BiCap: Bangla Image Captioning Using Attention-based Encoder-Decoder Architecture

Md Aminul Kader Bulbul
University of Dhaka
Dhaka-1000, Bangladesh
mdaminulkader-2017814980@cs.du.ac.bd

Abstract

Automatic image captioning has gained significant attention at the intersection of computer vision and natural language processing, yet research in low-resource languages such as Bangla remains limited. This work introduces BiCap, an attention-based encoder–decoder framework designed for Bangla image captioning. The model leverages a pretrained ResNet-50 as the encoder to extract rich visual features and a Long Short-Term Memory (LSTM) network as the decoder to sequentially generate Bangla captions. To overcome the fixed-length bottleneck of traditional encoder–decoder architectures, we integrate Bahdanau attention, enabling the decoder to dynamically focus on salient image regions while producing each word. The model is trained and evaluated on the Chitron dataset, with extensive preprocessing including vocabulary construction, tokenization, and word embedding. Experimental results demonstrate that BiCap achieves superior performance over the existing works (Masud et al., 2025; Hossain et al., 2024; Das et al., 2023; Humaira et al., 2021), yielding higher BLEU, METEOR, ROUGE, CIDEr scores. Improved fluency in human evaluation further confirms that the model generates more contextually accurate and semantically coherent captions, although occasional challenges remain with complex scenes. Recent advances in Vision–Language Models (VLMs), such as CLIP, BLIP, Flamingo, LLaVA, and MiniGPT-4, have redefined state-of-the-art captioning performance in high-resource settings. However, these models require large multimodal corpora and extensive pretraining that are currently unavailable for Bangla. BiCap therefore offers a resource-efficient, interpretable, and practically deployable solution tailored to low-resource multimodal learning.

1 Introduction

Image captioning, the task of automatically generating natural language descriptions for images, has become a central problem at the intersection of computer vision (CV) and natural language processing (NLP). It requires extracting high-level semantic information from visual inputs and mapping it into coherent textual sequences. Early captioning approaches relied on template-based methods (Kulkarni et al., 2013) or retrieval-based systems (Ordonez et al., 2011), which lacked flexibility and generalization. With the advent of deep learning, encoder-decoder architectures combining convolutional neural networks (CNNs) and recurrent neural networks (RNNs) achieved significant success (Vinyals et al., 2015). Subsequent advancements introduced attention mechanisms (Chorowski et al., 2015; Xu et al., 2015) to dynamically focus on salient image regions, alleviating the encoder-decoder bottleneck and improving semantic alignment between images and captions.

Meanwhile, the global landscape of image captioning has shifted toward Vision–Language Models (VLMs) such as CLIP (Radford et al., 2021), BLIP/BLIP-2 (Li et al., 2022, 2023), Flamingo (Alayrac et al., 2022), and multimodal LLM frameworks like LLaVA (Liu et al., 2024) and MiniGPT-4 (Zhu et al.). These systems leverage large-scale multimodal pretraining and transformer architectures to achieve near-human captioning performance. Their reliance on massive English-centric paired corpora, however, makes them impractical for Bangla, where annotated multimodal resources are extremely limited.

Despite these advancements in English, low-resource languages such as Bangla lacks large-scale multimodal corpora and pretrained vision-language models. From an NLP perspective, Bangla is morphologically rich, exhibits relatively free word order, and contains complex inflectional and derivational structures. These properties exacerbate challenges such as out-of-vocabulary (OOV) words, limited word embeddings, and poor generalization in sequence-to-sequence tasks. While efforts like

Proceedings of the Second Workshop on Bangla Language Processing (BLP-2025), pages 80–90

BNLIT (Rahman et al., 2019) represent initial attempts at Bangla captioning, generated sentences often suffer from repetition, limited lexical diversity, and semantic incompleteness. This highlights a critical need for models that can bridge the multimodal gap while generating fluent and contextually appropriate Bangla captions.

To address this gap, this study introduces BiCap, an attention-based encoder-decoder model designed for Bangla image captioning. BiCap is designed as a resource-efficient alternative to large transformer-based vision–language architectures, offering interpretable attention patterns and competitive captioning performance in the low-resource Bangla setting.

