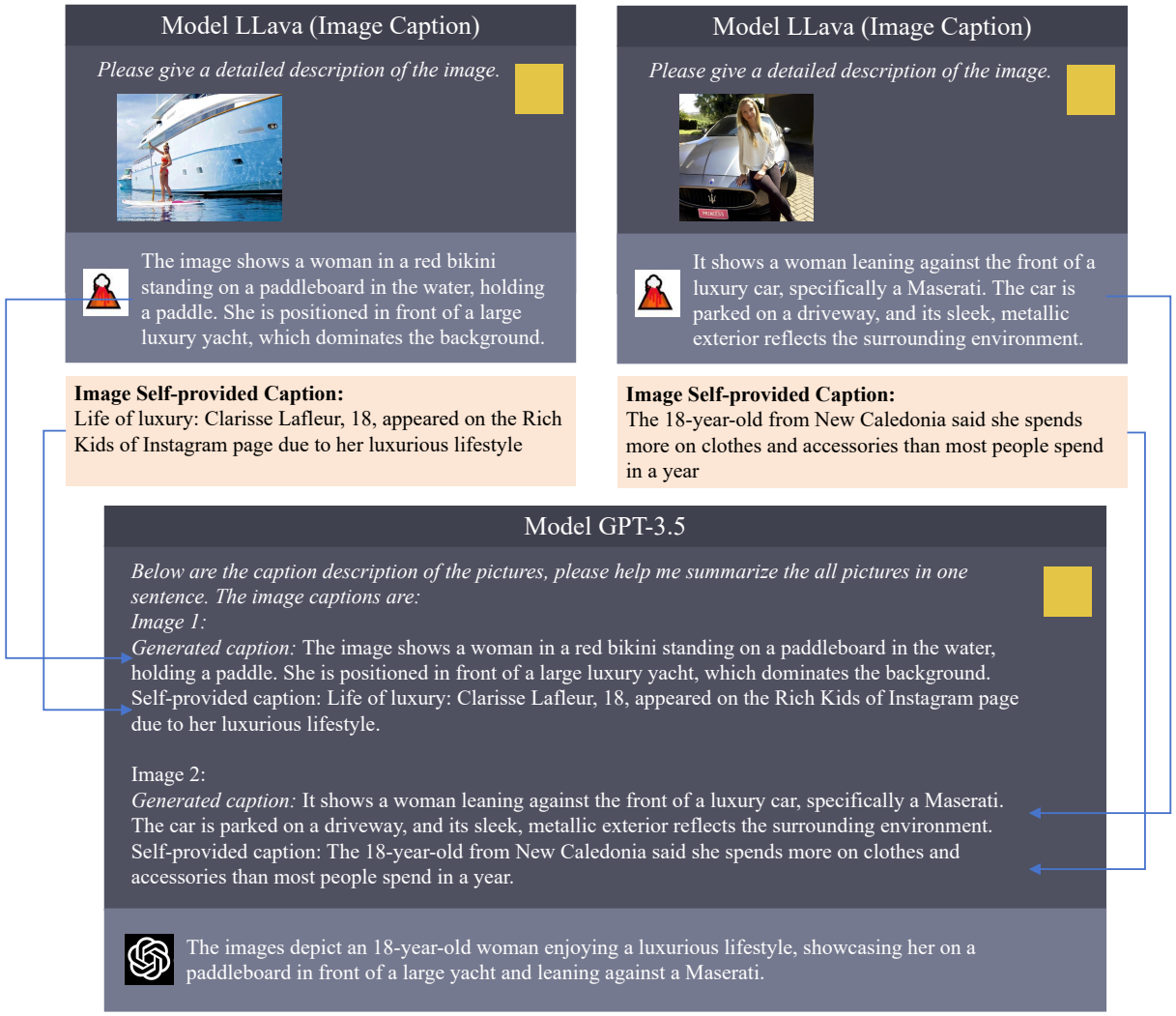
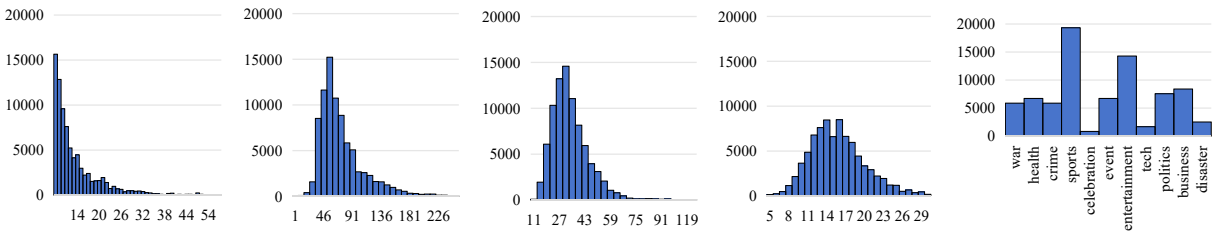


Figure 10: Step2: Visual summary converting.



(a) num of images (b) document summary words (c) visual summary words (d) visual summary abstrictively (e) scenario

Figure 11: Statistics of the NIMMS dataset distribution. The y-axis denotes the number of samples, and the x-axis denotes the corresponding attributes. (a-c) show: (a) the number of images per sample, (b) the number of words in the document summary, (c) the number of words in the visual summary, (d) the number of words in the visual summary that are not in the document or document summary, and (e) the distribution across different scenarios.