Visually-Aware Context Modeling for News Image Captioning

Anonymous ACL submission

Abstract

News Image Captioning aims to create captions from news articles and images, emphasizing the connection between textual context and visual elements. Recognizing the significance of human faces in news images and the face-name co-occurrence pattern in existing datasets, we propose a face-naming module for learning better name embeddings. Apart from names, which can be directly linked to an image area (faces), news image captions mostly contain context information that can only be found in the article. We design a retrieval strategy using CLIP to retrieve sentences that are semantically close to the image, mimicking human thought process of linking articles to images. Furthermore, to tackle the problem of the imbalanced proportion of article context and image context in captions, we introduce a simple yet effective method Contrasting with Language Model backbone (CoLaM) to the training pipeline. We conduct extensive experiments to demonstrate the efficacy of our framework. We outperform the previous state-of-the-art by 7.97/5.80 CIDEr scores on GoodNews/NYTimes800k.

1 Introduction

Online news consumption heavily relies on news images as a key source of supplementary information alongside articles. These images, paired with engaging and informative captions, play a crucial role in capturing readers' attention. Typically, a news image illustrates a portion of the article, with the caption linking the image content to the article. Ideally, readers should be able to grasp the essence of the news article by browsing through its images and their corresponding captions.

News Image Captioning, the task of generating a caption for an image using the contextual information derived from the corresponding article, contrasts with generic image captioning, where the image contains all necessary information for generating a descriptive sentence. Figure 1 shows a

Figure 1: Two types of image captions. The image contains all context needed for the generic image caption, while in the news image caption, we find more named entities, including the name of a celebrity whose face appears in the image, and context that is retrieved from the corresponding news article. Most of the context in the news image caption requires linking the image to the article.

generic image caption, and a news image caption from the GoodNews (Biten et al., 2019) dataset[1]. In the news image caption, Elizabeth Warren acts as a pivotal word. As a celebrity, Elizabeth Warren can also be recognized from the image. Furthermore, the news image caption contains context that is retrieved from the article. Moreover, all colored text in news image caption requires linking the image to the article, showing large imbalances in the proportion of article context and image context reflected in the caption. In contrast, the generic image caption simply serves as a descriptive sentence of the image without any additional information.

Given the distinct nature of news image captions compared to generic image captions, an important question arises: How can visual inputs in News Image Captioning be used more effectively? Current methods primarily incorporate visual features from pretrained image encoders through cross-attention modules (Tran et al., 2020; Yang et al., 2021) or visual prefixes (Zhang et al., 2022a) to pre-trained language models. This straightforward integration method is commonly applied in generic image captioning. However, as also indicated by Zhang and Wan (2023),there is a need for more effective uti-

[1]The generic caption is generated with BLIP-2 (Li et al., 2022). More examples can be found Appendix H.

lization of images in News Image Captioning.

We draw inspiration from studies on the human cognitive system, where studies indicate that faces uniquely capture human attention more than other objects in images (Ro et al., 2001). Additionally, recognizing familiar faces enhances the recall of detailed "person knowledge," like personal traits and intentions (Gobbini and Haxby, 2007). In News Image Captioning, this insight is particularly relevant, as news images frequently focus on human subjects. This understanding of how faces impact attention and memory guides our strategies for handling images centered around people. In two commonly used News Image Captioning datasets, GoodNews (Biten et al., 2019) and NYTimes800k (Tran et al., 2020), there is a notable pattern where over 56% of samples feature both faces and names, while about 32% have neither. All samples with significant faces in images also include names in their captions[2]. This pattern, aligned with cognitive science's emphasis on the importance of faces in image perception, motivates the differentiation of faces from other objects in images for distinct treatment. We design a face-naming module to help the model to selectively attend to relevant names from the accompanying article. The face-naming module includes a prefix-augmented attention module (Zhao et al., 2022) and is trained with a weakly supervised face-name alignment method (Qu et al., 2023).

Apart from the names, news image captions, unlike generic ones, often include contextual information (like "a Democrat running for the senate" in Figure 1) that cannot be directly linked to image areas. To generate these captions accurately, linking image content with relevant article segments is essential. We use a CLIP-based (Radford et al., 2021) sentence retrieval strategy to find article sentences closely related to the image, aiding our caption generation process.

Moreover, news captions typically emphasize more context derived from the articles to engage readers and abstractly illustrate the article's content. Instead of explicitly modeling the image context and the article context, which requires detailed annotations, we propose Contrasting with Language Model backbone (CoLaM) which implicitly guide the model to prioritize article context. We align the embedding space of the multimodal models to the embedding space of their frozen language model backbones using a margin loss. This approach ensures the model with multimodal inputs focus more effectively on article-related context.

[2]We provide more detailed statistics in Appendix C.

To sum up, we introduce a novel framework for News Image Captioning that utilizes visual inputs differently than previous works. Our main contributions include:

1. We are the first to introduce distinct modules tailored for different visual inputs in News Image Captioning, establishing the new state-of-the-art on two datasets.

2. For visual inputs like faces that can be directly linked to textual context, we design a face naming module to utilize the commonly-occurred pattern of face-name co-occurrence in News Image Captioning datasets. For visual inputs that cannot be directly visually grounded, we design a sentence retrieval strategy using CLIP to bridge the gap between the article segments and the images. The proposed modules result in significant improvement in performance.

3. Addressing the imbalance between article and image context in the captions, we propose CoLaM, a universal method using a margin loss to enhance article context learning, further improving captioning performance.

2 Related Work

In contrast to generic image captioning methods (Karpathy and Fei-Fei, 2015; Donahue et al., 2015; Vinyals et al., 2015; Xu et al., 2015; Anderson et al., 2018; Lu et al., 2017), which rely solely on images as input, News Image Captioning takes both images and news articles as input. It dictates that models should prioritize captions that not only depict image content, but also summarise the corresponding article segments.

Early approaches to the task focus on learning the representation of a news article and its connection with an image, including utilizing n-gram language models for extracting phrases seen in the article (Feng and Lapata, 2013), or building an encoder-decoder based architecture with VGG (Simonyan and Zisserman, 2015) for encoding the image, Word2Vec (Mikolov et al., 2013) for encoding the article, and an LSTM as caption decoder (Ramisa et al., 2018). They all fail to achieve satisfactory performance. Biten et al. (2019) propose

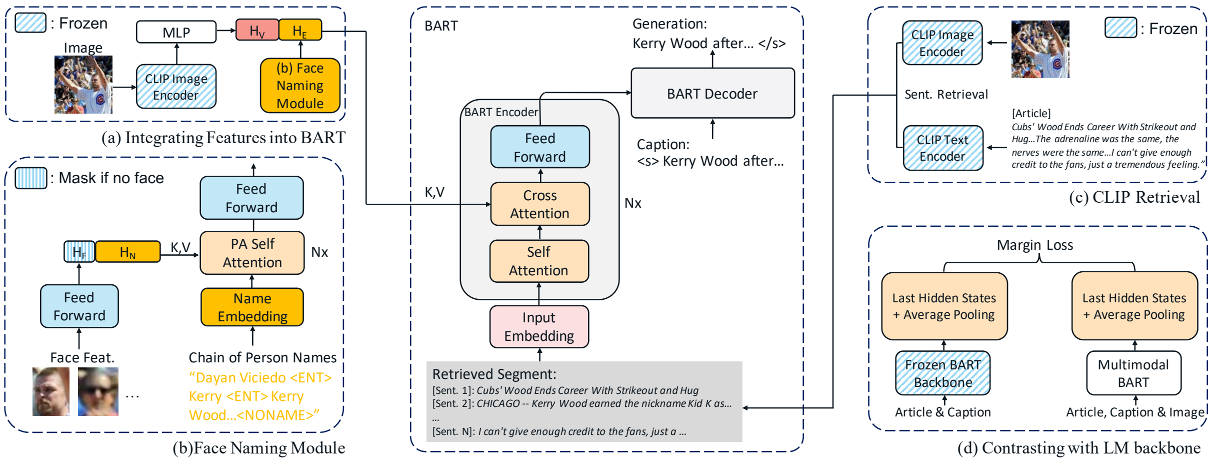

Figure 2: Method illustration. Our model is an encoder-decoder model built on BART (middle). Our method consists of: (a) Integrating Features into BART: In BART encoder, we concatenate visual (H_V) and name features (H_E) to obtain keys and values for the added cross-attention module; (b) Face Naming Module: We first get the embedding H_N of the chain of person names in the article. Then we prepend the face features H_F to H_N to obtain keys and values for the prefix-augmented self-attention module; (c) CLIP Retrieval: We conduct sentence retrieval using CLIP to learn from more accurate article context. (d) Contrasting with LM backbone (CoLaM): We contrast the multimodal BART with frozen pure-text BART to force the model to focus more on the article context.

the first large-scale dataset GoodNews for the task, and a two-stage template based captioning method. Following Biten et al. (2019), several works adopt transformer-based models with different types of features like Places 365 (Zhou et al., 2018) used by Yang and Okazaki (2020) or face and object features used by the Tell model (Tran et al., 2020).

