MUNIChus: Multilingual News Image Captioning Benchmark

Yuji Chen[1], Alistair Plum[2], Hansi Hettiarachchi[1], Diptesh Kanojia[3], Saroj Basnet[4], Marcos Zampieri[4] and Tharindu Ranasinghe[1]
[1]Lancaster University, UK [2]University of Luxembourg, Luxembourg
[3]University of Surrey, UK [4]George Mason University, USA
t.ranasinghe@lancaster.ac.uk

Abstract

The goal of news image captioning is to generate captions by integrating news article content with corresponding images, highlighting the relationship between textual context and visual elements. The majority of research on news image captioning focuses on English, primarily because datasets in other languages are scarce. To address this limitation, we **create the first multilingual news image captioning benchmark, MUNIChus**, comprising 9 languages, including several low-resource languages such as Sinhala and Urdu. We evaluate various state-of-the-art neural news image captioning models on MUNIChus and find that news image captioning remains challenging. We also make MUNIChus publicly available with over 20 models already benchmarked. MUNIChus opens new avenues for further advancements in developing and evaluating multilingual news image captioning models.

Keywords: News Image Captioning, Multimodal Large Language Models, Text Generation

1. Introduction

Image captioning is a task that bridges language and vision (Liu et al., 2021), where the goal is to generate a semantically accurate description of the visual content in a syntactically correct manner (Stefanini et al., 2023; Xu et al., 2023a). Image captioning can have profound impacts, particularly in assisting visually impaired individuals by enabling a better understanding of the content within an image (Daneshfar et al., 2024). Therefore, the task has attracted significant interest and achieved notable advancements in recent years (Ghandi et al., 2023). The recent introduction of multimodal large language models (MLLMs) (Liang et al., 2024; Caffagni et al., 2024; Xiao et al., 2025) has fuelled the development of image captioning models and provided state-of-the-art results (Agarwal and Verma, 2024; Stefanini et al., 2023). Furthermore, MLLMs have also been leveraged for multilingual image captioning, expanding their applicability across diverse linguistic contexts (Ramos et al., 2023; Farabi et al., 2024).

News image captioning, a variant of the image captioning task, requires generating an informative caption, rich in entities such as the names of people or news events, based on the provided news image and associated news article (Rajakumar Kalarani et al., 2023). Typically, a news image illustrates a portion of the article, with the caption linking the image content to the article. Ideally, readers should grasp the article's essence by simply browsing its images and their captions (Qu et al., 2024).

Generic image captions are different to news image captions as they are descriptive rather than interpretative, and referents to objects are generic rather than specific (Liu et al., 2021). For example, while a caption such as "A crowd of people standing around each other" properly describes the image in Figure 1a, it fails to capture the underlying context that is taking place in this picture, such as "Why are the people standing together, and who are they?". Similarly, in Figure 1b, the generic caption "A woman holding a trophy in front of a crowd" offers a basic description but omits the specific significance conveyed by the news caption "Maren Mjelde won the Women's Super League in her final season with Chelsea," such as the identity of the woman and the event's importance. Consequently, news image captioning models adopt distinct approaches compared to generic image captioning models, often integrating news content as an input to better contextualise the visual information.

News Image Caption: Michelle O'Neill attended the Belfast ceremony alongside Deputy First Minister Emma Little-Pengelly of the Democratic Unionist Party (DUP).

Generic Image Caption: A crowd of people standing around each other.

News Image Caption: Maren Mjelde won the Women's Super League in her final season with Chelsea.

Generic Image Caption: A woman holding a trophy in front of a crowd.

Figure 1: Comparison of news and generic image captions for two different images. The generic image captions were generated by BLIP (Li et al., 2022).

Over the years, researchers have introduced several benchmark datasets, such as VISUAL NEWS (Liu et al., 2021), NYTIMES800K (Tran et al.,

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9008–9017
11–16 May 2026.

	Arabic	Chinese	English	French	Hindi	Indonesian	Japanese	Sinhala	Urdu	All
Family	Afro-Asiatic (Semitic)	Sino-Tibetan*	Indo-European (Germanic)	Indo-European (Romance)	Indo-European (Indo-Aryan)	Austronesian	Japonic†	Indo-European (Indo-Aryan)	Indo-European (Indo-Aryan)	—
|Train|	5,119	9,389	79,195	10,247	12,566	12,137	7,641	2,418	6,602	145,314
|Test|	999	999	1,000	999	1,000	1,000	1,000	998	998	8,993
Unique Articles	2,289	2,922	38,558	2,853	2,760	1,952	3,805	1,046	2,478	58,663
Avg Images/Article	2.67	3.56	2.08	3.94	4.92	6.73	2.27	3.27	3.07	3.61
Avg Content Tokens	1,010	1,519	461	1,510	1,968	1,794	1,287	1,194	1,792	1,412
Avg Caption Tokens	12.70	17.28	13.66	17.23	12.32	16.24	20.63	16.39	15.68	15.98
Avg Title Tokens	10.85	14.38	7.76	14.23	14.03	14.35	18.12	10.16	18.28	13.68

