Exploiting Pseudo Image Captions for Multimodal Summarization

Chaoya Jiang[1][*], **Rui Xie**[1][*], **Wei Ye**[1], **Jinan Sun**[1,2][†], **Shikun Zhang**[1][†]

[1]National Engineering Research Center for Software Engineering, Peking University
[2]BIGO Technology
{sjn,zhangsk}@pku.edu.cn

Abstract

Multimodal summarization with multimodal output (MSMO) faces a challenging semantic gap between visual and textual modalities due to the lack of reference images for training. Our pilot investigation indicates that image captions, which naturally connect texts and images, can significantly benefit MSMO. However, exposure of image captions during training is inconsistent with MSMO's task settings, where prior cross-modal alignment information is excluded to guarantee the generalization of cross-modal semantic modeling. To this end, we propose a novel coarse-to-fine image-text alignment mechanism to identify the most relevant sentence of each image in a document, resembling the role of image captions in capturing visual knowledge and bridging the cross-modal semantic gap. Equipped with this alignment mechanism, our method easily yet impressively sets up state-of-the-art performances on all intermodality and intramodality metrics (e.g., more than 10% relative improvement on image recommendation precision). Further experiments reveal the correlation between image captions and text summaries, and prove that the pseudo image captions we generated are even better than the original ones in terms of promoting multimodal summarization.

1 Introduction

With the increase of multimedia data on the Web, multimodal summarization has drawn widespread attention from researchers in the communities of Web technologies(Messaoud et al., 2021; Jangra et al., 2021a), natural language processing (NLP) (UzZaman et al., 2011; Li et al., 2017, 2020b) and computer vision (CV) (Chen and Zhuge, 2018; Palaskar et al., 2019; Li et al., 2020a; Liu et al., 2020). More recently, many efforts (Zhu et al., 2018, 2020; Zhang et al., 2021b) have been ded-
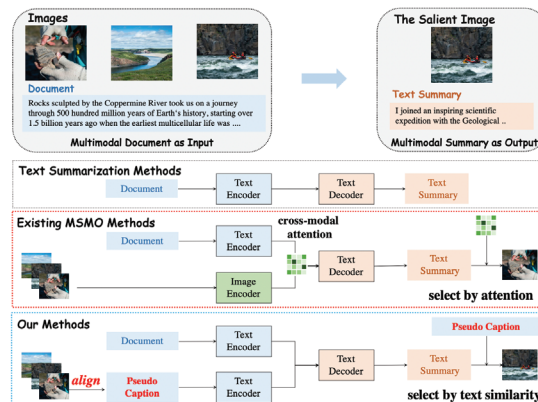

Figure 1: Overview of text summarization and MSMO. Compared with text summarization models, existing MSMO methods usually use an extra image encoder to project images into intermediate representations. They identify the salient image by cross-modal attention, which could be inaccurate due to the lack of golden images for training. We explicitly transform an image into a concrete caption by image-text alignment, capturing visual knowledge better and making text summarization and image selection more effective yet simpler.

icated to multimodal summarization with multimodal output (MSMO), the novel task of generating pictorial summaries given a Web document consisting of plain text and a collection of images. As shown in Figure 1, a pictorial summary generated by MSMO models consists of a text summary and a salient image, delivering more user-friendly information than single-modal text summaries, according to human judgments (Zhu et al., 2018, 2020).

MSMO faces two main challenges. (1) **There are no recommended image references available for training MSMO models.** Due to the lack of supervision signals from visual modality, it is non-trivial to optimize the cross-modal attention between texts and images, which is highly relied on by existing MSMO methods to pick salient images. According to previous best results(Zhang et al., 2021b), only about 60% of the predicted images are correct, indicating that image selection remains

---

[*]Equal contribution.
[†]Corresponding author.

a bottleneck. (2) **Visual knowledge is commonly underutilized to improve text summaries.** Existing MSMO efforts show no evident improvement or even negative impact on text summaries (e.g., decreased ROUGE scores) over typical single-modal text summarization methods. Previous literature(Zhu et al., 2018) explained that some images were noises and long text had contained enough information for text generation, while we conjecture that these methods may not sufficiently exploit visual knowledge to characterize salient text.

To summarize, previous efforts typically encode images and texts into the same semantic space, struggling with optimizing cross-modal interaction without training signals for image selection, as the red box in Figure 1 shows. In this dilemma, image captions, which naturally connect images and texts, can provide a cross-modal semantic bridge. Indeed, our preliminary experiments show the efficacy of introducing imageIn captions (see Section 4.4). Yet, exposure of image captions during training is inconsistent with MSMO's task settings, since MSMO excludes them to pursue better generalization of cross-modal semantic modeling(Zhu et al., 2018, 2020).On the other hand, however, it inspires us to identify a highly-relevant sentence for an image as its pseudo yet meaningful caption, providing us with a new perspective to improve MSMO. As shown in the blue box in Figure 1, *unlike current works that represent an image as an intermediate state, we transform it into a concrete sentence to better capture visual knowledge under MSMO settings*. This transformation presents an opportunity to incorporate pre-trained visual-language models more smoothly, while making further text summarization and image selection extremely simple.

