mathematics

MDPI

Article

Privacy-Preserving Image Captioning with Partial Encryption and Deep Learning

Antoinette Deborah Martin [1] and Inkyu Moon [1,2,*]

[1] Department of Robotics and Mechatronics Engineering, Daegu Gyeongbuk Institute of Science & Technology (DGIST), Daegu 42988, Republic of Korea; ekuama@dgist.ac.kr
[2] Department of Artificial Intelligence, Daegu Gyeongbuk Institute of Science & Technology (DGIST), Daegu 42988, Republic of Korea
* Correspondence: inkyu.moon@dgist.ac.kr

Abstract: Although image captioning has gained remarkable interest, privacy concerns are raised because it relies heavily on images, and there is a risk of exposing sensitive information in the image data. In this study, a privacy-preserving image captioning framework that leverages partial encryption using Double Random Phase Encoding (DRPE) and deep learning is proposed to address privacy concerns. Unlike previous methods that rely on full encryption or masking, our approach involves encrypting sensitive regions of the image while preserving the image's overall structure and context. Partial encryption ensures that the sensitive regions' information is preserved instead of lost by masking it with a black or gray box. It also allows the model to process both encrypted and unencrypted regions, which could be problematic for models with fully encrypted images. Our framework follows an encoder–decoder architecture where a dual-stream encoder based on ResNet50 extracts features from the partially encrypted images, and a transformer architecture is employed in the decoder to generate captions from these features. We utilize the Flickr8k dataset and encrypt the sensitive regions using DRPE. The partially encrypted images are then fed to the dual-stream encoder, which processes the real and imaginary parts of the encrypted regions separately for effective feature extraction. Our model is evaluated using standard metrics and compared with models trained on the original images. Our results demonstrate that our method achieves comparable performance to models trained on original and masked images and outperforms models trained on fully encrypted data, thus verifying the feasibility of partial encryption in privacy-preserving image captioning.

Keywords: double random phase encoding; deep learning; partial encryption; image captioning; privacy preserving

MSC: 68P27; 68P25; 68U10; 68T07

check for updates

Academic Editor: Lingfeng Liu

Received: 9 January 2025
Revised: 5 February 2025
Accepted: 6 February 2025
Published: 7 February 2025

Citation: Martin, A.D.; Moon, I. Privacy-Preserving Image Captioning with Partial Encryption and Deep Learning. *Mathematics* **2025**, *13*, 554. https://doi.org/10.3390/math13040554

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Image captioning involves accurately understanding an image's content and generating grammatically correct and contextually relevant sentences [1]. Thus, this involves two primary domains of Artificial Intelligence for completing the tasks: computer vision and natural language processing [2–4]. Image captioning must recognize image objects, their attributes, and relationships and translate them into semantically correct captions [1,5,6]. This task has gained considerable interest due to the advances in neural networks that generate human-like descriptions based on the input image [1] and its wide range of

applications, including content understanding, facilitating image search, and enhancing accessibility for visually impaired individuals [1,3,5,7,8].

The most popular approach proposed for image captioning is the encoder–decoder structure optimized end-to-end [9]. The encoder is utilized for visual cues; thus, a Convolutional Neural Network (CNN) is typically employed [10]. CNNs excel at computer vision tasks such as object detection and image classification due to effectively extracting salient features from images through convolutions, pooling, and activations into a feature vector. Architectures such as ResNet [11], which introduces residual connections, DenseNet [12], which uses a dense connection pattern, and Faster R-CNN [13] with its region proposal network, are primarily used in the encoder due to their ability to extract high-level visual features effectively [14]. Other works incorporate spatial attention mechanisms on CNN layers to focus on relevant features [15]. Another alternative is the vision transformer [16], which utilizes a self-attention mechanism to extract features and outperforms traditional CNNs in specific contexts. The decoder produces the output caption; thus, Recurrent Neural Networks (RNNs), including Long Short-Term Memory (LSTM), serve as a language model to process the extracted features in parallel with the text label [10,15]. RNNs can process entire data sequences and generate sentences word by word. Recently, transformers [17] have performed better than RNNs across natural language processing tasks such as machine translation and language generation [2,18,19].

Image captioning relies heavily on large amounts of image data, which raises concerns about data privacy as there is a risk of exposing sensitive information in the images if they are improperly handled or attacked [20]. Visual privacy protection has accumulated significant attention due to its demand in different fields, such as social networks, healthcare, and security. Researchers have explored methods using optics and algorithms to address these concerns [21–23]. Data encryption for cloud computing enables computations on encrypted data without decryption, enhancing security in social multimedia and medical applications [7,24,25]. Thus, privacy-preserving deep learning schemes that use techniques such as cryptographic algorithms, differential privacy, secure multi-party computation, and federated learning have been proposed to safeguard sensitive information [26].

In this paper, we propose a privacy-preserving image using partial encryption with Double Random Phase Encoding (DRPE) [27]. Full encryption, as used in [7], involves encrypting the entire image; however, partial encryption involves selecting specific regions or features containing sensitive information and encrypting them while preserving the image's overall structure and context. By using partial encryption, we can maintain privacy in the image while retaining other recognizable features, thereby allowing the generation of accurate captions. Thus, the information in the sensitive regions is not lost by using a gray box [19] but preserved using image encryption, and relevant features must be extracted from partially encrypted images' encrypted and unencrypted regions. We propose a dual-stream encoder based on ResNet50, which processes the real and imaginary parts of the encrypted region separately, resulting in the effective extraction of features. In the decoder, we employ a transformer to generate captions based on the extracted features from the decoder.

Additionally, we performed ablation studies to validate the structure of our proposed method in generating accurate captions for partially encrypted images. We evaluate our proposed method on the Flickr8k [28] dataset using standard metrics such as BLEU [29], ROUGE [30], METEOR [31], and CIDEr [32]. From the results, our proposed method achieves comparable metrics scores to models trained on original and masked images and outperforms models trained on fully encrypted images. The results verify the effectiveness of our approach in providing privacy in image captioning tasks.

The main contributions of this paper are summarized as follows:

- We performed image captioning using partial encryption with Double Random Phase Encoding (DRPE). The partially encrypted images ensure data privacy while retaining other recognizable information;
- We employed a dual-stream encoder based on ResNet50, which enables effective feature extraction from the encrypted regions and unencrypted regions as it processes the real and imaginary parts of the encrypted region separately;
- We used a transformer-based decoder with a 2-2-layer configuration and word embeddings trained from scratch to generate captions;
- Ablation studies were conducted to validate our proposed architecture's effectiveness, highlighting the advantages of the dual-stream encoder and comparing the performance across original, partially encrypted, fully encrypted, and partially blocked/masked images;
- We demonstrated that partial encryption provides a balance between privacy and usability, outperforming the performance of full encryption and masking with black/gray boxes.

The rest of the paper is organized as follows: Section 2 describes the related work; Section 3 describes the technique for partial encryption and the proposed method for privacy-preserving image captioning. Section 4 provides the ablation studies conducted, the experimental results, a comparison of performance across original, partially encrypted, fully encrypted, and partially blocked/masked images, and a comparison with state-of-the-art models. Finally, Section 5 concludes the paper by summarizing the proposed method and the results.

2. Related Work

Privacy preservation in machine learning and deep learning has emerged as a necessary research direction with various techniques proposed to protect privacy during training data collection, training, inference, and fine-tuning phases [26,33]. Differential privacy introduces noise to the output of the model to conceal sensitive data [33,34]. In [35], an end-to-end framework using Graph Transformer and Convolutional Networks is proposed to improve visual data classification and privacy preservation. They employ differential privacy-based graph construction and noise-induced graph transformation to protect the privacy of knowledge graphs and evaluate their framework on the MS-COCO dataset in a semi-supervised setting. They applied a General Data Protection Regulation compliant method to obfuscate sensitive information such as faces, passport numbers, and license plate numbers. In [36], a differentially private image captioner is trained, and unprecedented high-quality image features are obtained, which can be used for vision and vision–language downstream tasks. The trained model is used to caption images from the MS-COCO dataset. On the other hand, homomorphic encryption enables computations to be conducted on encrypted data [37]. A residual network implementation based on fully homomorphic encryption is proposed in [38] for the classification of encrypted images. They achieved an almost equivalent accuracy with the encrypted model to that of the plain model, demonstrating the feasibility of privacy preservation without performance degradation.