BiCap employs a pretrained ResNet-50 (He et al., 2016) to extract rich visual embeddings, while a Long Short-Term Memory (LSTM) decoder generates sequential captions. To mitigate the bottleneck issue in traditional encoder-decoder systems, Bahdanau additive attention (Chorowski et al., 2015) is integrated, allowing the decoder to attend to different spatial regions of the image at each time step. This design ensures stronger semantic alignment between visual content and textual descriptions, improving both adequacy and fluency in Bangla captions.

Experimental validation on the Chitron dataset (Sazzed, 2020) shows that BiCap outperforms the existing baselines in terms of BLEU, METEOR, ROUGE, CIDEr scores. Human evaluation further indicates significant improvements in fluency, relevance, and semantic adequacy, confirming the effectiveness of attention in Bangla captioning.

The contributions of this study are as follows:

- An attention-based encoder–decoder framework for Bangla image captioning using ResNet-50, Bahdanau attention, and an LSTM decoder.
- Advancement of Bangla multimodal NLP by demonstrating that attention-based architectures remain effective in low-resource scenarios.
- A thorough evaluation on the largest available Bangla captioning dataset (Chitron), along with human assessment of fluency, adequacy, and relevance.
- A reusable and extensible architecture, providing a foundation for future work involving transformer decoders, multilingual embeddings, or VLM-style pretraining.

2 Related Works

The field of image captioning has been significantly advanced by the encoder-decoder architecture, which links a visual input to a natural language description. A cornerstone of this research is the work of Vinyals et al. (Vinyals et al., 2015), who introduced a complete end-to-end system utilizing a deep Convolutional Neural Network (CNN) as an encoder to extract image features and a Long Short-Term Memory (LSTM) recurrent neural network as a decoder to generate the caption. This foundational approach was further refined by subsequent research. For instance, Rennie et al. (Rennie et al., 2017) developed Self-Critical Sequence Training (SCST), a reinforcement learning method that directly optimizes non-differentiable metrics like CIDEr and BLEU, leading to substantial performance improvements by better aligning generated captions with human evaluation. Other recent advancements include convolution-free models, such as the "Full-memory transformer for image captioning" by (Lu et al., 2023), replaces the CNN encoder with a Transformer to more effectively model global visual context.

Applying these models to low-resource languages, such as Bangla, presents a unique set of challenges, primarily stemming from the lack of large, high-quality, annotated datasets and the morphological complexity of the language itself. Automatic image captioning in low-resource languages, such as Bangla, remains underexplored due to dataset scarcity and modeling challenges; a CNN–BiLSTM hybrid encoder–decoder trained on the BNLIT dataset (Rahman et al., 2019) was the first notable work in this domain, though its generalizability across broader domains remains limited. (Humaira et al., 2021) employed hybrid CNN–RNN architectures on Flickr8k and BanglaLekha datasets, reporting improved BLEU scores, though the small dataset size constrains broader applicability. A recent study (Hossain et al., 2024) proposed an image captioning approach using EfficientNetB4 and ResNet-50 for feature extraction. Another recent study (Masud et al., 2025) introduced a human-annotated dataset and an attention-driven GRU-based end-to-end model, though scalability to larger and more diverse corpora is yet to be demonstrated.

2.1 Vision–Language Models (VLMs)

Modern vision–language models (VLMs) learn joint visual–textual representations through large-scale image–text pretraining. CLIP (Radford et al., 2021) trains on 400M paired samples using contrastive learning, while BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023) unify captioning and visual grounding via transformer encoders and Q-former modules. Flamingo (Alayrac et al., 2022) incorporates cross-attention layers between a vision encoder and a frozen large language model. LLaVA (Liu et al., 2024) aligns CLIP embeddings with Vicuna to enable multimodal dialogue, and MiniGPT-4 (Zhu et al.) links ViT-G/14 with LLaMA-based decoders using lightweight alignment layers. These architectures demonstrate that multimodal reasoning requires substantial data, carefully aligned encoders–decoders, and sophisticated cross-modal fusion.

However, their applicability to Bangla is limited. Large-scale paired datasets required by these models do not exist for Bangla. The pretraining pipelines are predominantly English-centric, which leads to weak Bangla generalization even when some multilingual text is included. Moreover, the absence of Bangla multimodal benchmarks prevents effective evaluation or pretraining at VLM scale.