Aligning a sentence with an image could be straightforward, but identifying sentences benefiting MSMO the most is non-trivial. The reasons are two-fold. (1) **A sentence well aligned with an individual image can not guarantee a suitable one for MSMO**. An intuitive way to select a sentence is to simply retrieve it from the document, with the image as the query of a pre-trained cross-modal retrieval model. Unfortunately, we find this manner yields unsatisfactory MSMO performance (see Section 4.3). (2) **A classical single-pass one-to-one alignment strategy may miss salient sentences for summarization** (see Section 4.2). There can be one-to-many and many-to-one relationships between images and sentences, and images can be similar in a document, so we need to synthesize yet distinguish image semantics from a global perspective to make better MSMO-oriented alignment.

To this end, we design a coarse-to-fine image-text alignment mechanism to produce pseudo image captions for MSMO. Firstly, a reference caption for an image is retrieved with a cross-modal retrieval model from the golden summary, rather than the whole document (Section 2.3), to capture more summary-friendly information. Since no golden summary exists at inference time, these reference captions are used to train a two-pass image-text alignment model (Section 2.4) that yields pseudo captions when making inferences (that's why "reference captions" are so named). Given a document with ten images, for example, we will first synthesize them as a whole to select ten sentences with many-to-many coarse-grained alignment, and then identify ten individual one-to-one fine-grained matchings by bipartite graph matching over the cross-modal attention matrix.

The pseudo image captions that imply visual knowledge are used as extra highlighted features for text summarization (Section 2.5), and the salient image is picked based on the ROUGE score between its pseudo captions and the generated summary (Section 2.6). Extensive experiments on an existing MSMO dataset not only verify the superiority of our method but also reveal the inner connection between image captions and summaries, demonstrating promising research opportunities for our novel perspective of bridging the cross-modal semantic gap by generating pseudo image captions.

2 Method

2.1 Problem Formulation

For MSMO task, the input is a multimodal document $\{T, V\}$ including a text document T with m sequential sentences, where $T = [t_1, t_2, \cdots, t_m]$, and a image collection V with n images, where $V = \{v_1, v_2, \cdots, v_n\}$. The output is a multimodal summary $\{S, \hat{v}\}$ where $S = [s_1, s_2, \cdots, s_l]$ is a text summary containing l generated sentences and $\hat{v}$ is the image selected from V.

2.2 Method Overview

Our method, named **SITA**, refers to a multimodal **S**ummarization model based on a coarse-to-fine **I**mage-**T**ext **A**lignment mechanism. SITA consists of four modules: (1) **Reference Caption**

(a) Overview of Image-Text Alignment
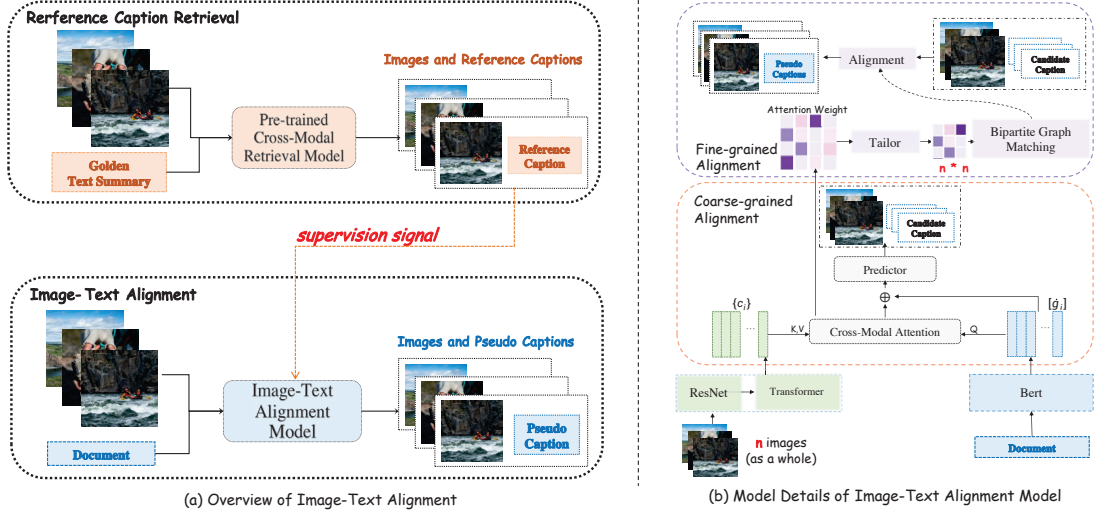
(b) Model Details of Image-Text Alignment Model

Figure 2: Coarse-to-fine Image-Text alignment. The left part (figure a) shows the overview of the whole image-text alignment mechanism. Reference captions are first retrieved from golden summaries based on a cross-modal retrieval model. We then train an image-text alignment model with reference captions as supervision signals, identifying a relevant sentence as a pseudo caption for each image. The right part (figure b) demonstrates how our two-pass coarse-to-fine alignment model works internally.

Retrieval, (2) **Image-Text Alignment**, (3) **Text Summarization**, and (4) **Image Selection**. See more illustrative details in Figure 2 (a).

2.3 Reference Caption Retrieval

Given the multimodal document $\{T, V\}$, we first retrieve reference captions from the golden text summary for each image in V, based on a pretrained cross-modal retrieval model consisting of an image encoder and a text encoder. The image encoder is ResNet152 (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) and the text encoder is a BERT-based sentence encoder for text summarization (Liu and Lapata, 2019). Following (Faghri et al., 2018), we train the model on the COCO dataset (Lin et al., 2014) by matching image representations and sentence representations.

We retrieve reference image captions from the golden summary rather than the whole document, to make the retrieval results more summary-friendly and narrower-focused (see Section 4.3). However, a new dilemma is the lack of golden summaries during inference. Therefore, we exploit them to train an image-text alignment model, which predicts pseudo captions during inference.

