Knowledge Completes the Vision: A Multimodal Entity-aware Retrieval-Augmented Generation Framework for News Image Captioning

Xiaoxing You[1,*], **Qiang Huang**[2,*], **Lingyu Li**[1], **Chi Zhang**[3], **Xiaopeng Liu**[3], **Min Zhang**[2], **Jun Yu**[2,4,†]

[1]Hangzhou Dianzi University, Hangzhou, China
[2]Harbin Institute of Technology (Shenzhen), Shenzhen, China
[3]People's Daily, Beijing, China
[4]Peng Cheng Laboratory, Shenzhen, China
{youxiaoxing, lilingyu0571}@hdu.edu.cn, {huangqiang, zhangmin2021, yujun}@hit.edu.cn,
{zhangchi, liuxiaopeng}@pdnews.cn

Abstract

News image captioning aims to produce journalistically informative descriptions by combining visual content with contextual cues from associated articles. Despite recent advances, existing methods struggle with three key challenges: (1) incomplete information coverage, (2) weak cross-modal alignment, and (3) suboptimal visual-entity grounding. To address these issues, we introduce **MERGE**, the first **M**ultimodal **E**ntity-aware **R**etrieval-augmented **GE**neration framework for news image captioning. MERGE constructs an entity-centric multimodal knowledge base (EMKB) that integrates textual, visual, and structured knowledge, enabling enriched background retrieval. It improves cross-modal alignment through a multistage hypothesis-caption strategy and enhances visual-entity matching via dynamic retrieval guided by image content. Extensive experiments on GoodNews and NYTimes800k show that MERGE significantly outperforms state-of-the-art baselines, with CIDEr gains of +6.84 and +1.16 in caption quality, and F1-score improvements of +4.14 and +2.64 in named entity recognition. Notably, MERGE also generalizes well to the unseen Visual News dataset, achieving +20.17 in CIDEr and +6.22 in F1-score, demonstrating strong robustness and domain adaptability.

Code — https://github.com/youxiaoxing/MERGE

1 Introduction

News articles typically include images accompanied by captions that blend visual elements with contextual details, enhancing reader comprehension and engagement. Unlike vanilla image captioning methods (Vinyals et al. 2016; Hossain et al. 2019; Yu et al. 2019; Xu et al. 2023), which primarily describe visible content, news image captioning demands both precise entity recognition and the incorporation of deeper contextual knowledge. Editors must analyze key elements–such as people, events, time, and location–and craft captions tailored to diverse journalistic contexts, where the same image may require entirely different descriptions (Nguyen et al. 2023).

Figure 1: Challenges in news image captioning: (a) Identifying entities absent from the article; (b) Aligning numerical details and visual objects across modalities; (c) Disambiguating entities in images with multiple subjects.

Automated news image captioning has been widely studied to assist editors. Early template-based systems (Ramisa et al. 2018; Biten et al. 2019; Hu, Chen, and Jin 2020) filled predefined templates with entities. While effective for structured output, these methods often yield rigid captions lacking nuanced context. Transformer-based models (Tran, Mathews, and Xie 2020; Zhao and Wu 2024) have introduced richer modeling of visual features–such as faces and objects–to improve entity-aware captioning. However, they often struggle to extract precise details from long or noisy articles, leading to incomplete or generic captions. Another prominent line of work (Zhou et al. 2022; Qu, Tuytelaars, and Moens 2024) focuses on extracting relevant textual contexts from articles. Techniques in this direction often leverage pre-trained or fine-tuned CLIP models (Radford et al. 2021) to retrieve salient sentences while minimizing redundancy. Yet, these methods typically fall short in establishing deep semantic connections between visual elements and textual narratives. More recently, Multimodal Large Language Models (MLLMs) (Xu et al. 2024a; Zhang, Zhang, and Wan 2024) have shown great promise by jointly modeling visual

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

and textual modalities. Their advanced reasoning capabilities and flexibility make them well-suited for the complex demands of news image captioning.

Despite significant progress, as illustrated in Figure 1, existing approaches continue to face three critical challenges:

- **Incomplete Information Coverage:** Generating accurate captions often requires referencing entities not explicitly mentioned in the accompanying article. For example, as shown in Figure 1(a), existing models fail to identify `Ruth Wilson` because she is absent from the text. Current methods (Biten et al. 2019; Tran, Mathews, and Xie 2020; Xu et al. 2024a; Zhang, Zhang, and Wan 2024) struggle to retrieve missing information and effectively integrate external knowledge (Xu et al. 2024b).
- **Limited Cross-modal Alignment:** At the sentence level, existing methods typically focus on either describing visual scenes (Hu, Chen, and Jin 2020; Qu, Tuytelaars, and Moens 2024) or extracting entity-rich sentences (Zhou et al. 2022; Zhang, Zhang, and Wan 2024), but struggle to holistically align visual objects with numerical details–for example, linking the `Toyota Tacoma` to its `2011` launch year, as depicted in Figure 1(b).
- **Suboptimal Visual-Entity Grounding:** Linking visual cues to correct named entities remains challenging, especially in images with multiple people or objects (Figure 1(c)). Implicit grounding methods (Qu, Tuytelaars, and Moens 2024) offer limited control, while fine-tuned models (Zhao and Wu 2024) struggle with unseen entities. Recent Retrieval-Augmented Generation (RAG) approaches (Xu et al. 2024b) use resources like ConceptNet (Speer, Chin, and Havasi 2017) for entity understanding but still lack robust visual-text integration, highlighting the need for multimodal RAG frameworks.