Since then, the focus of the community has shifted to learning better entity representations. VisualNews (Liu et al., 2021) adopts a multi-head attention-on-attention module (Huang et al., 2019) and visual selective gates. JoGANIC (Yang et al., 2021) brings external knowledge from a Wikipedia database to train an entity embedding. On top of Tell, Zhou et al. (2022) show that an entity-aware retrieval method can improve the performance further. Zhang et al. (2022a) propose a prompt-based model NewsMEP with pre-trained BART (Lewis et al., 2020) and CLIP features as the backbone. NewsMEP follows ClipCap (Mokady et al., 2021) to generate visual prompts from CLIP representations. Instead of differentiating between different types of entities like our method, all entities are treated equally by NewsMEP, where a bi-LSTM is trained to learn the most important entities from the article as part of prompts to the decoder of NewsMEP. We design different modules for different types of textual context information based on their connection to the various types of visual inputs, which yields superior performance than NewsMEP. We present a detailed comparison in Section 4.3.

Finally, in contrast to our work, where no additional paired datasets are used, recent works (Zhang et al., 2022b; Rajakumar Kalarani et al., 2023) explore the use of extra large-scale datasets on the task and obtain satisfactory results.

3 Methodology

3.1 Model Architecture

We present our model in Figure 2, which is an encoder-decoder model built upon the generative pre-trained language model BART (Lewis et al., 2020). We add a cross-attention module to the BART encoder to integrate the visual and name features, while keeping the BART decoder unchanged. To obtain the name features, we use a prefix-augmented self-attention module (Zhao et al., 2022) to softly select person names in the article that are similar to the detected faces. We also design a retrieval strategy using CLIP to retrieve article sentences that capture crucial context that cannot be directly inferred from the images. The retrieved segments serve as input to our model. Section 3.5 presents our training pipeline.

3.2 Integrating Features into BART

We add a cross-attention module to the BART encoder to incorporate visual (H_V) and name features (H_E). We use a simple MLP network as suggested by Mokady et al. (2021) to get visual representations denoted as H_V from the frozen CLIP image encoder. As shown in Figure 2 (a), We concatenate H_V and H_E as $[H_V; H_E]$, which is linearly transformed to get the keys K_A and values V_A. Together with the linearly transformed query Q_A from the article hidden states, in each encoder layer, we compute the cross-attention as softmax$(Q_A {K_A}^T/\sqrt{d_H})V_A$, with $1/\sqrt{d_H}$ as the scaling factor.

3.3 Face Naming Module

As stated before, there is a strong face-name co-occurrence pattern in the News Image Captioning datasets. We design a novel face naming module to learn a face-aware representation of person names, as shown in Figure 2 (b). Given a chain of person names in the article (e.g. "Dayan Viciedo ⟨ENT⟩ Kerry ⟨ENT⟩ ..." with ⟨ENT⟩ being a special token used as separator), we first compute the embedding H_N. We generate face embedding H_F by passing the face features [3] through a feed-forward layer. Then we prepend H_F to H_N as $[H_F; H_N]$, which is linearly transformed to keys K_N and values V_N. Together with the linearly transformed queries Q_N from H_N, we compute the prefix-augmented self attention (PA self attention) as softmax$(Q_N {K_N}^T/\sqrt{d_H})V_N$, with $1/\sqrt{d_H}$ as the scaling factor. Finally we use a feed-forward layer to generate the name features H_E of fixed length from the attended name embeddings.

With H_F, we can control the utilization of the contextual information from the faces. For images with no faces, we mask H_F, resulting in a conventional self attention. If faces occur, the chain of person names receives contextual information from the faces through PA self attention. We detail the learning of this module in the Section 3.5.

3.4 CLIP Retrieval

Unlike for the names and faces, where a clear connection between text and image can be found, it is difficult to find a direct connection between text and image for context that cannot be directly visually grounded. For such context, readers tend to use image contents to retrieve relevant information from the article. In an effort to simulate this cognitive process, we use CLIP to retrieve sentences that are semantically closest to the image representation generated by the CLIP image encoder (measured by cosine similarity). To make sure we include enough global information, we also add the first three sentences of each article segment if they are not part of the retrieved sentences, and keep the original sentence ordering.

[3] Face features are provided in the datasets.

3.5 Learning

In this section we detail the learning process of our model. Our full model is trained with a face naming loss, a margin loss (CoLaM) and a caption generation loss.

Face naming loss Inspired by Qu et al. (2023), we adopt a symmetric contrastive loss to align faces in images to names in captions during training. Given $m-1$ names in a caption, we add an additional ⟨NONAME⟩ token. We denote the name embedding of each name as $H^j_{N,gt}$, $j = 1, 2, \ldots, m$. Since we only use the ground truth names during training for learning better face representations, we apply stop gradient to the name embedding layer while computing the loss. We denote the name embeddings with stop gradient as $\tilde{H}^j_{N,gt}$, $j = 1, 2, \ldots, m$. For the corresponding image with n faces, we extract hidden states of faces H_F from the last layer of the face naming module in our model. We denote the representation of each face as H^i_F for $i = 1, 2, \ldots, n$. For face set F and name set N, we adopt the face-to-name contrastive loss as:

$$\mathcal{L}_{f,n} = -\log \frac{e^{sim_d(F,N)}}{\sum_{F_k \in batch} e^{sim_d(F_k,N)}} \quad (1)$$

where $sim_d(F, N) = \frac{1}{n}\sum_{i=1}^{n} \max_j A_{i,j}$, with $A_{i,j} = (H^i_F)^T \cdot \tilde{H}^j_{N,gt}$, for $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$.

Similarly, we obtain the name-to-face contrastive loss as:

$$\mathcal{L}_{n,f} = -\log \frac{e^{sim_d(N,F)}}{\sum_{N_k \in batch} e^{sim_d(N_k,F)}} \quad (2)$$

where $sim_d(N, F) = \frac{1}{m}\sum_{i=1}^{m} \max_i A_{j,i}$, with $A_{j,i} = (\tilde{H}^j_{N,gt})^T \cdot H^i_F$, for $j = 1, 2, \ldots, m$, $i = 1, 2, \ldots, n$.

Combining Equation 1 and 2, we obtain a symmetric face naming loss as:

$$\mathcal{L}_{f\leftrightarrow n} = \mathcal{L}_{f,n} + \mathcal{L}_{n,f} \quad (3)$$

CoLaM The key idea of our Contrasting with Language Model (LM) backbone (CoLaM) is to guide the multimodal LMs to learn to focus more on the context from the news articles through a margin loss by utilizing the LM backbones.

