CFSum: A Coarse-to-Fine Contribution Network for Multimodal Summarization

Min Xiao [1,2], Junnan Zhu[1,2], Haitao Lin[1,2], Yu Zhou[1,3]*, Chengqing Zong[1,2]
[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, haitao.lin, yzhou,
cqzong}@nlpr.ia.ac.cn,

Abstract

Multimodal summarization usually suffers from the problem that the contribution of the visual modality is unclear. Existing multimodal summarization approaches focus on designing the fusion methods of different modalities, while ignoring the adaptive conditions under which visual modalities are useful. Therefore, we propose a novel **C**oarse-to-**F**ine contribution network for multimodal **Sum**marization (CFSum) to consider different contributions of images for summarization. First, to eliminate the interference of useless images, we propose a pre-filter module to abandon useless images. Second, to make accurate use of useful images, we propose two levels of visual complement modules, word level and phrase level. Specifically, image contributions are calculated and are adopted to guide the attention of both textual and visual modalities. Experimental results have shown that CFSum significantly outperforms multiple strong baselines on the standard benchmark. Furthermore, the analysis verifies that useful images can even help generate non-visual words which are implicitly represented in the image[1].

1 Introduction

With the information explosion, the internet is flooded with various multimodal information. Multimodal summarization (MMS) can help generate more abundant and comprehensive summary information than unimodal based on extra visual information. Existing studies on multimodal summarization include multimodal sentence summarization (Li et al., 2018b), multimodal summarization with multimodal output (Zhu et al., 2018), multimodal meeting summarization (Li et al., 2019) and so on. In this paper, we focus on the task that generating a text summary based on the input of a text and an image. It has been proved that integrating multimodal data can help improve the quality of the summary (Li et al., 2018b; Jangra et al., 2020; Palaskar et al., 2019; Yu et al., 2021).

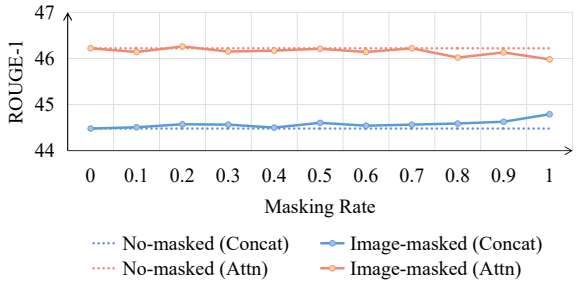

Figure 1: Experiments on existing mainstream multimodal summarization models. The performance is not affected by masking images. "Concat" is the concatenate fusion method, and "Attn" is the attention-based fusion method.

However, it is unclear whether the visual modality can indeed benefit the process of summarization. Thus, we conduct an experiment to explore the influence of masking images on the summary. As shown in Figure 1, the solid lines mean the performance of summary generated by masking portions of images, and the dashed lines indicate the origin performance. It can be observed that the dashed and the solid lines roughly coincide, which indicates that masking images do not affect the performance of the multimodal model. Some masking rates can even raise the ROUGE-1 value of the summary. It indicates that existing models do not make effective use of image information for

*Corresponding author.
[1]Code is available at https://github.com/xiaomin418/CFSum

the summary.

Existing approaches have two major problems. First, existing studies focus on multimodal fusion, such as concatenate, attention-based, and gate-based fusion (referring to Related Work). However, they ignore the adaptive conditions under which visual modalities are helpful. Thus they are poor at extracting useful visual information. Furthermore, all fusion methods do not explicitly model the image complementarity for the summary. Especially for the attention-based method, the inter-attention is not accurate enough, which leads to inefficient use of the image. Second, in many samples, the image may introduce noise, while existing fusion methods assume that all images are helpful for the summary without considering the interference of useless images. As analyzed above, we believe that: 1) It is essential to eliminate the influence of the useless image. 2) The contributions of the image to the summary need to be clarified. In particular, it is necessary to consider the complementarity of visual information relative to textual information.

Although we notice the lack of image contributions, it is difficult to detach various roles of images from a single fusion layer. Thus, in this work, we propose a novel Coarse-to-Fine contribution network for multimodal Summarization (CFSum) to extract the role of the image at different stages. First, we apply a pre-filter module to abandon useless images. It coarsely obtains helpful images for the summary. Specifically, the consistency of content between image and text is calculated. If the consistency is low, the image will be masked in subsequent encoding. Second, when the image is coarsely useful, the complement module is employed to finely guide the fusion of text with the image. To consider image contributions for text with different granularities, the complement module consists of two levels, word level and phrase level. For the word level complement module, to obtain the image complementarity over the text, the difference between bi-modal and uni-modal inputs is measured through a classification task. Then we add a loss to guide the attention between words and the image. For the phrase level complement module, similar to the word level, the image complementarity on phrases is acquired to guide the attention between phrases and the image. Through these modules, the model can acquire more explicit image contributions and provide better multimodal encoding for summary generation.

Our contributions are as follows:

(1) We propose a Coarse-to-Fine contribution network for multimodal Summarization (CFSum) to model different contributions of images for summarization.

(2) We innovatively design a pre-filter module to coarsely reduce the interference of the useless images and develop two visual complement modules to finely obtain image complementarity over the summary.

(3) Experimental results show that our model outperforms strong baselines. Besides, extensive analysis proves that useful image even contributes to non-visual words which are implicitly represented in the image.

2 Related Work

Multimodal Summarization Tasks. In the field of multimodal summarization, there are usually three steps. First, different feature extractor modules are adopted to extract the features of the text and the image, respectively. Second, the different features are fused at the fusion layer. Finally, the fused context features are fed into the text decoder to generate a summary. Existing studies focus on multimodal fusion. Specifically, the fusion methods consist of concatenate, attention-based, and gate-based. The concatenate fusion directly concatenates multimodal features into a fusion context (Li et al., 2018b, 2020a). It can fully extract high-level features of different modalities, but there is a large gap between high-dimensional spaces. The attention-based methods fuse all multimodal features with attention mechanism (Atri et al., 2021; Palaskar et al., 2019; Kitada et al., 2022), which can get the correlations between each unit of text and image. Gate-based methods take text as the central modality (Jangra et al., 2021) and exploit images to help focus on the core information (Liu et al., 2020; Li et al., 2020b). In summary, (1) all fusion methods do not explicitly model the image complementarity for the summary, which leads to inefficient use of the image. (2) concatenate and attention-based cannot eliminate the influence of useless images in the fusion layer.