To address these challenges, we propose **MERGE**, the first **M**ultimodal **E**ntity-aware **R**etrieval-augmented **GE**neration framework customized for news image captioning. By seamlessly integrating explicit multimodal knowledge with the implicit reasoning capabilities of MLLMs, MERGE introduces three key innovations:

- **Information Enhancement:** MERGE builds an Entity-centric Multimodal Knowledge Base (EMKB) that consolidates named entities, images, and structured background knowledge. This resource enables the model to supplement missing details absent from the article–for instance, identifying the actress `Ruth Wilson`, who might otherwise be overlooked (Figure 1(a)).
- **Fine-grained Cross-modal Alignment:** To improve sentence-level alignment, MERGE introduces Hypothesis Caption-guided Multimodal Alignment (HCMA), which employs a three-stage Chain-of-Thought (CoT) prompting mechanism. This structured reasoning process enables accurate matching between visual cues and textual details, including nuances like the `2011` launch year of the `Toyota Tacoma` (Figure 1(b)).
- **Precise Visual-Entity Alignment:** For robust entity-level grounding, MERGE incorporates Retrieval-driven Multimodal Knowledge Integration (RMKI), which dynamically retrieves multimodal evidence and constructs background knowledge graphs from EMKB. This allows MERGE to distinguish visually similar individuals and keep precise entity associations–for example, correctly identifying `Chloe`, `Luke`, and `Jason` in Figure 1(c).

We conduct extensive evaluations of MERGE on three real-world datasets: GoodNews, NYTimes800k, and Visual News. For the GoodNews and NYTimes800k datasets, MERGE achieves new state-of-the-art performance, improving CIDEr by +6.84 and +1.16, and boosting named entity recognition F1-scores by +4.14 and +2.64, respectively. Importantly, MERGE generalizes well to Visual News, outperforming prior methods with a +20.17 boost in CIDEr and +6.22 in F1-score, despite the dataset being excluded from EMKB construction. Ablation studies further validate MERGE's architecture: HCMA enhances cross-modal alignment, while RMKI and EMKB significantly improve both caption quality and entity-level precision. These results underscore MERGE's effectiveness in generating accurate, contextually grounded, and journalistically informative captions across diverse domains.

2 Related Work

News image captioning aims to generate captions by integrating visual and textual information to produce contextually relevant descriptions, which can be broadly divided into three categories: (1) processing original articles, (2) extracting relevant contexts, and (3) incorporating MLLMs.

Methods Processing Original Articles. These methods (Biten et al. 2019; Yang and Okazaki 2020; Zhang et al. 2022; Zhang and Wan 2023; Nguyen et al. 2023; Ajankar and Dutta 2024; Xu et al. 2024b) use original news articles as inputs, often truncated due to model input constraints (Zhou et al. 2022). Biten et al. (2019) introduced a two-stage template-based method using entity placeholders. Tran, Mathews, and Xie (2020) proposed a transformer-based framework incorporating face and object embeddings for end-to-end captioning. Yang et al. (2021) structured captions using six journalistic components, while Liu et al. (2021) developed the Visual News dataset to enhance image-article connections. Zhang et al. (2022) combined CLIP with BART (Lewis et al. 2020) to improve named entity recognition, and Kalarani et al. (2023) pre-trained OFA for tasks like visual entailment and keyword extraction. Zhao and Wu (2024) introduced an entity-matching module to build multimodal graphs linking faces, objects, and entities. Xu et al. (2024c) applied prefix-tuning to inject news-specific semantic rules into BART for guided captioning. MERGE advances this line of work by retrieving entities and background knowledge from a multimodal knowledge base, bridging visual cues with entities and dynamically handling new entities without additional fine-tuning.

Methods Using Extracted Relevant Contexts. To reduce irrelevant information, these methods use models like CLIP to retrieve sentences most relevant to image content, treating entity-relevant text as positives. ICECAP (Hu, Chen, and Jin 2020) fine-tuned VSE++ (Faghri et al. 2018) with coarse-to-fine attention for entity identification, while Zhou et al. (2022) adapted CLIP for visually grounded entity detection

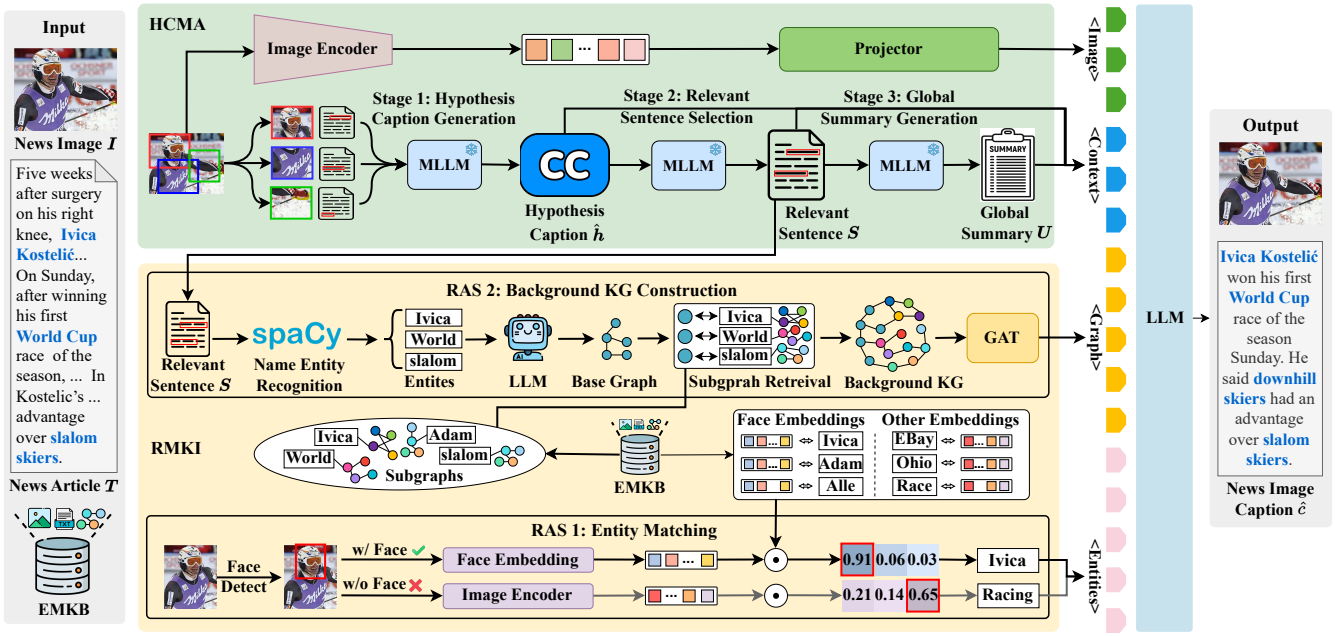

Figure 2: Overview of the MERGE framework.

and relation extraction. Qu, Tuytelaars, and Moens (2024) utilized CLIP for sentence retrieval and developed a Face Naming Module to associate faces with names. Building on these ideas, MERGE not only retrieves relevant context but also integrates background knowledge, producing captions that are accurate and journalistically informative.