Federated learning allows collaborative model training while protecting the privacy of data in distributed environments [39]. A federated learning framework is proposed [40] to improve the performance on a variety of vision-and-language grounding problems without sharing the downstream task data. Their centralized model converts the extracted features into fine-grained image representations. They validate their approach in three federated learning settings: horizontal federated learning, vertical federated learning, and federated transfer learning. Other approaches that allow performing image captioning

from privacy-protected images are the framework by [5], which addresses the problem from a hardware perspective, and [19], which focuses on generating captions for dietary assessment instead of original images, reducing the risk of privacy leakage from images. In [5], the approach comprises end-to-end learning of an optical system (reflective lens) for scene acquisition coupled with deep neural networks for generating captions. The reflective lens distorts the scene in a way that still allows extracting relevant features for image captioning, thus protecting sensitive information. In [19], when training their image captioning method, they use images of faces masked with gray boxes [19]. In this work, we focus on using cryptographic algorithms to preserve the privacy of data. We use the DRPE algorithm, which has been used by researchers to enhance image-based systems' privacy-preserving capabilities without compromising their utility [7,25,41]. In our previous work [7], we utilized full encryption; however, the performance of image captioning decreased as compared to non-privacy-preserving approaches. However, in the paper, we adopt partial encryption, which maintains privacy and allows for the generation of accurate captions.

3. Materials and Methods

3.1. Image Encryption

The optical cryptographic algorithm used in the study is the Double Random Phase Encoding (DRPE) [27] algorithm. DRPE uses parallel processing to encrypt extensive data, such as image data, at a high speed [41]. In DRPE, the original image is converted to stationary white noise using random phase masks and a 4f optical system to improve the difficulty of illegal attacks. The encrypted image does not reveal visual information without the appropriate decryption keys (two random phase masks). The input image and the random phase masks must be the same size to guarantee pixel-by-pixel multiplication [42]. The encryption process is mathematically expressed as follows:

$$g(x,y) = \text{IFT}(\text{FT}\{f(x,y).\exp[j2\pi t(x,y)]\}.\exp[j2\pi s(\mu,\nu)]) \quad (1)$$

where $g(x,y)$ is the encrypted image, FT and IFT are the Fourier Transform and Inverse Fourier Transform, respectively, $f(x,y)$ is the input image, and j is the mathematical notation for the imaginary symbol. The random phase masks are $p_1 = \exp[j2\pi t(x,y)]$ and $p_2 = \exp[j2\pi s(\mu,\nu)]$, respectively. $t(x,y)$ and $s(\mu,\nu)$ are random variables uniformly distributed on the interval [0, 1] and are independent and identically distributed. Sensitive visual regions of interest, such as faces and vehicle registration plate numbers, are manually selected from the images and encrypted using Equation (1) while the rest of the image remains unchanged. The same regions are selected in the random phase masks for multiplication. Since the resulting encryption pixel values are complex, the encryption regions are split into real and imaginary parts.

3.2. Proposed Image Captioning Method

3.2.1. Overall Process

The proposed framework for performing image captioning on partially encrypted images using DRPE, which follows the encoder–decoder architecture, is depicted in Figure 1. Since the encryption process of DRPE generates complex numbers in the encrypted regions, these regions are split into real and imaginary parts. A dual-stream encoder processes the images with real and imaginary parts separately and extracts features from the encrypted and unencrypted regions of the images, and a transformer-based decoder generates the corresponding predicted captions. The dual-stream encoder consists of two parallel encoders based on ResNet50, each initialized with pre-trained weights from ImageNet. The

outputs from both streams are concatenated to form a $14 \times 14 \times 4096$ and then flattened to a 196×4096 as input to the transformer encoder. The transformer-based decoder generates captions based on the 196×4096 features (and ground truth captions during training) one word at a time.

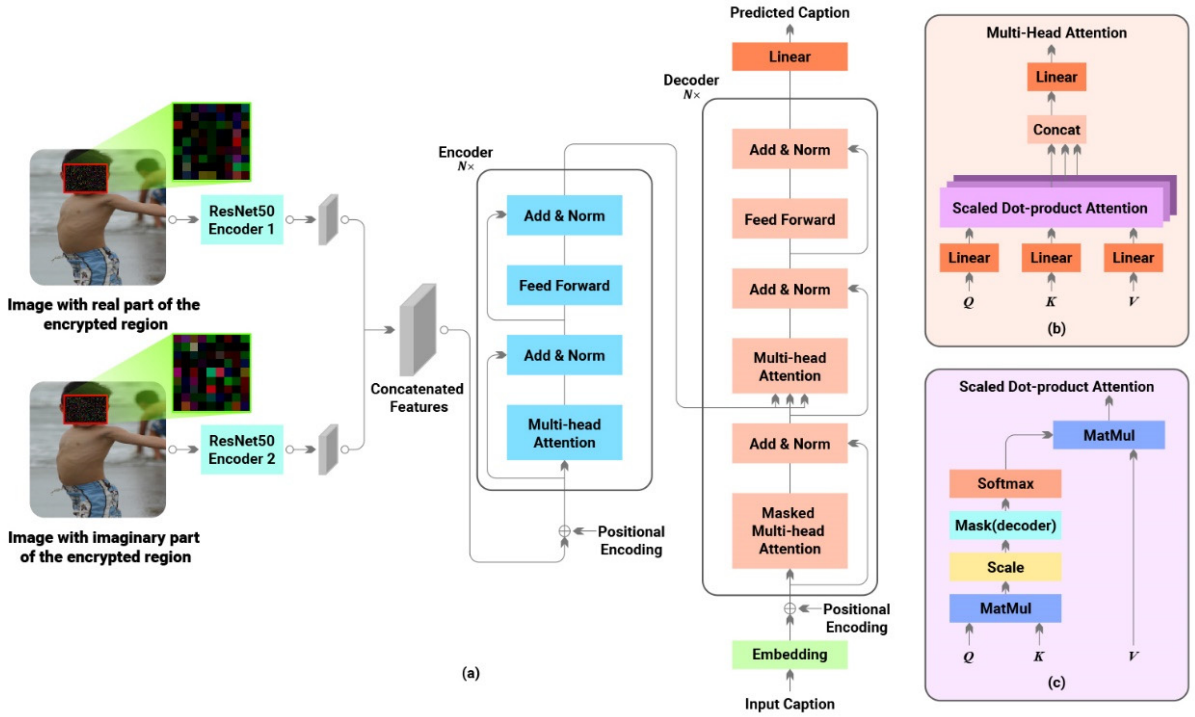

Figure 1. The proposed framework's architecture for image captioning of partially encrypted images using DRPE. (**a**) Overview of the proposed framework, which features a dual-stream encoder followed by a transformer-based decoder. (**b**) Detailed view of the multi-head attention block within the transformer. (**c**) Expanded representation of the scaled dot-product attention block found in (**b**).

3.2.2. Dual-Stream Encoder

The structure used for the dual-stream encoder, which consists of two parallel encoders, is the ResNet50 [11] architecture, commonly used in image classification and captioning tasks. The ResNet50 architecture consists of 50 layers with residual blocks of 1×1, 3×3, and 1×1 convolutional layers. The computational efficiency of ResNet50, depth, skip connections, performance, and transfer learning capabilities make it a compelling choice for various image-related tasks. The final pooling layer, fully connected layer, and Softmax layer of the ResNet50 architecture are removed in this task. We extract the features from the last convolutional layer, which has an output size of $B \times 8 \times 8 \times 2048$. An adaptive average pooling layer of size 14 is applied to the outputs of the last convolutional layer to obtain a final output size of $B \times 14 \times 14 \times 2048$, where B is the batch size. In the first encoder, the input to the encoder is the image with the real part of the encrypted region. In contrast, the second encoder takes the image (same) with the imaginary part of the encrypted region. The pre-trained weights of the ResNet50 (from ImageNet) layers are used to initialize the model layers. These are then fine-tuned during training (weights are updated with backpropagation), allowing the model to adapt to the partially encrypted images.