Table 1: Dataset statistics for MUNIChus across nine languages showing language family, number of images in train/test splits, unique articles, and average token counts for images per article, content, captions, and titles. Languages are categorized as high-resource, mid-resource, and low-resource following Joshi et al. (2020). *Chinese tokenised using Jieba. †Japanese tokenised using MeCab.

2020), and GoodNews (Biten et al., 2019) to perform news image captioning. However, these datasets are exclusively focused on English, limiting the scope of existing news image captioning models, which have been trained and evaluated solely in English, hindering progress in adapting such models to other languages. While techniques such as multilingual (Ramos et al., 2023; Kim et al., 2023; Xu et al., 2023b) and cross-lingual learning (Wu et al., 2023b; Chen et al., 2021; Zhang et al., 2024) have shown promise in improving generic image captioning, particularly for low-resource languages, they remain largely unexplored in the news domain due to the absence of multilingual datasets.

To address this gap, our work introduces **MUNIChus**, the first multilingual news image captioning benchmark. MUNIChus comprises more than $700,000$ news images, each accompanied by the news article, article headline and a corresponding caption. MUNIChus spans over 9 languages, including several low-resource languages such as Sinhala (De Silva, 2019; Hettiarachchi et al., 2024; Ranasinghe et al., 2025b) and Urdu . The **main contributions** of this work can be summarised as:

1. We **release MUNIChus**, the first multilingual news image captioning benchmark[1]. This is the largest publicly available news image captioning dataset, covering over 700,000 news images across diverse language families. MUNIChus compromises of 9 languages including three low-resource languages according to Joshi et al. (2020).

2. We **evaluate** more than 20 models on MUNIChus, including state-of-the-art MLLMs and generic image captioning models. We **show** that multilingual news image captioning remains a challenging task despite advances in MLLMs, even in high-resource languages.

[1]The dataset is available in HuggingFace following https://huggingface.co/datasets/tharindu/MUNIChus

2. MUNIChus: Multilingual News Image Captionning Benchmark

We first present the data collection methodology we used, followed by a detailed statistical analysis of MUNIChus.

MUNIChus comprises images, corresponding news articles, captions, news article headlines and associated metadata obtained from the British Broadcasting Corporation (BBC). We selected BBC as the data source due to its rigorous editorial charter. Moreover, the BBC's extensive international presence ensures comprehensive coverage across a wide range of topics and languages. Owing to these qualities, BBC news content has also been widely utilised in the construction of several NLP datasets such as XL-Sum (Hasan et al., 2021).

We developed a Python-based scraper to collect BBC news articles in the languages listed in Table 1, published before December 31, 2024. To ensure data quality, we first removed images with a height or width of less than 180 pixels. Next, we retained only those examples whose captions contain more than three words. While our experiments utilise only images and articles, the MUNIChus also provides additional metadata, such as article titles, which can be used in future studies.

As shown in Table 1, the final dataset consisted of 145,314 training images and 8,993 test images across nine languages, sourced from 58,663 unique BBC news articles. English contributes the most to the corpus, with over $79,000$ training images and $38,000$ unique articles. The average content length varies widely across languages, ranging from 461 tokens in English to $1,968$ in Hindi. Furthermore, the average number of images per article also shows considerable variation, with Indonesian having the highest density at 6.73 images per article, while English has the lowest at 2.08 images per article. The corpus exhibits diversity in caption and title lengths as well, with Japanese having the longest captions (20.63 tokens on average) and titles (18.12 tokens on average), while English has the shortest titles (7.76 tokens on average).

```
You are writing a caption for a newspaper image.
Given the image and this news article excerpt:
[News content]
Task: Write a concise, informative caption for this image in {language}.
Guidelines:
- Write in {language} language only
- Keep it brief
- Identify and include: people's names, locations, and organisations
- Connect what you see in the image to the news context
- Use journalistic style (factual, clear, objective)
- Focus on the main subject of the image
Caption in {language}:
```

Figure 2: Prompt template used for zero-shot image captioning. The {language} placeholder is replaced with the target language name (e.g., "English", "Arabic") at inference time.