Methods Incorporating MLLMs. The advent of GPT-4V (Achiam et al. 2023) has significantly advanced MLLMs, achieving impressive results across diverse tasks. For instance, InstructBLIP (Dai et al. 2023) employs Q-Former architecture for vision-language fusion and enhances cross-task generalization via instruction tuning. LLaVA (Liu et al. 2023) uses linear projection for alignment and excels on GPT-4-generated multimodal benchmarks. Advanced models like Qwen2-VL (Yang et al. 2024), InternVL2 (Chen et al. 2024b,a), and Llama3.2-Vision (Meta AI 2024) surpass GPT-4V in domains including medical diagnosis and autonomous driving. For news image captioning, Xu et al. (2024a) enhanced LLaVA with confidence-aware prompts and refined CLIP using matching-score comparative loss, while EAMA (Zhang, Zhang, and Wan 2024) adapted InstructBLIP with alignment tasks for improved multimodal understanding. MERGE goes further by integrating retrieval-augmented multimodal knowledge, addressing deeper challenges in entity grounding and contextual completeness.

3 The MERGE Framework

3.1 Overview

MERGE is a multimodal RAG framework developed to advance news image captioning by seamlessly integrating external multimodal data with structured knowledge. As depicted in Figure 2, MERGE contains three core components:

- **Entity-centric Multimodal Knowledge Base (EMKB):** Consolidates named entities, images, and background knowledge to bridge information gaps and strengthen contextual grounding.
- **Hypothesis Caption-guided Multimodal Alignment (HCMA):** Achieves fine-grained sentence-level alignment between visual and textual inputs through a three-stage Chain-of-Thought (CoT) prompting process.
- **Retrieval-driven Multimodal Knowledge Integration (RMKI):** Improves visual-entity grounding by matching visual cues to entities and dynamically constructing background knowledge graphs from EMKB.

3.2 EMKB

EMKB consolidates multimodal data and contextual knowledge as the foundation of MERGE.

Entity Extraction and Image Collection. We extract entities (e.g., celebrity names, locations, artworks, landmarks, and buildings) from the GoodNews (Biten et al. 2019) and NYTimes800k (Tran, Mathews, and Xie 2020) datasets using spaCy (Honnibal 2017), and expand this set via an LLM. For each entity, we collect a Wikipedia image and augment it with up to five images from Google Search. To cover less common or missing celebrity entities, we also incorporate four public datasets (see Appendix B), capping each entity at five images. This results in a visually rich, entity-centric dataset that supports precise visual-entity alignment.

Background Knowledge Acquisition. Beyond visual data, EMKB captures deeper contextual knowledge. Background information for each entity is extracted from Wikipedia and IMDb and structured into subgraphs using LLMs with a crafted prompt p_k (see Appendix A.1). Unlike static knowledge graphs (Alberts et al. 2021; Wang et al. 2019), these

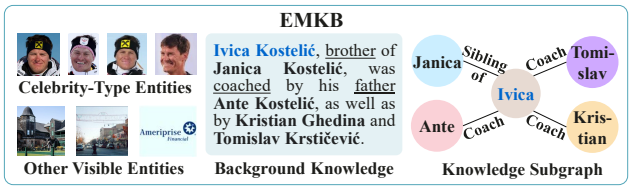

Figure 3: Architecture of EMKB, illustrating named entities, their images, background knowledge, and knowledge subgraphs, which support context-rich news image captioning.

subgraphs are dynamically retrieved during caption generation, enabling news-specific knowledge integration.

EMKB Formulation. Formally, the EMKB $\boldsymbol{B}$ is defined as:

$$\boldsymbol{B} = \{(\boldsymbol{e}_i, \{\boldsymbol{I}_j\}, \boldsymbol{b}_i, \boldsymbol{G}^i_{sub})\}^N_{i=1}, \tag{1}$$

where $\boldsymbol{e}_i$ is the i-th entity, $\{\boldsymbol{I}_j\}$ are its associated images, $\boldsymbol{b}_i$ denotes background knowledge, and $\boldsymbol{G}^i_{sub}$ is its structured knowledge subgraph. This formulation enhances contextual grounding, aids in entity disambiguation, and facilitates cross-modal alignment. Its structure and components are depicted in Figure 3. Implementation details and the updating mechanism for EMKB are provided in Appendix B.

3.3 HCMA

HCMA tackles sentence-level cross-modal alignment via a three-stage CoT prompting process.

Stage 1: Hypothesis Caption Generation. A hypothesis caption $\hat{\boldsymbol{h}}$ is generated via a crafted prompt p_h, encapsulating visual and textual cues from the input image $\boldsymbol{I}$ and article $\boldsymbol{T}$. The MLLM first extracts key sentences reflecting the central themes and visual content of $\boldsymbol{T}$. Let $\hat{h}_{<i}$ be the partial caption up to token i. The model then iteratively generates each word $\hat{h}_i$:

$$\hat{h}_i = \text{MLLM}(p_h, \hat{h}_{<i}, \boldsymbol{I}, \boldsymbol{T}). \tag{2}$$

Stage 2: Relevant Sentence Selection. HCMA refines context by selecting the most relevant sentences $\boldsymbol{S}$ from $\boldsymbol{T}$ using the hypothesis caption $\hat{\boldsymbol{h}}$ and image $\boldsymbol{I}$ as anchors. The selection uses a dedicated prompt p_s:

$$\boldsymbol{S} = \text{MLLM}(p_s, \hat{\boldsymbol{h}}, \boldsymbol{I}, \boldsymbol{T}), \tag{3}$$

where $|\boldsymbol{S}| \leq 5$ balances informativeness and efficiency, ensuring alignment with both visual and textual content.