Figure 2 (d) shows the simplified modeling process of our CoLaM. Specifically, let h_{lm} be a generative LM (e.g. BART (Lewis et al., 2020)) backbone, and h_{mm} be the generative multimodal LM built upon it. We extract the last hidden states C_{lm} and C_{mm} for the generated text from the decoders of h_{lm} and h_{mm}, respectively. We compute the margin loss as:

$$\mathcal{L}_m = \frac{1}{B}\sum_i \max\{0, \Delta - cos(pool(C^i_{lm}), pool(C^i_{mm}))\} \quad (4)$$

where B is the batch size, Δ is the margin hyperparameter, $cos(\cdot)$ denotes the cosine similarity, $pool(\cdot)$ is the average pooling operation which takes into account the masking.

By applying average pooling, we obtain the global representations from C_{lm} and C_{mm}, which is used to measure the cosine similarity between the two representations. As we freeze the text-only LM backbone, optimizing $\mathcal{L}_m$ is equivalent to adding a constraint to the multimodal LM. This constraint ensures the multimodal LM to put more emphasis on the news articles. As shown before, news image caption often contains more context from the article than from the image. Our CoLaM is a universal method for improving context modeling abilities of existing models, and can be seamlessly integrated into existing models. We further discuss the use of CoLaM in Section 4.3 and Appendix G.

Caption generation loss Given a news image and article pair, we minimize the negative log likelihood for caption generation as:

$$\mathcal{L}_{cap} = -\sum_{t=1}^{T} \log p(y_t|y_{<t};\theta) \quad (5)$$

where y_t denotes target caption token at time step t, $y_{<t}$ denotes the current token sequence and θ represents the learned parameters of the model.

Finally, we train our model with the loss as:

$$\mathcal{L} = \mathcal{L}_{cap} + \mathcal{L}_{f\leftrightarrow n} + \alpha\mathcal{L}_m \quad (6)$$

where α is the hyperparameter.

4 Experiments

4.1 Implementation Details

We conduct experiments on two large-scale News Image Captioning datasets, namely GoodNews(Biten et al., 2019) and NYTimes800k (Tran et al., 2020). The details of the datasets are presented in Appendix C. Following the same experimental settings as previous works (Tran et al., 2020; Yang et al., 2021), we train our full model for 16/9 epochs on GoodNews/NYTimes800k. We set the batch size to 32, the learning rate to 1e-5, and warm up for the first 5% steps. We adopt the AdamW optimizer (Loshchilov and Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e-8$, and apply weight decay of 0.01 to all weights as regularization. We clip the gradient norm at 0.1. Following Zhang et al. (2022a), a frozen CLIP-ViT-B/16 is used as image encoder. We set the length of the visual features and name features to be 20. We add two special tokens $\langle$ENT$\rangle$ and $\langle$NONAME$\rangle$, for separating names in chain of person names, and acting as a "NONAME" token as suggested by Qu et al. (2023), respectively. During training, we restrict the number of tokens in articles and captions to be 512 and 100, respectively. We set $\alpha = 2.0$ and $\Delta = 1.0$ for CoLaM. During inference, we use beam search with beam size of 5. We provide more implementation details in Appendix B.

4.2 Evaluation Metrics

We follow the same evaluation pipeline as in previous works (Biten et al., 2019; Tran et al., 2020; Yang et al., 2021; Zhao et al., 2021; Zhang et al., 2022a). To measure the overall quality of generated captions, we use BLEU-4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Denkowski and Lavie, 2014) and CIDEr scores (Vedantam et al., 2015). For BLEU-4 and ROUGE-L, every word contributes to the metric equally. METEOR focuses on synonym matching and lemmatization, which are seldomly found for named entities. CIDEr uses TF-IDF weighting to put more emphasis on rare words, e.g. named entities (Biten et al., 2019; Kilickaya et al., 2017; Elliott and Keller, 2014). So following previous works, we also consider CIDEr as the most suitable one for the task. We also use precision and recall to evaluate the quality of generated named entities.

4.3 Results

We report the overall performance and the main ablation studies in this part. We present more results on human evaluation, text summarization, additional ablation studies, in-depth study of CoLaM and qualitative analysis in Appendix D, E, F, G and H, respectively.

Overall Performance

	Method	Extra Data*	Caption Generation†				Named Entities†	
			B	M	R	C	P	R
GoodNews	Avg+CtxIns (Biten et al., 2019)	✗	0.89	4.37	12.20	13.10	8.23	6.06
	Tell (Tran et al., 2020)	✗	6.05	10.30	21.40	53.80	22.20	18.70
	VisualNews (Liu et al., 2021)	✗	6.10	8.30	21.60	55.40	22.90	19.30
	JoGANIC (Yang et al., 2021)	✓	6.83	11.25	23.05	61.22	26.87	22.05
	Tell + Focus! (Zhou et al., 2022)	•	6.30	⁄	23.00	60.30	24.20	20.90
	DiscExt CapGen (Zhang et al., 2022b)	✓	7.94	13.97	28.68	64.51	29.69	27.37
	Rajakumar Kalarani et al. (2023)	✓	7.14	11.21	24.30	72.33	24.37	20.09
	NewsMEP (Zhang et al., 2022a)	✗	<u>8.30</u>	<u>12.23</u>	<u>23.17</u>	63.99	23.43	<u>23.24</u>
	Ours$_{base}$ (w/ BART$_{base}$)	✗	7.20	11.00	21.97	<u>65.42</u>	<u>24.15</u>	22.18
	Ours$_{large}$, (w/ BART$_{large}$)	✗	**8.60**	**12.39**	**23.38**	**71.96**	**24.30**	**25.54**
NYTimes800k	Tell (Tran et al., 2020)	✗	6.30	10.30	21.70	54.40	24.60	22.20
	VisualNews (Liu et al., 2021)	✗	6.40	8.10	21.90	56.10	24.80	22.30
	JoGANIC (Yang et al., 2021)	✓	6.79	10.93	22.80	59.42	28.63	24.49
	Tell + Focus!(only CLIP) (Zhou et al., 2022)	•	6.40	⁄	⁄	57.50	25.70	22.70
	Tell + Focus! (Zhou et al., 2022)	•	7.00	⁄	22.90	63.60	29.80	25.90
	DiscExt CapGen (Zhang et al., 2022b)	✓	7.57	12.64	25.67	62.31	30.04	25.53
	Rajakumar Kalarani et al. (2023)	✓	7.54	11.27	23.28	66.41	28.21	23.25
	NewsMEP (Zhang et al., 2022a)	✗	**9.57**	**13.02**	**23.62**	<u>65.85</u>	26.61	<u>28.57</u>
	Ours$_{base}$ (w/ BART$_{base}$)	✗	7.87	11.19	21.95	64.64	**26.98**	25.33
	Ours$_{large}$ (w/ BART$_{large}$)	✗	<u>9.24</u>	<u>12.57</u>	<u>23.44</u>	**71.65**	<u>26.88</u>	**28.59**

Table 1: Performance comparison with state-of-the-art methods. We highlight the best scores and underline the second best scores of models that do not use (a) extra data, or (b) additional pre-trained models other than the language model&vision backbones. *: The use of extra data can be found in Appendix B. †: B: BLEU-4; R: ROUGE-L; M: METEOR; C: CIDEr; P: Precision; R: Recall. We adopt the same abbreviation in all tables.

We present the overall performance of our model in Table 1. With a much smaller language model (LM) BART$_{base}$ [4], we already achieve a competitive performance on both GoodNews and NYTimes800k datasets, when compared to the previous state-of-the-art (SOTA) model NewsMEP, which uses BART$_{large}$ as backbone. When we increase the LM size to BART$_{large}$, we establish a new SOTA in terms of CIDEr scores and outperform NewsMEP by a large margin (+6 points). Our model also yields new SOTA entity scores on both datasets leading to more trustworthy captions. Compared to our model, NewsMEP is constructed with the same vision backbone (CLIP-ViT-B/16) and LM (BART$_{large}$). NewsMEP also adopts a prefix-augmented attention module to integrate visual and entity information into the model. It learns to select entities through the interaction between image and article representations from the encoder in BART$_{large}$. However, NewsMEP fails to utilize the characteristics of different types of visual inputs, which should be treated differently as in our framework. By considering all visual inputs equally, NewsMEP lacks in linking rare words in articles or captions to visual inputs, resulting in much lower CIDEr scores when compared to our method. Unlike generic image captioning, where the goal is to make a simple descriptive caption to the image, News Image Captioning requires the generated captions to capture the essence of both images and articles. In that case, CIDEr as a evaluation metric which put more emphasis on rare words should be prioritized. Our method obtains the highest CIDEr scores, showing its efficacy.