2.1. Evaluation

We employed BLEU-4 (Papineni et al., 2002) and CIDEr (Oliveira dos Santos et al., 2021) as the primary evaluation metrics for assessing models on MUNIChus, comparing each generated caption against its corresponding reference caption. These metrics have been widely adopted in previous news image captioning studies (Liu et al., 2021; Qu et al., 2024). Although more advanced metrics such as BERTScore (Zhang* et al., 2020) and BLEURT (Sellam et al., 2020) have been proposed for text generation tasks, they lack support for Sinhala and Urdu. Since our goal was to provide a comprehensive evaluation across all languages in MUNIChus, we restricted our evaluation to BLEU-4 and CIDEr. Additionally, prior work (Liu et al., 2021; Qu et al., 2024) has proposed entity retrieval metrics that assess the overlap of named entities between generated and reference captions. While this approach offers valuable insights into factual consistency, its implementation requires robust named entity recognition (NER) models. Given the limited availability and suboptimal performance of NER systems for low-resource languages such as Sinhala and Urdu, particularly within the news domain, we did not incorporate this metric in our evaluation framework.

3. Methodology

Following recent advances in NLP, we evaluated several MLLMs for generating news image captions on MUNIChus using two approaches: (i) prompt-based generation (§3.1) and (ii) instruction fine-tuning (§3.2) described below.

For Chinese and Japanese, we applied language-specific word segmentation before computing BLEU-4 (Papineni et al., 2002) and CIDEr (Oliveira dos Santos et al., 2021) scores, as these languages lack explicit word boundaries. We used Jieba for Chinese word segmentation, and MeCab for Japanese morphological analysis. Both model predictions and reference captions were tokenised before metric calculation to enable proper word-level n-gram matching.

3.1. Prompt-based Generation

We used the following three different prompts to generate the image captions.

3.1.1. Zero-shot

In the zero-shot setting, as illustrated in Figure 2, models received only the task instruction and the news article context without any example captions. The prompt explicitly specified the target language and provided guidelines for generating journalistic captions, including instructions to keep captions concise, identify key entities (people, locations, organisations), maintain factual accuracy, and connect visual content to the news context. This approach tests the model's ability to perform multilingual image captioning solely from its pre-training, without task-specific data.

3.1.2. Random Few-shot

For the random few-shot condition, we augmented each query with three randomly sampled examples from the same language's training set. Each example consisted of an image, its corresponding news article excerpt, and the reference caption. These examples were prepended to the test query to provide the model with in-context learning examples. The random selection ensures that examples are representative of the overall distribution but may not be semantically related to the test instance.

3.1.3. Similar Few-shot

In the similar few-shot setting, we selected three examples that were most semantically similar to the test instance. We computed similarity by encoding images using the `nomic-embed-vision-v1.5` vision encoder (Nussbaum et al., 2024) and retrieving the three training instances with the highest cosine similarity to the test image embedding. This retrieval-augmented approach provides more contextually relevant examples than random selection, potentially enabling better adaptation to the specific characteristics of each test instance through in-context learning, and has been shown to produce better results in many tasks (Xia et al., 2024; Mathur et al., 2024; Basnet et al., 2025). The retrieved examples, along with their corresponding news articles and reference captions, were prepended to the test query following the same format as the random few-shot condition.

3.2. Instruction Fine-tuning

We fine-tuned `aya-vision-8b` and `Llama-3.2-11B-Vision-Instruct` using TRL's `SFTTrainer` under a QLoRA+LoRA setup. For both models, we load the base backbone in 4-bit `NF4` quantisation with double quantisation and `bf16` computation, then prepare it for k-bit training before injecting the LoRA adapters. We configure LoRA with rank $r = 64$, $\alpha = 32$, dropout $= 0.1$, and `bias=none`, and apply it to the standard attention and MLP projection modules (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`).

For data preparation, we use model-specific collators to construct chat-formatted multimodal training instances. These collators mask user-side prompt tokens and padding tokens with -100 so that the training loss is computed only over assistant-side target tokens. Training uses `bf16` precision with gradient checkpointing, no sequence packing, and a maximum sequence length of 4096 tokens.

We run optimisation for one epoch with a per-device batch size of 1 and gradient accumulation of 16. We use fused AdamW with a learning rate of $1.5e-4$, cosine learning-rate decay, a warmup ratio of 0.03, weight decay of $1e-6$, and gradient clipping at 0.3. We save checkpoints every 500 steps (retaining the two most recent checkpoints) and log training metrics every 50 steps. After training, we save the learned LoRA adapters and merge them into the corresponding base model, yielding a standalone checkpoint for deployment.