3.2.3. Transformer

The transformer [17] has an encoder–decoder structure, as depicted in Figure 1a. Given the features from the dual-stream encoder, the transformer encoder maps the features unto a continuous representation, z, and the transformer decoder generates an output sequence given z one word/element at a time. The transformer encoder takes inputs

of size 196 × 4096, where 196 represents the flattened 14 × 14 feature map, and 4096 is the dimension resulting from concatenating the results from the parallel encoders. The transformer decoder takes inputs of size 52 × 300, where 52 is the maximum sequence length (padded), and 300 is the embedding dimension. The transformer architecture relies heavily on the scaled dot-product attention function (Figure 1c). This function maps queries (Q), keys (K), and values (V) to an output and expressed as follows:

$$\text{Attention}(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

where d_k is the dimension of the keys, and $\frac{1}{\sqrt{d_k}}$ is a scaling factor to prevent the values of the dot products from being large in magnitude, causing minimal gradients. The multi-head attention layer (Figure 1b) in the transformer network ensures that the model learns to attend to different representations of the same input. Thus, multiple independent scaled dot-product attentions are computed over linear projected QKV vectors. The outputs are concatenated and linearly projected, resulting in final values. The formula for multi-head attention is given as follows:

$$\begin{gathered}\text{MultiHead } (Q, K, V) = \text{Concat}(\text{head}_\text{i}, \ldots, \text{head}_\text{n})W^O \\ \text{where head}_\text{i} = \text{Attention } \left(QW_i^Q,\ KW_i^K,\ VW_i^V\right)\end{gathered} \tag{3}$$

where W^O, W_i^Q, W_i^K, W_i^V are learnable weight metrics. The encoder consists of N identical layers in this study, N = 2. Each layer has two sub-layers: a multi-head attention layer and a feed-forward network. A residual connection is employed around each sublayer, followed by layer normalization. The decoder is also composed of N = 2 identical layers. Each layer has three sub-layers: first is the masked multi-head attention, which is modified to prevent the current token from attending to subsequent tokens. Therefore, predictions of the current token depend on past information instead of future information. The second sublayer performs multi-head attention over the output of the transformer encoder, and the third is a feed-forward network. The dimension for the feed-forward network for both transformer encoder and decoder is 4096. Likewise, residual connections are employed, followed by layer normalization. There are dropout layers after each sublayer to prevent overfitting.

4. Experiments

4.1. Dataset

In this study, we use one of the popular datasets, the Flickr8k dataset [28], which comprises 8092 images, each associated with five descriptions. The Karpathy splits configuration [43] is adopted, allocating 6000 images for training, 1000 for validation, and 1000 for testing. Images are resized to 256 × 256 for consistency using bilinear interpolation, and partial encryption is performed as discussed in Section 3.1. Each description text is standardized by converting to lowercase and removing punctuation. We remove words that occur less than five times, resulting in a 2633-word vocabulary. Additionally, each caption is prepended with <start> token and appended with <end> token, while all captions are padded to a fixed length of 52 to address variable caption lengths. Examples of partially encrypted images, original images, and corresponding captions are presented in Figure 2.

Figure 2. Illustration of partially encrypted images using DRPE. (**a**) Original image from the dataset, (**b**) partially encrypted images with the real part of the encrypted region of interest using DRPE, (**c**) partially encrypted images with the imaginary part of the encrypted region of interest using DRPE, and (**d**) the corresponding caption of the original image. The red boxes are the manually selected sensitive regions of interest in the original images and the encrypted regions in the partially encrypted images. The green boxes are the 9×9 center matrix of the selected regions.

4.2. Experimental Settings

The models were trained by a server with an Intel Xeon Silver 4214 CPU and NVIDIA RTX A5000 GPU. The models are constructed using Python 3.11.7 with the PyTorch framework, and the encryption process is performed using MATLAB R2024a. The optimizer used is the Adam Optimizer with a 32-batch size, a 0.0001 learning rate for the dual-stream encoder and transformer-based decoder, and categorical cross-entropy as a loss function. The encoders extracted a 2048 feature vector from their respective inputs. The maximum number of epochs is set to 50 for training, and a beam size of 3 is used during the inference stage. The number of encoder and decoder layers in the transformer is set to 2. The number of heads is 8, the dropout rate is 10%, and the embedding dimension is 300. We performed model evaluation using the standard metrics ROGUE [30], METEOR [31], BLEU [29], and CIDEr [32].

4.3. Ablation Study

4.3.1. Effect of Dual-Stream Encoder

We perform an ablation study to validate the effectiveness of the dual-stream encoder in our proposed framework. We trained four models with different encoders and evaluated them using the metrics. The encoders are Encoder 1 only (uses images with the real part of the encrypted region), Encoder 2 only (uses images with the imaginary part of the encrypted region), a single encoder that uses the concatenation of both images and our proposed dual-stream encoder. From Table 1, the proposed dual-stream encoder outperforms the other methods, achieving the highest scores for each metric. Interestingly, the model that used only Encoder 2 performed better than Encoder 1. We can assume that it may be more effective at capturing relevant features. Another is that the model with the concatenation of images as input had the lowest performance, which suggests that treating the image

parts separately and adopting the dual-stream encoder is more effective or beneficial. Furthermore, using a dual-stream encoder preserves information that might be lost if only one part of the encrypted region (complex representation) is considered, as utilized in Encoder 1 and Encoder 2. Based on the results we can assume that the dual-stream encoder possibly enables learning complementary feature representations. From Table 1, using the dual-stream encoder enhances the model's performance at achieving optimal results.

Table 1. Quantitative results on the effect of dual-stream encoder.

Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge	Cider
Using Encoder 1 only (Images with real part of the encrypted region)	62.41	45.85	32.47	22.54	22.24	48.64	60.54
Using Encoder 2 only (Images with imaginary part of the encrypted region)	67.70	49.86	35.56	24.96	22.40	49.03	62.41
Using one encoder (concatenation of both images)	59.64	40.43	26.96	17.84	20.06	43.66	42.57
Our proposed method (dual-stream encoder)	68.36	50.36	36.00	25.16	22.58	49.38	64.48

4.3.2. Transformer Architecture Analysis

The encoder and decoder layers are six each in the transformer architecture [17]. We evaluate four configurations of the transformer architecture, each defined by the number of encoder and decoder layers: (2–2) configuration, (2–4) configuration, (4–4) configuration, and (6–6) configuration. The number of layers is increased with the assumption that the deeper the model, the better the generalization; hence, a better performance would be achieved. Our proposed method ((2–2) configuration), which features two layers in both the encoder and decoder, achieved the best scores across most metrics except METEOR and CIDEr. The results in Table 2 suggest that a simple model architecture with fewer layers is more effective for this particular task and dataset. However, when the number of layers increases, the METEOR score also increases. The (6-6) configuration has the highest METEOR score (22.83) despite having the lowest scores for the other metrics. The METEOR score measures both precision and recall and considers stemming and synonyms.

Table 2. Quantitative results of transformer architecture analysis.

Encoder–Decoder Layers	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge	Cider
Our proposed method (2–2)	68.36	50.36	36.00	25.16	22.58	49.38	64.48
(2–4)	65.68	47.61	33.83	23.63	22.62	48.33	62.72
(4–4)	65.99	48.32	34.27	23.91	22.82	49.55	63.81
(6–6)	64.22	46.57	32.87	22.83	22.83	48.29	60.88

We can assume that the deeper models might better capture synonymy, although the performance for other metrics reduces. The (4–4) configuration performs well, with the highest ROUGE (49.55) and second-highest scores for the other metrics. This result also suggests that increasing the number of layers can improve specific evaluations. In summary, increasing the number of layers in the transformer encoder and decoder affects the model's performance. The (2–2) configuration proves to be most effective for our task.

4.3.3. GloVe Embeddings

Next, we experiment using GloVe (Global Vectors for Word Representation) [44] embeddings in the embedding layer for the transformer decoder. GloVe projects words onto a vector space of dimension, d. The numerical vector representation captures the semantic and syntactic information of the words [45,46]. If words have similar contexts, their representations are also very similar. We compare three configurations regarding the use of embeddings in our model: no GloVe embeddings, pre-trained GloVe embeddings, and fine-tuned GloVe embeddings. For no GloVe embeddings, the model uses randomly initialized embeddings, which are then learned from scratch during training. For pre-trained GloVe embeddings, the model uses GloVe embeddings, which are trained on a large corpus. In contrast, for fine-tuned GloVe embeddings, the pre-trained GloVe embeddings are further fine-tuned during training to allow the model to adapt the embeddings to the task.

Table 3 shows that the proposed method (without GloVe embeddings) outperforms the other GloVe-based approaches across all metrics discussed in this section. Notably, fine-tuning GloVe embeddings yields better results than pre-trained GloVe embeddings without fine-tuning, indicating that adapting the embeddings to a particular task is crucial. From the proposed model's performance, we can conclude that learning the embeddings from scratch may be more effective than using pre-trained or fine-tuning GloVe embeddings to achieve better performance in this task.

Table 3. Quantitative results on using GloVe Embeddings.

Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge	Cider
With GloVe embeddings	61.84	45.11	31.98	22.07	22.11	48.57	61.30
Fine-tuning GloVe embeddings	67.07	48.72	34.32	23.81	22.42	49.15	62.30
Our proposed method (no GloVe)	68.36	50.36	36.00	25.16	22.58	49.38	64.48

4.3.4. Fine-Tuning Strategies on the Encoder

In this section, we apply fine-tuning techniques to the ResNet50 encoder. With fine-tuning, we adapt the pre-trained model weights to the characteristics of our partially encrypted image dataset, which then improves the model's overall performance. We implemented three different fine-tuning approaches. The first is no fine-tuning (frozen weights); here, all the layers of the ResNet50 encoder are frozen, meaning that their weights remain unchanged during training. With this approach, the features from ImageNet are retained, but this limits the model to extract features from the encrypted patterns in our dataset. The second is partial fine-tuning (trainable weights in selected layers); here, we allow fine-tuning of the second to fourth convolutional blocks within the ResNet50 architecture as these layers capture mid-level and high-level features. The weights in these blocks are updated during training, while the remaining layers are kept frozen. Thus, the low-level features learned from ImageNet are retained, while the deeper layers are adapted to the characteristics of the partially encrypted images. The last approach that is used in our proposed model is full fine-tuning (all weights trainable). Here, all the weights of the ResNet50 encoder are set to trainable. Full fine-tuning allows the model to fully adapt to the partially encrypted images using DRPE, leveraging the entirety of the encoder's capacity to learn from the images. Figure 3 shows the fine-tuning schemes used in this section.

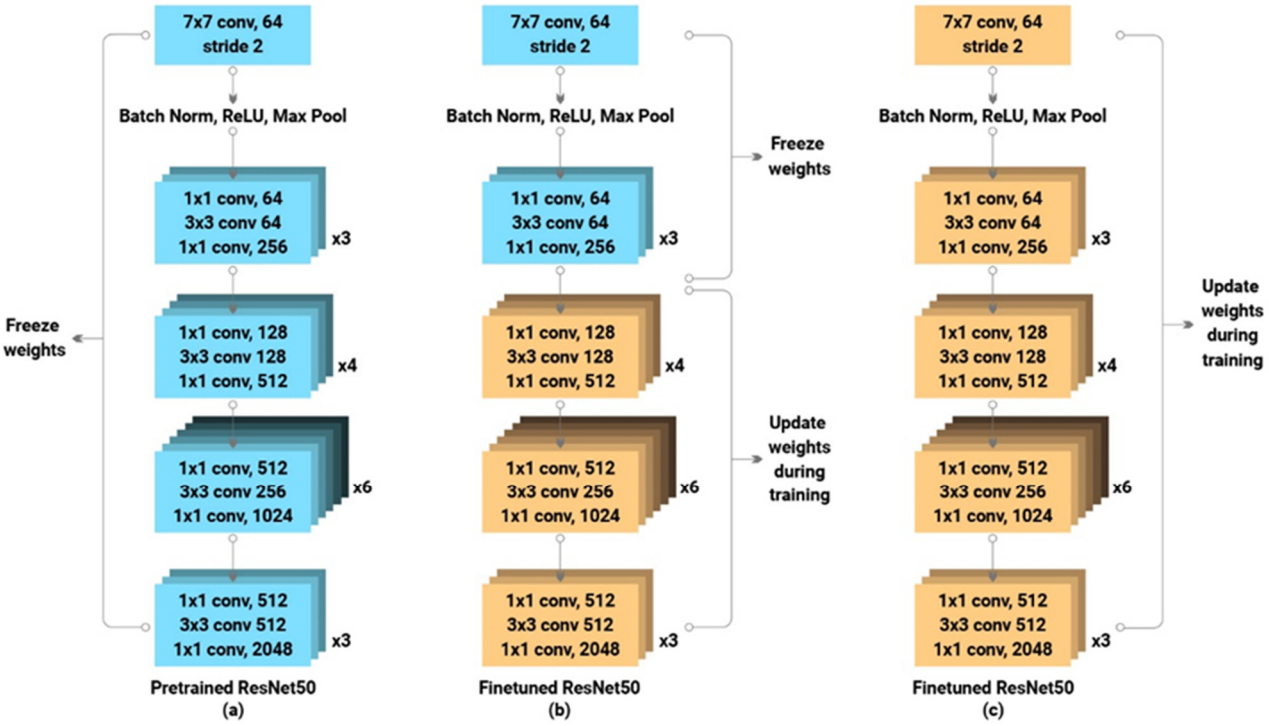

Figure 3. Fine-tuning strategies. (**a**) Fine-tuning without training weights. All layers are frozen; there is no update to the weights during backpropagation. (**b**) Fine-tuning with trainable weights for convolutional blocks 2 to 4 while the other layers are frozen. (**c**) Fine-tuning without frozen weights. All weights are trainable.

The results in Table 4 demonstrate the benefits of allowing more model parameters to be fine-tuned, as the more trainable weights, the higher the value of the metrics. The proposed method, which presumably allows all weights to be updated during training, consistently outperforms the other two approaches across all metrics. The approach with trainable weights in convolutional blocks 2 to 4 shows intermediate performance, better than the model without trainable weights but not as good as the fully trainable model. An average difference of 4.15 is calculated between our proposed method and training with frozen weights only.

Table 4. Quantitative results on different fine-tuning strategies.

Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge	Cider
Without trainable weights	63.42	45.17	31.31	21.20	21.98	47.23	57.17
With trainable weights in convolutional blocks 2 to 4	65.84	48.19	34.46	24.13	22.60	49.07	64.14
Our proposed method (without frozen weight)	68.36	50.36	36.00	25.16	22.58	49.38	64.48

4.4. Quantitative Results and Visualization

Our study involves a comparison among captioning models trained on original images, fully encrypted images, partially blocked images, and partially encrypted images. The models trained on partially blocked and original images utilized a ResNet50 encoder and transformer-based decoder. Additionally, the model trained on fully encrypted images used a modified ResNet50 encoder, in which the first layer is adjusted to take a tensor input with channel size 6 (the input is a concatenation of the real and imaginary parts of the encrypted images) and transformer-based decoder. We include a baseline model that comprises a

dual-stream encoder and attention-based LSTM decoder trained on partially encrypted images. The baseline model has an attention dimension and an LSTM dimension of 512. Table 5 presents the metrics results to evaluate the quality of the predicted captions on the Flickr8k dataset. As depicted in Table 5, the model trained on original images achieved the highest performance across all metrics. Notably, using partial encryption and deep learning, our proposed framework achieved comparable performance to the original one with a BLEU-4 of 25.16 and a CIDEr score of 64.48. The transformer-based decoder in our model outperformed the baseline model's decoder, which uses an attention-based LSTM; this can be attributed to the multi-head attention mechanism used in the transformer, which allows the model to attend to multiple aspects of the input features simultaneously.

Table 5. Quantitative results on the test dataset.

Images	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge	Cider
Original	69.59	51.56	37.36	26.43	22.88	49.68	65.48
Partially blocked	67.20	49.34	35.30	24.70	22.20	48.98	62.57
Fully encrypted (AES-CBC)	41.81	23.41	12.49	7.35	16.35	32.78	8.71
Fully encrypted (DRPE)	46.07	25.32	12.89	7.29	15.89	34.48	13.97
Partially encrypted (AES-CBC)	66.69	49.15	35.10	24.49	22.71	49.59	63.36
Partially encrypted (DRPE-Baseline)	64.64	46.81	32.78	22.34	22.35	48.71	58.96
Partially encrypted (DRPE-Proposed)	68.36	50.36	36.00	25.16	22.58	49.38	64.48

The model trained on partially blocked images performs slightly worse than our proposed framework, while the model trained on fully encrypted images showed the worst performance. The metrics scores of our proposed method being close to that of the model trained on original images suggest that our approach preserves essential visual information for caption generation while providing privacy protection. Also, comparing Encoder 2 only (Table 1) to partially blocked images in Table 5, the performance of Encoder 2 only (which uses images with the imaginary part of the encrypted region) is comparably close to that of the partially blocked approach. Encoder 2 only outperforms the partially blocked approach in most metrics except the CIDEr score. Therefore, by encrypting the sensitive regions, we preserve meaningful information rather than mask it (blocking). These results indicate that partial encryption offers a balance in providing privacy while maintaining captioning accuracy.