Focus! (Zhou et al., 2022) is a sentence retrieval method. Combining Focus! with Tell, relatively high CIDEr scores can be attained. Although no extra data sources are used, Focus! uses CLIP and OpenNRE (Han et al., 2019) (a pre-trained domain specific relation extraction model) to perform sentence retrieval, on top of the LM&vision backbones from Tell. Without OpenNRE, the CIDEr score of Tell + Focus! on NYTimes800k drops from 63.60 to 57.50, indicating the biggest gain of their method comes from the use of OpenNRE. However, we are more interested in the question: Without additional domain specific pre-trained models, how can we explore the connections between images, articles and captions of the given dataset? Experimental results show the merit of our method, which is also demonstrated in the ablation study of the different components of our model (see below).

There are also three methods, namely JoGANIC (Yang et al., 2021) DiscExt CapGen (Zhang et al., 2022b) and Rajakumar Kalarani et al. (2023), that use extra data sources in their framework. With-

[4]NewsMEP: BART$_{large}$; Tell& JoGANIC: RoBERTa$_{large}$

out using extra data, our method significantly achieves higher CIDEr scores as compared to the former two methods, and yields comparable or better performance than Rajakumar Kalarani et al. (2023), which uses the extra News Image Captioning dataset VisualNews (Liu et al., 2021) containing more than 1.2 million samples. Our method generates better captions by exploring the News Image Captioning datasets in a better way.

Ablation Study on Model Components

We present results for the ablation study on the different components of our model in Table 2[5].

	Model	VF	NF	RS	CoLaM	Caption Generation				Named Entities	
						B	M	R	C	P	R
GoodNews	⟨1⟩					6.14	9.69	19.51	55.24	21.17	18.81
	⟨2⟩	✓				6.59	10.17	20.33	58.55	22.20	20.46
	⟨3⟩	✓	✓			6.81	10.54	21.17	61.73	22.86	21.02
	⟨4⟩	✓	✓	✓		7.00	10.75	21.79	64.07	23.69	21.50
	⟨5⟩	✓	✓	✓	✓	**7.20**	**11.00**	**21.97**	**65.42**	**24.15**	**22.18**
NYTimes800k	⟨1⟩					6.63	9.89	19.14	51.76	22.30	21.63
	⟨2⟩	✓				6.75	10.12	19.74	54.45	24.04	22.47
	⟨3⟩	✓	✓			7.18	10.63	20.81	59.07	25.70	23.78
	⟨4⟩	✓	✓	✓		7.53	10.98	21.63	63.95	26.94	24.72
	⟨5⟩	✓	✓	✓	✓	**7.87**	**11.19**	**21.95**	**64.64**	**26.98**	**25.33**

Table 2: Effects of different components of our model on qualities of generated captions. Model ⟨1⟩: $\text{BART}_{\text{base}}$; VF: visual features; NF: name features from face naming; RS: retrieved segments. Model ⟨5⟩: $\text{Ours}_{\text{base}}$.

Visual Features When we discard image inputs, the task becomes a purely textual sequence-to-sequence problem. $\text{BART}_{\text{base}}$ (Model ⟨1⟩) can achieve fairly good results in this scenario on two datasets. However, it always generates the same caption for different images of an article. The addition of the visual features in Model ⟨2⟩ mitigates the problem. We observe consistent improvements in all evaluation metrics as shown in Table 2.

Face Naming Module On top of Model ⟨2⟩, when we add the name features learned from our face naming module (Model ⟨3⟩), we observe significant improvement on all the evaluation metrics from both datasets, especially regarding the CIDEr score (58.55→61.73 on GoodNews, 54.45→59.07 on NYTimes800k). When both the visual features and name features are added to $\text{BART}_{\text{base}}$, we already achieve CIDEr scores higher than some models based on $\text{RoBERTa}_{\text{large}}$ (e.g., Tell with 53.80/54.40 CIDEr on GoodNews/NYTimes800k), or comparably to models that use extra external data (e.g., JoGANIC with 61.22/59.42 CIDEr on GoodNews/NYTimes800k).

CLIP Retrieval We further improve the quality of the generated captions by retrieving sentences from the articles. In this way the model learns to focus on different segments for captioning different images. As shown in Table 2, we improve the CIDEr scores from 61.73 to 64.07 on GoodNews, 59.07 to 63.95 on NYTimes800k. Apart from the improvements in caption generation evaluation metrics, we also improve the precision of all entity names generated in the captions on two datasets after adding the retrieval component into our method.

CoLaM Finally, with the addition of CoLaM, the base version of our full model (Model ⟨5⟩) further improvements the performance of Model ⟨4⟩ on all metrics. It shows that the imbalanced proportion of context from articles and images in the captions can be a big problem for News Image Captioning models. We present more in-depth analyses of the behavior of CoLaM in Appendix G, together with the results of CoLaM with other model architectures to show its potential of being the universal add-on for News Image Captioning models.

Ablation Study for Entity Generation

The different components of our model also affect the generation of different types of entity names. We present the precision and recall scores of the three most commonly occurring entity types[6] in Table 3. We have designed the face naming module to force the model to focus on the correct PERSON-type entities (names) which can be visually grounded from the images. As shown in Table 3, by adding the name features, we observe a significant improvement in both precision (28.00 → 29.22 on GoodNews, 32.86 → 37.11 on NYTimes800k) and recall (24.13 → 25.91 on GoodNews, 29.41 → 33.17 on NYTimes800k) of PERSON-type entity names, which shows the effectiveness of our face naming module. Interestingly, we also observe improvements in entity scores for other types of entities after adding the name features, for instance the recall of GPE (27.69 → 28.34 on NYTimes800k). We think this is due to the large improvement in predicting PERSON-type entity names, which leads to more accurate context modeling in the articles, and to generating captions of higher quality. We also observe improvements in entity scores after adding the retrieval module, which helps the model learn better context that cannot be directly seen from the images (e.g. in some cases GPE and ORG are not clearly present in the images). And as expected, adding CoLaM to our training pipeline learns bet-

[5]Limited to resources, we conduct the ablation studies using $\text{BART}_{\text{base}}$ as the backbone LM.

[6]PERSON: people; GPE: countries, cites, states; ORG: companies, agencies, etc.

ter article context, which leads to improvements in entity scores.

	Model	VF	NF	RS	CoLaM	PERSON		GPE		ORG	
						P	R	P	R	P	R
GoodNews	$\langle 1 \rangle$					26.58	21.99	22.80	22.34	17.84	16.42
	$\langle 2 \rangle$	✓				28.00	24.13	24.17	24.37	19.97	19.47
	$\langle 3 \rangle$	✓	✓			29.22	25.91	24.62	24.38	21.31	20.24
	$\langle 4 \rangle$	✓	✓	✓		30.33	26.32	25.33	25.17	22.10	21.11
	$\langle 5 \rangle$	✓	✓	✓	✓	**31.00**	**27.42**	**25.99**	**25.62**	**22.21**	**21.51**
NYTimes800k	$\langle 1 \rangle$					29.65	29.47	25.84	25.77	18.38	17.75
	$\langle 2 \rangle$	✓				32.86	29.41	27.02	27.69	20.09	18.98
	$\langle 3 \rangle$	✓	✓			37.11	33.17	27.67	28.34	20.98	19.29
	$\langle 4 \rangle$	✓	✓	✓		38.53	34.22	28.28	29.00	22.66	20.42
	$\langle 5 \rangle$	✓	✓	✓	✓	**38.59**	**35.43**	**28.44**	**29.03**	**23.02**	**20.86**

Table 3: Effects of different modules on named entities. Same abbreviation applies as in Table 2.

