future internet

MDPI

Article

Detailed Image Captioning and Hashtag Generation

Nikshep Shetty and Yongmin Li *

Department of Computer Science, Brunel University London, Uxbridge UB8 3PH, UK; nikshep.shetty2199@gmail.com
* Correspondence: yongmin.li@brunel.ac.uk

Abstract: This article presents CapFlow, an integrated approach to detailed image captioning and hashtag generation. Based on a thorough performance evaluation, the image captioning model utilizes a fine-tuned vision-language model with Low-Rank Adaptation (LoRA), while the hashtag generation employs the keyword extraction method. We evaluated the state-of-the-art image captioning models using both traditional metrics (BLEU, METEOR, ROUGE-L, and CIDEr) and the specialized CAPTURE metric for detailed captions. The hashtag generation models were assessed using precision, recall, and F1-score. The proposed method demonstrates competitive results against larger models while maintaining efficiency suitable for real-time applications. The image captioning model outperforms the base Florence-2 model and favorably compares with larger models. The KeyBERT implementation for hashtag generation surpasses other keyword extraction methods in both accuracy and speed. This work contributes to the field of AI-assisted content analysis and generation, offering insights into the practical implementation of advanced vision-language models for detailed image understanding and relevant tag generation.

Keywords: image captioning; hashtag generation; vision-language models; AI-assisted content analysis

Citation: Shetty, N.; Li, Y. Detailed Image Captioning and Hashtag Generation. *Future Internet* **2024**, *16*, 444. https://doi.org/10.3390/fi16120444

Academic Editor: Paolo Bellavista

Received: 18 October 2024
Revised: 21 November 2024
Accepted: 26 November 2024
Published: 28 November 2024

1. Introduction

In today's digital world, visual content is prevalent on online platforms. While this abundance of images improves user engagement, it also creates challenges for content creators in effectively describing and categorizing this content. Current content management systems often lack built-in tools for generating detailed image descriptions and relevant hashtags, leading to time-consuming manual work and potential inconsistencies. Additionally, while assistive technologies like screen readers exist, they often provide limited descriptions of images, affecting both content creators' workflow and end users' experience.

Content creators, marketers, and social media managers often struggle to produce high-quality, engaging content quickly and consistently. This includes not only creating accurate image descriptions, but also generating relevant hashtags for improved content discoverability. There is a growing need for efficient content management tools in an online environment dominated by visual content.

This paper introduces CapFlow, a novel approach to detailed image captioning and hashtag generation, which combines a fine-tuned vision-language model for detailed image captioning with an efficient keyword extraction technique for hashtag generation, offering a useful solution for content analysis and creation. The key contributions of this work are as follows.

1. We present an integrated solution to both image captioning and hashtag generation.
2. A thorough evaluation was performed on numerous state-of-the-art image captioning and hashtag generation models, and model selection was achieved based on this evaluation.
3. We provide an application of the proposed model by developing a Chrome extension for direct image captioning and hashtag generation.

The rest of this paper is organized as follows. A literature review is provided in Section 2, which is followed by detailed discussions of the proposed model, including both image captioning and hashtag generation in Section 3. Experiments and result analysis are presented in Section 4, and the conclusions are drawn in Section 5.

2. Background

The literature review examines the current research in image captioning and hashtag generation, as well as their use in content creation and web accessibility. It covers the evolution of image captioning techniques, approaches to hashtag generation, applications in social media marketing, and considerations for making web content more accessible.

2.1. Image Captioning

Image captioning, the task of automatically generating natural language descriptions for visual content, has significantly progressed due to deep learning techniques [1]. Early approaches utilized encoder–decoder architectures with convolutional neural networks (CNNs) encoding visual features and recurrent neural networks (RNNs) generating captions [2]. These models, however, often produced simplistic captions lacking detail.

To address this limitation, attention mechanisms were introduced to focus on salient image regions when generating each word, improving caption quality [3]. The "bottom-up and top-down" method proposed by Anderson et al. [4] used Faster R-CNN to propose salient regions, implementing a "hard" attention mechanism. Graph-based methods also emerged, aiming to better capture the relationships between image elements by representing images as scene graphs [5]. The Scene Graph Auto-Encoder (SGAE) introduced by Yang et al. [5] incorporated the inductive bias of language generation into the encoder–decoder framework.

The introduction of Transformer architectures led to substantial improvements in image captioning. The Meshed-Memory Transformer [6] introduced a mesh-like connectivity between encoder and decoder layers, allowing for more nuanced caption generation. X-Linear Attention Networks [7] employ spatial and channel-wise bilinear attention to extract second-order interactions, facilitating more in-depth reasoning. Other Transformer-based models like Grid- and Region-based Image captioning Transformer (GRIT) [8] used both grid- and region-based features, while PureT [9] implemented a fully Transformer-based architecture for end-to-end training.

Developments in vision-language pre-training have shown promising results. Contrastive Language-Image Pre-Training (CLIP) [10] has emerged as a powerful pre-trained model, providing a shared representation for both image and text. ClipCap [11] leverages CLIP embeddings to produce a prefix for each caption, which is then used by a pre-trained language model (GPT-2) to generate captions. Object-Semantics Aligned Pre-training (OSCAR) [12] uses object tags as anchor points to align image and language modalities in a shared semantic space. VinVL [13] and mPLUG [14] further refined these approaches, with mPLUG introducing novel cross-modal skip-connections to improve computational efficiency and address information asymmetry.

Recent research has explored novel architectures and training paradigms. Expansion-Net v2 [15] introduces a Block Static Expansion layer to address potential performance bottlenecks related to input length. COS-Net [16] aims to unify semantic comprehension and ordering in the captioning process. VIVO (Visual VOcabulary pre-training) [17] learns a joint presentation of visual and text input using image-tag pairs for pre-training, allowing for zero-shot generalization to novel visual objects. Cross-modal Generative Pre-Training (XGPT) [18] uses a cross-modal encoder–decoder architecture that is directly optimized for generation tasks.

Despite these advancements, current captioning systems face several limitations. A key issue is the tendency to produce short, generic descriptions rather than detailed analyses of image contents [1]. The problem of object hallucination, where models detect objects not present in the input image, persists [19]. Additionally, most advanced models

require powerful hardware that is unavailable to average users, limiting their accessibility. Newer models like Florence-2 [20] aim to bridge this gap between speed and performance by training their model on high-quality data like FLD-5B, which consists of 5.4 billion comprehensive visual annotations on 126 million images.

The field of image captioning continues to evolve, with researchers exploring various techniques to improve caption quality, efficiency, and applicability. Potential directions for future research include the development of more efficient models that can run on consumer-grade hardware while still producing detailed, high-quality captions. Such advancements could bridge the gap between sophisticated academic models and practical, widely accessible captioning systems, potentially enabling new applications for automated image understanding in everyday scenarios.

2.2. Hashtag Generation

Hashtag generation is an important task for social media content creation and organization. Unlike image captioning, which aims to produce natural language descriptions, hashtag generation focuses on producing concise, relevant tags to categorize and make content discoverable.