2.4 Coarse-to-fine Image-Text Alignment

We design a coarse-to-fine **I**mage-**T**ext **A**lignment model (**ITA**) with training signals obtained from reference captions, to generate pseudo image captions. Since there can be one-to-many and many-

to-one relationships between images and sentences, employing a simple single-pass one-to-one alignment strategy tends to generate a limited set of aligned sentences repeatedly, incapable of recalling enough relevant sentences (see Section 4.2). To this end, we propose a novel two-pass coarse-to-fine mechanism to align sentences better.

Specifically, for the n images in V, we will take them as a whole to select n sentences from the document T with coarse-grained alignment, and then identify one-to-one matchings via fine-grained alignment. ITA consists of an image encoder, a sentence encoder, a coarse-grained alignment module, and a fine-grained alignment module.

2.4.1 Image Encoder

We first use ResNet152 to extract image features for each image in $\{v_1, v_2, \cdots, v_n\}$. These features are then fed into a Transformer-based encoder (Vaswani et al., 2017) as a whole to synthesize global knowledge and interaction information among all images. The position embeddings are not used here since image order information is unavailable. The final output of the image encoder is denoted as $\{c_1, c_2, \cdots, c_n\}$.

2.4.2 Sentence Encoder

The sentence encoder here is the same as the one used in reference caption retrieval. For m sentences denoted as $[t_1, t_2, \cdots, t_m]$, the sentence encoder generate a representation sequence $[g_1, g_2, \cdots, g_m]$.

163

2.4.3 Coarse-grained Alignment

To do coarse-grained alignment, we first apply a cross attention between sentences and images to refine sentence representations:

$$o_{i,j} = \frac{Q_i \cdot K_j}{\sum_{k=1}^{n} Q_i \cdot K_k} \quad (1)$$

$$a_{i,j} = \frac{exp(o_{i,j})}{\sum_{k=1}^{n} exp(o_{i,k})} \quad (2)$$

$$\dot{g}_i = Q_i + \sum_{k=1}^{n} a_{i,k} \times V_k \quad (3)$$

where $Q_i = W_q \times g_i$, $K_j = W_k \times c_j$, $V_j = W_v \times c_j$, $i \in \{1, 2, \cdots, m\}$, and $j \in \{1, 2, \cdots, n\}$. $W_q, W_k, W_v \in R^{D \times D}$ are learnable parameters, where D is 768 (the dimension of the image/text feature vectors). Noted that we have calculated an attention matrix $A \in R^{m \times n}$ based on the equation 1 and 2 where $a_{i,j}$ is the element in the i-th row and j-th column of A.

The refined representation $\dot{g}_i$ is then fed to a sigmoid classifier to predict whether sentence t_i will be selected:

$$p_i = \sigma(W_p \dot{g}_i + b) \quad (4)$$

where $W_p \in R^{D \times D}$ and $b \in R^D$ are learnable parameters.

To train the model, we need n recommended sentences as references for a multimodal document with n images. For each image v_i, we will calculate the ROUGE scores between sentences in the document and their reference captions generated in the first step, and the sentence with the highest score will be labeled as selected. If a sentence is selected more than once, we will pick another sentence with the next highest score. We use $y_i = 1$ to denote that sentence t_i is selected, and $y_i = 0$ otherwise. Then, for the m sentences in the document $T = [t_1, t_2, \cdots, t_m]$, we employ the binary cross-entropy loss to optimize the model as follow:

$$\mathcal{L}_{BCE} = -\frac{1}{m} \sum_{i=1}^{m} y_i log(p_i) + (1-y_i)log(1-p_i) \quad (5)$$

2.4.4 Fine-grained Alignment

Based on the coarse-grained alignment, we have calculated the an $m \times n$ cross-modal attention matrix (denoted as A), in which the element in the i-th row and j-th column is $a_{i,j}$. In this step, we want further to identify optimal one-to-one relationships between images and these sentences. Generally, the larger the attention weight between t_i and v_j, the more likely t_i and v_j match. Suppose we have obtained n selected sentences denoted as $t_{z_1}, t_{z_2}, \ldots, t_{z_n}$ and we extract the rows corresponding to these sentences from the matrix A and concatenate them as a new attention matrix $\dot{A}$:

$$\dot{A} = concat([A_{z_1}, A_{z_2}, \ldots, A_{z_n}]) \quad (6)$$

where $\dot{A} \in N^{n \times n}$, $A_{z_i} \in R^n, i \in \{1, 2, \ldots, n\}$. Based on the new cross-modal attention matrix $\dot{A}$, we can construct a complete weighted bipartite graph G containing two disjoint and independent vertice sets S and V, where $|S| = n$ and $|V| = n$. So there are $n \times n$ weighted edges in G. The vertice v_i in V represents an image, and vertice s_j in S represents a sentence. The weight of the edge in G between the vertice $v_i \in V$ and the the vertice $s_j \in S$ is the value $a_{ij} \in R$ in $\dot{A}$. Therefore, the fine-grained alignment of the sentences and images can be regarded as a maximum-weight perfect matching in the bipartite graphs G. We can easily utilize the bipartite graph matching algorithm (Kuhn-Munkres algorithm (Kuhn, 2010) in our implementation) to match the vertices in the two sets in the graph:

$$M = KM(\dot{A}) \quad (7)$$

where $M = [I_1, I_2, \ldots, I_n], I_i \in \{1, 2, \ldots, n\}$ represents the index list of selected sentences(e.g., the first image is aligned with the I_1-th sentence in the selected sentences), and KM represents the Kuhn-Munkres algorithm.