Cross-modal tasks. Some studies have noted the contributions of modalities and explored the cross-modal influence in other multimodal tasks. Zeng et al. (2021) propose loss modulation to explore the contribution of individual modalities and devise a modality filter to reduce modality noise, which con-

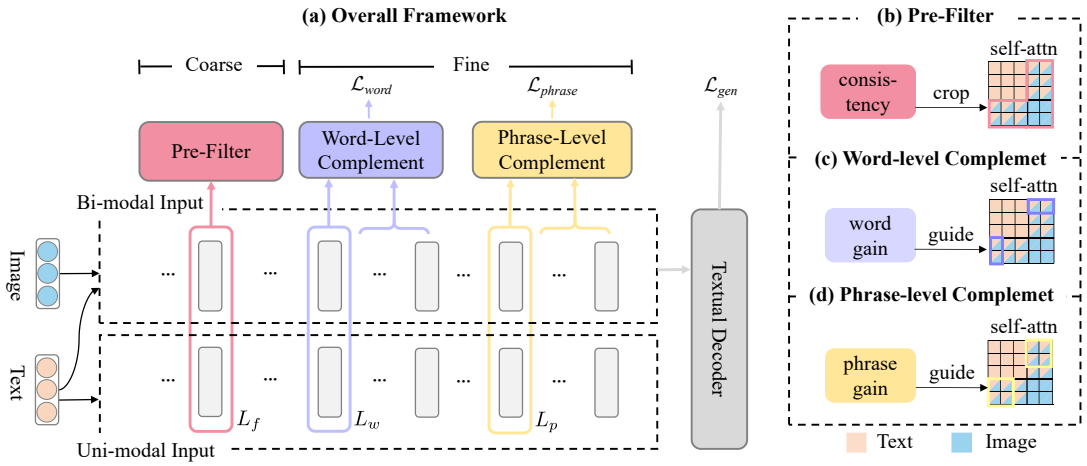

Figure 2: CFSum framework. L_f, L_w, L_p denote the starting layer of the pre-filter, the word-level complement, and the phrase-level complement modules, respectively.

siders consistency and complementarity between different modalities. Zhu et al. (2018) propose multi-task summarization: the method also selects the image that best matches the summary when generating a text summary. It guarantees the positive effect of images on the summary. Li et al. (2022) exploit ReLu-based cross-attention to align visual features to textual representation, which abandons low-value attention scores for those unaligned visual features. Inspired by the above studies, we propose CFSum, which considers various image contributions for better encoding input text and generating the final summary.

3 Proposed Methods

3.1 Overview

In this section, we introduce the details of CFSum. Given a dataset consisting of n triplets $(t_i, v_i, s_i)_{i\in[1,n]}$ with a text t_i, an image v_i, and a summary s_i, the multimodal summarization task aims at generating s_i based on t_i and v_i.

As depicted in Figure 2, the CFSum takes bi-modal and uni-modal streams as input parallelly. It builds coarse and fine image contributions with three modules (Coarse-to-Fine Structure). First, the pre-filter module coarsely filters the images inconsistent with texts (Pre-filter Module). Second, two levels of visual complement modules consisting of word level (Word-level Complement) and phrase level (Phrase-level Complement) make accurate use of useful images.

3.2 Coarse-to-Fine Structure

We build our model based on the multimodal transformer UNITER (Chen et al., 2020) and GRU (Chung et al., 2014) encoder-decoder architectures. We refer the model to UniG. As shown in Figure 2(a), in order to evaluate the complementarity of different modalities, the bi-modal and uni-modal inputs are operated parallelly with the same encoder. The two parallel streams can catch the gain of the image. Additionally, we generate a summary relying on bi-modal encoding. Uni-modal encoding assists in measuring various contributions and guiding the bi-modal encoding.

Specifically, the multimodal encoder consists of $L = 12$ multimodal transformer layers. We serve the L layers as a hierarchical structure and divide L layers into three parts as shown in Figure 2(a). L_f, L_w, L_p mark as the starting layer of the pre-filter, the word-level complement, and the phrase-level complement modules, respectively. Existing studies assume all images benefit summary generation or input text encoding, resulting in damage from unnecessary images. The pre-filter module is utilized to eliminate the interference of misleading images in advance. Next, the word-level complement module is developed to model the gain of the image on input words for the summary. Then the image gain guides the subsequent attention between words and the image. Finally, similar to the word level, the phrase-level complement module concentrates on phrases at higher layers. Each component will be elaborated in the following subsections.

3.3 Pre-filter Module

The bi-modal and uni-modal features from the i^{th} layer are encoded as $m^i \in \mathbb{R}^{C\times H}, u^i \in \mathbb{R}^{T\times H}$, where $i \in [1, L]$, and C, T denote the lengths of bi-modal and uni-modal tokens. H denotes the hidden dimension. The bi-modal self-attention matrix in the i^{th} layer is $A^i = (a^i_{r,s}) \in \mathbb{R}^{C\times C}$.