3.3. Baselines

We also employed two general image captioning baselines. We fed the image and generated English captions using BLIP (Li et al., 2022), then translated them into the required language using NLLB (Costa-Jussà et al., 2022). As the second baseline, we used `PaliGemma-3b`, a state-of-the-art multilingual image captioning model (Beyer et al., 2024).

4. Results

Table 2 shows the results for the approaches above when evaluated in the `MUNIChus` test sets. Overall, instruction fine-tuning substantially outperforms all prompting-based strategies. Fine-tuned `Llama-3.2-11B` achieves the highest overall BLEU-4 score of 8.40, while fine-tuned `Aya-vision-8b` achieves the highest overall CIDEr score of 56.34, both representing more than a twofold improvement over the best prompting results. The gains are particularly good for certain languages: Hindi reaches a CIDEr score of 100.12 and Japanese 92.56 under fine-tuned `Aya-vision-8b`, compared to 91.74 and 21.83, respectively, for the best prompting approach. Fine-tuning also yields the best results across the majority of individual languages in both metrics.

Among prompting-based strategies, `GPT-4o` with random few-shot learning performs best, with an average BLEU-4 score of 3.57 and CIDEr score of 36.17, delivering the highest prompting-based results in five out of nine languages across both metrics. The zero-shot approach using `Aya-vision-32b` follows as the second-best prompting method, performing best in three languages but showing weaker results in low-resource settings. However, both approaches fall considerably short of the fine-tuned models. For qualitative comparison, Figures 3a and 3b present several example captions generated by `GPT-4o` under the random few-shot setting with the corresponding image and reference caption.

The overall results fall within ranges consistent with those observed in previous news image captioning research (Liu et al., 2021), where BLEU-4 scores typically range from 2 to 4 and CIDEr scores from 20 to 50. However, these modest scores indicate that news image captioning remains a challenging task, particularly given its domain-specific nature.

We describe six key findings (F) next.

> **F1:** *Traditional Image Captioning Models Demonstrate Dramatic Performance Gap*

Both of our traditional image captioning models performed poorly in multilingual news image