Stage 3: Global Summary Generation. While $\boldsymbol{S}$ and $\hat{\boldsymbol{h}}$ capture localized context, they can miss broader connections, as shown in Table 2. To capture a global perspective and manage input length, HCMA generates a concise global summary $\boldsymbol{U}$ from $\boldsymbol{T}$ using prompt p_g:

$$\boldsymbol{U} = \text{MLLM}(p_g, \boldsymbol{T}), \tag{4}$$

with $\boldsymbol{U}$ limited to 100 words for brevity and clarity.

Remarks. By integrating local context from $\boldsymbol{S}$ and global context from $\boldsymbol{U}$, HCMA produces captions that are precise and contextually rich. Details of prompts p_h, p_s, and p_g are provided in Appendix A.2.

Algorithm 1: Background KG Construction

Input: EMKB $\boldsymbol{B}$, relevant sentences $\boldsymbol{S}$
Output: Background knowledge graph $\boldsymbol{G}$

$\boldsymbol{E}_{sen} \leftarrow$ Spacy($\boldsymbol{S}$); $\boldsymbol{R} \leftarrow$ LLM(p_r, $\boldsymbol{E}_{sen}$);
$\boldsymbol{G}_{base} \leftarrow$ ConstructBaseGraph($\boldsymbol{E}_{sen}$, $\boldsymbol{R}$);
$\boldsymbol{\Phi} \leftarrow \varnothing$; ▷ Store knowledge subgraphs
foreach $\boldsymbol{e} \in \boldsymbol{E}_{sen}$ **do**
  $\boldsymbol{G}^i_{sub} \leftarrow$ Retrieve($\boldsymbol{e}$, $\boldsymbol{B}$); $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} \cup \{\boldsymbol{G}^i_{sub}\}$;
$\boldsymbol{G} \leftarrow$ IntegrateGraph($\boldsymbol{G}_{base}$, $\boldsymbol{\Phi}$);
return $\boldsymbol{G}$;

3.4 RMKI

At the entity level, RMKI strengthens visual-entity grounding through two Retrieval-Augmented Strategies (RAS) within the EMKB $\boldsymbol{B}$: entity matching and background knowledge graph construction.

RAS 1: Entity Matching. RMKI matches visual cues in image $\boldsymbol{I}$ to entities stored in EMKB $\boldsymbol{B}$ via two pathways:

- **Face Images:** Faces detected in $\boldsymbol{I}$ are encoded as feature vectors $\boldsymbol{F}$ using InsightFace.[1] For each vector $\boldsymbol{y} \in \boldsymbol{F}$, RMKI computes cosine similarity against face vectors $\boldsymbol{x}_j$ from EMKB images $\boldsymbol{I}_j \in \boldsymbol{B}$:

$$j^* = \arg\max_j \cos(\boldsymbol{x}_j, \boldsymbol{y}). \tag{5}$$

  The matched entities from $\boldsymbol{I}_{j^*}$ are then extracted to form the entity set $\boldsymbol{E}$.
- **Non-Face Images:** For images without faces, RMKI leverages CLIP's image encoder to generate visual embeddings. Cosine similarity is used to identify the closest matching images $\boldsymbol{I}_j \in \boldsymbol{B}$, yielding the entity set $\boldsymbol{E}$.

RAS 2: Background Knowledge Graph (KG) Construction. To enrich contextual understanding, RMKI constructs a background KG for entities identified in the relevant sentences $\boldsymbol{S}$. This process unfolds in four steps:

- **Named Entity Recognition (NER):** Identify named entities $\boldsymbol{E}_{sen}$ within $\boldsymbol{S}$ using spaCy.
- **Relation Extraction:** Use LLMs with a dedicated prompt p_r to extract relations $\boldsymbol{R}$ among entities in $\boldsymbol{E}_{sen}$, forming the base relation graph $\boldsymbol{G}_{base}$.
- **Subgraph Retrieval:** For each entity $\boldsymbol{e} \in \boldsymbol{E}_{sen}$, retrieve the knowledge subgraph $\boldsymbol{G}^i_{sub}$ from the EMKB $\boldsymbol{B}$ and aggregate them into a set $\boldsymbol{\Phi}$.
- **Graph Integration:** Integrate subgraphs $\boldsymbol{\Phi}$ into $\boldsymbol{G}_{base}$, deduplicating overlapping nodes and edges, to produce the final knowledge graph $\boldsymbol{G}$.

This dynamic retrieval and graph integration enables MERGE to enrich generated captions with precision, entity-specific knowledge, improving both contextual coherence and factual accuracy. Algorithm 1 details this process, with prompt p_r provided in Appendix A.3.