The pre-filter module aims at filtering images that are unnecessary to the summary. As shown in Figure 2(a), given two encoded features m^{L_f} and u^{L_f} from the L^{th}_f layer, the goal of the filtering module is to select those useless images and guide the self-attention of all subsequent layers. We believe that if the bi-modal feature has low consistency with the uni-modal feature, the image may introduce interferential information. Specifically, we first calculate the consistency Δ^C between uni-modal feature u^{L_f} and bi-modal feature m^{L_f} as follows:

$$pu = \text{MeanPool}(u^{L_f}), \tag{1}$$

$$pm = \text{MeanPool}(m^{L_f}), \tag{2}$$

$$\Delta^C = \text{Sign}(\text{cosine}(pu, pm) - \alpha) \tag{3}$$

We define the indicator function as:

$$I_{r,s} = \begin{cases} 1, & r \geq T, s \geq T \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

which represents the text attending to the image, the image attending to the text, and the image attending to itself shown in Figure 2(a). Then we calculate the new subsequent self-attention $na^i_{r,s}$ with:

$$\begin{aligned} na^i_{r,s} =& a^i_{r,s} \times (1 - I_{r,s}) + a^i_{r,s} \times I_{r,s} \times \Delta^C, \\ & i \in [L_f + 1, L] \end{aligned} \tag{5}$$

By correcting the attention matrix, the image with a large deviation in content is cropped out. In other words, the multimodal inconsistency features degenerate into text-only features through this process. The simple method has been shown to be effective in our experiments.

3.4 Word-level Complement

This section introduces a word-level complement module, considered as an auxiliary task during the training process. First, we measure the image gain on input words for the summary. Then the image gain is applied to guide the attention between words and the image (as shown in Figure 2(b)).

Image gain measurement. Intuitively, the text tokens should concern the image which is helpful for the summary. In previous attention-based studies, inter-modality correlation can be modeled as softmax$(\frac{QK}{\sqrt{D}})V$. Q, K, V are the projected features from the bi-modal input. However, it does not explicitly model the image complementarity for the summary, which leads to inefficient use of the image.

Following the motivation above, we hope to calculate the image gain on the summary with mutual information. In other words, we want to measure whether generating summaries based on bi-modal feature m^L is more deterministic than generating summaries based on uni-modal feature u^L. Thus, we expect to calculate the image gain on the k-th word of the reference summary:

$$GI_k = Gain(s_k/u^L, s_k/m^L) \tag{6}$$

However, we intend to obtain GI_k before generating summary S and encoding m^L. Thus GI can be beneficial for generating S and encoding m^L. To this end, we define Copy Classification task Y to approximate the summary task S: for each input text token t_j, the target is to binary categorize whether it appears in the reference summary. If the token appears in the reference summary, it is classified as $\hat{y_j} = 1$; otherwise, $\hat{y_j} = 0$. Next, the GI_j is given by:

$$GI_j = Gain(y_j/u^{L_w}, y_j/m^{L_w}) \tag{7}$$

where u^{L_w}, m^{L_w} denote the uni-modal and the bi-modal feature acquired by L^{th}_w layer. Finally, we measure the gain that the image brings to predict whether a word appears correctly in the summary as follows:

$$\begin{aligned} GI_j &= Gain(y_j/m^{L_w}, y_j/u^{L_w}) \\ &= logP(y_j = \hat{y_j}/m^{L_w}) - logP(y_j = \hat{y_j}/u^{L_w}) \end{aligned} \tag{8}$$

Derivation details can refer to the Appendix B. In addition, to ensure the correct gain direction, we add a binary cross-entropy loss to train the Copy Classification Task Y:

$$\mathcal{L}_{copyc} = \text{BCE}(y_j, \hat{y_j}/m^{L_w}) + \text{BCE}(y_j, \hat{y_j}/u^{L_w}) \tag{9}$$

Image gain application. We introduce divergence loss to restrain that the image with greater gain should receive more textual attention. In successive $i^{th}{}_{i\in[L_w+1, L_w+3]}$ layer, the average inter-attention between each text token t_j and the image is:

$$T2V_j^i = \frac{1}{2(C-T)}\left(\sum_{s=T+1}^{s=C} a_{j,s}^i + \sum_{s=T+1}^{s=C} a_{s,j}^i\right) \quad (10)$$

where $a_{j,s}^i, a_{s,j}^i$ represent the attention of image-to-text and text-to-image, respectively.

Finally, an attention divergence loss is added to restrain the inter-attention scores $T2V_j^i$ with GI_j:

$$\mathcal{L}_{word} = \text{KL}(\text{Softmax}(GI_j)||\text{Avg}(T2V_j^i)) \quad (11)$$

By minimizing the divergence loss, the text token attends to the image according to the gain it brings. Interaction between word gain and inter-attention learns to pay attention to the useful image. Appendix C provides examples to figure out the word-level complement.

3.5 Phrase-level Complement

Considering the image contribution to text of different granularities, we put forward a phrase-level complement module similar to the word level (as shown in Figure 2(c)).

Image gain measurement. Different from copy classification task at the word level, we define Copy Scorer task to measure the image gain on phrases: We obtain phrases $\{p_1, ..., p_k...\}$ from the text with StandfordNLP[2]. $\{l_1, ..., l_k...\}$ is the number of words in the phrases. The task targets scoring the proportion of words that appear in both the phrase and the reference summary:

$$R_{p_k}^u = \text{Scorer}(u^{L_p}) \quad (12)$$
$$R_{p_k}^m = \text{Scorer}(m^{L_p}) \quad (13)$$

where Scorer is a MLP. The ground truth proportion is obtained with the following:

$$\hat{R_{p_k}} = \frac{Count_{t_{j'} \in p_k}(t_{j'})}{l_k} \quad (14)$$

where $Count_{t_{j'} \in p_k}$ denotes the number of words that appear in both the phrase p_k and the reference summary. Therefore, the image gain on phrase can be acquired as:

$$GS_{p_k} = |R_{p_k}^u - \hat{R_{p_k}}| - |R_{p_k}^m - \hat{R_{p_k}}| \quad (15)$$

Similarly, to guarantee the correctness of phrase gain, we add a squared loss for the Copy Scorer task:

$$\mathcal{L}_{copys} = \text{MSE}(R_{p_k}^m, \hat{R_{p_k}}) + \text{MSE}(R_{p_k}^u, \hat{R_{p_k}}) \quad (16)$$

[2] https://github.com/stanfordnlp

Especially, for the convenience of applying phrase gain GS_{p_k}, we project it to token gain GS_j as:

$$GS_j = \max\{GS_{p_k}, t_j \in p_k\} \quad (17)$$

Image gain application. Second, we introduce a phrase attention divergence loss to restrain that the image with greater phrase gain should receive more textual attention. We obtain the inter-attention score $T2V_j^i$ from $i \in [L_p+1, L_p+3]$ layers as formula 10. Finally, we restrain it with the following:

$$\mathcal{L}_{phrase} = \text{KL}(\text{Softmax}(GS_j)||\text{Avg}(T2V_j^i)) \quad (18)$$

The phrase-level restraint guarantees the image contributing to the text of phrase granularity.