The DRPE algorithm is compared to the Cipher Block Chaining (CBC) mode of the Advanced Data Encryption Standard (AES) [47]. For this task, we used the 128-bit block size with a 256-bit key, which requires 14 rounds to encrypt the data. In [48], images are encrypted with AES and encryption algorithms for the classification of encrypted images. We chose the CBC mode with the assumption that it would be significantly secure as compared to the other modes. The models trained on fully encrypted and partially encrypted AES-CBC images consist of a single ResNet50 encoder and a transformer-based decoder. The fully encrypted AES-CBC model achieved higher BLEU-4 (7.35) and METEOR (16.35) scores than the fully encrypted DRPE but achieved lower scores for the other metrics. The results for partial encryption with AES-CBC were comparable to our DRPE-based approach with a relatively small gap (CIDEr: 63.36 vs. 64.48). This indicates that both encryption methods can be used to protect the privacy of sensitive information and achieve fairly accurate captions.

Figure 4 illustrates the captioning results of different models and images. The model trained on original images usually provides fairly accurate captions. However, there are some inaccuracies, such as in Figure 4a (a woman instead of a man), Figure 4c (a <unk>

variable), and Figure 4e (repetition of black hat). Our proposed model performs comparably well as the captions are close to the ground truth. However, it also has inaccuracies, such as in Figure 4c (a group of people instead of a man and woman) and Figure 4e (building instead of a red wall). The captions by the model trained on partially blocked images are also reasonably accurate. There are also some inaccuracies, such as the repetition of words (Figure 4a,e) and Figure 4c (subway instead of bench). However, the model trained on fully encrypted images struggles to provide accurate captions (Figure 4a,b,d,e), but the predicted caption for Figure 4c matches the ground truth caption.

Figure 5 illustrates the caption results of models trained on original images and partially encrypted images using DRPE and AES-CBC. It also includes the captions generated by our baseline model for partially encrypted images using DRPE. For Figure 5a, the baseline line model could not identify the bicycle as it represents it with a <unk> variable. All the models identified the overcoat as a different color and referred to the overcoat as a jacket. For Figure 5b, the captions generated fairly describe the picture as the key words "black and white dog", "ball", and "mouth" were in the captions, but the "grass" was omitted in all predicted captions. The predicted captions for Figure 5c also had some omissions ("on a tripod is smiling for another camera"). Both the baseline and proposed model identify two cameras, but the caption fails to describe the scene. The model trained on partially encrypted images using AES-CBC included words that are not visible in the picture. The predicted captions for Figure 5d identify two out of the three girls in the picture, and the captions are generalized with missing words.

Based on Table 6, which compares the time complexity and inference time for the 1000 images in the test dataset, key observations can be made. Masking the sensitive regions in the images takes the shortest time (14.06 s) as compared to encrypting the regions with either DRPE (60.21 s) or AES (66.38 s). Masking involves minimal computation as it involves pixel overwriting. Notably, encryption with DRPE is approximately 6 s faster than encryption with AES. For the time complexity, AES and masking have a linear complexity of O(N), while DRPE has a higher theoretical complexity of O(nlogn). However, we notice a difference between theoretical complexity and the encryption or masking time, which can be attributed to operations per step. AES involves complex mathematical operations and multiple encryption rounds for each block of data, which resulted in a longer time (66.38 s) than DRPE and masking. Due to the additional encoder in the proposed method, the inference time (208.66 s) and GPU memory (1534.83 MB) are the highest compared to the other methods; however, from Tables 1 and 5, using the dual-stream encoder enhances the model's performance at achieving optimal results that are close to that of the model trained on original images.

Table 6. Comparison of time complexity and inference time for 1000 images in the test dataset.

Method	Encryption Time (s)	Inference Time (s)	GPU Allocated Memory (MB)	Time Complexity
Single encoder with original images	--	182.01	882.19	--
Single encoder with partially blocked images	14.06	184.65	882.19	$O(N)$
Single encoder with partially AES-CBC-encrypted images	66.38	181.90	1098.98	$O(N)$
Dual-stream encoder with partially DRPE-encrypted images	60.21	208.66	1534.83	$O(n \log n)$

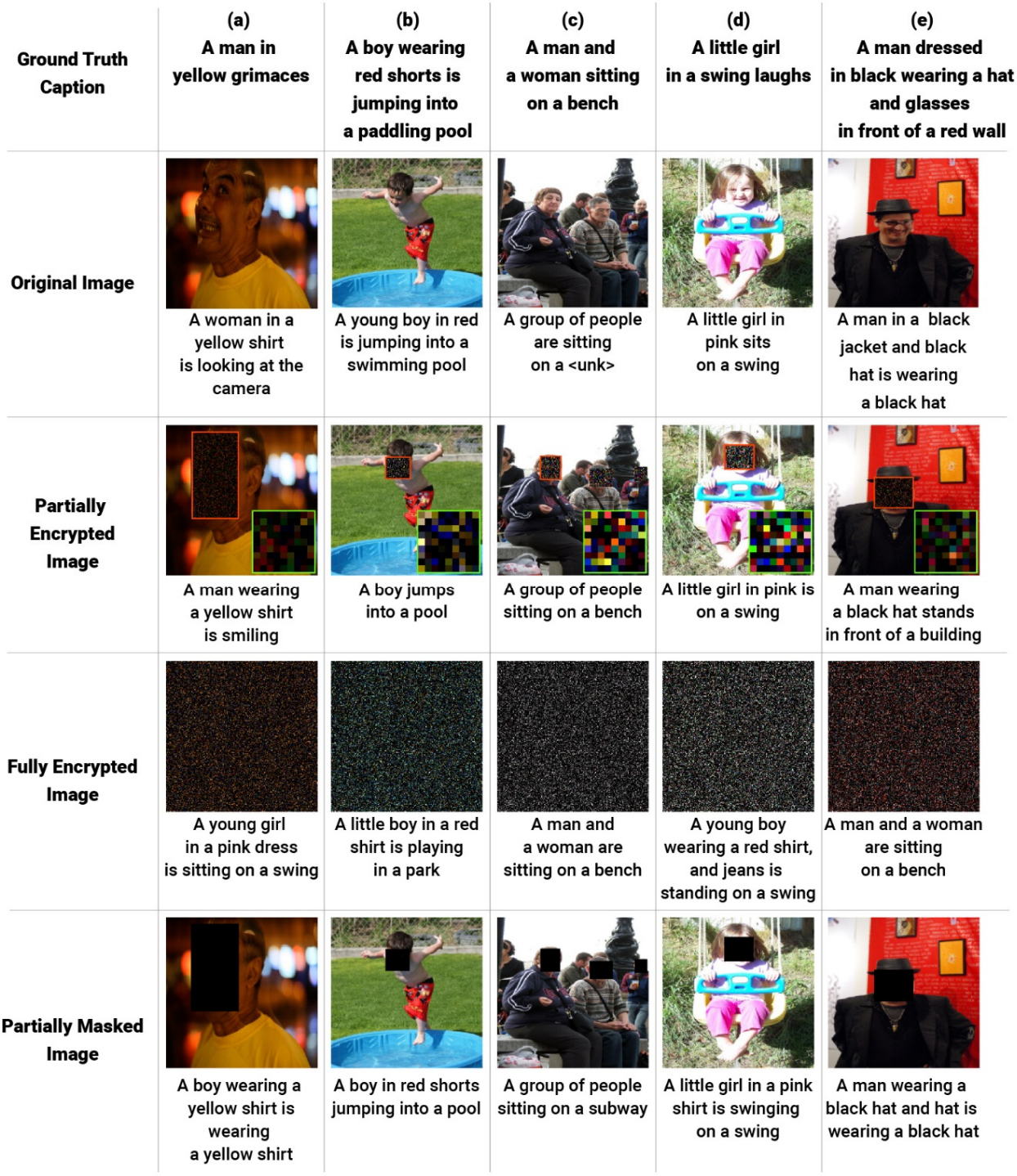

Figure 4. Visualization of generated captions from the test dataset. The predicted captions for each model trained on the four different images (original images, partially encrypted images using DRPE, fully encrypted images using DRPE, and partially masked images) are below each image.