	High-resource			Mid-resource			Low-resource				High-resource			Mid-resource			Low-resource			
	En	Fr	Zh	Ar	Hi	Ja	Id	Si	Ur	All	En	Fr	Zh	Ar	Hi	Ja	Id	Si	Ur	All
§3.1.1 Zero-shot																				
Aya-vision-8b	2.52	2.19	2.19	1.53	2.20	4.39	2.75	0.33	1.30	2.16	28.13	14.71	17.45	16.78	18.23	24.61	16.89	6.72	12.84	17.37
Aya-vision-32b	3.13	**3.05**	**2.47**	1.99	2.52	6.83	3.02	0.13	1.95	2.78	35.55	18.12	25.13	22.42	25.17	46.16	24.94	9.00	20.28	25.20
GPT-4o	2.54	1.71	1.48	2.52	3.54	3.26	2.36	1.15	2.18	2.31	39.64	11.31	22.29	29.21	39.13	28.18	24.21	16.42	26.19	26.25
Llama-3.2-11B	2.36	2.41	0.94	1.09	1.77	3.41	2.16	0.24	1.64	1.78	19.37	12.41	8.69	11.59	17.19	16.95	13.39	4.62	15.61	13.31
Phi-3.5-vision	2.27	2.21	1.88	1.83	2.63	4.03	2.55	0.49	1.97	2.21	25.21	13.92	19.98	18.76	22.29	26.94	18.45	8.15	17.12	18.87
Qwen2.5-VL-7B	2.62	2.32	2.15	1.98	2.88	5.69	2.70	0.62	2.44	2.61	32.69	14.08	28.68	24.25	28.71	48.46	18.05	12.59	22.46	25.72
Qwen3-VL-8B	3.30	2.30	2.35	2.00	2.99	4.61	2.49	0.78	2.43	2.59	41.75	15.92	**33.80**	24.46	32.18	50.60	23.46	14.58	27.65	29.38
§3.1.2 Random Few-shot																				
Aya-vision-8b	2.53	1.91	2.02	1.50	2.13	4.20	2.93	0.56	1.44	2.10	27.93	12.51	17.86	16.04	18.23	23.79	17.55	9.25	16.99	16.90
Aya-vision-32b	2.02	1.84	1.63	1.56	1.84	3.35	2.20	0.34	1.66	1.82	32.75	13.59	21.92	18.12	20.25	31.92	19.03	9.84	21.92	21.04
GPT-4o	3.90	2.19	1.20	**2.55**	**10.28**	3.05	5.08	**1.24**	2.68	3.57	52.59	18.90	17.11	31.61	91.74	21.83	37.92	19.04	32.82	36.17
Llama-3.2-11B	2.14	1.97	0.47	0.48	1.01	1.99	2.29	0.10	0.68	1.23	23.42	12.17	3.27	6.11	10.29	5.90	13.91	2.53	6.97	8.78
Phi-3.5-vision	2.75	2.10	1.77	1.94	2.81	4.37	2.84	0.70	2.11	2.38	29.17	14.85	18.97	20.42	24.31	29.45	19.98	10.52	19.15	20.54
Qwen2.5-VL-7B	2.94	2.45	1.99	1.86	3.35	4.90	2.66	0.67	2.59	2.77	38.76	16.97	25.37	22.21	32.04	43.76	21.29	11.13	24.62	24.86
Qwen3-VL-8B	3.62	2.48	1.42	2.32	3.74	4.96	3.16	0.96	2.80	2.83	44.71	17.69	27.05	26.52	35.04	47.98	25.51	15.42	29.08	27.78
§3.1.3 Similar Few-shot																				
Aya-vision-8b	2.70	2.17	1.94	1.56	2.30	4.19	2.86	0.64	1.66	2.25	28.63	14.58	19.08	16.60	20.08	26.14	19.16	9.45	16.69	18.93
Aya-vision-32b	3.30	2.70	2.23	2.28	1.92	5.03	4.77	0.43	2.17	3.11	21.86	18.36	25.22	24.05	46.22	38.47	32.75	12.12	23.11	21.13
GPT-4o	2.95	1.94	1.30	2.25	4.65	2.53	2.64	0.89	1.85	2.33	46.57	15.33	20.30	25.21	48.38	19.12	23.19	16.68	23.17	26.44
Llama-3.2-11B	2.24	2.07	0.80	1.15	1.84	3.37	2.42	0.36	1.73	1.80	22.84	13.30	8.99	12.65	18.16	17.61	14.76	6.24	16.60	14.79
Phi-3.5-vision	2.84	2.39	1.86	2.04	3.00	4.51	2.99	0.80	2.27	2.56	31.29	16.14	21.82	21.92	26.05	30.45	21.56	12.02	20.63	21.77
Qwen2.5-VL-7B	3.10	2.58	2.14	1.99	3.68	5.11	2.87	0.83	2.76	2.87	40.19	18.16	26.50	23.95	34.13	44.49	23.11	13.02	25.49	25.78
Qwen3-VL-8B	3.74	2.65	1.55	2.49	4.06	5.16	3.48	1.12	3.04	3.07	46.13	19.08	28.89	27.34	37.79	49.35	26.69	16.32	30.09	29.87
§3.2 Instruction Fine-tuning																				
Aya-vision-8b	**7.40**	2.99	**6.68**	2.61	8.50	20.18	6.23	0.66	2.21	8.37	**78.27**	**23.60**	**67.52**	**32.66**	**100.12**	**92.56**	**39.30**	10.19	**37.92**	**56.34**
Llama-3.2-11B	**7.40**	**3.20**	5.10	**2.63**	4.15	**21.43**	**6.57**	0.74	**3.73**	**8.40**	71.59	22.57	39.26	23.82	42.40	78.21	29.92	11.50	34.83	44.77
§3.3 Baselines																				
BLIP + NLLB	0.16	0.12	0.07	0.14	0.26	0.61	0.07	0.04	0.46	0.20	5.90	3.34	2.78	3.29	5.18	6.41	3.49	1.39	4.50	4.03
Paligemma-3b	0.17	0.20	0.15	0.03	0.15	0.56	0.02	0.01	0.00	0.03	3.48	2.67	1.75	0.66	1.78	5.33	0.40	0.07	0.00	1.25

Table 2: BLEU-4 (left-side) (Papineni et al., 2002) and CIDEr (right-side) (Oliveira dos Santos et al., 2021) evaluation scores across different languages. The best result for each language (any method) is in **bold**.

News Image Caption:
Patricia Tomlinson and her friends arranging flowers at Oak Court in Blaby.

Generated Caption:
Residents at Oak Court enjoy a flower-arranging class in Blaby.

News Image Caption:
Joseph and Thomas are walking because their mum used to be a nurse at the hospice.

Generated Caption:
Joseph and Thomas tackle Val De Terres hill for Les Bourgs Hospice fundraiser.