3.6 Training and Inference

In the training phase, to ensure the accuracy of the information difference between bi-modal and uni-modal, we initialize the model only with the summary generation loss. We apply negative log-likelihood for the target word sequence as the overall loss:

$$\mathcal{L}_{gen} = \frac{1}{T}\sum_{t=1}^{T}(-\log P(s_t)) \quad (19)$$

Then the model is finetuned with the hierarchical modules' objectives:

$$\mathcal{L} = \mathcal{L}_{gen} + \mathcal{L}_{word} + \mathcal{L}_{phrase} + \mathcal{L}_{copyc} + \mathcal{L}_{copys} \quad (20)$$

In the inference phase, we only maintain the prefilter module. $\mathcal{L}_{word}$ and $\mathcal{L}_{phrase}$ are added to let the model learn how to fuse multimodal information. Hence, differences in training and inference phases would not hurt the generation.

Dataset	Size	Src. Length (Min/Avg/Max)	Ref. Length (Min/Avg/Max)
train	62,000	11/21.68/63	2/7.72/25
dev	2,000	11/24.35/47	3/7.68/17
test	2,000	11/22.97/51	3/7.67/24

Table 1: Statistical information about the dataset. "Src. Length" and "Ref. Length" denote the number of words in the source sentence and reference summary.

4 Experiment

4.1 Settings

We experiment with the multimodal sentence summarization dataset[3] (Li et al., 2018a). It contains

		ROUGE-1	ROUGE-2	ROUGE-L	BLEU	BERTScore	MoverScore
Lead△		33.64	13.40	31.84	-	-	-
Compress△		31.56	11.02	28.87	-	-	-
ABS△		35.95	18.21	31.89	-	-	-
SEASS△		44.86	23.03	41.92	-	-	-
Multi-Source△		39.67	19.11	38.03	-	-	-
Doubly-Attention△		41.11	21.75	39.92	-	-	-
MAtt△		47.28	24.85	44.48	-	-	-
MSE△		45.63	23.68	42.97	-	-	-
UniG (T)		45.90	24.08	42.98	47.09	86.54	31.06
UniG		46.22	24.28	43.47	46.85	86.57	30.95
K1	CFSum-F_3	47.39*	25.42*	44.35*	48.51*	86.90*	31.89*
	CFSum-W_6	47.33*	25.38*	44.26*	48.43*	86.91*	31.84*
	CFSum-P_9	47.28*	25.13*	44.18*	48.19*	86.91*	31.67
K2	CFSum-W_6F_9	47.53*	25.37*	44.41*	48.48*	86.94*	32.24*
	CFSum-F_3W_6	47.66*	25.33*	44.54*	48.45*	86.95*	31.88*
	CFSum-F_3P_9	47.72*	25.51*	44.58*	48.66*	86.96*	32.03*
K3	CFSum-$F_3W_6P_9$	**47.86***	**25.64***	**44.64***	**48.83***	**86.98***	**32.36***
	CFSum-$F_9W_3P_6$	47.58*	25.42*	44.49*	48.35*	86.95*	32.10*

Table 2: Automatic evaluation results of CFSum. "△" marks the results from Li et al. (2018b) and Li et al. (2020b)[4]. "K1/2/3" denotes one/two/three kind(s) of contribution(s). "*" indicates the model performs significantly better than the UniG by the 95% confidence interval (p<0.05).

66,000 samples in total. And each sample is a triplet of <sentence, image, summary>. Some statistical information is shown in Table 1. Appendix D gives the categories of test images.

We set both the text embedding dimension and hidden dimension as 768. We apply "bert-base-uncased" (Devlin et al., 2019) vocabulary with 28,996 tokens. The dropout (Srivastava et al., 2014) rate is set to 0.1. Besides, the batch size is set to 8. For texts, we use the max text encoding length of 60, and the minimum text decoding length is 8. For images, the object detection tool BUTD (Anderson et al., 2018) is applied to extract the image feature, with the maximum boxes as 36. We use the Adam (Kingma and Ba, 2014) optimizer and set the learning rate as $5e-05$, momentum parameters as $\beta_1 = 0.9, \beta_2 = 0.98$. The model is initially trained with the summary generation loss for 35 epochs. To obtain our final model, we train for a further 15 epochs with the hierarchical framework. In the test phase, we employ beam search and set the beam size as 4 to generate the summary. The parameter α in the pre-filter module is set as $\alpha = 0.65$.

4.2 Comparative Methods

Lead: Exploiting the first eight words as the summary.

Compress (Clarke and Lapata, 2008): It uses integer linear programming to infer global optimal compressions.

ABS (Rush et al., 2015): It utilizes an attention-based model to generate words of summary conditioned on the input text.

SEASS (Zhou et al., 2017): It constructs a second-level sentence representation with a sentence encoder and a selective gate for summarization.

Multi-Source (Libovický and Helcl, 2017): It combines multiple source modalities based on the hierarchical attention mechanisms over each modality for solving the multimodal machine translation.

Doubley-attentive (Calixto et al., 2017): It uses two separate attention mechanisms to incorporate the visual feature, which minified the gap between the image and the translation.

MAtt (Li et al., 2018b): It proposes modality attention and image filtering for multimodal summarization.

MSE (Li et al., 2020b): It proposes to apply the visual selective gates to multimodal summarization.