Evaluation on Different Subsets of the Test Data Because the design of our face naming module is guided by the face-name co-occurrence patterns found in News Image Captioning dataset, we split the dataset into three mutually exclusive subsets[7] to explore the effectiveness of our face naming module: 1. F✗,N✗subset with no faces in images and no names in the captions; 2. F✗,N✓subset with no faces in the images, but has names in the captions and 3. F✓,N✓subset with faces in the images and names in the captions. We show the results of our model, with or without face naming, on the NYTimes800k dataset in Table 4.

Model	Subset	PERSON		GPE		ORG	
		P	R	P	R	P	R
$\langle 2 \rangle$	F✗,N✗	⁄	⁄	**28.71**	28.75	21.56	19.84
$\langle 3 \rangle$	F✗,N✗	⁄	⁄	28.66	**29.54**	**23.43**	**20.78**
$\langle 2 \rangle$	F✗,N✓	30.11	**18.94**	22.05	23.17	**15.03**	**16.16**
$\langle 3 \rangle$	F✗,N✓	**35.43**	16.65	**24.05**	**23.46**	14.30	14.99
$\langle 2 \rangle$	F✓,N✓	**42.34**	31.24	26.36	27.53	19.96	18.78
$\langle 3 \rangle$	F✓,N✓	41.77	**36.06**	**27.42**	**28.11**	**20.37**	**18.83**

Model	Subset	Caption Generation				Named Entites	
		B	M	R	C	P	R
$\langle 2 \rangle$	F✗,N✗	5.44	8.93	17.11	41.16	17.99	20.57
$\langle 3 \rangle$	F✗,N✗	**5.45**	**9.04**	**17.54**	**43.11**	**19.91**	**20.92**
$\langle 2 \rangle$	F✗,N✓	**5.64**	**8.75**	**17.74**	**47.03**	20.91	**17.11**
$\langle 3 \rangle$	F✗,N✓	4.84	8.14	16.39	41.04	**22.03**	16.33
$\langle 2 \rangle$	F✓,N✓	7.72	11.18	22.11	64.95	28.44	24.39
$\langle 3 \rangle$	F✓,N✓	**8.68**	**12.15**	**24.16**	**73.75**	**29.46**	**26.55**

Table 4: Effects of the face naming module on different subsets of NYTimes800k (e.g. F✗,N✗: subset with no faces in images and no names in captions. See text for details.). Model $\langle 2 \rangle$: $\text{BART}_{\text{base}}$ + visual features; Model $\langle 3 \rangle$: $\langle 2 \rangle$ + face naming.

Because the design of the prefix-augmented self attention in our face naming module provides a strong signal of face-name co-occurrence, we expect that our model with face naming (Model $\langle 3 \rangle$) would perform much better on PERSON-type entities than the model with only the visual features (Model $\langle 2 \rangle$) on F✓,N✓subset, and much better overall on both F✗,N✗and F✓,N✓subsets. As shown in Table 4, on F✓,N✓subset, we increase the recall from 31.24 to 36.06, while maintaining the same level of precision after adding the face naming module into Model $\langle 2 \rangle$. And the CIDEr scores of the generated captions from Model $\langle 3 \rangle$ are significantly higher than the counterpart. On top of that, the correctness of the generated entities overall improves on both F✗,N✗and F✓,N✓subsets.

The F✗,N✓subset contains samples without face-name co-occurrence pattern as we modeled in our face naming module. It covers around 11% data in each dataset, as shown in Table 2 in the appendix. It can be seen from Table 4, the trade-off of significantly increasing the model performance on F✗,N✗and F✓,N✓subsets is that the model would generate worse captions on the F✗,N✓subset. Model $\langle 3 \rangle$ performs worse on most of the evaluation metrics than Model $\langle 2 \rangle$. However, one interesting finding is that Model $\langle 3 \rangle$ achieves a much higher precision while reaching a lower recall with PERSON-type entity names on F✗,N✓subset. Meanwhile, an opposite trend can be found on F✓,N✓subset. With the face naming module, our model tends to generate less PERSON-type entity names when there is no face, and to generate more PERSON-type entity names otherwise, showing its effectiveness. The same ablation study on GoodNews dataset is presented in Appendix F.

5 Conclusion

In this paper, we introduce a new framework for utilizing visual inputs in News Image Captioning. Inspired by human attention mechanisms, we developed a face naming module for aligning names with faces in images, based on face-name co-occurrence patterns. For context that cannot be visually grounded in the images, we utilize CLIP for sentence retrieval from articles, aiding comprehension. To address the imbalance between article and image context in captions, we introduce CoLaM, guiding the model to focus more on article content. Our extensive experiments demonstrate the effectiveness of our method, which achieves more than 6-point improvement in CIDEr scores over the previous state-of-the-art on two commonly-used News Image Captioning datasets.

[7]The distribution of samples on the full datasets can be seen in Appendix C. There are no samples without names in the caption that have faces in the corresponding image.

6 Limitations

Our face naming module effectively aligns faces in images to names in articles/captions, which can be directly visually grounded from the images and trigger higher human attention priority. However, for contexts like time or organizations that typically cannot be directly visually grounded, we depend on CLIP retrieval to infer links between articles and images. A potential improvement involves designing specific modules for these types of contexts. Additionally, our CoLaM approach currently treats all image-caption-article triplets equally, applying the same constraints during training. A valuable area for future research would be to investigate a weighting mechanism that selectively adjusts the margin loss computation for these triplets.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ali Furkan Biten, Lluis Gomez, Marcal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 452–457, Baltimore, Maryland. Association for Computational Linguistics.

Yansong Feng and Mirella Lapata. 2013. Automatic caption generation for news images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):797–812.

M. Ida Gobbini and James V. Haxby. 2007. Neural systems for recognition of familiar faces. *Neuropsychologia*, 45(1):32–41. The Perception of Emotion and Social Cues in Faces.

Xu Han, Tianyu Gao, Yuan Yao, Deming Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. 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): System Demonstrations*, pages 169–174, Hong Kong, China. Association for Computational Linguistics.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. 2017. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

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

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

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. 2021. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. 2023. Weakly supervised face naming with symmetry-enhanced contrastive loss. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3505–3514.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Abisek Rajakumar Kalarani, Pushpak Bhattacharyya, Niyati Chhaya, and Sumit Shekhar. 2023. "let's not quote out of context": Unified vision-language pretraining for context assisted image captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 695–706, Toronto, Canada. Association for Computational Linguistics.

A. Ramisa, F. Yan, F. Moreno-Noguer, and K. Mikolajczyk. 2018. Breakingnews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 40(05):1072–1085.

Tony Ro, Charlotte Russell, and Nilli Lavie. 2001. Changing faces: A detection advantage in the flicker paradigm. *Psychological Science*, 12(1):94–99. PMID: 11294237.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *ICLR*.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France. PMLR.

Xuewen Yang, Svebor Karaman, Joel Tetreault, and Alejandro Jaimes. 2021. Journalistic guidelines aware news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5162–5175, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhishen Yang and Naoaki Okazaki. 2020. Image caption generation for news articles. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1941–1951, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jingjing Zhang, Shancheng Fang, Zhendong Mao, Zhiwei Zhang, and Yongdong Zhang. 2022a. Fine-tuning with multi-modal entity prompts for news image captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4365–4373, New York, NY, USA. Association for Computing Machinery.

Junzhe Zhang and Xiaojun Wan. 2023. Exploring the impact of vision features in news image captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12923–12936, Toronto, Canada. Association for Computational Linguistics.

Zhengkun Zhang, Huibin Zhang, Jun Wang, Zhe Sun, and Zhenglu Yang. 2022b. Generating news image captions with semantic discourse extraction and contrastive style-coherent learning. *Computers and Electrical Engineering*, 104:108429.