2.2 Large Language Models (LLMs)

Large-language-model–based captioning frameworks extend multimodal generation by leveraging powerful text decoders. OFA (Wang et al., 2022) provides a unified architecture for captioning and visual question answering, while PaLI (Chen et al.) and PaLI-X (Chen et al., 2024) scale this approach using mixture-of-experts vision–language components. GIT (Wang et al.) adopts a GPT-style decoder trained on 800M image–text pairs, enabling fluent caption synthesis at scale. InstructBLIP (Dai et al., 2023) advances this line of work by introducing instruction-following capabilities, allowing open-ended multimodal reasoning and adaptable task formats.

These systems, however, face significant constraints for Bangla captioning. Their training requires massive multimodal corpora, which are unavailable for Bangla, and multilingual tokenization struggles with Bangla's morphology, reducing downstream accuracy. Fine-tuning LLMs at this scale is computationally out of reach for most Bangla research environments, and the ecosystem lacks any open instruction-tuned multimodal model specifically optimized for Bangla tasks.

2.3 Bangla Transformer-Based Captioning

Bornon (Muhammad Shah et al., 2022) introduced a transformer-based decoder for Bangla captioning. Despite being an early effort in Bangla captioning, Bornon has several limitations that restrict its effectiveness. It is not a true multimodal Transformer and relies on global Inception-v3 features without region-level attention, resulting in weak visual grounding. The Transformer decoder also requires large datasets, but Bangla resources are limited, leading to overfitting and unstable training. In addition, Bornon uses basic word-level tokenization that poorly handles Bangla's rich morphology, causing frequent OOV and fluency issues. Its evaluation is limited to a small dataset and lacks human assessments or strong baseline comparisons.

2.4 Positioning of our proposed BiCap Architecture

Our proposed approach - BiCap is positioned as a practical and resource-efficient captioning model tailored for Bangla's low-resource setting. In contrast to VLM and LLM-based captioners that depend on massive multimodal corpora and large transformer architectures, BiCap achieves strong performance using a lightweight design. It combines ResNet-50 for efficient visual feature extraction, Bahdanau attention for fine-grained image–text alignment, and an LSTM decoder well-suited to small datasets and Bangla's morphological complexity. This makes BiCap more interpretable, more stable during training, and more effective under limited data conditions, while still remaining extensible for future integration with multilingual transformers or VLM-style modules.

3 Methodology

The BiCap model is an end-to-end multimodal architecture designed for the task of image captioning in the Bangla language. It operates on an attention-based encoder-decoder framework, meticulously crafted to synchronize the extraction of visual features with the generation of natural language sentences. The encoder, a robust convolutional neural network (CNN), processes the input image to produce a rich, spatially-distributed feature map. Subsequently, a dynamic attention mechanism acts as a spotlight, selectively highlighting

the most salient visual regions pertinent to each word being generated. The decoder, a Long Short-Term Memory (LSTM) network, leverages these attended visual features and the linguistic context of previously generated words to construct fluent and semantically accurate Bangla sentences. This unified pipeline ensures a deep fusion of visual and textual information, essential for grounded and contextually relevant caption generation. The overall architecture is illustrated in Figure 1.

3.1 Encoder

The encoder component serves as the visual backbone of the BiCap model, tasked with transforming raw pixel data from an input image into a high-level, semantic representation. For this purpose, we employ a ResNet-50 (He et al., 2016) architecture, a deeply-layered CNN pre-trained on the vast ImageNet dataset. The selection of ResNet-50 is motivated by its proven efficacy in various computer vision tasks, particularly its ability to learn hierarchical features and mitigate the vanishing gradient problem through residual connections. This pre-training allows the model to leverage a vast reservoir of learned visual knowledge, which is then fine-tuned for the specific task of image captioning.

Given an input image $I \in R^{H \times W \times 3}$, where H and W are the height and width, the encoder generates a spatial feature map F:

$$F = ResNet50(I), \quad F \in R^{H' \times W' \times D} \tag{1}$$

where $H' \times W'$ are the reduced spatial dimensions of the feature map, and $D = 2048$ represents the channel depth of the final convolutional layer's output. The resulting feature map F is then reshaped into a sequence of $n = H' \times W'$ feature vectors:

$$F = \{f_1, f_2, \ldots, f_n\}, \quad f_i \in R^D \tag{2}$$

Each vector f_i corresponds to a specific region of the original image, capturing localized visual information such as objects, textures, and background cues. This sequence of feature vectors retains the spatial granularity that is crucial for the subsequent attention mechanism, thereby avoiding the loss of fine-grained details that would occur if the features were condensed into a single global vector.