Several approaches have been proposed for hashtag generation from visual and textual content. In Gong and Zhang [21], a method for hashtag recommendation for multimodal microblog posts was developed. Their approach uses both visual features from images and textual features from post content to generate relevant hashtags. They employed a learning-to-rank framework to recommend hashtags based on their relevance scores. In Hachaj and Miazga [22], a voting deep neural network combined with associative rules mining for image hashtag recommendations was proposed. Their method uses an ensemble of convolutional neural networks to extract visual features, which are then used to predict relevant hashtags through a voting mechanism. Association rule mining is applied to discover the relationships between hashtags. Some approaches leverage image captioning as an intermediate step. For example, the method proposed by AL-Sammarraie et al. [23] first generates image captions using a CNN-RNN model, then applies natural language processing techniques to extract keywords from the captions and convert them into hashtags. This two-step approach allows leveraging advances in image captioning while tailoring the output for hashtag generation.

Keyword extraction techniques can be adapted for hashtag generation from textual content like captions. As discussed by Nadim et al. [24], there are several approaches to keyword extraction. Statistical methods like TF-IDF, KPMiner, and YAKE use statistical features of the text to identify important keywords. These methods are computationally efficient but may miss semantic relationships. Graph-based methods such as TextRank and PositionRank construct graphs representing word relationships and use graph algorithms to extract key terms. These can capture some semantic information but may struggle with longer texts. More recent deep learning methods like KeyBERT [25] leverage pre-trained language models to extract keywords based on semantic similarity. These can potentially capture deeper semantic relationships but require more computational resources.

There are several limitations to current hashtag generation approaches. Many methods struggle to incorporate broader contextual information beyond the immediate visual or textual content, which can lead to generic or irrelevant hashtags. Hashtag usage on social media evolves rapidly, and static models may fail to capture emerging trends or platform-specific hashtag cultures. User preferences for hashtag style and content vary widely, but most methods do not account for individual user preferences or posting history.

The evaluation of hashtag generation models presents its own challenges. As noted by Nadim et al. [24], evaluating the quality of generated keywords (which can be applied to hashtags) is challenging as relevance can be subjective and context-dependent. Traditional metrics like precision, recall, and F1-score may not fully capture the quality of generated hashtags in a real-world social media context.

Data quality and bias are also significant concerns. Training data for hashtag generation often comes from social media platforms, which can introduce biases and quality issues that affect model performance. Additionally, some of the more advanced deep learning approaches require significant computational resources, limiting their applicability in real-time or mobile scenarios [23]. Addressing these limitations presents opportunities for future research in hashtag generation, particularly in developing more context-aware, personalized, and computationally efficient approaches.

2.3. Use in Content Creation

AI-powered tools for image captioning and hashtag generation can significantly enhance social media content creation, particularly in producing "snackable content"—short, easily digestible pieces designed for social media platforms [26]. This aligns with findings that AI-powered automation in social media marketing can lead to more engaging, relevant, and personalized content [27].

AI technologies can streamline the process of creating snackable content by assisting with watching material, selecting clips, editing content, and uploading for review [26]. By leveraging algorithms to analyze images and generate captions and hashtags, AI can improve efficiency and engagement metrics across multiple stages of content creation.

Research indicates that AI-powered content often outperforms human-created content in terms of likes, shares, comments, and click-through rates [27]. Such tools could help content creators achieve similar improvements, while also facilitating the creation of "interactive content" that is important for audience engagement [26].

However, the implementation of AI in content creation must consider potential challenges such as algorithmic biases and privacy concerns [27]. Additionally, care must be taken to balance the creation of quick, digestible content with the need for depth and meaningful engagement [26].

While AI can enhance efficiency in content creation, human creativity remains crucial. The role of a content creator is to provide engaging, interesting content [26]. Therefore, AI tools should be designed to augment human creativity rather than replace it entirely, ensuring ethical use of AI technology in social media marketing.

Recent advancements in image captioning and hashtag generation have improved content creation, but challenges remain in efficiency, context understanding, and hashtag relevance. Integrating these technologies offers opportunities to enhance accessibility and content optimization, while presenting challenges in model efficiency and ethics. In this work, we introduce CapFlow, an integrated model that employs customized models for real-time image captioning and hashtag generation.

3. Methods

The proposed approach aims to provide integrated solution to both image captioning and hashtag generation, addressing limitations in current content management systems. As shown in Figure 1, our system consists of two main components: the Florence-2 detailed captioning model and the KeyBERT hashtag generator. We start by discussing the image captioning model, which is based on Microsoft's Florence-2. Florence-2 utilizes a DaViT image encoder and multi-modal encoder–decoder structure to process visual and textual information. This section explores our implementation of Low-Rank Adaptation (LoRA) for fine tuning of Florence-2, as illustrated in the diagram's LoRA Adapter block, and it also examines how various caption dataset combinations were evaluated to optimize model performance.

Next, we discuss the text-to-hashtag generation process. This uses KeyBERT to analyze the captions and generate the top three relevant hashtags, as illustrated in the KeyBERT Hashtags section of the diagram. This process includes text pre-processing, BERT word embedding, and cosine similarity calculation. Finally, we combine these components into a working Chrome extension to showcase the capability and use case of our models.

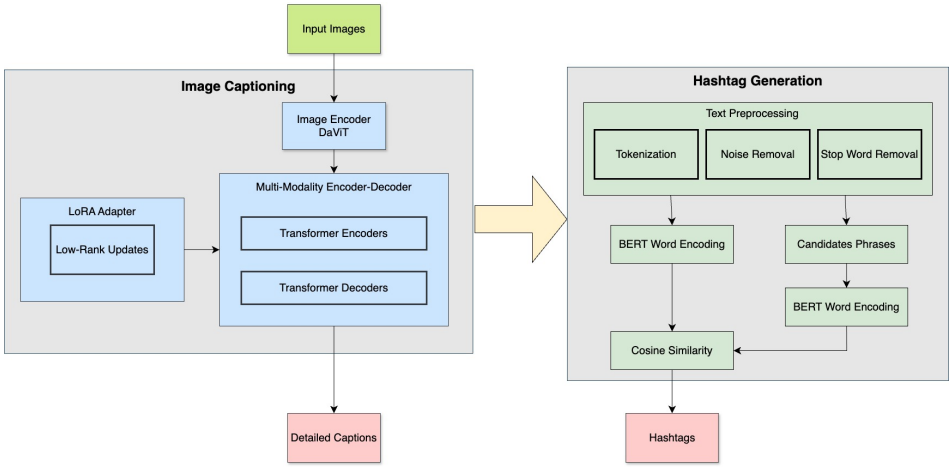


Figure 1. Block diagram: CapFlow image captioning and Hashtag generator.

3.1. Detailed Image Captioning

For image captioning, we experimented with and evaluated a number of state-of-the-art models. These models are discussed below.

Florence-2 is a vision foundation model capable of handling various visual tasks across different spatial hierarchies and semantic granularities. The architecture of Florence-2 consists of two main components: an image encoder and a multi-modality encoder–decoder. The model is trained on a collection of 5.4 billion comprehensive visual annotations on 126 million images, which Xiao et al. [20] call FLD-5B.

The image encoder in Florence-2 utilizes a DaViT (Dual-Attention Vision Transformer) model. Developed by Ding et al. [28], DaViT is a vision transformer that incorporates two complementary types of self-attention to efficiently capture both local and global visual dependencies.

The key components of DaViT's architecture include spatial window attention, channel group attention, alternating attention blocks, and a hierarchical structure. Spatial window attention divides the image into separate windows and applies self-attention within each window, allowing the model to efficiently capture detailed local features. Channel group attention operates on "channel tokens" instead of spatial tokens. Each channel token contains a summary of the entire image, enabling the model to capture overall context naturally. DaViT's design alternates between spatial window attention and channel group attention blocks throughout its structure. This approach helps the model refine local details while maintaining an understanding of the whole image. Additionally, DaViT employs a hierarchical structure with multiple stages. As the image progresses through these stages, the spatial resolution decreases while the feature dimension increases [28].