UniG: It is our base model with multimodal transformer UNITER and GRU decoder.

UniG (T): UniG fed only with textual modality.

4.3 Automatic Evaluation Results

Our methods are reported with six automatic metrics, including ROUGE-1, ROUGE-2, ROUGE-L (Lin and Hovy, 2002), BLEU (Papineni et al., 2002), BERTScore (Zhang* et al., 2020), and MoverScore (Zhao et al., 2019). More details of

[3] http://www.nlpr.ia.ac.cn/cip/dataset.htm

[4] Because there is no output from these systems, we only report ROUGEs in papers. In addition, BLEU, BERTScore, and MoverScore cannot be recalculated.

Model	Informativeness	Fluency	Non-Redundancy
UniG (T)	3.63	3.48	2.91
UniG	3.69	3.66	3.05
CFSum	**3.91**	**3.90**	**3.31**

Table 3: Human evaluations. 1 stands for the worst, and 5 stands for the best for three metrics.

evaluation scripts are given in Appendix A.

Comparisons with Baselines. We compare our work with our baselines and other work on the multimodal sentence summarization dataset. Table 2 shows the results of different models. The results show that **UniG** performs comparably with **UniG (T)**. **CFSum**s build on **UniG**, and introduces coarse-to-fine contribution network. "F", "W", and "P" represent the pre-filter, the word-level complement, and the phrase-level complement modules contained in the CFSum. The footnote is the location of the corresponding module. For example, CFSum-F_3 contains a pre-filter module with $L_f = 3$. Generally, our methods **CFSum**s outperform the baselines **UniG (T)** and **UniG**. The best methods is **CFSum-$F_3W_6P_9$**. And it achieves 1.64 higher points on ROUGE-1 than **UniG**. We also conduct ablation experiments by applying one or two kinds of contributions. The results demonstrate that each image contribution benefits the model. In addition, combining all image contributions brings greater gains than a single contribution. Therefore, it can be concluded that different contributions are complementary to the summary. Besides, we conduct ablation studies by placing the pre-filter module at the beginning ($L_f = 3$) or the end of the hierarchical layers ($L_f = 9$). In comparison, placing the pre-filter module at the beginning (**CFSum-$F_3W_6P_9$**) yields better performance.

4.4 Human Evaluation Results

We randomly select 50 samples from the test dataset and invite three postgraduates to score 1-5 for the summary quality. 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 reference summary? Table 3 shows the human evaluation results. We run the inter-annotator agreement study on three volunteers' scores and achieve reasonable scores, 0.47, 0.39, and 0.43 on informativeness, fluency, and non-redundancy, respectively. The results show that our method **CFSum-$F_3W_6S_9$** achieves the best performance on all three aspects over **UniG (T)** and **UniG** baselines. Thus we conclude that our method is also effective through human evaluation.

4.5 Further Analysis

4.5.1 Complement Modules Analysis

In other multimodal tasks such as image captioning and multimodal translation, their models learn to attend to the image more for visual words like "red", "rose" and "woman" (Lu et al., 2017; Calixto et al., 2017). Since our proposed complement modules aim at extracting complementary information relative to textual modality, we want to know which word or phrase the image provides gains on. As shown in Figure 3, we visualize the complement gain value for the input words. We manually align the reference summary and the input text. The word highlighted with a red box indicates that it appears in the reference summary generatively[5] or extractively.

First, we find that words with positive image gain can basically cover the reference summary information. It proves that our calculated gain helps in generating the target summary words. Second, it can be observed that different complement modules bring positive gains in different areas, which means different levels of complement modules are complementary. It further explains that multiple contributions are better than a single contribution in the experimental results. At last, it is worth noting that some words are gained from the image but are not visible in the image, *i.e.*, "relatives" and "victims". Therefore, we believe the image brings gain in both visible and invisible words. We explain further in Gainable Images.

4.5.2 Pre-filter Module Analysis

Since we believe that images should provide meaningful contributions instead of robustness enhancements in multimodal summarization, we wonder whether unpaired multimodal data may affect the performance of our model. Therefore, we try generating the summary based on the unpaired image and text.

In the test set, most of the images are highly similar in theme and content. Generating unpaired data

[5]"Generatively" means that the summary word is obtained by paraphrasing or synonymous substitution of the input word.

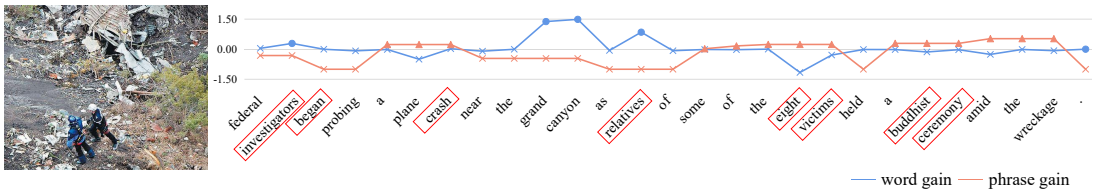

Figure 3: Visualization of word-complement gain and phrase-complement gain produced by our model. ▲/● indicates that the value is greater than 0.

Figure 4: Visualization of gainable images.

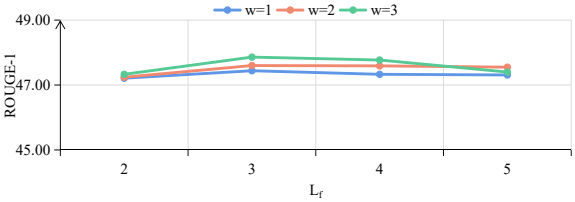

Figure 5: Ablation studies of layer setting.

with automatic shuffling is not significant for analysis. Therefore, we manually exchange v_i and v_j in pairs <t_i, v_i>, <t_j, v_j>, where v_i, v_j have different themes or contents.