In this work, the encoder weights were initialized from the ImageNet-pretrained ResNet-50 model and partially fine-tuned during training. Specifically, the top convolutional block was updated, while the lower layers were kept frozen to preserve general visual representations and prevent overfitting on the relatively small Chitron dataset. This hybrid fine-tuning strategy ensures that the model adapts to the captioning domain without compromising training stability. We also observed that fully fine-tuning all layers led to faster overfitting, whereas keeping the encoder entirely frozen reduced caption diversity. The chosen partial fine-tuning configuration provides an effective trade-off between adaptability and generalization.

It is acknowledged that BiCap relies solely on a ResNet-50 encoder, which, as a pure convolutional network, primarily captures local and hierarchical visual cues. Such representations are sometimes insufficient for modeling complex global relationships among multiple objects or spatial arrangements in highly detailed scenes—an aspect crucial for generating fully contextualized captions. However, this design reflects a deliberate trade-off between representational power and resource efficiency. In the low-resource Bangla setting, transformer-based or hybrid vision encoders require substantially larger datasets and massive multimodal corpora, which are currently unavailable. The Bahdanau attention layer in BiCap helps compensate for this limitation by dynamically highlighting the most relevant regions, thereby enriching spatial context without incurring the computational cost of vision transformers. This choice ensures training stability, faster convergence, and interpretable attention maps. In the future work, we will explore lightweight CNN–ViT hybrids or multi-scale attention mechanisms to further enhance global scene understanding once larger Bangla multimodal datasets become accessible.

3.2 Attention Mechanism

A common challenge in encoder-decoder architectures is the bottleneck problem, where compressing all visual information into a single fixed-size vector can lead to the loss of fine-grained spatial and semantic details. To circumvent this, the BiCap model integrates an attention mechanism, which allows the decoder to dynamically focus on the most relevant parts of the image at each step of the caption generation process. We adopt the Bahdanau additive attention (Chorowski et al., 2015) mechanism due to its effectiveness in aligning visual and textual modalities.

At each decoding time step t, the attention mechanism computes an alignment score $e_{t,i}$ between

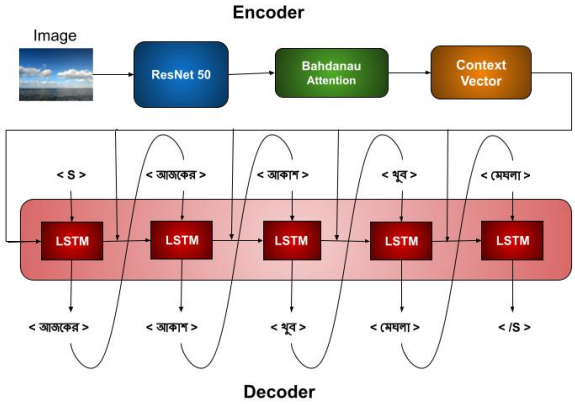

Figure 1: Overview of the BiCap architecture: ResNet-50 encoder extracts spatial features, Bahdanau attention computes context vectors, and an LSTM decoder generates captions sequentially.

the previous decoder hidden state h_{t-1} and each encoder feature vector f_i. This score quantifies the relevance of the i-th visual region to the next word to be generated. The alignment score is calculated as follows:

$$e_{t,i} = v_a^\top \tanh(W_h h_{t-1} + W_f f_i) \tag{3}$$

where $W_h \in R^{k \times d}$ and $W_f \in R^{k \times D}$ are learnable weight matrices, and $v_a \in R^k$ is a trainable vector. Here, d is the hidden dimension of the decoder and D is the dimension of the encoder features.

The alignment scores are then normalized using a softmax function to obtain a set of attention weights $\alpha_{t,i}$:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{n} \exp(e_{t,j})} \tag{4}$$

These weights, which sum to one, represent a probability distribution over the encoder's feature vectors. They indicate how much attention the decoder should pay to each visual region. The context vector c_t is then computed as a weighted sum of the encoder features, with the attention weights acting as coefficients:

$$c_t = \sum_{i=1}^{n} \alpha_{t,i} f_i \tag{5}$$

The context vector c_t is a dynamically created representation of the most salient visual information, specifically tailored for the generation of the word at time step t. This mechanism effectively simulates a human-like visual focus, ensuring that the decoder's output is not only syntactically correct but also semantically grounded in the visual content.