The DaViT component of Florence-2 transforms an input image $I \in \mathbb{R}^{H \times W \times 3}$ into a set of visual token embeddings $V \in \mathbb{R}^{N_v \times D_v}$, where H and W represent the height and width of the image, N_v is the number of visual tokens, and D_v is their dimensionality [20]. After the image encoding process, the multi-modality encoder–decoder processes both the visual information from DaViT and any text input. This component is based on a standard Transformer architecture. The multi-modality encoder–decoder operates as follows.

1. The visual embeddings V from DaViT undergo linear projection and normalization to obtain $V' \in \mathbb{R}^{N_v \times D}$.
2. Any text input, such as prompts or questions, is converted into embeddings $T_{prompt} \in \mathbb{R}^{N_t \times D}$ using a tokenizer and embedding layer.
3. These visual and text embeddings are combined to form the input for the encoder: $X = [V', T_{prompt}]$.
4. The multi-modality encoder processes this combined input through several Transformer layers to capture relationships between visual and textual elements.

5. Finally, the decoder generates output text based on the encoder's output and any previously generated tokens.

This design allows Florence-2 to handle a wide range of tasks that involve both images and text, such as image captioning, visual question answering, etc. The combination of DaViT's efficient image encoding and the flexible multi-modality processing enables Florence-2 to achieve high performance across various vision-language tasks.

To adapt Florence-2 for detailed image captioning, we employed Low-Rank Adaptation (LoRA) [29], an efficient fine-tuning technique. LoRA allows for the update of a small number of model parameters while maintaining overall model performance. Our implementation of LoRA focuses on adapting the attention mechanisms within the model, facilitating more nuanced caption generation without significantly increasing computational requirements.

The LoRA technique works by representing weight updates during fine tuning as low-rank decompositions. For a pre-trained weight matrix $W_0 \in \mathbb{R}^{d \times k}$, LoRA constrains its update as

$$W = W_0 + \Delta W = W_0 + BA, \tag{1}$$

where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$, and the rank r is much smaller than $\min(d, k)$.

We developed three LoRA adapters through fine tuning the original Florence model using three different datasets: Descriptions of Connected and Contrasting Images (DOCCI), Pixelprose, and Recap DataComp. These specific datasets were selected due to their variety and their capability to strengthen the model's performance in producing detailed, contextually nuanced captions.

Besides Florence-2, we also experimented with and evaluated the alternative models for image captioning.

1. Large Language and Vision Assistant (LLaVA) [30] stands out as a significant development in multimodal modeling. At its core, LLaVA combines a CLIP ViT-L/14 image encoder with a large language model, which is linked through a linear projection layer. The latest iteration, LLaVA-1.6, incorporates the Mistral 7B model as its language component, resulting in a robust 7 billion parameter system. LLaVA's use of Low-Rank Adaptation (LoRA) for fine tuning allows it to efficiently adapt to detailed captioning tasks without compromising overall performance. LLaVA makes use of GPT-4 to generate multimodal language-image instruction following data, enabling impressive chat abilities that sometimes mirror multimodal GPT-4's performance on novel images and instructions. When applied to specific tasks like Science QA, the combination of LLaVA and GPT-4 has set new benchmarks in accuracy.
2. PaliGemma [31] takes a different approach, focusing on versatility and efficiency for deployment on resource-constrained devices. The base PaliGemma-3b-mix-448 model, with its 3 billion parameters, builds on the SigLIP-So400m vision encoder and Gemma-2B language model. Despite its relatively compact size, PaliGemma holds its own against much larger models like Mixtral 8x7B and GPT-3.5. This impressive performance stems from a training regimen involving 3.3 trillion tokens, which is drawn from carefully filtered web data and synthetic sources. PaliGemma's Llama-2-like architecture facilitates easy integration with existing tools. The model family includes variants optimized for multilingual, multimodal, and long-context scenarios. Particularly noteworthy is PaliGemma's ability to run locally on modern phones when quantized to 4 bits, while still delivering strong performance.
3. The Phi-3 model series [32] represents another leap forward in efficient, high-performance language models. The base phi-3-mini model, at just 3.8 billion parameters, competes with much larger models on key benchmarks like MMLU and MT-bench. The Phi-3-vision-128k-instruct variant, slightly larger at 4.2 billion parameters, specializes in multimodal tasks and can handle impressively long context windows of up to 128K tokens. Phi-3's success lies in its "data optimal regime" training approach, which carefully

balances knowledge and reasoning ability. This methodology allows Phi-3 to achieve top-tier performance while remaining compact enough for local deployment on mobile devices. Extensive post-training, including supervised fine tuning and direct preference optimization, further enhances Phi-3's capabilities across various tasks and improves its safety features.

4. CLIP-based models, such as ClipCap [11], offer strong visual–textual alignment but often require separate image encoders and text decoders. While powerful, this architecture can increase computational complexity, potentially limiting real-time applications.
5. BLIP-2 [33], though not directly compared in our results, offers robust performance but with a larger parameter count. This makes it less suitable for our goal of real-time processing in browser environments.

Each of these approaches brings unique strengths to the table. Larger models often excel across a wide range of tasks but at the cost of increased computational demands. CLIP-based models shine in visual–textual alignment but may struggle with generating fluent, detailed captions. Our approach with Florence-2 and LoRA aims to strike a balance between performance and efficiency, which are tailored for real-time, browser-based applications.

3.2. Hashtag Generation Model

To generate hashtags, we utilized KeyBERT [25], which is an unsupervised approach for extracting keywords that makes use of BERT embeddings. The KeyBERT methodology involves three primary phases: extracting candidate keywords, word embedding using BERT, and computing cosine similarities.

In the candidate keyword extraction phase, we utilize Scikit-Learn's Count Vectorizer to obtain a list of n-gram candidates from the input text (in our case, the generated image caption). This step includes the removal of stop words and allows for the adjustment of keyword length to transform them into key phrases.

The BERT embedding step transforms both the input text and the n-gram candidates into numeric data using a pre-trained BERT model. This results in contextual word embeddings that capture semantic relationships within the text.

In the final similarity calculation step, we compute the cosine similarity between the embeddings of the n-gram candidates and the full text. Candidates with the highest similarity scores are selected as the most representative hashtags. The similarity is defined as

$$\text{Similarity} = \cos(w \cdot s), \tag{2}$$

where w is the word's word embedding vector and s is the sentence embedding vector.

This approach differs from traditional frequency-based methods by focusing on the relevance between words in the context of the sentence, utilizing the semantic and contextual information of words and phrases in the extraction process [25].

Several alternative methods for hashtag generation were considered in this study.

1. The Regular BERT [34] approach uses the standard BERT model for keyword extraction without the optimizations introduced in KeyBERT. While it leverages BERT's powerful language understanding capabilities, it can be less efficient and may not be optimized for the specific task of hashtag generation.
2. YAKE (Yet Another Keyword Extractor) [35] is a statistical approach to keyword extraction that does not rely on external corpora or training data. It is often faster than embedding-based methods but may miss semantic relationships captured by more advanced models.
3. Gemma-2-2B-it [36] represents a large language model approach to hashtag generation. While potentially more flexible and capable of understanding complex contexts, such large models can be computationally expensive and slower, making them less suitable for real-time applications.