We exchange 20 pairs from 100 pairs of test samples. And we conduct experiments with different sampling for three times. The mean and standard deviation reports as Table 4. "Paired" represents ROUGE-1 on test set, "Unpaired" represents ROUGE-1 on the unpaired set. "CFSum (filter-off)" represents turning down the pre-filter mechanism.

Model	Paired	Unpaired
UniG	46.22	46.20(±0.012)
CFSum	47.86	47.46(±0.007)
CFSum (filter-off)	47.77	47.12(±0.011)

Table 4: Performance of unpaired multimodal data for the baseline and our methods.

The results show different trends in the two models. For UniG, unpaired multi-modalities do not affect the performance. We guess UniG does not exploit meaningful image information while relying only on text to generate the summary. In contrast, CFSum hurt more severely from unpairing. The difference exists because CFSum depends on the image and text. Thus, the unpaired image would reduce the correct information that CFSum gets. However, CFSum still performs better than UniG, proving that it is fault-tolerant. Furthermore, CFSum (filter-off) significantly suffers from unpaired data, showing that pre-filter can eliminate useless images.

4.5.3 Ablation Study

One of the most important hyperparameters in CFSum is the location of different contribution modules. Because the three modules' order in the network is fixed, we change their absolute position in the encoder layers and report the corresponding performance in Figure 5. w denotes the number of layers between two modules, and the X axis denotes the starting layer of the pre-filter module. The results show that the different layer settings achieve comparable performance. It is noticeable that $w = 2$ weakens the model. This is due to the fact that the network with small w loses the advantage of a hierarchical structure in the encoder.

4.5.4 Gainable Images

We select three gained words and corresponding gainable images to show in Figure 4. Consistent with our perception, images bring gains on visual words, such as "earthquake". More importantly, they bring gains on non-visual words such as "celebrate" and "victims". For example, "celebrate" may be used in competitions, events, and diplomacy as shown in Figure 4. Multimodal tasks such as image captioning or multimodal question answering focus on establishing associations between visual words and images. However, multimodal summarization also needs to pay attention to the associations between non-visual words and images. In other words, image contributes to both visual and non-visual words.

5 Conclusion

Based on the observation that existing multimodal summary models do not take full advantage of useful image information, this paper focuses on modeling different contributions of images for summarization. Therefore, we propose a novel framework CFSum consisting of pre-filter, word-level complement, and phrase-level complement modules. The pre-filter coarsely eliminates the impact of useless images. The two-level visual complement modules measure different aspects of image gains and guide the fusion of different modalities. Experimental results have shown that CFSum can significantly improve the summary. More importantly, the complement modules make images contribute to visual words and non-visual words.

Limitations

Since our method constructs on the multimodal transformer, it cannot be migrated to the dual-stream model. Experiment results show that CFSum can achieve comparable performance with strong baselines. But it still cannot surpass the SOTA of some dual-stream large models.

Acknowledgements

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

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086.

Yash Kumar Atri, Shraman Pramanick, Vikram Goyal, and Tanmoy Chakraborty. 2021. See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization. *Know.-Based Syst.*, 227(C).

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *Computer Vision – ECCV 2020*, pages 104–120, Cham. Springer International Publishing.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, pages 399–429.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anubhav Jangra, Adam Jatowt, Mohammad Hasanuzzaman, and Sriparna Saha. 2020. Text-image-video summary generation using joint integer linear programming. In *Advances in Information Retrieval*, pages 190–198, Cham. Springer International Publishing.

Anubhav Jangra, Adam Jatowt, Sriparna Saha, and Mohammad Hasanuzzaman. 2021. A survey on multimodal summarization. *CoRR*, abs/2109.05199.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Shunsuke Kitada, Yuki Iwazaki, Riku Togashi, and Hitoshi Iyatomi. 2022. Dm2s2: Deep multimodal sequence sets with hierarchical modality attention. *IEEE Access*, 10:120023–120034.

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, Tianshan Liu, Jiajun Zhang, and Chengqing Zong. 2018a. Multi-modal sentence summarization with modality attention and image filtering. In *International Joint Conference on Artificial Intelligence*.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018b. 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.

Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020b. Multimodal sentence summarization via multimodal selective encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiangfeng Li, Zijian Zhang, Bowen Wang, Qinpei Zhao, and Chenxi Zhang. 2022. Inter- and intra-modal contrastive hybrid learning framework for multimodal abstractive summarization. *Entropy*, 24(6).

Manling Li, Lingyu Zhang, Heng Ji, and Richard J. Radke. 2019. Keep meeting summaries on topic: Abstractive multi-modal meeting summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2190–2196, Florence, Italy. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, pages 45–51, Phildadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3242–3250.

Shruti Palaskar, Jindřich Libovický, 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, Florence, Italy. 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.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

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.

Ying Zeng, Sijie Mai, and Haifeng Hu. 2021. Which is making the contribution: Modulating unimodal and cross-modal dynamics for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1262–1274, 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.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Vancouver, Canada. 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.

A Experiment details

Here, we will introduce some detailed settings for our experiments. All methods are run on NVIDIA GeForce RTX 3090. UniG has 139M parameters. When the batch size is 8, it takes 20 hours to train for 50 epochs with a single GPU.

We also provide evaluation scripts for reproduction. For ROUGE score, we use file2rouge[6] with default settings. For BERTScore[7], 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. For MoverScore, we use moverscore_v2[8], which leverages BERT and Earth Mover Distance to measure the similarity.

B Derivation details

Derivation detail of formula 8 is:

$$\begin{aligned} GI_j &= Gain(y_j/m^{L_w}, y_j/u^{L_w}) \\ &= \mathrm{KL}(\hat{y_j}||y_j/m^{L_w}) - \mathrm{KL}(\hat{y_j}||y_j/u^{L_w}) \\ &= P(\hat{y_j}=1)\cdot logP(y_j=1/m^{L_w}) \\ &+ P(\hat{y_j}=0)\cdot logP(y_j=0/m^{L_w}) \\ &- P(\hat{y_j}=1)\cdot logP(y_j=1/u^{L_w}) \\ &- P(\hat{y_j}=0)\cdot logP(y_j=0/u^{L_w}) \\ &= P(y_j=\hat{y_j})\cdot logP(y_j=\hat{y_j}/m^{L_w}) \\ &- P(y_j=\hat{y_j})\cdot logP(y_j=\hat{y_j}/u^{L_w}) \\ &= logP(y_j=\hat{y_j}/m^{L_w}) - logP(y_j=\hat{y_j}/u^{L_w}) \end{aligned} \quad (21)$$

Thus the gain is simplified to entropy difference.