3.3 Decoder

The decoder's primary role is to translate the multimodal representation into a coherent and linguistically fluent Bangla sentence. For this, we utilize a Long Short-Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997), a type of recurrent neural network particularly well-suited for processing sequential data. LSTMs are capable of capturing long-range dependencies, which is critical for handling the rich morphology and flexible syntactic structure of the Bangla language.

At each time step t, the decoder takes as input the concatenation of the word embedding of the previous token $E(y_{t-1})$ and the attention-derived context vector c_t:

$$x_t = [E(y_{t-1}); c_t] \tag{6}$$

where $E(\cdot)$ is a word embedding function that maps each word in the vocabulary to a dense vector space. This combined input provides the LSTM with both the linguistic context (from the previous word) and the relevant visual context (from the attention mechanism). The LSTM then updates its hidden state h_t and cell state s_t based on the current input and the previous states:

$$h_t, s_t = LSTM(x_t, h_{t-1}, s_{t-1}) \tag{7}$$

The final hidden state h_t is then used to predict the next word in the sequence. This is achieved by passing the hidden state through a fully connected layer followed by a softmax function to produce a probability distribution over the entire vocabulary:

$$p(y_t \mid y_{1:t-1}, I) = Softmax(W_o h_t + b_o) \tag{8}$$

where W_o and b_o are learnable output parameters.

During training, we employ the teacher forcing technique, where the ground-truth word y_{t-1} is fed as input at each time step. This method stabilizes the training process and accelerates convergence by providing the model with correct historical context. In contrast, during inference, the model operates autoregressively, feeding its own predicted word from the previous step as input. This process continues until an end-of-sequence token is generated, at which point the complete Bangla caption is formed. The tight synchronization between the encoder's visual understanding, the attention mechanism's dynamic focus, and the decoder's linguistic generation is what enables the BiCap model to produce high-quality, descriptive Bangla captions.

We acknowledge that the LSTM-based decoder represents a sequential, Recurrent Neural Network (RNN) architecture. Such models process inputs step by step, which can limit their ability to capture very long-range dependencies and make them slower to train than fully parallelized Transformer decoders. Moreover, this architecture does not directly exploit the large-scale multimodal representations available in modern Transformer-based vision–language models such as BLIP or CLIP.

However, this design choice is intentional and well-motivated for Bangla captioning under low-resource conditions. Transformer-based decoders typically require millions of image–text pairs and powerful GPUs to train effectively, while the available Bangla datasets (e.g., Chitron) are relatively small. In such cases, RNN-based models like LSTM remain more stable, data-efficient, and interpretable. The combination of Bahdanau attention and LSTM allows BiCap to focus on semantically relevant regions of the image while maintaining grammatical and morphological accuracy in Bangla output. Furthermore, the sequential structure enables explicit attention visualization, offering interpretability that is often opaque in Transformer architectures. As Bangla multimodal resources expand, BiCap can be adapted to integrate Transformer or hybrid decoding modules to improve scalability and contextual reasoning in its future extensions.

4 Experimental Setup

4.1 Dataset

All experiments are conducted on the **Chitron** dataset (Sazzed, 2020), the largest publicly available Bangla image captioning corpus. It consists of 15,438 images with human-annotated Bangla captions, collected from diverse sources including news portals, Wikipedia Commons, and social media. The images span a variety of domains such as people, objects, scenes, and events, making it a suitable benchmark for Bangla captioning.

For this study, the dataset is partitioned into training (80%), validation (10%), and test (10%) splits. Table 1 summarizes the dataset statistics.

	Images & Captions	Vocabulary Size	Max Caption Length
Training	12350	9800	20
Validation	1544	—	20
Test	1544	—	20

Table 1: Chitron dataset statistics and experimental split.

4.2 Preprocessing

All images were resized to 299×299 pixels before feature extraction to ensure consistent input dimensions for the ResNet-50 encoder. This resolution is a standard preprocessing step inherited from the ImageNet training configuration, which ResNet-50 expects for optimal performance. Resizing maintains a uniform aspect ratio across images, reduces computational overhead, and enables efficient batch processing without affecting semantic content. Using 299×299 therefore ensures compatibility with pretrained convolutional filters while balancing accuracy and efficiency during training.