[1] https://github.com/deepinsight/insightface

Dataset	Method	Caption Quality				Named Entity Accuracy		
		BLEU-4 ↑	METEOR ↑	ROUGE ↑	CIDEr ↑	Precision ↑	Recall ↑	F1-score ↑
GoodNews	Biten et al. (2019)	0.89	4.37	12.20	13.10	8.23	6.06	6.98
	ICECAP (Hu, Chen, and Jin 2020)	1.96	6.01	15.70	26.08	-	-	12.03
	Tell (Tran, Mathews, and Xie 2020)	6.05	10.30	21.40	53.80	22.20	18.70	20.30
	JoGANIC (Yang et al. 2021)	6.83	11.25	23.05	61.22	26.87	22.05	24.22
	Liu et al. (2021)	6.10	8.30	21.60	55.40	22.90	19.30	20.95
	Zhou et al. (2022)	6.30	-	22.40	60.30	24.20	20.90	22.43
	NewsMEP (Zhang et al. 2022)	8.30	12.23	23.17	63.99	23.43	23.24	23.33
	Kalarani et al. (2023)	7.14	11.21	24.30	72.33	24.37	20.09	22.02
	Zhao and Wu (2024)	8.31	12.32	23.22	64.15	-	-	23.39
	Xu et al. (2024c)	8.18	12.50	23.56	71.58	25.51	23.68	24.56
	Qu, Tuytelaars, and Moens (2024)	8.60	12.39	23.38	71.96	24.30	25.54	24.90
	Xu et al. (2024a)	8.49	12.88	26.22	83.52	30.19	26.57	28.26
	EAMA (Zhang, Zhang, and Wan 2024)	10.04	13.95	27.06	87.70	27.58	28.92	28.23
	MERGE	**10.19**	**14.31**	**28.02**	**94.54**	**32.89**	**31.93**	**32.40**
NYTimes800k	Tell (Tran, Mathews, and Xie 2020)	6.30	10.30	21.70	54.40	24.60	22.20	23.34
	JoGANIC (Yang et al. 2021)	6.79	10.93	22.80	59.42	28.63	24.49	26.40
	Liu et al. (2021)	6.40	8.10	21.90	56.10	24.80	22.30	23.48
	Zhou et al. (2022)	7.00	-	22.90	63.60	29.80	25.90	27.71
	NewsMEP (Zhang et al. 2022)	9.57	13.02	23.62	65.85	26.61	28.57	27.56
	Kalarani et al. (2023)	7.54	11.27	23.28	66.41	28.11	23.25	25.45
	Zhao and Wu (2024)	9.53	13.30	23.89	66.43	-	-	27.71
	Xu et al. (2024c)	9.41	13.10	24.42	72.29	28.15	28.80	28.47
	Qu, Tuytelaars, and Moens (2024)	9.24	12.57	23.44	71.65	26.88	28.59	27.71
	Xu et al. (2024a)	9.07	13.17	26.48	83.72	**32.38**	30.08	31.19
	EAMA (Zhang, Zhang, and Wan 2024)	11.03	14.22	27.15	87.00	29.79	32.24	30.97
	MERGE	**11.47**	**14.94**	**27.51**	**88.16**	31.87	**36.04**	**33.83**
Visual News	Tell (Tran, Mathews, and Xie 2020)	9.60	-	22.80	83.80	23.70	19.20	21.21
	Visual News Captioner (Liu et al. 2021)	5.30	8.20	17.90	50.50	19.70	17.60	18.59
	Zhou et al. (2022)	11.60	-	25.00	107.60	26.20	21.20	23.44
	Kalarani et al. (2023)	6.91	10.54	21.29	65.14	19.31	19.90	19.60
	MERGE	**14.77**	**15.72**	**28.26**	**127.77**	**29.88**	**29.45**	**29.66**

Table 1: Experimental results on GoodNews, NYTimes800k, and Visual News, comparing MERGE with baseline methods.

3.5 Caption Generation

Given an image $\boldsymbol{I}$ and a news article $\boldsymbol{T}$, MERGE proceeds as follows: First, HCMA generates a hypothesis caption $\hat{\boldsymbol{h}}$, selects relevant sentences $\boldsymbol{S}$, and creates a global summary $\boldsymbol{U}$. Second, RMKI matches $\boldsymbol{I}$ to entities $\boldsymbol{E}$ and builds a background knowledge graph $\boldsymbol{G}$. Finally, InstructBLIP (Dai et al. 2023) is employed for caption generation, enhanced with a 4-layer Graph Attention Network (GAT) (Veličković et al. 2018) to encode $\boldsymbol{G}$, integrating multimodal inputs: $\boldsymbol{X} = \{\boldsymbol{I}, \hat{\boldsymbol{h}}, \boldsymbol{S}, \boldsymbol{U}, \boldsymbol{E}, \boldsymbol{G}\}$ to produce the final caption $\hat{\boldsymbol{c}}$:

$$\hat{\boldsymbol{c}} = \text{MLLM}(\boldsymbol{X}; \boldsymbol{\theta}), \tag{6}$$

where $\boldsymbol{\theta}$ denotes the parameters of MLLM. The model is trained to minimize the Cross-Entropy (CE) loss, which measures the negative log-likelihood of the ground truth caption $\boldsymbol{c}$ conditioned on the multimodal inputs:

$$\mathcal{L}_{CE} = -\sum_{i=1}^{|\boldsymbol{c}|} \log P(c_i \mid c_{<i}, \boldsymbol{X}). \tag{7}$$

MERGE combines textual, visual, and structured knowledge, resulting in comprehensive, contextually aligned news image captions, as validated in Section 4.2.

4 Experiments

4.1 Experiment Setup

Datasets. We assess MERGE on GoodNews (Biten et al. 2019), NYTimes800k (Tran, Mathews, and Xie 2020), and Visual News (Liu et al. 2021). Visual News is excluded from EMKB construction and used to evaluate MERGE's generalization. Additional details are in Appendix C.

Metrics. We assess caption quality using BLEU-4 (Papineni et al. 2002), METEOR (Denkowski and Lavie 2014), ROUGE (Lin 2004), and CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). Additionally, named entity accuracy is measured via Precision, Recall, and F1-score using spaCy (Honnibal 2017).

Baselines. We compare MERGE (implementation details in Appendix D) with state-of-the-art baselines below.

- **Methods Processing Original Articles:** Biten et al. (2019), Tell (Tran, Mathews, and Xie 2020), JoGANIC (Yang et al. 2021), NewsMEP (Zhang et al. 2022), alongside recent approaches like Kalarani et al. (2023), Xu et al. (2024c) and Zhao and Wu (2024).
- **Methods Using Extracted Relevant Contexts:** ICECAP (Hu, Chen, and Jin 2020), Zhou et al. (2022), and Qu, Tuytelaars, and Moens (2024).