C Examples of Complement Modules

We will provide some examples to explain further Word-level Complement. For one of the input words t_j, we assume that it appears in the reference summary. Then the ground truth of the copy classification is $\hat{y_j}=1$. We list hypothetical classification results of bi-modal and uni-modal in Table 5.

	P($y_j=1$)	P($y_j=0$)
u^{L_w}	0.4	0.6
m^{L_w}	0.6	0.4

Table 5: Copy classification task results.

Then, the GI_j is calculated as:

$$\begin{aligned} GI_j &= Gain(y_j/m^{L_w}, y_j/u^{L_w}) \\ &= logP(y_j=1/m^{L_w}) - logP(y_j=1/u^{L_w}) \\ &= log0.6 - log0.4 \\ &= 0.405 \end{aligned} \quad (22)$$

which means the image may give the input word t_j a gain of 0.405. Furthermore, the image brings a positive gain. Thus in the attention layer, the text word t_j should give the image a higher attention score.

D Impact of image category

To further analyze the impact of our approach on different categories of images. We categorize the test images with VGG19 and show the performance of each type of image. As shown in Figure 6, there are 380 categories in the test images, and we list the top 10 categories with the highest proportion. It can be seen that the image is evenly distributed. The line charts also show that CFSum is superior to UniG in all categories. Therefore there is no category bias in our method.

E Guided Attention

We visualize (1) the attention matrix from the 8^{th} encoder layer of CFSum-$F_3W_6S_9$, whose layer is under the word-level guidance. (2) the attention matrix from the 11^{th} encoder layer of CFSum-$F_3W_6S_9$, whose layer is under the phrase-level guidance. The attention matrix is renormalized after removing [CLS] and [SEP]. They are shown in Figure 7 and Figure 8.

From the attention under the word-level guidance, we can observe that some input words which generatively or extractively occur in the reference summary will attend to the image, such as "crash" and "relatives". From the attention under the phrase-level guidance, we can observe that some input phrases which generatively or extractively occur in the reference summary attend to the image more. Above all, it also proves that two visual complement modules succeed in providing better encoding to generate summaries.

[6] https://github.com/pltrdy/files2rouge
[7] https://pypi.org/project/bert-score/0.2.1
[8] https://github.com/AIPHES/emnlp19-moverscore

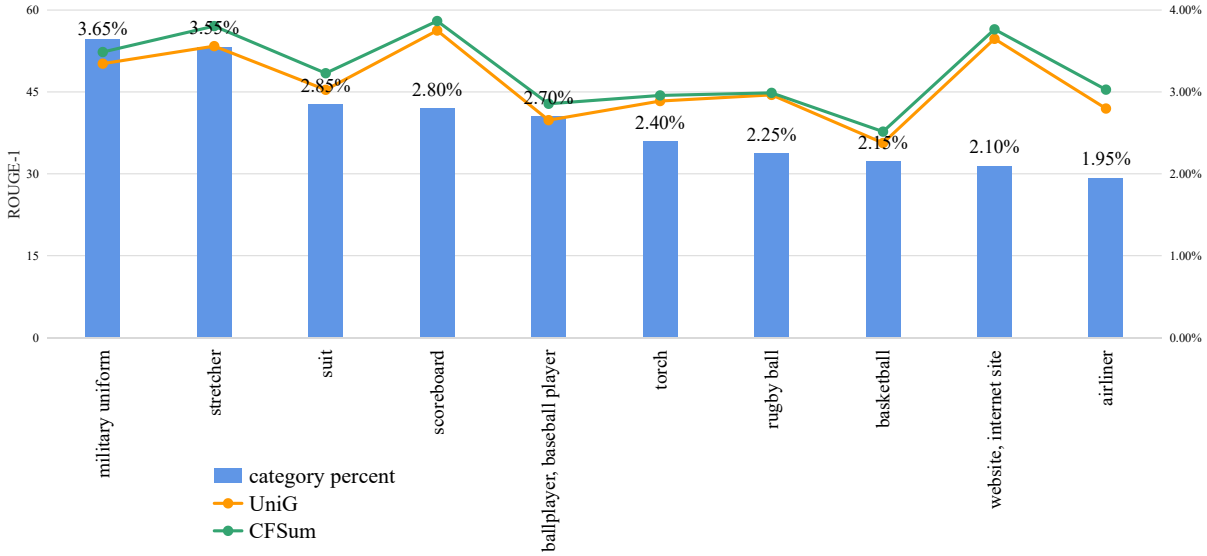

Figure 6: Top10 categories of test images and their corresponding performance.