Figure 3: Comparison of the actual news image caption and the generated caption by the best model - `GPT-4o` random few-shot approach.

captioning, demonstrating that they cannot generate contextual captions for news images. All languages achieved BLEU-4 scores below 0.7 in both baselines, with most below 0.3. The `BLIP + NLLB` approach averages only 0.20 BLEU-4 overall, while `Paligemma-3b` performs even worse at 0.03 BLEU-4, including complete failures on languages like Urdu. These results suggest the need for specific models and architectures for news image captioning.

F2: *Low-Resource Languages Exhibit High Cross-Model Variability*

Low-resource languages like Indonesian, Sinhala, and Urdu exhibit highly inconsistent performance across models in news image captioning, with some architectures handling them reasonably well while others struggle largely. For instance, `GPT-4o` achieves 5.08 BLEU-4 for Indonesian in random few-shot, while `Llama-3.2-11B` manages only 0.10 for Sinhala in the same approach, and performance gaps of 5–10x between best and worst models are common for these languages. Instruction fine-tuning reduces this variability to some extent, with both fine-tuned models achieving comparable scores for most languages, though the gap between high-resource and low-resource languages persists. This high variability under prompting suggests that model-specific factors such as training data composition, tokenisation strategies, architectural inductive biases, and the presence of news content in specific languages during pretraining play crucial roles in determining low-resource language performance.

F3: *Sinhala Shows Consistently Lowest Performance*

Sinhala performs the worst across all models

News Image Caption:
The Elizabeth line has opened up more direct journey options across London since it opened in May 2022.

News Image Caption:
TfL says there is an increase in housing growth and employment near stations.

Similarity:
0.973

News Image Caption:
Tube services will run as originally planned after industrial action was suspended.

Similarity:
0.953

News Image Caption:
The new Metros have linear-style seating and more space for pushchairs and wheelchairs.

Similarity:
0.934

Figure 4: Comparison of similar images retrieved for the similar shot approach. The first image is the test instance, and the rest of the images are the most similar images retrieved from the training set.

and experimental conditions on `MUNIChus`, with BLEU-4 scores consistently below 1.2 and CIDEr scores rarely exceeding 17 under prompting-based approaches. Even with instruction fine-tuning, Sinhala remains the weakest language by a large margin, achieving only 0.66–0.74 BLEU-4 and 10.19–11.50 CIDEr, while other languages see dramatic improvements. This substantial and persistent performance gap suggests severe underrepresentation of Sinhala in the pre-training data of the experimented MLLMs, particularly in news-related content, which may include domain-specific vocabulary and cultural references. The consistently poor results across diverse model architectures, prompting strategies, and even after fine-tuning indicate that Sinhala requires targeted intervention, potentially through a dedicated collection of a multimodal news corpus.

F4: *Model Size Does Not Guarantee Superior Performance*

The relationship between model size and performance on news image captioning appears non-linear and context-dependent. The larger `Aya-vision-32b` model actually underperforms its smaller 8b counterpart in the majority of the languages, most notably in the random few-shot scenarios. This pattern is further reinforced by the instruction fine-tuning results, where `Llama-3.2-11B` achieves comparable or superior BLEU-4 scores to fine-tuned `Aya-vision-8b` despite being among the weakest models under prompting. This finding has important practical implications for deploying vision-language models in this task, as it suggests that smaller, more efficient models may sometimes be preferable to larger alternatives, particularly when task-specific fine-tuning is feasible.

F5: *No Clear Winner from the Prompting Approaches*

The three prompting strategies yielded mixed results, with no clear indication that any is beneficial for news-image-captioning. Particularly, previous research in MLLMs has shown that a similar few-shot approach improves the results (Ramos et al., 2023). We investigated this issue with the English dataset in `MUNIChus`. As illustrated in Figure 4, although the retrieved training images are visually similar to the test instance, they provide limited contextual information relevant to caption generation. These observations indicate that few-shot prompting does not confer substantial benefits in news image captioning.

F6: *Instruction Fine-tuning Dramatically Outperforms Prompting Strategies*

Instruction fine-tuning yields substantial improvements over all prompting-based approaches, with fine-tuned `Llama-3.2-11B` and `Aya-vision-8b` achieving average BLEU-4 scores of 8.40 and 8.37, and CIDEr scores of 44.77 and 56.34, respectively, more than doubling the best prompting results. The gains are consistent across resource levels, with particularly striking improvements in mid-resource languages such as Hindi (100.12 CIDEr) and Japanese (92.56 CIDEr) under fine-tuned `Aya-vision-8b`. Notably, even relatively small models benefit considerably from fine-tuning: `Llama-3.2-`

11B, which ranked among the weakest models under prompting, becomes highly competitive after fine-tuning, reinforcing finding F4 that model size alone is not the determining factor. However, the improvements for Sinhala remain modest (10.19–11.50 CIDEr), consistent with finding F3, suggesting that fine-tuning alone cannot fully compensate for severe underrepresentation in pre-training data. These results demonstrate that task-specific adaptation through instruction fine-tuning is far more effective than in-context learning for multilingual news image captioning.