The various keyword extraction approaches each come with their own advantages and disadvantages when considering three key factors: how well they understand meaning (se-

mantic comprehension), how quickly and efficiently they can process data (computational performance), and how straightforward they are to put into practice (implementation simplicity). After careful consideration, we chose KeyBERT as our solution because it strikes the right balance between deep understanding of text meaning and efficient processing speed. These characteristics make it especially well suited for applications that need to work instantly in web browsers, where users expect quick responses and accurate results.

4. Experiments and Results

4.1. Datasets

In this study, the datasets used for training include the following.

- **DOCCI (Descriptions of Connected and Contrasting Images)** [37]: This dataset consists of 15,000 images with long, human-written descriptions. DOCCI was specifically designed to advance image-to-text tasks, including detailed captioning, by providing contextually rich annotations.
- **Pixelprose** [38]: PixelProse is a comprehensive dataset of over 16 million synthetically generated captions, leveraging cutting-edge vision-language models for detailed and accurate descriptions.
- **Recap DataComp** [39]: Recap-DataComp-1B is a large-scale image–text dataset that has been recaptioned using an advanced LLaVA-1.5-LLaMA3-8B model to enhance the alignment and detail of textual descriptions.

The following datasets were used for evaluation.

- **DetailCaps-4870** [40]: This dataset contains 4870 images from various datasets, which are accompanied by ground truth detail captions generated by GPT-4V, Gemini-1.5-Pro, and GPT-4O for evaluation.
- **Tech Keywords Topics Summary** [41]: This dataset comprises a collection of technical articles with associated keywords, serving as the ground truth for assessing the relevance and accuracy of generated hashtags.

4.2. Experimental Setup

The fine-tuning process for the Florence-2 model was implemented using PyTorch and the Transformers library. The experiments were conducted on Kaggle's platform, utilizing two NVIDIA T4 GPUs. The model was fine tuned separately on the UCSC-VLAA/Recap-DataComp-1B, DOCCI, and Pixelprose datasets, creating three distinct LoRA adapters.

For each dataset, the pre-trained Florence-2-base-ft model and its corresponding processor were initialized. The weights of the vision tower were frozen to focus the fine tuning on the language model component. The LoRA configuration was applied to the key, value, query, and output projection matrices of the attention mechanism. The following hyperparameters were consistently used across all three fine-tuning runs: LoRA rank: 8; LoRA alpha value: 32; Learning rate: 1×10^{-4}; Batch size: 8; Number of epochs: 5.

A custom StreamingDataset class was implemented for efficient handling of the large-scale datasets. This class processed each item by extracting the image URL and caption, downloading the image and applying necessary transformations. Error handling was incorporated to manage the potential issues with data loading or processing.

The training loop utilized the AdamW optimizer with a linear learning rate schedule. Gradient accumulation and mixed precision training were implemented to optimize memory usage and training speed. The model's performance was evaluated on both training and validation sets after each epoch, tracking metrics such as loss.

To ensure reproducibility and facilitate continued training or deployment, the training process was integrated with Hugging Face's model hub. After each epoch, the updated model and processor were pushed to a specified repository for each dataset.

4.3. Evaluation Metrics

Image captioning research employs various metrics to assess the quality of generated captions. The most commonly used metrics include BLEU [42], ROUGE [43], METEOR [44], CIDEr [45], and SPICE [46]. BLEU, a pioneer in evaluating machine-generated texts, compares n-grams in the generated caption with reference captions. It is widely used due to its simplicity, language independence, and comparability with human judgment [42]. ROUGE, originally designed for text summarization, measures the overlap of n-grams, word sequences, and word pairs between generated and reference captions [43]. METEOR attempts to address some limitations of BLEU by incorporating stemming and synonymy matching, aiming to correlate better at the sentence or segment level [44]. CIDEr, which is designed specifically for image captioning, uses term frequency-inverse document frequency (TF-IDF) weighting to encode how often n-grams in the candidate sentence are present in the reference sentences [45]. SPICE employs scene graphs to evaluate semantic propositional content, measuring how well objects, attributes, and the relations between them are covered in image captions [46].

These metrics, while useful for comparative evaluation, have several limitations when assessing detailed captions. Key issues include their lack of semantic understanding, bias towards common phrases, inability to assess factual accuracy, limited evaluation of descriptive richness, and difficulty in assessing coherence and fluency [1]. Most metrics focus on lexical similarity rather than semantic meaning, potentially missing the nuances of detailed descriptions and frequently favoring occurring phrases in the training data.

The limitations of these metrics highlight the need for more sophisticated evaluation methods that can better assess the quality, accuracy, and richness of detailed image captions. As the field of image captioning continues to advance, with models capable of generating increasingly detailed and nuanced descriptions, the development of more robust evaluation metrics becomes crucial [1]. Future research could focus on developing metrics that incorporate deeper semantic understanding and can better evaluate the nuanced aspects of detailed descriptions, potentially bridging the gap between current evaluation methods and the complexities of human-like image understanding and description. One attempt at solving these limitations was proposed by Dong et al. [40] and is called CAPTURE (CAPtion evaluation by exTracing and coUpling coRE information), which is specially designed for detailed caption evaluation and works by extracting visual elements such as objects, attributes, and relations for better feature representation of the image.

CAPTURE calculates precision and recall for each type of visual element, combines these into F1 scores, and produces a final weighted score:

$$CAPTURE Score = \frac{5 \cdot F1_{obj} + 5 \cdot F1_{attr} + 2 \cdot F1_{rel}}{5 + 5 + 2}, \quad (3)$$

where $F1_{obj}$, $F1_{attr}$, and $F1_{rel}$ are the F1 scores for objects, attributes, and relations, respectively.

For the hashtag generation model, performance was evaluated using precision, recall, and F1-score. These metrics assess the relevance and accuracy of the generated hashtags.

- Precision: Measures the proportion of generated hashtags that are relevant.
- Recall: Measures the proportion of relevant hashtags that were successfully generated.
- F1-score: The harmonic mean of precision and recall, providing a balanced measure of the model's performance.

Additionally, the average processing time for each method was measured to assess their suitability for real-time applications.

4.4. Results

4.4.1. Image Captioning Performance

The performance of various configurations of the image captioning model, which were fine tuned using different LoRA adapters, was evaluated and compared to the base Florence-2 model. The results, presented in Table 1, demonstrate the effectiveness of the

LoRA adaptation technique in improving the performance of the Florence-2 model for detailed image captioning.

Table 1. Performance metrics for various configurations of the image captioning model.

Model Configuration	METEOR	BLEU	ROUGE-L	CIDEr	CAPTURE
Base Florence 2 Model	0.2128	0.1100	0.2753	0.0312	0.5458
DOCCI Adapter	**0.2671**	**0.1850**	0.2874	**0.0863**	**0.5757**
Pixelprose Adapter	0.2501	0.1552	**0.2982**	0.0388	0.5554
Recap DataComp Adapter	0.2397	0.1501	0.2942	0.0348	0.5530

The DOCCI adapter showed the most significant improvements across all metrics, with particularly notable gains in the CAPTURE score, which is specifically designed to evaluate detailed image captions. The DOCCI adapter improved the CAPTURE score from 0.5458 (base model) to 0.5757—a 5.48% increase. This improvement suggests that the DOCCI dataset, with its focus on detailed and contrasting image descriptions, provides valuable training data for generating more nuanced captions. The Pixelprose and Recap DataComp adapters also showed improvements over the base model, but to a lesser extent than the DOCCI adapter.