Lulu Zhao, Fujia Zheng, Weihao Zeng, Keqing He, Weiran Xu, Huixing Jiang, Wei Wu, and Yanan Wu. 2022. Domain-oriented prefix-tuning: Towards efficient and generalizable fine-tuning for zero-shot dialogue summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 4848–4862, Seattle, United States. Association for Computational Linguistics.

Wentian Zhao, Yao Hu, Heda Wang, Xinxiao Wu, and Jiebo Luo. 2021. Boosting entity-aware image captioning with multi-modal knowledge graph. *CoRR*, abs/2107.11970.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464.

Mingyang Zhou, Grace Luo, Anna Rohrbach, and Zhou Yu. 2022. Focus! relevant and sufficient context selection for news image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6078–6088, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Overview of the Appendix

In the appendix, we first provide more implementation details and dataset statistics in Appendix B and C, respectively. This is followed by human evaluation of the generated captions form our method in Appendix D. Following that, to showcase the connection and difference between News Image Captioning and Text Summarization, we conduct additional experiments on text summarization as presented in Appendix E. Further, we provide more ablation studies on different subsets of the GoodNews test data in Apendix F. Then we provide a in-depth study of our CoLaM in Appendix G. We conclude the appendix with a qualitative analysis on the generated captions in Appendix H.

B More Implementation Details

Extra Data regard the external data and include: JoGANIC (Yang et al., 2021): Wikipedia database; DiscExt CapGen (Zhang et al., 2022b): 2.755 million caption-style pairs. Rajakumar Kalarani et al. (2023): 1.2 million paired News Image Captioning data from VisualNews (Liu et al., 2021). •: Apart from the language model&vision backbones, Focus! (Zhou et al., 2022) uses CLIP and domain specific relation extraction model OpenNRE (Han et al., 2019) for context retrieval.

The MLP network in visual feature generation module consists of two linear layers with hyperbolic tangent activation in between. The two linear layers are of shape Linear(dim_{model}, $dim_{model} \times 10$) and Linear($dim_{model} \times 10$, $dim_{model} \times 20$), where dim_{model} is the model dimension of the BART backbone. We reshape the mapped visual feature from $(batch_{size}, dim_{model} \times 20)$ to $(batch_{size}, 20, dim_{model})$.

To obtain the name embedding in our face naming module, the same embedding layer structure as in BART is adopted. Given a chain like ”name1 ⟨ENT⟩ name2 . . .”, with ⟨ENT⟩ being the added token in vocabulary,we compute the word embedding H_N of the chain.

We use the transformers package to build our models. For $\text{BART}_{\text{base}}$, we adopt the "facebook/bart-base" checkpoint; while for $\text{BART}_{\text{large}}$, we adopt the "patrickvonplaten/bart-large-fp32" checkpoint. The default vocabulary size is 50265. The training of $\text{Ours}_{\text{base}}$ and $\text{Ours}_{\text{large}}$ takes roughly 1 and 2 days on 1×A100, respectively. For the GoodNews dataset, the full articles are used. While for NYTimes800k, we follow the standard protocol (Tran et al., 2020) to use the 512 tokens surrounding the images. For our full model, we apply length penalty of 2 during decoding. Following Tran et al. (2020), we adopt pycocoevalcap package and spacy package (ver. 2.1.9) for evaluating generated captions and entity scores, respectively.

C Dataset Statistics

In this section, we provide dataset statistics of GoodNews (Biten et al., 2019) and NYTimes800k (Tran et al., 2020). The overall statistics of two datasets are provided in Table 5.

	GoodNews	NYTimes800k
Number of images	462642	792971
Average article length	451	974
Average caption length	18	18
% of captions with named entities	97%	96%
% of captions with person names	68%	68%
% of images with faces	56%	57%

Table 5: Dataset statistics for GoodNews and NYTimes800k

Table 6 presents the statistics of face-name co-occurrence patterns in two datasets.

Dataset	F ✓, N ✓	F ✗, N ✗	F ✓, N ✗	F ✗, N ✓
GoodNews	56.30%	31.91%	0%	11.79%
NYTimes800k	56.91%	32.05%	0%	11.04%

Table 6: Statistics of face-name co-occurrence patterns in two News Image Captioning datasets.[8]"F ✓, N ✗" refers to samples with faces in images, but no names in captions.

D Human Evaluation

We present human evaluation in Table 7. We hire three graduate students with domain knowledge to rank 50 randomly sampled captions on correctness and fluency from 1 to 5, and pick their preferred caption. Captions generated by our method better align with human judgement (C=3.83), and are preferred by humans in **67%** of the cases.

Model	Correctness(C)	Fluency(F)	Preferred by
Baseline	$3.15_{\pm 0.13}$	$4.67_{\pm 0.25}$	$16\%_{\pm 4\%}$
Ours	$\mathbf{3.83}_{\pm 0.19}$	$\mathbf{4.86}_{\pm 0.10}$	$\mathbf{67\%}_{\pm 5\%}$

Table 7: Human evaluation for generated captions.

E Additional Experiments on Text Summarization

Since our work is also closely related to text summarization, in this section, we present more experiments on text summarization with BART.

E.1 Experiments with frozen BART

By keeping the BART backbone frozen during training, we can have a better idea of whether the modules we designed can guide the caption generation effectively. We present the results in Table 8. The frozen $\text{BART}_{\text{base}}$ works poorly by only achieving 6.56 CIDEr score. While after adding our modules, even without training the BART backbone, we achieve fairly good results with CIDEr=56.66. This significant improvement in performance shows that our added modules can effectively guide the generation process of BART.

Model	BLEU-4	METEOR	ROUGE-L	CIDEr
Frozen $\text{BART}_{\text{base}}$ (pure text)	1.68	8.33	11.43	6.56
Ours + frozen $\text{BART}_{\text{base}}$	5.85	9.93	20.86	56.66

Table 8: Performance comparison with frozen BART on GoodNews dataset

E.2 Experiments with summarization with retrieval

Our retrieval method aims to locate the sentences that are semantically closer to the images. Without adding any visual information into the model, we perform text summarization on the retrieved segments only as shown in Table 9. The large improvements in performance (CIDEr=55.24 → 59.21) by changing the inputs from the full articles to the retrieved segments prove that our retrieval component can locate more accurate semantic information from the articles.

[8]Calculation based on face features provided in the datasets.

Model	BLEU-4	METEOR	ROUGE-L	CIDEr
$\text{BART}_{\text{base}}$	6.14	9.69	19.51	55.24
$\text{BART}_{\text{base}}$ + our retrieval	6.43	10.03	20.50	59.21

Table 9: Text summarization with $\text{BART}_{\text{base}}$ on GoodNews with our retrieved article segments.

F Additional Ablation Studies

Ablation Study on Number of Retrieved Sentences

We evaluate the impact of the number of retrieved sentences, as outlined in Table 10. Here we do not apply CoLaM to show a clear image of how the number of retrieved sentences can affect the performance of our model. Our results indicate consistent performance across the range of 7-10 retrieved sentences on both GoodNews and NYTimes800k datasets. It's worth noting that while the top CIDEr score doesn't consistently align with the highest achievements in other evaluation metrics, such as BLEU-4 (7.02) and METEOR (10.77) which are attained with retrieving 9 sentences in the case of GoodNews, models with the highest CIDEr score generally maintain strong performance in other metrics.