2.5 Text Summarization

We build the text summarization module based on BERTSum, a recent simple yet robust summarization model (Liu and Lapata, 2019). We concatenate all pseudo image captions as a new text document denoted as T_s. The origin text document T and the new text document T_s are fed into the encoder of BERTSum separately, generating two representation sequences R and R_s. Then, unlike the traditional Transformer decoder, we have two individual cross attention modules—corresponding to the two documents—after the self-attention module in each Transformer block. The outputs of the two cross attention modules are simply summed, leaving other components in the Transformer block unchanged.

Model	R-1	R-2	R-L	IP	M_{sim}	MR_{max}	MMAE++
ATG	40.63	18.12	37.53	59.28	25.82	56.54	67.63
ATL	40.86	18.27	37.75	62.44	13.26	55.67	67.26
HAN	40.82	18.30	37.70	61.83	12.22	55.29	66.93
MOF	41.20	18.33	37.80	65.45	26.38	58.38	69.66
UniMS	42.94	20.50	40.96	69.38	29.72	-	-
SITA (Ours)	**43.64**	**20.53**	**41.03**	**76.41**	**33.47**	**65.38**	**77.91**

Table 1: Main results of different metrics. R-$\{1, 2, L\}$ refers to ROUGE-$\{1, 2, L\}$.

2.6 Image Selection

Given the generated summary denoted as S and pseudo captions $\{t_{z_1}, t_{z_2}, \ldots, t_{z_n}\}$, the image $\hat{v}$ whose pseudo caption $\hat{t}$ generates the highest ROUGE-L with the summary S, is selected as the most salient image, where:

$$\hat{t} = \arg\max_{t_k}(R(t_k, S)) \qquad (8)$$

$k \in \{z_1, z_2, \ldots, z_n\}$ and $R(t_k, S)$ represent the function which calculates the ROUGE-L socre between t_k and S.

Please refer to appendix A and our released code for more architecture and implementation details[4].

3 Experiment Settings

3.1 Dataset

We use the dataset build by Zhu et al. (2018), which is constructed from the Daily Mail website[1], and contains 293,965 articles for training, 10,355 articles for validation, and 10,261 articles for testing. Please refer to appendix B for more dataset details.

Model	R-1	R-2	R-L
PGN	41.11	18.31	37.74
ATL	40.86(↓0.05)	18.27(↓0.04)	37.75(↓0.01)
MOF	41.20(↑0.09)	18.33(↑0.02)	37.80(↑0.06)
BERTSum	41.51	19.43	38.85
SITA	**43.64**(↑2.13)	**20.53**(↑1.10)	**41.03**(↑2.18)
BART	41.83	19.83	39.74
UniMS	**42.94**(↑1.11)	**20.50**(↑0.67)	**40.96**(↑1.22)

Table 2: Comparison of text summary quality in terms of Rouge scores. The numbers in parentheses represent relative performance improvements of multimodal models over their single-modal ones (e.g., PGN, BERTSum and BART). ↑ indicates a positive effect, and ↓ indicates a performance decrement.

3.2 Evaluation Metrics

Following Zhu et al. (2018, 2020), we choose the following metrics. (1) **ROUGE-**$\{1, 2, L\}$ is the

standard text summarization evaluation metric. (2) **IP** is the abbreviation of **I**mage **P**recision and used to evaluate image selection. It is defined by dividing the size of the intersection between the recommended images rec_{img} and the reference images ref_{img} by the number of recommended images. (3) **M_{sim}** evaluates the image-text relevance by calculating the maximum similarity between the image and each sentence in the model summary. (4) **MR_{max}** evaluates the information integrity of the multimodal summary. It exploits a joint multimodal representation to calculate the similarity between model outputs and multimodal references. (5) **MMAE++** evaluates the overall quality of multimodal summaries. It projects both the candidate multimodal summary and the reference summary into a joint semantic space with a trained neural network. For the details of MMAE++, please check subsection 3.3 in Zhu et al. (2018)'s work.

Meanwhile, we propose **Caption-ROUGE-L**, a metric specific to SITA and its variants by calculating ROUGE-L between a generated pseudo caption and the golden caption.

3.3 Baselines

We compare our method with the five multimodal summarization methods. (1) **ATG** (Zhu et al., 2018) is a multimodal attention model, which measures image salience by the visual attention distribution over the global image features. (2) **ATL** is an ATG variant using attention distributions over image patches. (3) **HAN** is an ATL variant by adding a hierarchical attention mechanism on image patches. (4) **MOF** (Zhu et al., 2020) introduces a multimodal objective function into ATG. Among the four MOF variants, we choose the one having the best performance in five of the seven metrics we used. (5) **UniMS** (Zhang et al., 2021b) is a recent unified framework for multimodal summarization.

We also compare our method with the three text summarization methods. (1) **PGN** (See et al., 2017) is the Pointer-Generator Network for abstractive

---

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

text summarization model. (2) **BERTSum** is a recent robust BERT-based summarization model proposed by Liu and Lapata (2019), upon which our SITA is built. (3) **BART** (Lewis et al., 2020) is a pretrained seq2seq model consisting of a bidirectional encoder and an auto-regressive decoder.

4 Experiment Results

4.1 Main Results

Table 1 and 2 show the performance of the baseline models and our method. By investigating the results, we have the following observations.