To further assess the performance of the proposed model, it was compared to other state-of-the-art vision-language models (VLMs). The results of this comparison are presented in Table 2.

Table 2. Performance comparison with other vision-language models.

Model	METEOR	BLEU	ROUGE	CIDEr	CAPTURE	AvgTime(s)
Paligemma-3b-mix-448	0.1242	0.0530	0.1947	0.0157	0.4038	1.201
Llava-v1.6-mistral-7b-hf	**0.3346**	**0.2754**	**0.3231**	0.0883	0.5674	17.983
Phi-3-vision-128k-instruct	0.3144	0.2359	0.3230	**0.0955**	0.5509	7.358
Florence-2-base-ft	0.2128	0.1100	0.2753	0.0312	0.5458	**0.523**
Our Model	0.2671	0.1850	0.2874	0.0863	**0.5757**	0.688

These results demonstrate that, despite having significantly fewer parameters (232 million) compared to models like Llava-v1.6-mistral-7b-hf (7 billion) and Phi-3-vision-128k-instruct (4.2 billion), our model achieves competitive performance, particularly in the CAPTURE metric. The proposed model demonstrates remarkable efficiency, with an average processing time of 0.688 s, which is second only to the base Florence-2 model. This is significantly faster than larger models like Llava-v1.6-mistral-7b-hf (17.983 s) and Phi-3-vision-128k-instruct (7.358 s).

Figure 2 provides a visual comparison of the processing times for different models. Both the base Florence-2 model and our adapted model demonstrated remarkably quick processing times below one second. The PaliGemma model, while quick on average, shows some inconsistency with occasional processing times reaching 3 s. The Phi-3-Vision model performed decently in terms of speed but lagged behind our model. The Llava-mistral model, being the largest, was significantly slower and would require more powerful hardware for deployment.

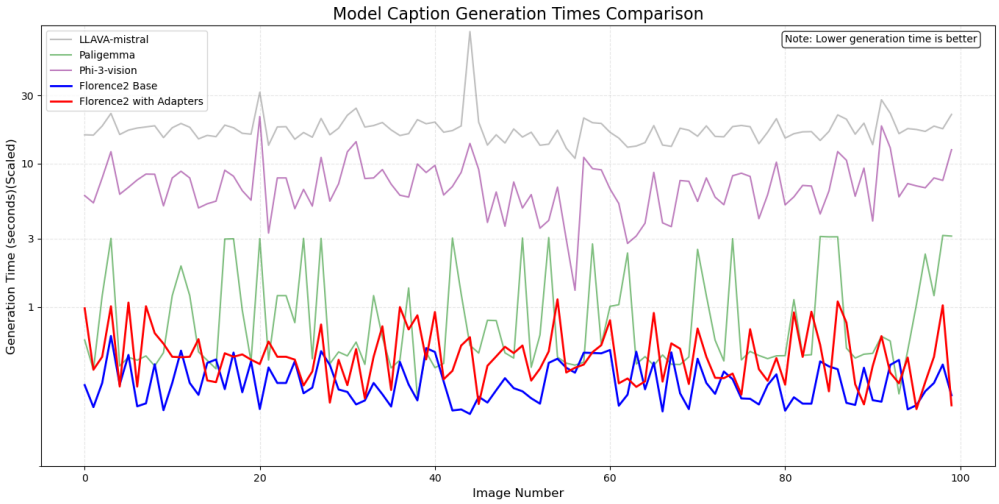


Figure 2. Speed comparison of the different vision-language models.

4.4.2. Hashtag Generation Performance

The performance of the hashtag generation model using KeyBERT was evaluated and compared to other techniques: Regular BERT, YAKE (Yet Another Keyword Extractor), and Gemma-2-2B-it. The results of this comparative analysis are presented in Table 3.

Table 3. Performance comparison of the Hashtag Generation Methods.

Method	Precision	Recall	F1-Score	Avg Time (s)
KeyBERT	**0.6132**	**0.6105**	**0.5932**	0.0201
Regular BERT	0.3392	0.3222	0.3199	0.0107
YAKE	0.5126	0.5068	0.4939	**0.0064**
Gemma-2-2B-it	0.4808	0.4239	0.4537	2.6293

KeyBERT demonstrated the highest precision at 0.6132, indicating that approximately 61% of the hashtags it generated were relevant and accurate (based on the test dataset). It also achieved the best recall score of 0.6105, suggesting that it successfully identified a significant portion of the relevant hashtags. The combination of these metrics resulted in an F1-score of 0.5932, the highest among the four methods. KeyBERT's average processing time was 0.0201 s per input, making it the second fastest method after YAKE.

The precision recall curves in Figure 3 compares these four methods for automated hashtag generation. KeyBERT shows the best performance with an average precision (AP) of 0.791, maintaining consistent precision across recall values. Being specifically optimized for keyword extraction, KeyBERT effectively translates this strength to hashtag generation. Gemma (AP = 0.765) achieves strong precision in early recall regions but shows noticeable decline after mid-range recall. YAKE (AP = 0.721) demonstrates reliable performance as a non-neural method, benefiting from its keyword extraction optimization. Regular BERT (AP = 0.608) performed notably worse due to not being optimized for such a task. The results indicate that methods optimized for keyword extraction perform particularly well for hashtag generation, with KeyBERT providing the most reliable performance across all operating points.

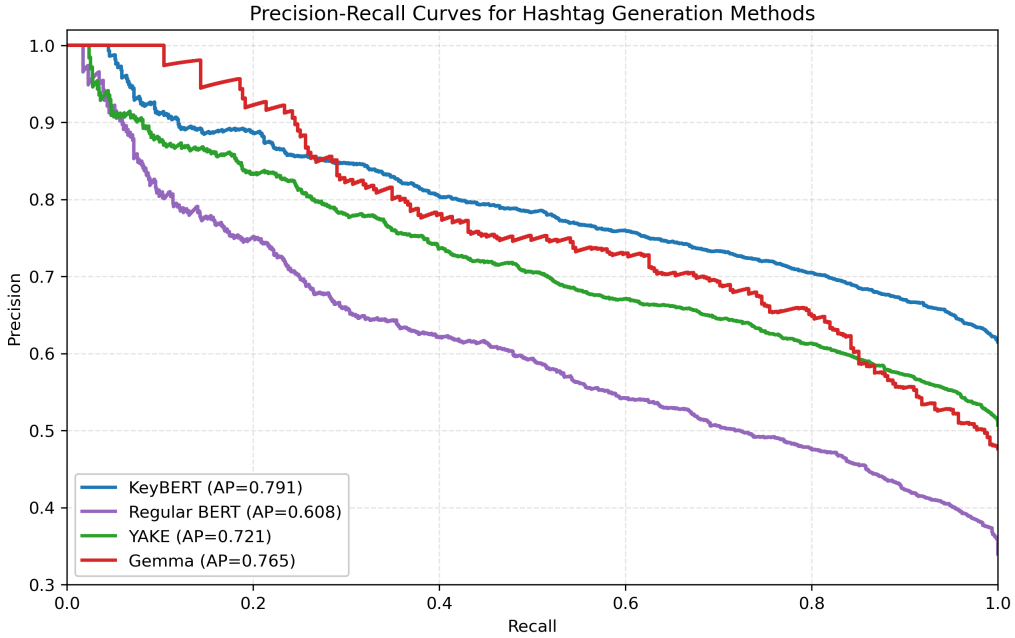
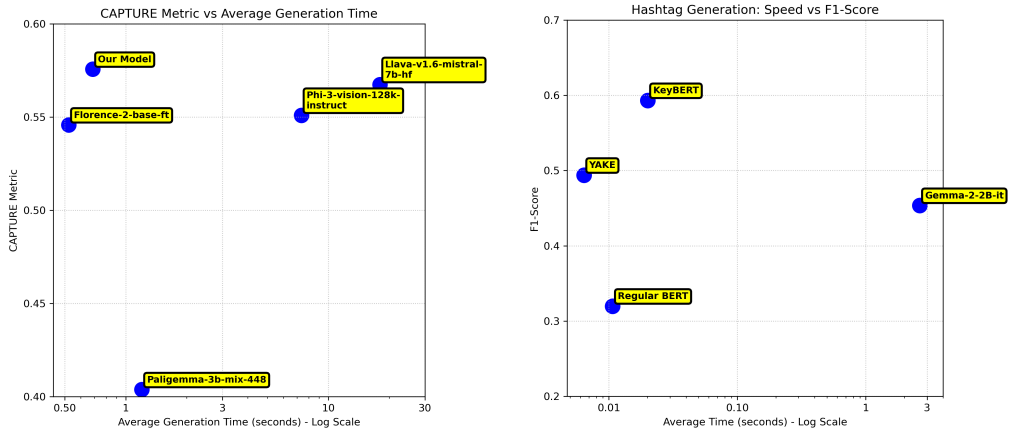