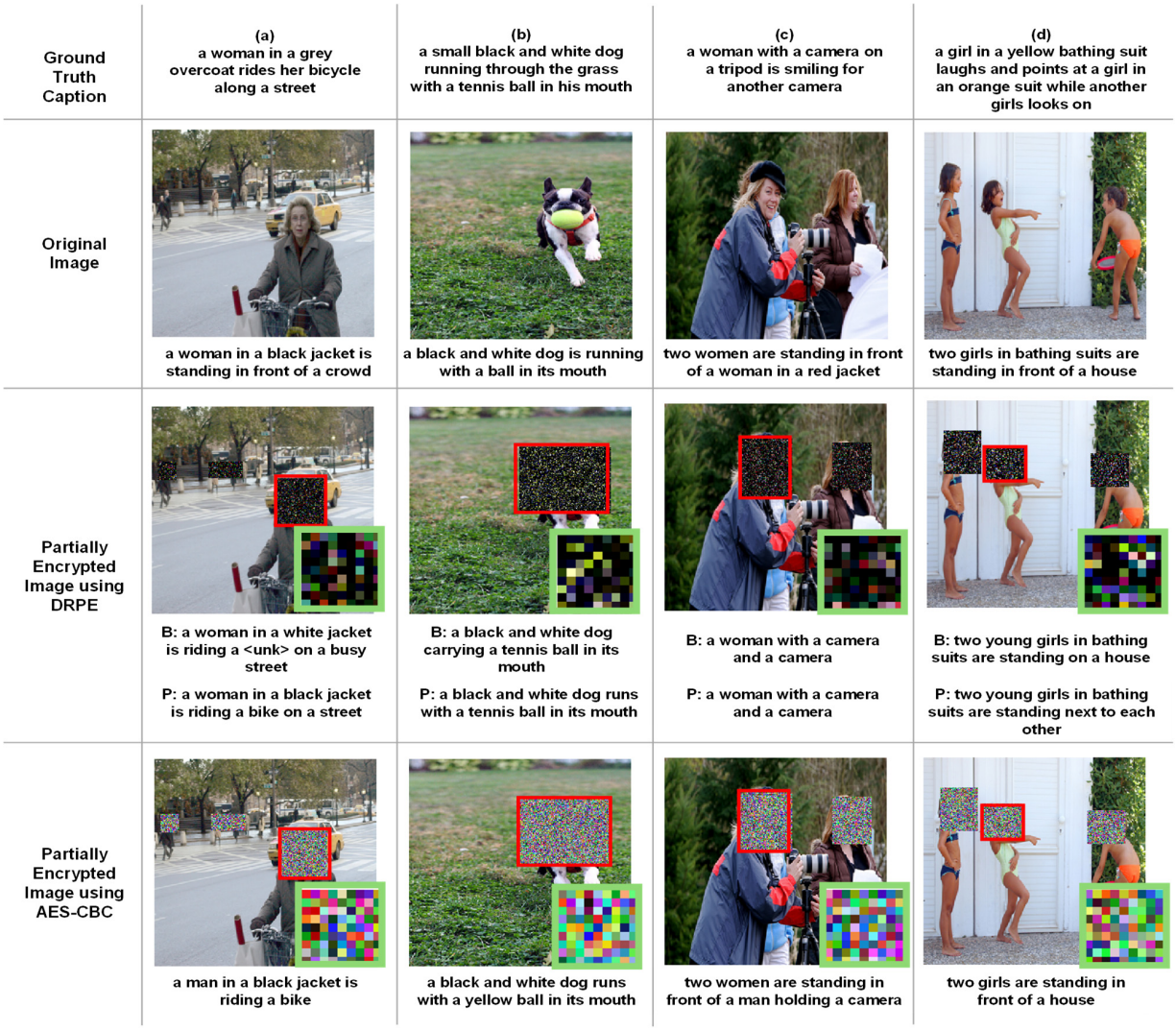

Figure 5. Visualization of generated captions from the test dataset. The predicted captions for each model trained on the three different images (original images, partially encrypted images using DPRE, and partially encrypted images using AES-CBC) are below each image. For the partially encrypted image using DRPE, B denotes the baseline model caption, and P denotes the proposed model caption.

Figure 6 illustrates the captioning results of 5 sample images with different types of encrypted objects. For Figure 6a, only the human faces are encrypted. The predicted caption remains relatively close to the ground truth caption except for the mention of the genders in the original images; therefore, the words "two men and a woman" are replaced with "a group of people" in the predicted caption. For Figure 6b, though both the faces of the children and the dogs are encrypted, the model failed to identify the dogs in the image as the caption focuses on the children in the pool. Thus, we can assume that the model struggles when there is a mixture of human and animal faces encrypted in the image. For Figure 6c, where the vehicle's license plate and the driver's face are encrypted, the predicted caption accurately describes the presence of the car, though it misses the color information about the car. In Figure 6d, the faces of the protestants and the text on the signboards are encrypted, but the predicted caption turned out to be more generic. This is also noticed in Figure 6e, where the restaurant signs and other objects are encrypted. We can assume that the model might be more sensitive to the encryption of human faces and probably animal faces that encrypted text information. Thus, a more diverse dataset would improve the performance of the model.

Figure 6. Impact of different types of encrypted objects on image captioning. The first row shows the original images, while the second shows the partially encrypted images. The encrypted regions are (**a**) only human faces, (**b**) human and dog faces, (**c**) vehicle license plate and human face, (**d**) human faces and text on signboards, and (**e**) restaurant signs and other objects.

Figure 7 illustrates an analysis of caption generation errors highlighting key failure cases such as subject misidentification, omission of details, and action mismatch. From the figure, the model struggles to identify some attributes, such as in Figure 7c (brown-eyed boy syrup on his lips); instead, it introduces details such as "wearing a blue shirt" and "smiling". Though the young boy is wearing a blue shirt, the word "smiling" in the predicted caption is an action mismatch. Another example of an action mismatch is seen in Figure 7a, where "standing" is predicted instead of "walking". Also, the model misidentifies the gender of one lady as the predicted caption states "a man and a woman" instead of "two ladies". A subject misidentification occurs in Figure 7d, where the prediction has "a black dog" instead of "a bull". This can be attributed to the fact that the dataset contains more images of dogs than other animals. In Figure 7b, we noticed the model focused on the more salient object (the man riding the bike) while ignoring secondary elements (the two others following). Therefore, refining the dual-stream encoder to better handle the encrypted regions during training could minimize the generation error.

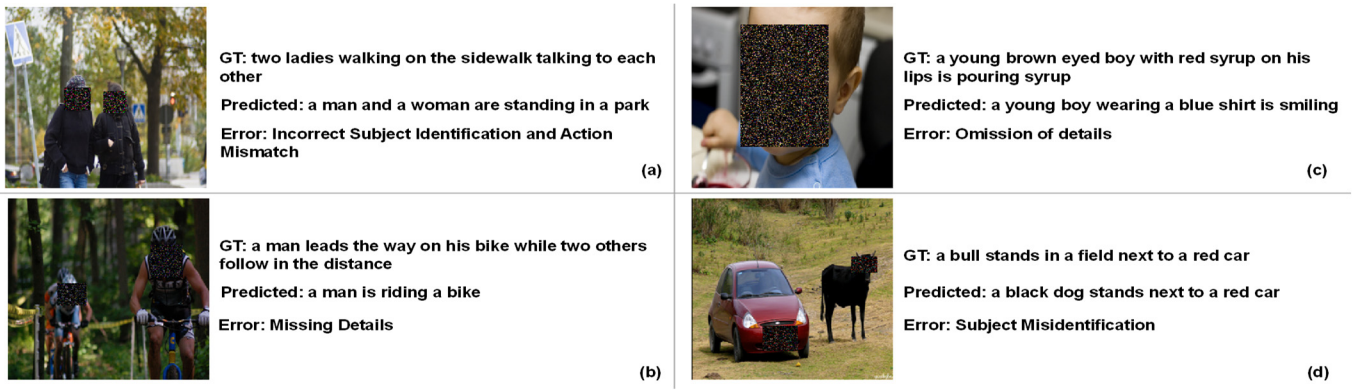

Figure 7. Analysis of caption generation errors highlighting key failure cases such as (**a**,**d**) subject misidentification, (**b**,**c**) omission of details, and (**a**) action mismatch.

4.5. Comparison with State of the Art

We compare the performance of our approach with several proposed models for image captioning. The models we compare include Google NIC [49], which established early foundations in image captioning, and Soft-Attention and Hard-Attention [50], which use VGG for image extraction and LSTM for sentence generation in the decoder. However, in [50], attention mechanisms are utilized in the decoder to enhance performance. More recent approaches include SwinCaption [51], which uses a Swin transformer to extract image features and a feature enhancement technique to capture more information about the features, and 2PSC-w and 2PSC [5], which uses ResNet101 to extract image features, and LSTM with an attention mechanism for generating sequences. 2PSC-w is trained on original images, while 2PSC is trained on images from the proposed optical encoder for scene distortion to protect privacy. Additionally, we compare to the models in [52–54] that introduce advanced techniques. Reference [52] employs an ensemble learning strategy, which combines eight CNN models via a voting process to fine-tune the ideal caption for every image. Reference [54] utilizes wavelet decomposition, a visual attention prediction network, and a contextual spatial relation extractor for effective feature extraction [53] proposed a network that perceives object-level information from inter-layer fusion and intra-layer interaction in the transformer-based decoder.