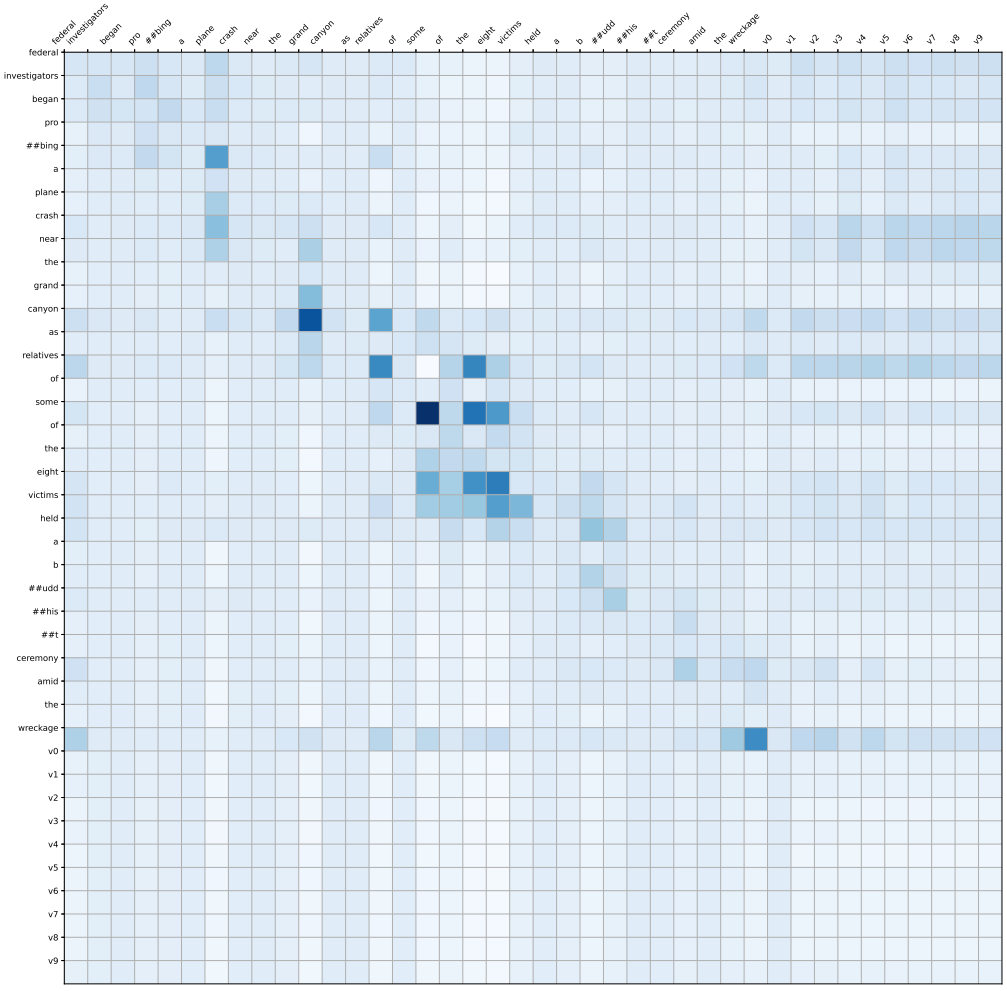

Figure 7: Visualize 8^{th} layer's attention under the word-level guided module. The reference summary is "crash investigation begins relatives mourn eight victims". $v_0 \sim v_9$ is the image object detected feature.

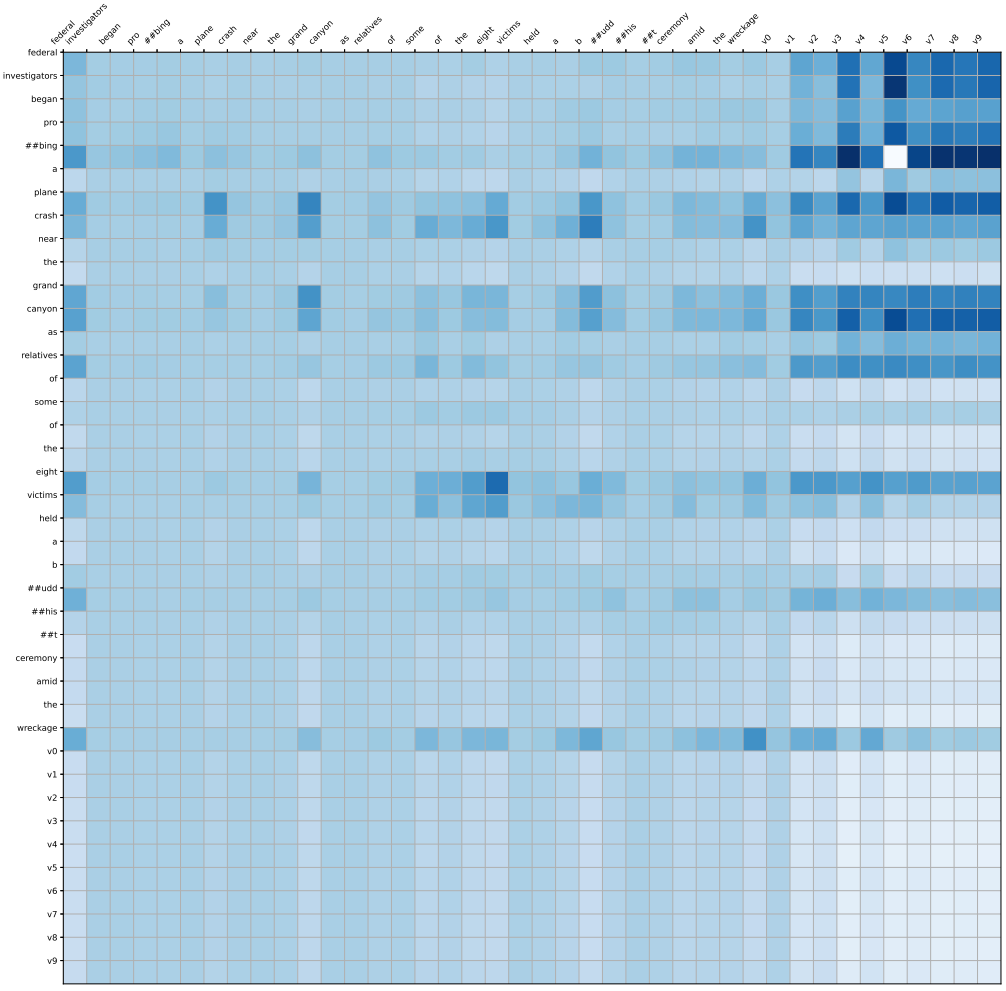

Figure 8: Visualize 11^{th} layer's attention under the word-level guided module. The reference summary is "crash investigation begins relatives mourn eight victims". $v_0 \sim v_9$ is the image object detected feature.