Figure 3. Precision vs. recall curve for Hashtag Generation Models.

To better illustrate the balance between performance and speed, Figure 4 presents both the CAPTURE score against the average processing time for image captioning and the accuracy against speed for hashtag generation. Specifically, Figure 4a clearly demonstrates that our model achieves the best balance between performance (as measured by the CAPTURE score) and processing speed. It outperforms significantly larger models in detailed image captioning tasks while taking only a fraction of the time to process each image.

Figure 4b demonstrates that KeyBERT offers the best combination of accuracy (as measured by F1-score) and speed. While YAKE is slightly faster, it achieves lower accuracy. Gemma-2-2B-it, despite its more complex architecture, did not outperform KeyBERT and was significantly slower, making it less suitable for real-time applications. Regular BERT's position on this graph underscores its suboptimal performance, offering neither a speed nor an accuracy advantage.



(**a**) CAPTURE metric vs. average generation time. (**b**) Hashtag F1 score vs. generation time

Figure 4. Performance evaluation for both image captioning (**a**) and hashtag generation (**b**).

These results indicate that the KeyBERT-based approach provides an effective solution for hashtag generation, balancing accuracy and computational efficiency in a manner well-suited for real-time applications like the proposed browser extension.

4.5. Browser Extension Implementation

As an application of the CapFlow model, we developed a Chrome browser extension integrating its image captioning and hashtag generation capabilities directly into the user's web browsing experience. It utilizes HTML, CSS, and JavaScript for the frontend, while the backend API for the captioning model is implemented using Python with Flask.

To address the limitations of Chrome's default extension popup, which closes when the user interacts outside it, we developed a custom popup solution. Upon clicking the extension button, our custom interface appears and can be moved freely on the screen. This approach enhances usability by allowing users to simultaneously interact with both the webpage and the extension.

The extension's main functionality is triggered when the user clicks a "Start" button in the popup. This action applies a semi-transparent overlay to dim the entire webpage, signaling to the user that the extension is active and waiting for input. Users can then select any image on the webpage for processing.

Upon image selection, the chosen image is visually highlighted against the dimmed background. The extension then sends the image data to the backend API, which hosts our image captioning and hashtag generation models. These models process the image and return a detailed caption along with relevant hashtags. The results are then displayed within the extension's popup interface. If the user does not select an image after activating the extension, an error message is shown, and the process terminates, returning the webpage to its normal state. Users can restart the process at any time by clicking a "Start Again" button, which resets the interface and allows for a new image selection.

The frontend of the extension uses vanilla JavaScript for DOM manipulation and event handling, with CSS for styling the popup and creating visual effects like the page dimming. To capture the selected image, we employed Chrome's content scripts feature, which allows us to inject JavaScript into the webpage to handle image selection and communicate with our custom popup.

The communication between different parts of the extension (popup, background scripts, and content scripts) utilizes Chrome's messaging APIs, such as chrome.runtime.sendMessage() and chrome.tabs.sendMessage(). For interactions with the backend server, we use asynchronous HTTP requests, allowing data transfer without reloading the extension popup or the webpage. The backend API, built with Python and Flask, is containerized using Docker to ensure consistency across different environments and simplify deployment. We use a docker-compose.yml file to define and manage the multi-container Docker application, coordinating the services required for the extension's operation.

This implementation separates the user interface concerns from the complex processing of image captioning and hashtag generation. It allows for independent updates and maintenance of different extension components, contributing to the overall modularity and scalability of the system.

A video demonstration of our browser extension, CapFlow, is available on YouTube at https://youtu.be/9vuyXFFfgok accessed on 27 November 2024. Screenshots from the video, which showcase the extension's workflow, can be seen in Figure 5.

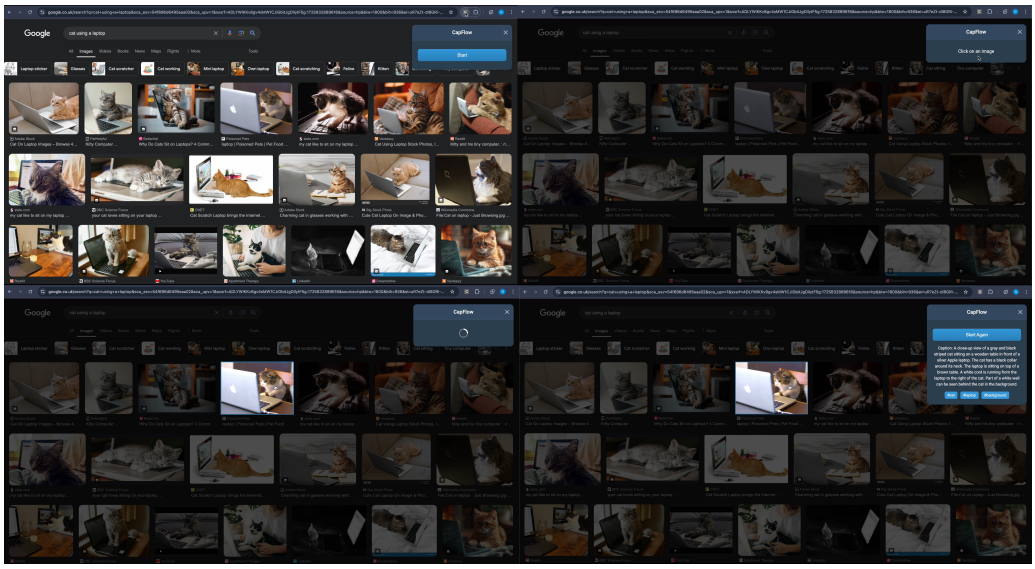


Figure 5. Introducing CapFlow.

4.6. Discussions and Limitations

The results of this study demonstrate significant advancements in both detailed image captioning and hashtag generation, with important implications for AI-assisted content analysis and creation. The performance of the image captioning model, particularly its efficiency and high CAPTURE score, suggests that it is well suited for applications requiring detailed image descriptions in real-time scenarios. The model's ability to generate detailed captions quickly could be particularly valuable in improving web accessibility, aiding content creators, and enhancing image search capabilities.