Table 7 shows that our proposed method, trained on original images using a single encoder, achieved the highest results for the BLEU-2 (51.56), BLEU-3 (37.36), ROUGE (49.68), and CIDEr (65.48) metrics. Meanwhile, our proposed method, trained on partially encrypted images using DRPE, achieved the second-highest results on the aforementioned metrics. The relatively small difference in performance (1-3%) between our models trained on original and partially encrypted images indicates that our model effectively captures semantic information even under partial encryption. Our proposed model performed better than training on the fully encrypted DRPE model proposed in [7], highlighting the challenge of generating accurate captions from fully encrypted images. It further indicates the effectiveness of partial encryption.

Table 7. Comparison results on the test dataset.

Images	Model	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	Rouge	Cider
Original	Google NIC [49]	63	41	27	-	-	-	-
	Soft-Attention [50]	67	44.8	29.9	19.5	18.93	-	-
	Hard-Attention [50]	67	45.7	31.4	21.3	20.3	-	-
	2PSC-w [5]	65.7	47.6	33.9	23.8	**25.5**	35.6	57.9
	Our proposed method	69.59	**51.56**	**37.36**	<u>26.43</u>	22.88	**49.68**	**65.48**
	SwinCaption [51]	67.7	46.8	32.9	22.9	22.7	-	-
	[52]	**72.8**	49.5	32.3	20.8	23.5	43.2	60.4
	[53]	67.4	-	-	24.3	21.5	44.8	63.6
	[54]	<u>70.5</u>	50.2	37.3	**28.6**	24.5	-	-
Fully encrypted using DRPE Privacy	[7]	48.3	29	17.1	10.1	13.6	36.1	22.5
Partially encrypted	2PSC [5]	63.5	45.2	31.4	21.5	<u>24.7</u>	34.7	51.8
	AES—CBC	66.69	49.15	35.10	24.49	22.71	49.57	63.36
	Baseline (ours) (DRPE)	64.64	46.81	32.78	22.34	22.35	48.71	58.96
	Our proposed method (DRPE)	<u>68.36</u>	<u>50.36</u>	<u>36.00</u>	<u>25.16</u>	22.58	<u>49.38</u>	<u>64.48</u>

Bold results are the highest, and underlined results are the second highest.

Our proposed model performed comparably better than some methods trained on original images. Thus, our model can learn from the images' encrypted and unencrypted parts. These results demonstrate that our proposed architecture offers a balance between privacy and utility. However, we note that the model [54] achieved the highest BLEU-4 score (28.6), which indicates that the use of a visual attention predictor network that consists

of atrous convolution, channel attention, and spatial attention further allows extraction of relevant features that are needed to acquire the best captions. The ensemble approach in [52] achieved the highest BLEU-1 score (72.8), which denotes that different encoders can help capture different aspects of an image as we use a dual-stream encoder for our proposed framework (partial encryption). Likewise, Table 1 confirms this as the dual-stream encoder outperformed the single encoder for the partially encrypted images. Future work could explore incorporating attention mechanisms such as that in [54] for the encoder to improve the model's performance.

5. Conclusions

This paper demonstrated the feasibility of performing image captioning on partially encrypted images using Double Random Phase Encoding. Specific regions of interest said to have sensitive information are encrypted using DRPE, resulting in a stationary white noise rendering no visual information to unauthorized parties. By partially encrypting the images, the structure of the image is preserved, and enough background information is retained. We used a dual-stream encoder based on the ResNet50 architecture and a transformer-based decoder to generate captions for the images. The dual-stream encoder processes the encrypted regions' real and imaginary parts separately, and we fine-tuned all the model layers, leading to richer feature extraction. We trained the embeddings in the embedding layer of the transformer from scratch, and the number of encoder and decoder layers is set to 2–2. We performed ablation studies to validate our architecture. We evaluated the proposed framework on the Flickr8k dataset and achieved a BLEU-4 score of 25.16, demonstrating a comparable performance to models trained on original images, and the captions generated fairly describe the provided images. Also, our method outperforms the image captioning method that utilized fully encrypted images, though using fully encrypted images provides more robust security. From this, we can assume that encrypting the whole image makes it difficult for the model to extract meaningful features for captioning.

Additionally, comparing our proposed method (partially encrypted images) with partially blocked images suggests that encryption preserves and keeps the image information rather than losing it by blocking or masking. Thus, partially encrypted images offer advantages over fully encrypted and partially blocked images (masked with black or gray boxes) as they balance privacy (encryption of sensitive information) and retain other information, which allows the model to process both areas and generate captions. Our results suggest that our method is a viable solution for privacy-preserving image captioning using deep learning and DRPE. From our results, we can assume that the proposed model helps use machine learning as a service as it ensures data privacy and information security. Furthermore, incorporating attention mechanisms in the encoder or using a vision transformer could further enhance the model's performance in order to minimize caption generation error. An extension of this work would be to use complex and diverse datasets such as the COCO dataset.

Author Contributions: Conceptualization, I.M.; methodology, A.D.M.; software, A.D.M.; validation, A.D.M. and I.M.; writing—original draft preparation, A.D.M.; writing—review and editing, I.M.; visualization, A.D.M.; supervision, I.M.; funding acquisition, I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No. 2020-0-00126, Research on AI-based Cryptanalysis and Security Evaluation). This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded

by the Korean government (MSIT) (RS-2024-00400368, Development of Image Integrity Verification Technology for Forgery Prevention and Original Proof).

Data Availability Statement: The dataset used in this paper is a public dataset, which has been referenced in the manuscript.

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

References

1. Sharma, D.; Dhiman, C.; Kumar, D. Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey. *Expert Syst. Appl.* **2023**, *221*, 119773. [CrossRef]
2. 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, Seattle, WA, USA, 14–19 June 2020. [CrossRef]
3. Ke, L.; Pei, W.; Li, R.; Shen, X.; Tai, Y.W. Reflective Decoding Network for Image Captioning. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV 2019), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8887–8896. [CrossRef]
4. Lu, J.; Xiong, C.; Parikh, D.; Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2017; pp. 3242–3250. [CrossRef]
5. Arguello, P.; Lopez, J.; Sanchez, K.; Hinojosa, C.; Rojas-Morales, F.; Arguello, H. Learning to Describe Scenes via Privacy-Aware Designed Optical Lens. *IEEE Trans. Comput. Imaging* **2024**, *10*, 1069–1079. [CrossRef]
6. Wu, Q.; Shen, C.; Wang, P.; Dick, A.; Van Den Hengel, A. Image Captioning and Visual Question Answering Based on Attributes and External Knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 1367–1381. [CrossRef] [PubMed]
7. Martin, A.D.; Ahmadzadeh, E.; Moon, I. Privacy-Preserving Image Captioning with Deep Learning and Double Random Phase Encoding. *Mathematics* **2022**, *10*, 2859. [CrossRef]
8. Alzubi, J.A.; Jain, R.; Nagrath, P.; Satapathy, S.; Taneja, S.; Gupta, P. Deep image captioning using an ensemble of CNN and LSTM based deep neural networks. *J. Intell. Fuzzy Syst.* **2021**, *40*, 5761–5769. [CrossRef]
9. Omri, M.; Abdel-Khalek, S.; Khalil, E.M.; Bouslimi, J.; Joshi, G.P. Modeling of Hyperparameter Tuned Deep Learning Model for Automated Image Captioning. *Mathematics* **2022**, *10*, 288. [CrossRef]
10. Mokady, R.; Hertz, A.; Bermano, A.H. ClipCap: CLIP Prefix for Image Captioning. November 2021. Available online: https://arxiv.org/abs/2111.09734v1 (accessed on 16 August 2024).
11. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. Available online: http://image-net.org/challenges/LSVRC/2015/ (accessed on 19 August 2021).
12. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, 21–26 July 2016; pp. 2261–2269. [CrossRef]
13. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *39*, 1137–1149. [CrossRef]
14. Singh, H.; Sharma, A.; Pant, M. Pixels to Prose: Understanding the Art of Image Captioning. *arXiv* **2024**, arXiv:2408.15714.
15. Rennie, S.J.; Marcheret, E.; Mroueh, Y.; Ross, J.; Goel, V. Self-critical Sequence Training for Image Captioning. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Las Vegas, NV, USA, 27–30 July 2016; pp. 1179–1195. [CrossRef]
16. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the ICLR 2021—9th International Conference on Learning Representations, Virtual, 3–7 May 2020. Available online: https://arxiv.org/abs/2010.11929v2 (accessed on 16 August 2024).
17. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
18. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020. [CrossRef]
19. Qiu, J.; Lo, F.P.; Gu, X.; Jobarteh, M.L.; Jia, W.; Baranowski, T.; Steiner-Asiedu, M.; Anderson, A.K.; McCrory, M.A.; Sazonov, E.; et al. Egocentric Image Captioning for Privacy-Preserved Passive Dietary Intake Monitoring. *IEEE Trans. Cybern.* **2024**, *54*, 679–692. [CrossRef]
20. Arguello, P.; Lopez, J.; Hinojosa, C.; Arguello, H. Optics Lens Design for Privacy-Preserving Scene Captioning. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 6–19 October 2022. [CrossRef]