	# of sent	Caption Generation				Named Entites	
		B	M	R	C	P	R
GoodNews	7	6.92	10.74	21.68	62.62	23.39	21.27
	8	7.00	10.75	**21.79**	**64.07**	**23.69**	**21.50**
	9	**7.02**	**10.77**	21.61	62.64	23.19	21.30
	10	6.83	10.66	21.64	63.43	23.49	21.21
NYTimes800k	7	7.38	10.90	21.61	63.13	26.62	24.39
	8	7.43	10.92	21.52	62.55	26.85	24.28
	9	7.50	**10.98**	21.53	62.84	26.58	24.60
	10	**7.53**	**10.98**	**21.63**	**63.95**	**26.94**	**24.72**

Table 10: Influence of the number of retrieved sentences

Ablation on Different Subsets of the Test Data (GoodNews)

We present the ablation study on different subsets of the GoodNews test data in Table 11. For F✗,N✗and F✓,N✓subsets, we observe similar trend in performance improvements when adding entity prefix into the model as the case for the subsets of GoodNews test data. The biggest improvements in entity scores can be seen in recall for PERSON type entities on F✓,N✓subset when we add entity prefix into the model (25.15→27.77). And the quality of the generated captions is drastically enhanced on F✓,N✓subset, as demonstrated by an approximate 7.5-percentage-point improvements (from 68.11 to 73.24).

Interestingly, on F✗,N✓subset of GoodNews test data, we observe uniformly decreasing in all metrics when adding name features to the model. While on on F✗,N✓subset of NYTimes800k test data, we observe a small improvement in entity precision. Moreover, a slightly different pattern in entity scores can be observed on the F✗,N✓subset from two datasets. It shows that for subsets without the face-name co-occurrence pattern we modeled, the performance of our model is somewhat dependent to the data distribution. Notably, the F✗,N✓subset constitutes approximately 11% of the entire dataset. Consequently, the substantial performance improvements observed in the remaining 89% of the data contribute to generating superior captions in the aggregate.

Model	Subset	PERSON		GPE		ORG	
		P	R	P	R	P	R
(2)	F✗,N✗	/	/	25.85	26.67	20.58	18.91
(3)	F✗,N✗	/	/	**26.43**	**27.51**	**21.95**	**19.39**
(2)	F✗,N✓	26.90	**18.03**	18.18	**18.99**	14.11	14.93
(3)	F✗,N✓	**29.83**	14.88	**18.52**	18.22	**16.02**	**16.28**
(2)	F✓,N✓	**33.54**	25.15	24.03	**23.60**	20.71	20.63
(3)	F✓,N✓	31.50	**27.77**	**24.31**	23.10	**21.91**	**21.48**

Model	Subset	Caption Generation				Named Entites	
		B	M	R	C	P	R
⟨2⟩	F✗,N✗	5.48	9.10	17.70	44.30	17.96	19.92
⟨3⟩	F✗,N✗	**5.58**	**9.34**	**18.42**	**46.66**	**19.69**	**20.05**
⟨2⟩	F✗,N✓	**5.48**	**9.00**	**18.18**	**48.02**	**19.08**	**15.89**
⟨3⟩	F✗,N✓	4.81	8.48	17.47	44.28	18.96	14.61
⟨2⟩	F✓,N✓	7.35	10.96	22.26	68.11	24.79	21.54
⟨3⟩	F✓,N✓	**7.82**	**11.58**	**23.49**	**73.24**	**24.82**	**22.62**

Table 11: Effectiveness of face naming module with $\text{Ours}_{\text{base}}$ on different subsets of GoodNews. Model ⟨2⟩: $\text{BART}_{\text{base}}$ + visual features; Model ⟨3⟩: ⟨2⟩ + name features from face naming module. F✗,N✗: subset with no faces in images and no names in captions.

Ablation on Lead3 Sentences and Varying Feature Length

We present the ablation study on Lead3 sentences in CLIP retrieval and varying feature length in Table 12. As expected, removing Lead3 sentences from the retrieved segments harms the performance of our method, due to lack of global context from the articles. And different feature lengths yield similar performance, while feature length=20 achieves the best CIDEr score.

Additional Ablation on Face Naming Module

Table 13 shows results on replacing the face naming module with feature concatenation (w/ retrieved segments (RS)). As expected, by replacing our face naming module with simple concatenation features, we observe significant degradation in performance.

	Lead3	Length	BLEU-4	METEOR	ROUGE-L	CIDEr
GoodNews	✗	20	6.80	10.67	21.43	61.37
	✓(Ours)	20	7.00	10.75	21.79	64.07
	✓	16	6.90	10.70	21.50	62.04
	✓	20(Ours)	7.00	10.75	21.79	64.07
	✓	24	6.88	10.65	21.45	61.86

Table 12: Performance comparison w/ or w/o Lead3 & w/ varying feature length using $\text{Ours}_{\text{base}}$ w/o CoLaM

Dataset	RS	Feature Integration	BLEU-4	METEOR	ROUGE-L	CIDEr
GoodNews	✓	concatenation	6.76	10.46	21.02	60.87
GoodNews	✓	face naming module	**7.00**	**10.75**	**21.79**	**64.07**

Table 13: Ablation study on face naming module and feature concatenation ($\text{Ours}_{\text{base}}$ w/o CoLaM).

G In-depth Study of CoLaM

We conduct ablation studies of CoLaM using $\text{BART}_{\text{base}}$ as the LM backbone. Limited to resources, the batch size is set to 24 for all ablation studies in this section.

Impact of the Margin Values Δ

We present the results with varying values for the margin parameter Δ in Table 14. Since the range of cosine similarity is within $[-1, 1]$, with $\Delta = 1.0$, the optimization of CoLaM affects all samples in the datasets. Our model mainly promote the visual inputs during generation, which makes the consistently added constraint from our CoLaM more favorable. As expected, the model reaches the best performance when we set $\Delta = 1.0$.

	Δ	α	Caption Generation				Named Entities	
			B	M	R	C	P	R
GoodNews	✗	✗	<u>6.93</u>	<u>10.75</u>	<u>21.69</u>	<u>62.94</u>	<u>23.41</u>	<u>21.46</u>
	0.4	1.0	6.95	10.75	21.74	63.87	23.38	21.45
	0.6	1.0	7.00	10.81	21.73	63.48	23.25	21.39
	0.8	1.0	7.14	10.90	21.75	63.65	23.19	21.75
	1.0	1.0	**7.19**	**10.94**	**21.96**	**65.06**	**23.78**	**21.81**
NYTimes800k	✗	✗	<u>7.63</u>	<u>11.00</u>	<u>21.40</u>	<u>62.03</u>	<u>25.44</u>	<u>23.74</u>
	0.4	1.0	7.59	11.00	21.35	62.76	26.14	24.48
	0.6	1.0	7.66	11.02	21.52	63.09	26.46	24.57
	0.8	1.0	7.52	10.96	21.48	62.90	**26.55**	24.62
	1.0	1.0	**7.73**	**11.14**	**21.66**	**63.44**	26.40	**24.84**

Table 14: Impact of the choice of the margin (Δ) on the performance. ✗: model trained without CoLaM.

Impact of the Loss Weights α

Table 15 shows the impact of the weight α for $\mathcal{L}_m$ in CoLaM. We obtain similar results with different values of α, showing that CoLaM is less sensitive to the weights. Setting α to 1.0 or 2.0 yields similar performance. In practice, we suggest to select $\alpha = 1.0$ to avoid unnecessary hyperparameter tuning.

	α	Δ	Caption Generation B	M	R	C	Named Entities P	R
GoodNews	✗	✗	6.93	10.75	21.69	62.94	23.41	21.46
	0.5	1.0	7.15	10.90	21.87	64.54	23.51	21.79
	1.0	1.0	7.19	10.94	21.96	**65.06**	**23.78**	21.81
	2.0	1.0	**7.27**	**11.02**	**21.97**	64.53	23.60	**22.13**
NYTimes800k	✗	✗	7.63	11.00	21.40	62.03	26.26	24.46
	0.5	1.0	**7.79**	**11.15**	21.65	63.15	26.56	24.86
	1.0	1.0	7.73	11.14	21.66	63.44	26.40	24.84
	2.0	1.0	7.73	11.14	**21.72**	**63.78**	**26.61**	**24.96**

Table 15: Impact of weights α for our $\mathcal{L}_m$. ✗: model trained without CoLaM.