(1) Our SITA achieves improvements over baselines across all evaluation metrics of image precision, text summary quality, image-text relevance, and multimodal information integrity, clearly setting up a new state-of-the-art performance.

(2) Regarding the visual modality metric (IP), MOF generally outperforms its predecessor baselines by a slight margin due to its auxiliary training objective of image selection. UniMS further gain a notable improvement over MOF by distilling knowledge in a vision-language pre-trained model. *Our SITA impressively improves more than 10% over UniMS in the precision of recommended images (e.g., 76.41 of SITA v.s. 69.38 of UniMS on the IP metric).* The reason is that the pseudo captions identified by our coarse-to-fine alignment mechanism provide much more informative clues for image selection. We will provide more detailed analyses in the following experiments.

(3) Regarding textual modality metrics, more comprehensive comparisons are shown in Table 2, which consists of three groups of results. In the first group, existing multimodal methods (ATL and MOF) demonstrate no superiority over the single-modal text summarization model they used (PGN). These efforts concluded that too many images could bring noise, and the long document had contained enough information for text generation (Zhu et al., 2018, 2020). In contrast, our SITA (in the second group) gains a much more remarkable improvement, e.g., of 2.18 ROUGE-L, on text summaries, even based on a more robust base model (BERTSum). The latest state-of-the-art UmiMS (in the third group), built upon BART, also achieves performance improvements (e.g., +1.22 ROUGE-L) on text summarization, but not as evident as ours. Note that BART performs better than BERTSum on text summarization (e.g., 39.74 v.s. 38.85 of ROUGE-L), but SITA still outperforms

UmiMS. *These results suggest that visual information actually benefits text generation, and our method exploits it more effectively.*

(4) M_{sim}, MR_{max}, and MMAE++ are used to check image-text relevance, image-text integrity, and the overall effectiveness of pictorial summaries. As expected, SITA maintains dominance over baselines on the three intermodality metrics. These superiorities come from remarkable improvements on intramodality metrics and SITA's inherent capabilities of bridging the cross-modal semantic gap.

Note that IP and all intermodality metrics depend on the selected salient images, hence indirectly relying on the generated text summaries. Rigorously, baseline methods and our SITA utilize different text summarization models (e.g., PGC, BART, and BERTSum), so these metrics will be more friendly to methods with better-performed base text summarization model. However, this fact has minor impacts on our above analyses, since image selection improvements of SITA mainly benefit from pseudo captions but not the text summaries.

4.2 Effects of the Coarse-to-fine Mechanism

Model	R-1	R-2	R-L	IP	CR$-L$
SITA	**43.64**	**20.53**	**41.03**	**76.41**	**39.39**
-*w/o* ITA	41.79	19.54	38.97	72.95	38.23
One-pass	40.83	18.32	37.98	57.28	12.31
One-pass(Dedup)	41.67	18.98	38.63	64.32	33.21

Table 3: Performance of SITA and its variants. CR$-L$ refers to Caption-ROUGE-L. -*w/o* ITA directly retrieves pseudo captions in a document without image-text alignment, One-pass does image-text alignment in a single-pass manner, and One-pass (Dedup) adds an sentence deduplication mechanism over One-pass.

4.2.1 Results of One-pass Alignment Strategy.

To investigate how the coarse-to-fine alignment strategy boosts performance, we replace it with a single-pass alignment method, which is trained to select a pseudo caption for only one single image at a time. The results of this method variant (named **One-pass**) are shown in Table 3, from which we see notable performance degradation on all metrics. Through further qualitative exploration on its prediction results, we find this method tends to generate a small set of sentences repeatedly among different images, incapable of recalling enough relevant sentences. The low Caption-ROUGE-L of **One-pass** (e.g., 12.31) also verifies this observation. One possible reason is that images in a document can sometimes be similar, making the single-

pass strategy fail to characterize the correlation and difference among these images. *In contrast, by introducing the coarse-to-fine mechanism, our alignment model synthesizes multiple images from a global perspective in the coarse-grained pass, recalling more sentences more accurately and hence facilitating further fine-grained alignment.*

4.2.2 Comparison with Simple Deduplication

To avoid recalling repeated sentences in one-pass alignment, one simple alternative strategy is introducing a deduplication mechanism. We hence implement **One-pass (Dedup)**, which will select another sentence with the next highest score if the current sentence has been chosen. As shown in Table 3, we can see that the deduplication mechanism over one-pass image text alignment brings improvements (e.g., +0.65 on R-L and +7.04 on IP). But the performance of **One-pass (Dedup)** is still far from our full SITA with the coarse-to-fine alignment strategy (e.g., with a significant gap of 2.4 on R-L and 12.09 on IP). The main reason is that one image may align with multiple semantically rich sentences. For such an image, even with the deduplicating mechanism, one-pass alignment can only recall a single sentence, potentially missing critical information, especially when other images do not semantically overlap with it. That roughly explains the performance gaps. This comparison further verifies the necessity and soundness of the technical design of the two-pass coarse-to-fine alignment.
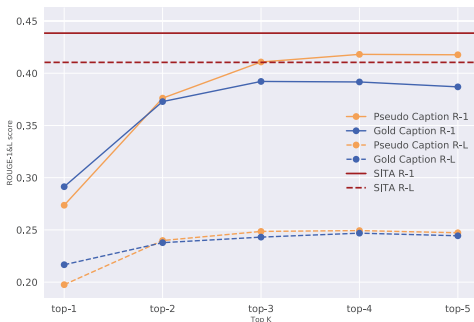

Figure 3: ROUGE-1 and ROUGE-L scores of simple summaries generated by simply concatenating pseudo captions (orange) or golden captions (blue) of a document's first k images. The scores are calculated by matching them against the reference summaries. The horizontal red (dashed) lines represent the text summaries generated by SITA. ROUGE-2 is similar to Rouge-1, which is not shown for better visualization.