Dataset	Method	Caption Quality				Named Entity Accuracy		
		BLEU-4 ↑	METEOR ↑	ROUGE ↑	CIDEr ↑	Precision ↑	Recall ↑	F1-score ↑
GoodNews	InstructBLIP (w/o FT)	3.46	8.41	13.82	24.42	16.28	14.21	15.17
	InstructBLIP (w/ FT)	8.57	12.64	24.53	84.8	32.23	27.64	29.76
	+ HCMA (w/ Stage 1)	9.33	13.50	26.31	84.83	30.10	29.16	29.62
	+ HCMA (w/ Stage 1+2)	9.49	13.33	26.63	85.24	30.73	29.19	29.94
	+ HCMA (w/ Stage 1+2+3)	9.32	12.53	26.59	86.08	30.77	29.30	30.02
	+ RMKI & EMKB (w/ RAS 1)	9.62	13.74	25.75	91.52	**33.72**	30.98	32.29
	+ RMKI & EMKB (w/ RAS 2)	9.20	13.12	26.52	85.79	28.91	30.94	29.89
	+ RMKI & EMKB (w/ RAS 1+2)	9.73	13.79	25.76	91.36	33.37	31.27	32.29
	MERGE	**10.19**	**14.31**	**28.02**	**94.54**	32.89	**31.93**	**32.40**
NYTimes800k	InstructBLIP (w/o FT)	3.77	8.14	13.16	22.33	13.66	15.68	14.60
	InstructBLIP (w/ FT)	9.16	12.68	24.70	73.68	28.72	29.69	29.20
	+ HCMA (w/ Stage 1)	9.15	13.07	25.04	75.40	30.72	30.92	30.82
	+ HCMA (w/ Stage 1+2)	10.27	14 15	25.51	76.81	27.74	32.91	30.10
	+ HCMA (w/ Stage 1+2+3)	10.01	13.62	25.79	80.14	30.23	31.83	31.01
	+ RMKI & EMKB (w/ RAS 1)	11.09	14.48	26.53	81.04	30.53	34.82	32.53
	+ RMKI & EMKB (w/ RAS 2)	10.39	14.07	25.16	74.43	27.80	32.59	30.01
	+ RMKI & EMKB (w/ RAS 1+2)	10.75	14.17	26.61	82.99	31.76	34.28	32.97
	MERGE	**11.47**	**14.94**	**27.51**	**88.16**	**31.87**	**36.04**	**33.83**
Visual News	InstructBLIP (w/o FT)	1.67	3.85	5.92	7.63	5.57	6.29	5.91
	InstructBLIP (w/ FT)	14.00	15.10	27.47	103.48	26.20	28.04	27.09
	+ HCMA (w/ Stage 1)	13.39	15.12	25.60	103.95	25.81	29.33	27.46
	+ HCMA (w/ Stage 1+2)	14.37	15.34	27.33	105.28	28.77	26.61	27.65
	+ HCMA (w/ Stage 1+2+3)	14.57	15.44	27.53	108.97	27.32	**29.56**	28.40
	+ RMKI & EMKB (w/ RAS 1)	13.33	14.90	27.37	116.84	29.74	25.17	27.26
	+ RMKI & EMKB (w/ RAS 2)	**14.95**	15.46	27.83	113.41	27.06	28.99	27.99
	+ RMKI & EMKB (w/ RAS 1+2)	14.12	15.03	27.30	123.50	26.36	28.01	27.11
	MERGE	14.77	**15.72**	**28.26**	**127.77**	**29.88**	29.45	**29.66**

Table 2: Ablation results on GoodNews, NYTimes800k, and Visual News, highlighting the impact of MERGE's components.

- **Methods Incorporating MLLMs:** Xu et al. (2024a) and EAMA (Zhang, Zhang, and Wan 2024).

4.2 Comparison Results of Different Baselines

Table 1 summarizes how MERGE performs against baseline methods on GoodNews, NYTimes800k, and Visual News.

Caption Quality. MERGE achieves state-of-the-art results across nearly all metrics. Compared to the strongest baseline EAMA, MERGE improves CIDEr by +6.84 on GoodNews and +1.16 on NYTimes800k, while remaining competitive on other metrics. The smaller margin on NYTimes800k reflects its higher complexity, with articles being nearly twice as long as GoodNews, hindering relevant content extraction.

These gains highlight prior methods' limitations in cross-modal alignment and visual-entity grounding. Baselines like Qu, Tuytelaars, and Moens (2024) and Zhang, Zhang, and Wan (2024) rely on sentence retrieval but focus on either image descriptions or entity-rich text, causing suboptimal alignment. Moreover, existing methods often fail to robustly link entities across modalities, degrading caption quality.

In contrast, MERGE's HCMA module effectively selects relevant context, while RMKI and EMKB integrate multimodal RAG to balance visual and textual information, achieving more precise entity alignment. These advantages are further demonstrated in Section 4.4.

Named Entity Accuracy. MERGE also sets new benchmarks for NER. On GoodNews, MERGE surpasses Xu et al. (2024a) by +2.70 precision and +4.14 F1-score, and exceeds EAMA by +3.01 recall, validating EMKB and RMKI effectiveness. On NYTimes800k, MERGE achieves +2.64 F1-score improvement, though precision slightly trails Xu et al. (2024a) due to their additional training data (20% training set plus full validation) and knowledge distillation for fine-tuning. In contrast, MERGE's HCMA requires no extra training, ensuring adaptability.

Generalization Test. Table 1 also shows MERGE's generalization on Visual News. MERGE outperforms the second-best method, Zhou et al. (2022), by +20.17 in CIDEr and +6.22 in F1-score. This significant improvement underscores how EMKB provides broad coverage for news image captioning, even for datasets not included in its construction.

4.3 Ablation Study

Table 2 reports the ablation results, quantifying the contributions of MERGE's core components.