Overall, our results suggest that news image captioning is a challenging task even for state-of-the-art MLLMs. Interestingly, current multimodal benchmarks do not contain news image captioning (Li et al., 2024). With the release of MUNIChus, we encourage the research community to develop specialised architectures and training strategies that can better leverage contextual information from news articles to generate accurate, entity-rich captions across diverse languages. The persistent challenges observed across all models, particularly for low-resource languages and the limited effectiveness of few-shot prompting, highlight the need for novel approaches that go beyond general-purpose vision-language modelling to address the unique requirements of news image captioning.

5. Conclusion

This research presented MUNIChus, the first multilingual news image captioning benchmark, comprising 145,314 training images and 8,993 test images across nine languages. Unlike existing news image captioning datasets that focus primarily on English, MUNIChus addresses the critical gap in multilingual news image captioning by providing diverse linguistic representation, including several low-resource languages such as Sinhala, Indonesian, and Urdu. We evaluated several state-of-the-art multimodal large language models on MUNIChus using prompt-based generation and instruction fine-tuning approaches.

The results show that news image captioning remains a challenging task across all languages. Notably, traditional image captioning models demonstrate dramatic performance gaps, highlighting the domain-specific nature of news image captioning. Furthermore, low-resource languages, particularly Sinhala, exhibit consistently poor performance across all models, suggesting severe underrepresentation in pre-training corpora and highlighting the need for multilingual resources.

In future work, we plan to expand MUNIChus to include additional languages and explore specialised architectures tailored for news image captioning. We also plan to create a comprehensive benchmark suite for multilingual news understanding tasks, incorporating MUNIChus alongside other news-related challenges. We release MUNIChus as an open-access, publicly available dataset alongside trained models[2] and evaluation scripts to facilitate ongoing research in multilingual multimodal understanding.

6. Acknowledgement

Hansi Hettiarachchi is partially supported by the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

We acknowledge the EuroHPC Joint Undertaking for awarding us access to Leonardo at CINECA, Italy, to run some LLM experiments.

Some of the experiments reported in this paper were conducted on the MeluXina high-performance computing infrastructure, an allocation granted by the University of Luxembourg on the EuroHPC supercomputer hosted by LuxProvide.

7. Ethics Statement

MUNIChus was collected from publicly available BBC news articles and images, and none of the records were edited in the process. The BBC content is publicly accessible, and our scraping adhered to the site's terms of service. For every instance in MUNIChus, we provide the URL to the original news article, ensuring proper attribution and enabling verification of the source material. We release MUNIChus under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License, which prevents users from altering the instances in the dataset and ensures appropriate attribution to the original sources.

The dataset may reflect inherent biases present in news media, including potential geographical, cultural, or topical imbalances in coverage across languages. While we selected the BBC as our data source due to its editorial charter enforcing principles of political impartiality and balanced reporting, we acknowledge that no news source is entirely free from bias. Researchers using MUNIChus should be mindful of these potential biases and consider their implications for downstream applications, particularly when deploying models trained on this dataset in real-world contexts.

[2]Fine-tuned Aya-vision-8b model is available at https://huggingface.co/alita9/xl-munichus-CohereLabs-aya-vision-8b and the fine-tuned Llama-3.2-11B model is available at https://huggingface.co/alita9/xl-munichus-meta-llama-Llama-3.2-11B-Vision-Instruct

8. Bibliographical References

Lakshita Agarwal and Bindu Verma. 2024. From methods to datasets: A survey on image-caption generators. *Multimedia Tools and Applications*, 83(9):28077–28123.

Saroj Basnet, Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, and Marcos Zampieri. 2025. Evaluating open-source vision-language models for multimodal sarcasm detection. *arXiv preprint arXiv:2510.11852*.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai. 2024. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*.

Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12467.

Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, and Rita Cucchiara. 2024. The revolution of multimodal large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand. Association for Computational Linguistics.

Aozhu Chen, Xinyi Huang, Hailan Lin, and Xirong Li. 2021. Towards annotation-free evaluation of cross-lingual image captioning. In *Proceedings of the 2nd ACM International Conference on Multimedia in Asia*, MMAsia '20, New York, NY, USA. Association for Computing Machinery.