4.3 Effects of Cross-modal Retrieval

To investigate the effect of the cross-modal retrieval, we directly retrieve pseudo captions in a document (rather than a summary), obtaining another method variant (named *w/o* **ITA**) requiring no image-text alignment training anymore.

As shown in Table 3, *w/o* **ITA** bring modest enhancement on text summaries (e.g., 38.85 of BERTSum v.s. 38.97 of *w/o* **ITA** on ROUGE-L), while achieving more impressive image salience (e.g., 72.28 on IP). Compared with our full SITA, this method variant (named *w/o* **ITA**) demonstrates significant performance degeneration on both text and image salience (e.g., -1.06 on ROUGE-L and -3.04 on IP). These results reveal that *(1) the knowledge in the pre-trained cross-modal retrieval model mainly helps image selection, and the image-text alignment over retrieval results is more critical for the overall performance; and (2) retrieving reference captions from summaries instead of documents is a key design of SITA.*

Note that our cross-modal retrieval model is pre-trained with 113K image-text pairs. Though UniMS distills knowledge from a vision-language model pre-trained by more than 400M image-text pairs, SITA demonstrates significant superiority.

4.4 Quality of Pseudo Captions

Model	R-1	R-2	R-L	IP	CR$-L$
Caption-train	42.22	19.70	39.29	73.59	**39.54**
Caption-input	42.71	20.04	39.85	75.33	-
SITA	**43.64**	**20.53**	**41.03**	**76.41**	39.39

Table 4: Comparisons of SITA with models using golden captions. Caption-train use golden captions to train image-text alignment model, and Caption-input directly use golden captions as input for text summarization and image selection. CR$-L$ refers to Caption-ROUGE-L. Caption-input does not generate pseudo captions, so its CR-L is unavailable.

We further analyze the effectiveness of our method from the perspective of pseudo captions' quality. We are interested in the relation between golden captions and our pseudo captions. In the MSMO's task settings, golden image captions are excluded. To perform this study, we allow the compared models to use golden captions in training under a easier task setting. Here we build another two baselines.

The first one, named **Caption-train**, **train**s the image-text alignment model with golden captions instead of the reference sentences retrieved in the first step. We compare SITA with it on the metrics

of ROUGE-$\{1, 2, L\}$, IP, and Caption-ROUGE-L. Looking into the empirical results shown in Table 4, the Caption-ROUGE-L of SITA and **Caption-train** are generally similar. Hence, from the perspective of recovering image captions, the quality of aligned sentences generated by **Caption-train** and SITA are identical. However, SITA generates better text summaries and salient images than **Caption-train** (e.g., with improvements of 0.74 on ROUGE-L and 2.82 on IP), suggesting that our aligned sentences benefit more MSMO. The reason is that the *reference captions used for alignment training are retrieved from text summaries, inherently making predicted pseudo captions imply better summary features.*

The second one, named **Caption-input**, directly utilizes golden captions instead of pseudo captions as **input**s for text summarization and image selection. We find that SITA also outperforms **Caption-input** on all metrics. The performance enhancement is less evident but still impressive, considering that SITA uses a more restricted task setting. This observation proves that *the pseudo captions we generated are even better than the original image captions for MSMO.*

The above analyses verify that pseudo captions are not only semantically consistent with images but also informative for text summarization.

4.5 Correlation between Image captions and Text Summaries

We also investigate the correlation between image captions and text summaries. Specifically, we construct a simple summary by concatenating golden (or pseudo) captions of the first k images in a document. Then, we calculate the ROUGE scores of those simple summaries. The results are shown in Figure 3, and we have the following observations:

(1) Simply aggregating some (pseudo) image captions can generate generally good summaries. For example, when selecting more than three captions, the resulting summaries even have a better ROUGE-1 than MOF. The observation verifies the inherent capabilities of image captions on the briding cross-modal semantic gap.

(2) The upward trend of the ROUGE-L with the increase of k is not as notable as that of ROUGE-1. The reason is that text generated by sentence concatenation (in random order) may lack coherence. ROUGE-L is calculated based on the longest common substring, the length of which will be limited

in this situation. This phenomenon suggests that an individual text summarization component is still required given these high-quality image captions.

(3) Generally, the red line is above the blue line most of the time, indicating that *simple summaries constructed by pseudo captions are even better than their counterparts consisting of golden captions*. The observation, again, verifies that pseudo captions generated by our image-text alignment mechanism are more informative than the original ones, in terms of improving MSMO performance.

5 Related Work

Existing text summarization approaches can be roughly categorized into extractive summarization (Narayan et al., 2018; Xiao and Carenini, 2019; Zhong et al., 2020; Wang et al., 2020) and abstractive summarization(Syed et al., 2021; Paulus et al., 2018; Zhang et al., 2020; Lewis et al., 2020; Tan et al., 2017). Classical abstractive summarizaition model such as Pointer Generator Network(See et al., 2017) and BERTSum(Liu and Lapata, 2019) serve as fundamental components for previous MSMO works.