Impact of MLLMs. MERGE adopts InstructBLIP as its MLLM backbone (Section 3.5). **InstructBLIP (w/o FT)** shows general-purpose MLLMs fall short on domain-specific news captioning when used zero-shot. After task-specific tuning, **InstructBLIP (w/ FT)** achieves substantial

Case	Image	Caption
(a)		**Ground-Truth Caption:** Meryl Streep and Clint Eastwood in a scene from "The Bridges of Madison County" (1995). **Tell (Tran et al., 2020):** Robert James Waller, left, with his daughter, Rachael, in an undated photo. (Missing: The Bridges… and 1995) **Kalarani et al. (2023):** Robert James Waller with Francesca Kincaid on the Roseman Covered Bridge in Iowa. (Missing: 1995) **MERGE:** Meryl Streep and Clint Eastwood on the set of "The Bridges of Madison County" in 1995.
(b)		**Ground-Truth Caption:** Facebook CEO Mark Zuckerberg appeared before the Senate Commerce and Judiciary committees on Tuesday. **Tell (Tran et al., 2020):** Senator Jeff Flake of Arizona said on Tuesday that he would not seek re-election. (Missing: Senate Commerce…) **Kalarani et al. (2023):** Speaker Paul Ryan on Capitol Hill on Tuesday He announced that he would... (Missing: Senate Commerce…) **MERGE :** Mark Zuckerberg, Facebook's chief executive, testifying before the Senate Judiciary and Commerce Committees on Tuesday.
(c)		**Ground-Truth Caption:** The complex houses 11,232 units across 80 acres east of First Avenue, between 14th and 23rd Streets. **Tell (Tran et al., 2020):** A dog near the site of the Stuyvesant Town-Peter Cooper Village. (Missing: 11,232, 80 acres, 14th and 23rd) **Kalarani et al. (2023):** Stuyvesant Town-Peter Cooper Village is owned by Blackstone Group. (Missing: 11,232, 80 acres, 14th and 23rd) **MERGE :** Stuyvesant Town-Peter Cooper Village has 11,232 units across 80 acres east of First Avenue, between 14th and 23rd Streets.
(d)		**Ground-Truth Caption:** From left: Jacquemus; Salvatore Ferragamo; Stella McCartney. **Tell (Tran et al., 2020):** From left: Salvatore Ferragamo, Marco Giorgio, spring 2015; (Missing: Jacquemus and Stella McCartney) **Kalarani et al. (2023):** From left Salvatore Ferragamo, Stella McCartneys. (Missing: Jacquemus) **MERGE:** From left: Salvatore Ferragamo; Jacquemus; Stella McCartney.
(e)		**Ground-Truth Caption:** Moxie Marlinspike, founder of Open Whisper Systems, a maker of the encryption app Signal. **Tell (Tran et al., 2020):** Mark Zuckerberg, the chief executive of Facebook, which has been accused of using the Signal app. **Kalarani et al. (2023):** Mark Zuckerberg chief executive of Facebook. (Missing: Signal) **MERGE:** Moxie Marlinspike, a co-founder of Open Whisper Systems, which makes the Signal messaging app.

Figure 4: Case study on GoodNews. Entities correctly identified by MERGE are depicted in blue, while errors are shown in red.

gains over conventional baselines (Table 1), validating fine-tuning effectiveness. We further evaluate other advanced MLLMs and compare efficiency in Appendices F–H, confirming MERGE's flexibility and efficiency.

Impact of HCMA. Introducing the HCMA component incrementally improves both generation quality and entity recognition. **HCMA (w/ Stage 1)** introduces hypothesis captioning, offering notable gains in both CIDEr and F1-scores. Incorporating sentence selection in **HCMA (w/ Stage 1+2)** further enhances alignment by focusing on semantically relevant context. Finally, including global summary generation in **HCMA (w/ Stage 1+2+3)** achieves the best overall performance, highlighting the importance of progressively structured textual grounding.

Impacts of RMKI and EMKB. The RMKI and EMKB jointly drive significant improvements. **RMKI & EMKB (w/ RAS 1)**, which supports visual-entity alignment through entity matching, enhances precision in NER. **RMKI & EMKB (w/ RAS 2)** leverages background knowledge graphs, improving contextual grounding and recall. Combining both in **RMKI & EMKB (w/ RAS 1+2)** leads to the highest F1 and CIDEr scores, confirming the complementary benefits of structured retrieval and visual grounding.

Overall Impact. These findings confirm that MERGE's performance stems from the synergy of its key components: a fine-tuned MLLM backbone, multistage cross-modal alignment, and dual-stream retrieval from structured and unstructured knowledge. Together, they enable MERGE to effectively tackle the core challenges of news image captioning.

4.4 Case Study

Figure 4 showcases MERGE's outputs on GoodNews examples, highlighting three key capabilities:

- **Information Enhancement (via EMKB):** In case (a), MERGE correctly identifies `Clint Eastwood` using EMKB, although his name is absent from the article.
- **Fine-grained Cross-modal Alignment (via HCMA):** Cases (b) and (c) demonstrate accurate alignment between visual and textual cues, successfully incorporating details like `Senate Commerce and Judiciary committees`, `11,232 units`, and `80 acres`.
- **Precise Visual-Entity Alignment (via RMKI):** Cases (d) and (e) demonstrate how RMKI enables MERGE to distinguish among multiple individuals and visually similar subjects, ensuring correct entity grounding.

These results highlight MERGE's ability to enrich missing context, align cross-modal details, and resolve complex visual references, which are crucial for real-world news captioning. Additional examples from NYTimes800k and Visual News are provided in Appendix I.

5 Conclusions

In this work, we introduced MERGE, a novel multimodal entity-aware RAG framework for news image captioning. By building an entity-centric multimodal knowledge base and integrating structured, visual, and textual information, MERGE effectively tackles key challenges in contextual grounding, cross-modal alignment, and visual-entity association. Extensive experiments on GoodNews, NYTimes800k, and Visual News show that MERGE consistently surpasses strong baselines, achieving state-of-the-art results in both caption generation and named entity recognition. These findings underscore the power of multimodal RAG for complex, knowledge-intensive vision-language tasks and mark an important step forward in bridging multimodal understanding and real-world journalistic applications.

Acknowledgements

We sincerely thank the anonymous reviewers for their insightful and constructive feedback, which greatly improved this paper. This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62125201 and U24B20174.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Ajankar, S.; and Dutta, T. 2024. Image-Relevant Entities Knowledge-Aware News Image Captioning. *IEEE MultiMedia*, 31(1): 88–98.

Alberts, H.; Huang, N.; Deshpande, Y.; Liu, Y.; Cho, K.; Vania, C.; and Calixto, I. 2021. VisualSem: A High-quality Knowledge Graph for Vision and Language. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, 138–152.