The fine-tuned Florence-2 model with the DOCCI adapter demonstrates significant improvements in generating detailed and contextually rich image captions. The CAPTURE metric results, in particular, highlight the model's ability to capture nuanced visual elements and their relationships within images. This improvement is crucial for applications requiring in-depth image understanding, such as content analysis for visually impaired users or detailed content indexing.

The competitive performance against larger models like Llava-v1.6-mistral-7b-hf and Phi-3-vision-128k-instruct is particularly noteworthy, especially considering the proposed model's significantly lower computational requirements. This efficiency makes the approach more suitable for real-time applications and deployment on consumer-grade hardware, potentially enabling new applications for automated image understanding in everyday scenarios.

For hashtag generation, the KeyBERT-based model shows promising results in terms of both accuracy and computational efficiency. The high precision indicates that the majority of generated hashtags are relevant, while the strong recall suggests that the model is effective at identifying a wide range of appropriate hashtags. The balanced F1-score demonstrates the model's overall effectiveness in generating relevant and descriptive hashtags.

Compared to other methods, the KeyBERT implementation stands out for its balance of performance and speed. While YAKE offers faster processing, its lower precision and recall make it less suitable for applications where accuracy is crucial. The regular BERT approach, despite its popularity, fell significantly behind in all metrics, justifying the optimized KeyBERT approach. The Gemma-2-2B-it model, while showing moderate performance, is hampered by its long processing time, making it impractical for real-time applications.

The model's fast processing time is particularly noteworthy as it enables real-time hashtag generation. This speed, combined with its accuracy, makes it well suited for integration into content creation workflows, allowing for the immediate suggestion of relevant hashtags as content is being produced.

The combination of detailed image captioning and efficient hashtag generation presents significant opportunities for enhancing content analysis and creation workflows. Detailed captions and relevant hashtags can significantly improve the searchability and categorization of visual content across various platforms. The detailed image descriptions generated by the model can greatly improve the web browsing experience for visually impaired users, providing rich contextual information about image content.

By automating the process of generating detailed image descriptions and relevant hashtags, these models can streamline content creation workflows, particularly for social media managers and digital marketers. The efficiency of the models makes them suitable for integration into various platforms, from web browsers to mobile applications, providing consistent functionality across different content management systems.

However, there are limitations and areas for future research. While the image captioning model demonstrates strong performance, there is still room for improvement in handling complex scenes or rare objects. The hashtag generation model, while effective, could benefit from more context-aware approaches that consider user preferences and platform-specific trends.

Future work could explore several directions to address these limitations and further enhance the capabilities of the system.

1. Multimodal Context Integration: Enhancing the image captioning model to incorporate broader contextual information from surrounding text or user history could lead to even more relevant and contextually appropriate captions.
2. Multilingual Support: Expanding the capabilities of both the image captioning and hashtag generation models to support multiple languages would greatly increase their utility in global content creation and management scenarios.
3. Dynamic Adaptation: Implementing mechanisms for continuous learning and adaptation based on user feedback could further improve the relevance and accuracy of generated captions and hashtags over time.
4. Ethical Considerations: As with any AI system processing and describing visual content, ongoing research into potential biases and strategies for their mitigation is crucial to ensure fair and inclusive content analysis.
5. Hashtag Generation Improvements: While KeyBERT provides an efficient solution for hashtag generation, there is room for improvement in terms of relevance and diversity of generated hashtags. Exploring more sophisticated approaches, such as fine-tuned language models specifically for hashtag generation, could yield better results.

5. Conclusions

In this paper, we have presented CapFlow, an integrated approach to detailed image captioning and hashtag generation for automated content analysis and management. The key contributions of this work can be summarized as follows.

First, we presented an efficient detailed image captioning model. By fine tuning the Florence-2 model using Low-Rank Adaptation (LoRA) with carefully selected datasets, particularly DOCCI, state-of-the-art performance in generating detailed image captions was achieved. Model selection was achieved through comprehensive evaluation of numerous state-of-the-art image captioning models. The selected model outperformed larger, more complex models in the CAPTURE metric (which is specifically designed to evaluate detailed captions), while maintaining computational efficiency suitable for real-time applications.

Second, we developed an effective hashtag generation method. The use of KeyBERT for hashtag generation from detailed captions demonstrated superior performance compared to alternative methods. Despite the inherent challenges in hashtag generation tasks, the proposed approach achieves a balance between relevance and diversity in generated hashtags, outperforming other tested methods in precision, recall, and F1-score.

Lastly, we demonstrated the implementation of the CapFlow into a Chrome extension by integrating detailed image captioning and hashtag generation into a unified, browser-

based system. This application demonstrates the potential for enhancing content creation workflows and improving web accessibility without requiring powerful hardware.

The experimental results highlight several important findings. The DOCCI-adapted Florence-2 model achieved a CAPTURE score of 0.5757, surpassing both the base model and larger, more complex vision-language models (VLMs). This performance was achieved with an average processing time of only 0.688 s, demonstrating a remarkable balance between accuracy and efficiency. The KeyBERT-based hashtag generation method achieved a precision of 0.6132 and an F1-score of 0.5932, outperforming other tested methods. While these scores might seem modest in traditional text classification contexts, they represent strong performance in the more subjective and open-ended task of hashtag generation.

These contributions have several important implications for the field. The ability to generate detailed image captions quickly and accurately can significantly improve web accessibility for visually impaired users, providing richer context for screen readers and other assistive technologies. The integrated system can streamline content creation workflows for marketers, social media managers, and content creators, automating the process of generating detailed image descriptions and relevant hashtags. The efficiency of the proposed models demonstrates the potential for deploying advanced AI capabilities in resource-constrained environments, such as web browsers or mobile devices, making these technologies more widely accessible.

Author Contributions: Conceptualization, N.S. and Y.L.; Methodology, N.S.; Software, N.S.; Validation, N.S.; Formal analysis, N.S.; Investigation, N.S. and Y.L.; Data curation, N.S.; Writing—original draft, N.S.; Writing—review & editing, N.S. and Y.L.; Visualization, N.S.; Supervision, Y.L.; Project administration, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ghandi, T.; Pourreza, H.; Mahyar, H. Deep Learning Approaches on Image Captioning: A Review. *ACM Comput. Surv.* **2024**, *56*, 1–39. [CrossRef]
2. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and Tell: A Neural Image Caption Generator. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
3. Xu, K.; Ba, J.L.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.S.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *arXiv* **2015**, arXiv:1502.03044.
4. Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; Zhang, L. 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), Salt Lake City, UT, USA, 18–23 June 2018.
5. Yang, X.; Tang, K.; Zhang, H.; Cai, J. Auto-Encoding Scene Graphs for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
6. Cornia, M.; Baraldi, L.; Cucchiara, R. Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019.
7. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-Linear Attention Networks for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
8. Nguyen, V.Q.; Suganuma, M.; Okatani, T. *GRIT: Faster and Better Image Captioning Transformer Using Dual Visual Features*; Springer: Cham, Switzerland, 2022; pp. 167–184. [CrossRef]
9. Wang, Y.; Xu, J.; Sun, Y. End-to-End Transformer Based Model for Image Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 2585–2594. [CrossRef]
10. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; Meila, M., Zhang, T., Eds.; PMLR, Proceedings of Machine Learning Research; 2012; Volume 139, pp. 8748–8763.
11. Mokady, R.; Hertz, A.; Bermano, A.H. ClipCap: CLIP Prefix for Image Captioning. *arXiv* **2021**, arXiv:2111.09734. [CrossRef]

12. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*; Springer: Cham, Switzerland, 2020; pp. 121–137. [CrossRef]
13. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. VinVL: Revisiting Visual Representations in Vision-Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 5579–5588.
14. Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. *arXiv* **2022**, arXiv:2205.12005. [CrossRef]
15. Hu, J.C.; Cavicchioli, R.; Capotondi, A. Exploiting Multiple Sequence Lengths in Fast End to End Training for Image Captioning. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2022. [CrossRef]
16. Li, Y.; Pan, Y.; Yao, T.; Mei, T. Comprehending and Ordering Semantics for Image Captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 17990–17999.
17. Hu, X.; Yin, X.; Lin, K.; Zhang, L.; Gao, J.; Wang, L.; Liu, Z. VIVO: Visual Vocabulary Pre-Training for Novel Object Captioning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 1575–1583. [CrossRef]
18. Xia, Q.; Huang, H.; Duan, N.; Zhang, D.; Ji, L.; Sui, Z.; Cui, E.; Bharti, T.; Zhou, M. *XGPT: Cross-modal Generative Pre-Training for Image Captioning*; Springer: Cham, Switzerland, 2021; pp. 786–797. [CrossRef]
19. Rohrbach, A.; Hendricks, L.A.; Burns, K.; Darrell, T.; Saenko, K. Object Hallucination in Image Captioning. *arXiv* **2018**, arXiv:1809.02156. [CrossRef]
20. Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; Yuan, L. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 4818–4829.
21. Gong, Y.; Zhang, Q. Hashtag Recommendation Using Attention-Based Convolutional Neural Network. In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16), New York, NY, USA, 9–15 July 2016; Volume 16, pp. 2782–2788.
22. Hachaj, T.; Miazga, J. Image Hashtag Recommendations Using a Voting Deep Neural Network and Associative Rules Mining Approach. *Entropy* **2020**, *22*, 1351. [CrossRef] [PubMed]
23. AL-Sammarraie, Y.Q.; AL-Qawasmi, K.; AL-Mousa, M.R.; Desouky, S.F. Image Captions and Hashtags Generation Using Deep Learning Approach. In Proceedings of the 2022 International Engineering Conference on Electrical, Energy, and Artificial Intelligence (EICEEAI), Zarqa, Jordan, 29 November–1 December 2022; pp. 1–5. [CrossRef]
24. Nadim, M.; Akopian, D.; Matamoros, A. A Comparative Assessment of Unsupervised Keyword Extraction Tools. *IEEE Access* **2023**, *11*, 144778–144798. [CrossRef]
25. Khan, M.Q.; Shahid, A.; Uddin, M.I.; Roman, M.; Alharbi, A.; Alosaimi, W.; Almalki, J.; Alshahrani, S.M. Impact analysis of keyword extraction using contextual word embedding. *PeerJ Comput. Sci.* **2022**, *8*, e967. [CrossRef] [PubMed]
26. Ferdinandus, D.D.; Alvin, S. Snackable Content Creation In The Digital Age: A Case Study Of Social Media Content Production at Net TV. *Int. J. Econ. Bus. Account. Agric. Manag. Sharia Adm. (IJEBAS)* **2023**, *3*, 669–680. [CrossRef]
27. Manoharan, A. Enhancing audience engagement through ai-powered social media automation. *World J. Adv. Eng. Technol. Sci.* **2024**, *11*, 150–157. [CrossRef]
28. Ding, M.; Xiao, B.; Codella, N.; Luo, P.; Wang, J.; Yuan, L. *DaViT: Dual Attention Vision Transformers*; Springer: Cham, Switzerland, 2022; pp. 74–92. [CrossRef]
29. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv* **2021**, arXiv:2106.09685. [CrossRef]
30. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In *Proceedings of the Advances in Neural Information Processing Systems*; Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S., Eds.; Curran Associates, Inc.: New Orleans, LA, USA, 2023; Volume 36, pp. 34892–34916.
31. Beyer, L.; Steiner, A.; Pinto, A.S.; Kolesnikov, A.; Wang, X.; Salz, D.; Neumann, M.; Alabdulmohsin, I.; Tschannen, M.; Bugliarello, E.; et al. PaliGemma: A versatile 3B VLM for transfer. *arXiv* **2024**, arXiv:2407.07726. [CrossRef]
32. Abdin, M.; Aneja, J.; Awadalla, H.; Awadallah, A.; Awan, A.A.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; Behl, H.; et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv* **2024**, arXiv:2404.14219. [CrossRef]
33. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S., Eds.; PMLR, Proceedings of Machine Learning Research; ML Research Press: Cambridge, MA, USA, 2022; Volume 162, pp. 12888–12900.
34. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
35. Campos, R.; Mangaravite, V.; Pasquali, A.; Jorge, A.; Nunes, C.; Jatowt, A. YAKE! Keyword extraction from single documents using multiple local features. *Inf. Sci.* **2020**, *509*, 257–289. [CrossRef]
36. Team, G.; Riviere, M.; Pathak, S.; Sessa, P.G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv* **2024**, arXiv:2408.00118. [CrossRef]
37. Onoe, Y.; Rane, S.; Berger, Z.; Bitton, Y.; Cho, J.; Garg, R.; Ku, A.; Parekh, Z.; Pont-Tuset, J.; Tanzer, G.; et al. DOCCI: Descriptions of Connected and Contrasting Images. *arXiv* **2024**, arXiv:2404.19753.

38. Singla, V.; Yue, K.; Paul, S.; Shirkavand, R.; Jayawardhana, M.; Ganjdanesh, A.; Huang, H.; Bhatele, A.; Somepalli, G.; Goldstein, T. From Pixels to Prose: A Large Dataset of Dense Image Captions. *arXiv* **2024**, arXiv:2406.10328. [CrossRef]
39. Li, X.; Tu, H.; Hui, M.; Wang, Z.; Zhao, B.; Xiao, J.; Ren, S.; Mei, J.; Liu, Q.; Zheng, H.; et al. What If We Recaption Billions of Web Images with LLaMA-3? *arXiv* **2024**, arXiv:2406.08478. [CrossRef]
40. Dong, H.; Li, J.; Wu, B.; Wang, J.; Zhang, Y.; Guo, H. Benchmarking and Improving Detail Image Caption. *arXiv* **2024**, arXiv:2405.19092. [CrossRef]
41. Ilsilfverskiold. Tech Keywords Topics Summary. 2024. Available online: https://huggingface.co/datasets/ilsilfverskiold/tech-keywords-topics-summary (accessed on 17 October 2024).
42. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 6–12 July 2002; Isabelle, P., Charniak, E., Lin, D., Eds.; Association for Computational Linguistics: St. Stroudsburg, PA, USA, 2002; pp. 311–318. [CrossRef]
43. Lin, C.Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Text Summarization Branches Out, Barcelona, Spain, 25–26 July 2004; pp. 74–81.
44. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*; Association for Computational Linguistics: Ann Arbor, MI, USA, 2005; pp. 65–72.
45. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based Image Description Evaluation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
46. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. *SPICE: Semantic Propositional Image Caption Evaluation*; Springer: Cham, Switzerland, 2016; pp. 382–398. [CrossRef]