Multimodal summarization takes data of more than one modalities as input and synthesizes information across different modalities to generate the output (UzZaman et al., 2011; Li et al., 2018; Sanabria et al., 2018; Fu et al., 2020; Im et al., 2021; Yu et al., 2021; Zhu et al., 2018, 2020; Li et al., 2020b; Jangra et al., 2020a,b, 2021b; Zhang et al., 2021a). Zhu et al. (2018) first propose generating pictorial summaries given a document and an image collection. Zhu et al. (2020) further introduced a extra cross-entropy loss for image selection. Recently, Zhang et al. (2021b) proposed to utilize knowledge distillation with a vision-language pre-trained model to help image selection, but the image precision was still far from ideal.

6 Conclusion

We have presented SITA, a multimodal **S**ummarization method based on coarse-to-fine **I**mage-**T**ext **A**lignment. SITA introduces a novel perspective of bridging the semantic gap between visual and textual modality by exploiting pseudo image captions. Our cross-modal alignment mechanism effectively generates high-quality pseudo image captions, enabling SITA to set up state-of-the-art performance easily. We discuss the feasibility and potential of leveraging pseudo

image captions , and release code[2], to inspire more future studies from our proposed perspective.

Limitations

By retrieving pseudo captions from summaries, one limitation is that the most relevant sentence for a specific image may not be in the summary. However, it has a trivial impact on the overall MSMO performance. If this happens, most of the time, the image will not be the salient image to select, and its caption will provide no helpful information for the text summary. In this situation, selecting a pseudo caption from summary sentences will not hinder the overall performance, though it may not be the best for the specific image.

Besides, even though our task setting (including the dataset and all evaluation metrics we used) strictly follows three previous works (Zhu et al., 2018, 2020; Zhang et al., 2021b), another possible limitation is that only one MSMO benchmark is used (no other dataset exists). We believe providing more diversified datasets and investigating more about the rationale under the task setting are critical to pushing forward the multimodal summarization community, although they are out of the scope of this work.

Acknowledgements

References

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2020. Multi-modal summarization for video-containing documents. *ArXiv*, abs/2009.08018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 388–403, Online. Association for Computational Linguistics.

Anubhav Jangra, Adam Jatowt, Md. Hasanuzzaman, and Sriparna Saha. 2020a. Text-image-video summary generation using joint integer linear programming. *Advances in Information Retrieval*, 12036:190 – 198.

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Md. Hasanuzzaman. 2020b. Multi-modal summary generation using multi-objective optimization. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammed Hasanuzzaman. 2021a. 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.

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammed Hasanuzzaman. 2021b. 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*, pages 818–828.

D. Kingma and J. Ba. 2014. Adam: A method for stochastic optimization. *Computer Science*.

H. W. Kuhn. 2010. The hungarian method for the assignment problem. *Naval Research Logistics*, 52(1-2):7–21.

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. *ArXiv*, abs/1910.13461.

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

---

[2]https://github.com/sitaProject/SITA

169

Haoran Li, Junnan Zhu, Tianshan Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*.

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.

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.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1845.

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.

Safa Messaoud, Ismini Lourentzou, Assma Boughoula, Mona Zehni, Zhizhen Zhao, Chengxiang Zhai, and Alexander G. Schwing. 2021. Deepqamvs: Query-aware hierarchical pointer networks for multi-video summarization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1389–1399, New York, NY, USA. Association for Computing Machinery.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Ranking sentences for extractive summarization with reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1747–1759.

Shruti Palaskar, Jindřich Libovickỳ, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596.

A. Paszke, S. Gross, F. Massa, A. Lerer, and S. Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *International Conference on Learning Representations*.

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. In *NeurIPS*.

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.

Ayesha Ayub Syed, Ford Lumban Gaol, and Tokuro Matsuo. 2021. A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access*, 9:13248–13265.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *ACL*.

Naushad UzZaman, Jeffrey P Bigham, and James F Allen. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th international conference on Intelligent user interfaces*, pages 43–52.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv*.

Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. Heterogeneous graph neural networks for extractive document summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6209–6219.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. 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 3011–3021.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. *ArXiv*, abs/2109.02401.

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

Litian Zhang, Xiaoming Zhang, Junshu Pan, and Feiran Huang. 2021a. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. *ArXiv*, abs/2112.12072.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2021b. Unims: A unified framework for multimodal summarization with knowledge distillation. *ArXiv*, abs/2109.05812.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuan-Jing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208.

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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9749–9756.

A Implementation Details

We use Pytorch-Transformers[3] to implement the Bert-base model. We use the Adam optimizer (Kingma and Ba, 2014) and set the learning rate to 0.0001. We limit the text length to 512 tokens and resize the resolution of each image to 224×224. The overall process is implemented with PyTorch(Paszke et al., 2019). We run our experiment using 2 NVIDIA V100 GPUs. The maximum number of training iterations is set to 200k, and we save the checkpoint every 2k iterations. We select the best checkpoints according to the validation loss and report the results on the test set. When training the image text alignment model, we freeze the weight of ResNet152 and use a maximum batch size of 512. When training the text summarization model, we use beam search in decoding and set the beam search size to 5. The batch size is set to 512, and each input in the batch contains a text article with 512 tokens and a pseudo caption set with 128 tokens. For more implementation details, please refer to our released code at Github[4].

B Dataset

	Train	Valid	Test
#Documents	293965	10355	10261
#AvgImgsNum	6.56	6.62	6.97
#AvgTokensNum(Document)	720.87	766.08	730.80
#AvgTokensNum(Summary)	70.12	70.02	72.16

Table 5: MSMO Dataset statistics.