Marta R Costa-Jussà, James Cross Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Fatemeh Daneshfar, Ako Bartani, and Pardis Lotfi. 2024. Image captioning by diffusion models: A survey. *Engineering Applications of Artificial Intelligence*, 138:109288.

Ali Daud, Wahab Khan, and Dunren Che. 2017. Urdu language processing: a survey. *Artificial Intelligence Review*, 47(3):279–311.

Nisansa De Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, IJCAI '24.

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. 2023. Deep learning approaches on image captioning: A review. *ACM Comput. Surv.*, 56(3).

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Hansi Hettiarachchi, Damith Premasiri, Lasitha Randunu Chandrakantha Uyangodage, and Tharindu Ranasinghe. 2024. NSina: A news corpus for Sinhala. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12307–12312, Torino, Italia. ELRA and ICCL.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Yongil Kim, Yerin Hwang, Hyeongu Yun, Seunghyun Yoon, Trung Bui, and Kyomin Jung.

2023. PR-MCS: Perturbation robust metric for MultiLingual image captioning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12237–12258, Singapore. Association for Computational Linguistics.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, CAICE '24, page 405–409, New York, NY, USA. Association for Computing Machinery.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham. Springer International Publishing.

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.

Puneet Mathur, Zhe Liu, Ke Li, Yingyi Ma, Gil Karen, Zeeshan Ahmed, Dinesh Manocha, and Xuedong Zhang. 2024. DOC-RAG: ASR language model personalization with domain-distributed co-occurrence retrieval augmentation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5132–5139, Torino, Italia. ELRA and ICCL.

Zach Nussbaum, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed vision: Expanding the latent space. *arXiv preprint arXiv:2406.18587*.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. 2021. CIDEr-R: Robust consensus-based image description evaluation. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 351–360, Online. Association for Computational Linguistics.

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

Tingyu Qu, Tinne Tuytelaars, and Marie-Francine Moens. 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 (Volume 1: Long Papers)*, pages 2927–2943, Mexico City, Mexico. Association for Computational Linguistics.

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.

Rita Ramos, Bruno Martins, and Desmond Elliott. 2023. LMCap: Few-shot multilingual image captioning by retrieval augmented language model prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1635–1651, Toronto, Canada. Association for Computational Linguistics.

Tharindu Ranasinghe, Hansi Hettiarachchi, Constantin Orasan, and Ruslan Mitkov. 2025a. MUSTS: MUltilingual semantic textual similarity benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 331–353, Vienna, Austria. Association for Computational Linguistics.

Tharindu Ranasinghe, Hansi Hettiarachchi, Nadeesha Chathurangi Naradde Vidana Pathirana, Damith Premasiri, Lasitha Uyangodage, Isuri Nanomi Arachchige, Alistair Plum, Paul Rayson, and Ruslan Mitkov. 2025b. Sinhala encoder-only language models and evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 8623–8636, Vienna, Austria. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Silvia Cascianelli, Giuseppe Fiameni, and Rita Cucchiara. 2023. From Show to Tell: A Survey on Deep Learning-Based Image Captioning . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 45(01):539–559.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13032–13042.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023a. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256.

Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. 2023b. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, Toronto, Canada. Association for Computational Linguistics.

Peng Xia, Kangyu Zhu, Haoran Li, Hongtu Zhu, Yun Li, Gang Li, Linjun Zhang, and Huaxiu Yao. 2024. RULE: Reliable multimodal RAG for factuality in medical vision language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1081–1093, Miami, Florida, USA. Association for Computational Linguistics.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. 2025. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117:102888.

Liming Xu, Quan Tang, Jiancheng Lv, Bochuan Zheng, Xianhua Zeng, and Weisheng Li. 2023a. Deep image captioning: A review of methods, trends and future challenges. *Neurocomputing*, 546:126287.

Yueyuan Xu, Zhenzhen Hu, Yuanen Zhou, Shijie Hao, and Richang Hong. 2023b. Cite: Compact interactive transformer for multilingual image captioning. In *Proceedings of the 2023 6th International Conference on Image and Graphics Processing*, ICIGP '23, page 175–181, New York, NY, USA. Association for Computing Machinery.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Jing Zhang, Dan Guo, Xun Yang, Peipei Song, and Meng Wang. 2024. Visual-linguistic-stylistic triple reward for cross-lingual image captioning. *ACM Trans. Multimedia Comput. Commun. Appl.*, 20(4).

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

9. Language Resource References

Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12458–12467.

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.

Alasdair Tran, Alexander Mathews, and Lexing Xie. 2020. Transform and tell: Entity-aware news image captioning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13032–13042.