Using Encoder or Decoder Hidden States

Since we learn the multimodal interaction only in the encoder, comparing the performance using the last encoder hidden states and the last decoder hidden states provides insights into whether the additional information from the caption influences CoLaM. As shown in Table 16, we obtain similar results with two types of hidden states, indicating the extra information from the caption does not have a significant impact on the functioning of CoLaM in the training pipeline.

Dataset	Hidden States	α	Δ	Caption Generation B	M	R	C	Named Entities P	R
GoodNews	Encoder	1.0	1.0	7.16	10.93	22.05	65.24	23.78	21.81
GoodNews	Decoder	1.0	1.0	7.19	10.94	21.96	65.06	23.78	21.81
NYTimes800k	Encoder	1.0	1.0	7.65	11.04	21.44	62.81	26.43	24.58
NYTimes800k	Decoder	1.0	1.0	7.73	11.14	21.66	63.44	26.40	24.84

Table 16: Impact of using the last hidden states from the encoder or the decoder.

Generalization Ability of CoLaM

CoLaM presents extraordinary generalization ability to other model architectures. We implemented the prefix-based method as proposed by Zhang et al. (2022a), which is the baseline version of the previous SOTA NewsMEP. NewsMEP utilizes visual and entity prefixes to guide the learning of the multimodal language model. Following Zhang et al. (2022a), we use the visual prefix to guide the language model, and term this model as NewsMEP'$_{\text{base}}$ for clarity.

Method	Dataset	Caption Generation B	M	R	C	Named Entities P	R
NewsMEP'$_{\text{base}}$	GoodNews	6.45	9.99	20.15	57.78	22.45	20.47
NewsMEP'$_{\text{base}}$ + CoLaM	GoodNews	**6.73**	**10.42**	**20.86**	**60.63**	**22.89**	**21.05**
NewsMEP'$_{\text{base}}$	NYTimes800k	6.56	9.95	19.50	53.36	23.79	22.09
NewsMEP'$_{\text{base}}$ + CoLaM	NYTimes800k	**7.20**	**10.48**	**20.54**	**57.31**	**24.86**	**23.72**

Table 17: Results for integrating CoLaM into the training pipeline of NewsMEP'$_{\text{base}}$

Just as the experiments with our model, without changing any training or architectural designs of NewsMEP'$_{\text{base}}$ and by simply adding our CoLaM to its training pipeline, we obtain significant performance gain over the original NewsMEP'$_{\text{base}}$. It shows that our CoLaM can be a valuable addition to the field, and possibly the standard method in any other News Image Captioning model's training pipeline in the future.

H Qualitative Analysis

H.1 Qualitative examples without CoLaM

Table 18 shows two examples of generated captions. In the first one, our proposed model generates a caption that matches the ground truth with the exception of a missing quote. In the second one, models with face naming module manage to capture important context from the news article, while after adding the sentence retrieval component, more precise context is generated (e.g. Mayor Nan Whaley of Dayton), showing the merit of our method.

We also provide the example generic image captions for news images in Table 18, Figures 3 and 4.

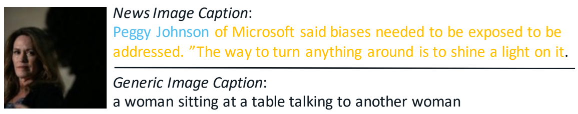

Figure 3: Comparison between two types of image captions for image in Table 18 (1)

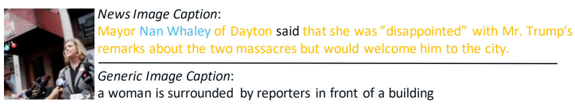

Figure 4: Comparison between two types of image captions for image in Table 18 (2)

H.2 Qualitative examples with the addition of CoLaM

We present multiple examples of captions generated using Ours$_{\text{base}}$ with or without CoLaM in Figure 5. With the addition of CoLaM, our model is able to capture the correct context from the article, which leads to improved captioning performance.

Image	Article	Caption
	Exposing the Bias Embedded in Tech... ...said Peggy Johnson, executive vice president of business development at Microsoft... Ms. Johnson told an audience at the New Rules Summit, ... Meredith Whittaker, a founder and a director of the AI Now Institute at New York University...said Ms. Whittaker, who also works at Google. "It's almost never white men... historical marginalization."... *(Johnson occurs 4 times, Whittaker occurs 7 times in total.)*	**I GT:** Peggy Johnson of Microsoft said biases needed to be exposed to be addressed. "The way to turn anything around is to shine a light on it." **II (1) $\text{BART}_{\text{base}}$:** "It's almost never white men who are discriminated against by these systems," said Meredith Whittaker, a founder and a director of the AI Now Institute at New York University. **III (2): (1) + visual features:** "It's almost never white men who are discriminated against by these systems," said Meredith Whittaker, a founder and a director of the AI Now Institute at New York University. **IV (3): (2) + face naming:** Peggy Johnson, executive vice president of business development at Microsoft, spoke at the New Rules Summit. **V (4): (3) + retrieval:** Peggy Johnson, executive vice president of business development at Microsoft.
	President Plans Visits to Places... Dayton, Ohio, and El Paso on Wednesday... The Democratic mayor of Dayton, Nan Whaley, said on Tuesday that she had been "disappointed" with Mr. Trump's remarks the day before about the two massacres, which left a combined 31 people dead... Representative Veronica Escobar, and her predecessor,... But like Ms. Whaley in Dayton, El Paso's mayor, Dee Margo, a Republican ...	**I GT:** Mayor Nan Whaley of Dayton said that she was "disappointed" with Mr. Trump's remarks about the two massacres but would welcome him to the city. **II (1) $\text{BART}_{\text{base}}$:** President Trump spoke at the National Rifle Association convention in Louisville, Ky., on Monday. **III (2): (1) + visual features:** Representative Veronica Escobar, Democrat of El Paso, has urged President Trump not to visit the city. "He should not come here while we're in mourning," she said. **IV (3): (2) + face naming:** Nan Whaley, the mayor of Dayton, Ohio, said on Tuesday that she was "disappointed" with President Trump's remarks the day before about the two massacres. **V (4): (3) + retrieval:** Mayor Nan Whaley of Dayton, Ohio, said on Tuesday that she had been "disappointed" with President Trump's remarks the day before about the two massacres.

Table 18: Generated captions with the models discussed in the ablation. We mark the correct/wrong content in green/red. Here we adopt model w/o CoLaM for illustration. Qualitative examples for model w/ CoLaM can be found in Figure 5.

Ground Truth Caption:
MORE SPACE The Oakes Center, Dreamcatcher's destination.
w/o CoLaM:
MORE SPACE The Oakes Center in Summit will host spring's play readings.
w/ CoLaM:
MORE SPACE The Oakes Center in Summit is Dreamcatcher's new home.

Ground Truth Caption:
Rupert Murdoch, the chief of News Corporation, and his son James, in 2010. Both have testified before a parliamentary panel.
w/o CoLaM:
Rupert Murdoch, left, and his son James have been largely unscathed because much of News Corporation's business and most of its profits lie in London.
w/ CoLaM:
Rupert Murdoch, left, the chairman of News Corporation, and his son James waited for law enforcement officials to take out employees of the company's British newspaper division accused of engaging in criminal conduct.

Ground Truth Caption:
Richard Blumenthal issued a civil investigative demand.
w/o CoLaM:
Richard Blumenthal asked Google to provide more information about its data collection.
w/ CoLaM:
Richard Blumenthal issued a civil investigative demand.

Ground Truth Caption:
New York Philharmonic, back from California, was led by Alan Gilbert on Saturday at Avery Fisher Hall.
w/o CoLaM:
New York Philharmonic, with Glenn Dicterow as soloist, at Avery Fisher Hall on Saturday.
w/ CoLaM:
New York Philharmonic, with Alan Gilbert conducting, at Avery Fisher Hall on Saturday.

Figure 5: Qualitative comparison w/ or w/o CoLaM. We mark the context from the news articles that is captured by our method in blue. Here we use the models with $\text{BART}_{\text{base}}$ as backbone LM for comparison.