We use the MSMO dataset build by Zhu et al. (2018), which is the largest benchmark dataset. This dataset is constructed from the Daily Mail website[5], containing 293,965 articles for training, 10,355 articles for validation, and 10,261 articles for testing. Each article contains a text document, and approximately seven images on average. The manually written highlights offered by Daily Mail are taken as a reference text summary. Noted that the pictorial summaries are only available on the test set, so there is no label information about the salient images during training. Image captions are excluded from the dataset for generalization.

C Case Study

To qualitatively verify our proposed method's effectiveness, we conduct a case study on generated
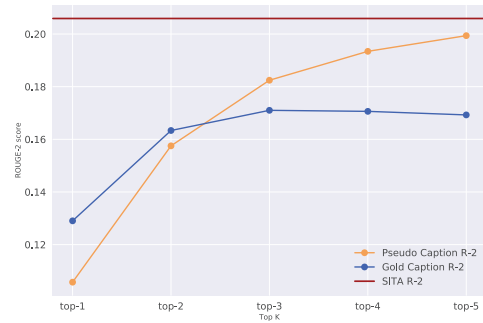

Figure 4: ROUGE-2 of simple summaries generated by simply concatenating pseudo captions (red) or golden captions (blue) of a document's first k images. The scores are calculated by matching them against the reference summaries. The horizontal red dashed lines represent the text summaries generated by our SITA model.

pseudo image captions and multi-modal summaries. As illustrated in Figure 5, the pseudo captions generated by our model can express image semantics appropriately. For the critical entities in the images, we can find the corresponding descriptions in the high-quality pseudo captions we produce. Compared with the text summary generated by single-modal and alternative multi-modal models, SITA's output captures the article's main point better, thanks to the effective incorporation of pseudo image captions implying visual knowledge. For example, the descriptions of "A robed figure" and "M16" are missing in the text summaries of compared models. In contrast, our SITA model generates a more accurate summary with the help of pseudo captions containing these essential facts, which also assists in identifying the salient image correctly.

D Rouge-2 of Simple Summaries

We only plot Rouge-1 and -L scores of simple summaries in Figure 3 for better visualization in limited space. The trend of Rouge-2 is similar to that of Rouge-1, as shown in Figure 4

---

[3]https://pytorch.org/hub/huggingface_pytorch-transformers/

[4]https://github.com/sitaProject/SITA

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

Images	Pseudo Captions	Gold Image Captions	Pictorial Summary by SITA	Pictorial Summary by MOF	Single-modal Text Summary
	Nebraska on Friday night , **after a robed figure carrying a duffel bag** and what appeared to be **an M16** strode into the Monster Mash Bash . *(R-L with SITA Sum :0.463)*	**A robed figure carrying a duffel bag** and what appeared to be an M16 strode into the Monster Mash Bash at ;the Oak View Mall in Omaha, Nebraska on Friday night	**A robed figure carrying a duffel bag** and what appeared to be **an M16** walked into the Omaha , Nebraska on Friday night. Families fled in panic from the Oak View Mall in Omaha. The costume was from the 2010 movie ` The Town ' , which features bank robbers disguised as nuns .	Families fled in panic from the Oak View Mall in Omaha. An embarrassed man explained he didn't intend to look like an ISIS gunman in a clumsy Halloween costume. The masked figure was a toy , and the costume was from the 2010 movie ` the town ' .	Families fled in terror from the Oak view mall in Omaha , Nebraska on Friday night. Police said he didn't intend to look like an ISIS gunman. The masked figure in long black robes left as quickly as he appeared .
	A man who sparked a panic in a crowded shopping mall has explained that he didn't intend to look like an ISIS gunman with his bumbling Halloween costume. *(R-L with SITA Sum: 0.197)*	**Hugo Mendoza, the father** of two young daughters, shows the nun mask he was wearing as he explains the costume was from the 2010 film 'The Town'	*(R-1 with Gold Sum:0.547 R-2 with Gold Sum:0.224 R-L with Gold Sum:0.357)*	*(R-1 with Gold Sum:0.441 R-2 with Gold Sum:0.142 R-L with Gold Sum:0.337)*	*(R-1 with Gold Sum:0.421 R-2 with Gold Sum:0.151 R-L with Gold Sum:0.289)*
	The embarrassed dad shared photos of **the toy gun** he was toting on Friday , which show that **the orange tip** is visible looking down the barrel , but not from the sides. *(R-L with SITA Sum : 0.191)*	**The orange plug in the tip** is only visibly looking down the barrel of **the mock M16** assault rifle			

Gold Summary	Reference Silent Image	Predicted Silent Image	Predicted Silent Image
Robed man with mock M16 and duffel bag appeared on Friday in Omaha , Nebraska . Families attending Halloween event at the Oak View mall fled in panic . Dad Hugo Mendoza explains he was dressed as nun bank robber from 'The Town' .			

Figure 5: An example of multimodal summarization generation. The text summary in the last column is generated by the single-modal text summarization model BertSum (Liu and Lapata, 2019). The pictorial summary in the second last column is generated by MOF (Zhu et al., 2020) re-implemented by ourselves. The orange text in the pseudo and gold image captions corresponds to the semantically important entities in red boxes of images.

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☒ A2. Did you discuss any potential risks of your work?
Left blank.

☒ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

☐ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

☐ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

☐ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D   ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.