Biten, A. F.; Gomez, L.; Rusinol, M.; and Karatzas, D. 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)*, 12466–12475.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; Ma, J.; Wang, J.; Dong, X.; Yan, H.; Guo, H.; He, C.; Shi, B.; Jin, Z.; Xu, C.; Wang, B.; Wei, X.; Li, W.; Zhang, W.; Zhang, B.; Cai, P.; Wen, L.; Yan, X.; Dou, M.; Lu, L.; Zhu, X.; Lu, T.; Lin, D.; Qiao, Y.; Dai, J.; and Wang, W. 2024a. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *Science China Information Sciences*, 67(12): 220101.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2024b. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 24185–24198.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 49250–49267.

Denkowski, M.; and Lavie, A. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, 376–380.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *Proceedings of the British Machine Vision Conference (BMVC)*, 1–12.

Honnibal, M. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Hossain, M. Z.; Sohel, F.; Shiratuddin, M. F.; and Laga, H. 2019. A Comprehensive Survey of Deep Learning for Image Captioning. *ACM Computing Surveys (CSUR)*, 51(6): 1–36.

Hu, A.; Chen, S.; and Jin, Q. 2020. ICECAP: Information Concentrated Entity-aware Image Captioning. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 4217–4225.

Kalarani, A. R.; Bhattacharyya, P.; Chhaya, N.; and Shekhar, S. 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)*, 695–706.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 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 (ACL)*, 7871–7880.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, 74–81.

Liu, F.; Wang, Y.; Wang, T.; and Ordonez, V. 2021. Visual News: Benchmark and Challenges in News Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6761–6771.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 34892–34916.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

Nguyen, K.; Biten, A. F.; Mafla, A.; Gomez, L.; and Karatzas, D. 2023. Show, Interpret and Tell: Entity-aware Contextualised Image Captioning in Wikipedia. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 1940–1948.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 311–318.

Qu, T.; Tuytelaars, T.; and Moens, M. F. 2024. Visually-Aware Context Modeling for News Image Captioning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, 2927–2943.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.;

Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, 8748–8763.

Ramisa, A.; Yan, F.; Moreno-Noguer, F.; and Mikolajczyk, K. 2018. BreakingNews: Article Annotation by Image and Text Processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 40(5): 1072–1085.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 4444–4451.

Tran, A.; Mathews, A.; and Xie, L. 2020. Transform and Tell: Entity-Aware News Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13035–13045.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4566–4575.

Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; and Bengio, Y. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2016. Show and Tell: Lessons learned from the 2015 MSCOCO Image Captioning Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39(4): 652–663.

Wang, M.; Qi, G.; Wang, H.; and Zheng, Q. 2019. Richpedia: A Comprehensive Multi-modal Knowledge Graph. In *Joint International Semantic Technology Conference*, 130–145.

Xu, N.; Gao, Y.; Zhang, T.-T.; Tian, H.; and Liu, A.-A. 2024a. Cross-Modal Coherence-Enhanced Feedback Prompting for News Captioning. In *Proceedings of the 32nd ACM International Conference on Multimedia (ACM MM)*, 9369–9377.

Xu, N.; Wang, Y.; Zhang, T.; Tian, H.; Kankanhalli, M.; and Liu, A.-A. 2024b. How to Understand Named Entities: Using Common Sense for News Captioning. *arXiv preprint arXiv:2403.06520*.

Xu, N.; Zhang, T.; Tian, H.; and Liu, A.-A. 2024c. Rule-Driven News Captioning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 34(11): 11657–11667.

Xu, X.; Wang, Z.; Zhang, G.; Wang, K.; and Shi, H. 2023. Versatile Diffusion: Text, Images and Variations All in One Diffusion Model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 7754–7765.

Yang, A.; Yang, B.; Hui, B.; Zheng, B.; Yu, B.; Zhou, C.; Li, C.; Li, C.; Liu, D.; Huang, F.; Dong, G.; Wei, H.; Lin, H.; Tang, J.; Wang, J.; Yang, J.; Tu, J.; Zhang, J.; Ma, J.; Yang, J.; Xu, J.; Zhou, J.; Bai, J.; He, J.; Lin, J.; Dang, K.; Lu, K.; Chen, K.; Yang, K.; Li, M.; Xue, M.; Ni, N.; Zhang, P.; Wang, P.; Peng, R.; Men, R.; Gao, R.; Lin, R.; Wang, S.; Bai, S.; Tan, S.; Zhu, T.; Li, T.; Liu, T.; Ge, W.; Deng, X.; Zhou, X.; Ren, X.; Zhang, X.; Wei, X.; Ren, X.; Liu, X.; Fan, Y.; Yao, Y.; Zhang, Y.; Wan, Y.; Chu, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; Guo, Z.; and Fan, Z. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yang, X.; Karaman, S.; Tetreault, J.; and Jaimes, A. 2021. Journalistic Guidelines Aware News Image Captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5162–5175.

Yang, Z.; and Okazaki, N. 2020. Image Caption Generation for News Articles. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 1941–1951.

Yu, J.; Li, J.; Yu, Z.; and Huang, Q. 2019. Multimodal Transformer With Multi-View Visual Representation for Image Captioning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 30(12): 4467–4480.

Zhang, J.; Fang, S.; Mao, Z.; Zhang, Z.; and Zhang, Y. 2022. Fine-tuning with Multi-modal Entity Prompts for News Image Captioning. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM)*, 4365–4373.

Zhang, J.; and Wan, X. 2023. Exploring the Impact of Vision Features in News Image Captioning. In *Findings of the Association for Computational Linguistics: ACL 2023*, 12923–12936.

Zhang, J.; Zhang, H.; and Wan, X. 2024. Entity-Aware Multimodal Alignment Framework for News Image Captioning. *arXiv preprint arXiv:2402.19404*.

Zhao, W.; and Wu, X. 2024. Boosting Entity-Aware Image Captioning With Multi-Modal Knowledge Graph. *IEEE Transactions on Multimedia (TMM)*, 26: 2659–2670.

Zhou, M.; Luo, G.; Rohrbach, A.; and Yu, Z. 2022. Focus! Relevant and Sufficient Context Selection for News Image Captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6078–6088.