21. Sitzmann, V.; Diamond, S.; Peng, Y.; Dun, X.; Boyd, S.; Heidrich, W.; Heide, F.; Wetzstein, G. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Trans. Graph.* **2018**, *37*, 114. [CrossRef]
22. Hinojosa, C.; Niebles, J.C.; Arguello, H. Learning Privacy-preserving Optics for Human Pose Estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021. [CrossRef]
23. Hinojosa, C.; Marquez, M.; Arguello, H.; Adeli, E.; Fei-Fei, L.; Niebles, J.C. PrivHAR: Recognizing Human Actions from Privacy-Preserving Lens. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer Nature: Cham, Switzerland, 2022. [CrossRef]
24. Zhang, Z.; Zhou, F.; Qin, S.; Jia, Q.; Xu, Z. Privacy-Preserving Image Retrieval and Sharing in Social Multimedia Applications. *IEEE Access* **2020**, *8*, 66828–66838. [CrossRef]
25. Yi, F.; Jeong, O.; Moon, I. Privacy-Preserving Image Classification with Deep Learning and Double Random Phase Encoding. *IEEE Access* **2021**, *9*, 136126–136134. [CrossRef]
26. Tanuwidjaja, H.C.; Choi, R.; Baek, S.; Kim, K. Privacy-preserving deep learning on machine learning as a service-a comprehensive survey. *IEEE Access* **2020**, *8*, 167425–167447. [CrossRef]
27. Javidi, B.; Refregier, P. Optical image encryption based on input plane and Fourier plane random encoding. *Opt. Lett.* **1995**, *20*, 767–769. [CrossRef]
28. Hodosh, M.; Young, P.; Hockenmaier, J. Framing Image Description as a Ranking Task: Data, Models and Evaluation Metrics. *J. Artif. Intell. Res.* **2013**, *47*, 853–899. [CrossRef]
29. 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; pp. 311–318. [CrossRef]
30. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. 2004, Volume Text Summa, pp. 74–81. Available online: https://aclanthology.org/W04-1013/ (accessed on 14 April 2022).
31. Denkowski, M.; Lavie, A. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, MD, USA, 26–27 June 2014; pp. 376–380. [CrossRef]
32. 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, Boston, MA, USA, 7–12 June 2015; pp. 4566–4575. [CrossRef]
33. Feretzakis, G.; Papaspyridis, K.; Gkoulalas-Divanis, A.; Verykios, V.S. Privacy-Preserving Techniques in Generative AI and Large Language Models: A Narrative Review. *Information* **2024**, *15*, 697. [CrossRef]
34. Yang, Y.; Zhang, B.; Guo, D.; Du, H.; Xiong, Z.; Niyato, D.; Han, Z. Generative AI for Secure and Privacy-Preserving Mobile Crowdsensing. *IEEE Wirel. Commun.* **2024**, *31*, 29–38. [CrossRef]
35. Vu, X.-S.; Le, D.-T.; Jiang, L.; Nguyen, H.D. Privacy-Preserving Visual Content Tagging using Graph Transformer Networks. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020. [CrossRef]
36. Sander, T.; Yu, Y.; Sanjabi, M.; Durmus, A.; Ma, Y.; Chaudhuri, K.; Guo, C. Differentially Private Representation Learning via Image Captioning. *Proc. Mach. Learn. Res.* **2024**, *235*, 43255–43275.
37. Aziz, R.; Banerjee, S.; Bouzefrane, S.; Le Vinh, T. Exploring Homomorphic Encryption and Differential Privacy Techniques towards Secure Federated Learning Paradigm. *Future Internet* **2023**, *15*, 310. [CrossRef]
38. Rovida, L.; Leporati, A. Encrypted Image Classification with Low Memory Footprint Using Fully Homomorphic Encryption. *Int. J. Neural Syst.* **2024**, *34*, 2450025. [CrossRef]
39. Padate, R.; Kalla, M.; Gupta, A.; Sharma, A. Federated Learning for Image Captioning: A Comprehensive Review of Privacy-Preserving Collaborative Model Training in Distributed Environments. In Proceedings of the 2023 2nd International Conference on Edge Computing and Applications, Namakkal, India, 19–21 July 2023; pp. 857–865.
40. Liu, F.; Wu, X.; Ge, S.; Fan, W.; Zou, Y. Federated Learning for Vision-and-Language Grounding Problems. In Proceedings of the AAAI Conference on Artificial Intelligence, New York City, NY, USA, 7–8 February 2020; Volume 34, pp. 11572–11579. [CrossRef]
41. Jeong, O.; Moon, I. Adaptive transfer learning-based cryptanalysis on double random phase encoding. *Opt. Laser Technol.* **2024**, *168*, 109916. [CrossRef]
42. Zhang, L.; Zhou, Y.; Huo, D.; Li, J.; Zhou, X. Multiple-image encryption based on double random phase encoding and compressive sensing by using a measurement array preprocessed with orthogonal-basis matrices. *Opt. Laser Technol.* **2018**, *105*, 162–170. [CrossRef]
43. Karpathy, A.; Fei-Fei, L. Deep Visual-Semantic Alignments for Generating Image Descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *39*, 664–676. [CrossRef]
44. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; pp. 1532–1543. [CrossRef]

45. Baziotis, C.; Pelekis, N.; Doulkeridis, C. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, BC, Canada, 3–4 August 2017; pp. 747–754. [CrossRef]
46. Aka Uymaz, H.; Kumova Metin, S. Vector based sentiment and emotion analysis from text: A survey. *Eng. Appl. Artif. Intell.* **2022**, *113*, 104922. [CrossRef]
47. Nechvatal, J.; Barker, E.; Bassham, L.; Burr, W.; Dworkin, M.; Foti, J.; Roback, E. Report on the Development of the Advanced Encryption Standard (AES). *J. Res. (NIST JRES) Natl. Inst. Stand. Technol.* **2001**, *106*, 511–577. [CrossRef]
48. Lidkea, V.M.; Muresan, R.; Al-Dweik, A. Convolutional Neural Network Framework for Encrypted Image Classification in Cloud-Based ITS. *IEEE Open J. Intell. Transp. Syst.* **2020**, *1*, 35–50. [CrossRef]
49. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2014; pp. 3156–3164.
50. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R.; Bengio, Y. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In Proceedings of the 32nd International Conference on International Conference on Machine Learning (ICML 2015), Lille, France, 6–11 July 2015; Volume 3, pp. 2048–2057. [CrossRef]
51. Liu, L.; Jiao, Y.; Li, X.; Li, J.; Wang, H.; Cao, X. Swin Transformer-based Image Captioning with Feature Enhancement and Multi-stage Fusion. In Proceedings of the ICNC-FSKD 2023—2023 International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery, Harbin, China, 29–31 July 2023. [CrossRef]
52. Al Badarneh, I.; Hammo, B.H.; Al-Kadi, O. An ensemble model with attention based mechanism for image captioning. *Comput. Electr. Eng.* **2025**, *123*, 110077. [CrossRef]
53. Ma, Y.; Ji, J.; Sun, X.; Zhou, Y.; Ji, R. Towards local visual modeling for image captioning. *Pattern Recognit.* **2023**, *138*, 109420. [CrossRef]
54. Sasibhooshan, R.; Kumaraswamy, S.; Sasidharan, S. Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction. *J. Big Data* **2023**, *10*, 18. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.