Captions are tokenized using a rule-based Bangla tokenizer that handles whitespace, punctuation, and compound words. Rare words (frequency < 5) are replaced with an <UNK> token, resulting in a vocabulary size of approximately 9.8K unique tokens. Special tokens <START> and <END> are appended to mark caption boundaries, and all sequences are padded or truncated to a maximum length of 20 tokens. Here, the maximum caption length was truncated to 20 tokens to standardize sequence lengths for training and reduce computational complexity. Analysis of the Chitron dataset revealed that most Bangla captions fall below this length, so truncation affects very few instances while allowing efficient batch processing and stable LSTM training. Limiting the token size also mitigates the risk of overfitting on rare, excessively long sequences and ensures that the attention mech-

anism focuses on the most semantically relevant words in the caption. Future work may explore dynamic sequence lengths or subword tokenization to better accommodate longer captions without compromising efficiency.

We acknowledge that BiCap uses a simple frequency-based preprocessing strategy, where rare words (frequency < 5) are replaced with a generic token, yielding a vocabulary of approximately 9.8K words. This approach may limit the model's ability to generate captions containing rare or morphologically complex Bangla words, potentially reducing linguistic diversity. However, this design choice was made to stabilize training and prevent sparsity in the embedding layer given the small dataset size. The model's focus is on producing semantically accurate and fluent captions rather than exhaustive lexical coverage. Despite the reduced vocabulary, BiCap achieves strong fluency and adequacy scores in both automatic and human evaluations, indicating that the simplified preprocessing did not significantly harm overall descriptive quality. Future work will explore subword-based tokenization and morphological segmentation to better capture Bangla's rich morphology and enhance lexical diversity.

BiCap incorporates several Bangla-specific adaptations across preprocessing and model design to better handle the language's rich morphology and script characteristics. During preprocessing, Unicode normalization and punctuation cleaning were applied to accommodate compound and conjunct characters in Bangla text. Tokenization was performed using a custom rule-based segmenter designed for Bangla word boundaries, as existing multilingual tokenizers often split characters incorrectly. In the decoder, Bangla word embeddings were trained from scratch to capture morphological variations such as inflection and postpositions, which differ from English syntax. The attention-based decoder further aids in aligning Bangla sentence structure with image regions, improving grammatical fluency. These tailored adjustments ensure that BiCap learns robust visual–linguistic associations specific to Bangla rather than relying on multilingual defaults.

4.3 Model Training

The encoder uses a pretrained ResNet-50 truncated before the final classification layer, while the decoder is a single-layer LSTM with hidden size 512 and embedding dimension 300. The attention module employs a hidden alignment size of 256. Dropout ($p = 0.5$) is applied to embeddings and LSTM outputs. The optimizer is Adam with learning rate 5×10^{-4}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and batch size 64.

4.4 Implementation Details

The model is implemented in `PyTorch` 1.11 and trained on an NVIDIA Tesla P100 GPU with 16 GB memory. Each training epoch requires approximately 10 minutes, and full training converges within 30-35 epochs.

5 Results and Discussion

This section provides a rigorous, data-driven evaluation of the proposed BiCap framework. A comprehensive set of established metrics (BLEU, METEOR, ROUGE, CIDEr), coupled with human judgments, is employed to quantify performance. The results of BiCap are systematically benchmarked against several state-of-the-art baseline models (Masud et al., 2025; Hossain et al., 2024; Das et al., 2023; Humaira et al., 2021). This comparative analysis is specifically designed to isolate and highlight the significant performance gains attributable to the incorporation of the Bahdanau attention mechanism in encoder-decoder model within the context of Bangla image captioning.

We acknowledge that the model has been trained and evaluated exclusively on the Chitron dataset, which, despite being the largest resource available, remains relatively small compared with datasets used in high-resource languages. This limitation naturally raises concerns regarding generalizability to unseen or cross-domain image distributions. However, this experimental focus is deliberate: BiCap is explicitly designed and optimized for low-resource multimodal settings, where large-scale Bangla datasets are unavailable. By operating effectively within such constraints, BiCap demonstrates the feasibility of attention-based captioning under realistic data scarcity conditions. Moreover, the model's attention maps and human evaluation results indicate that it learns meaningful visual–textual correspondences rather than overfitting to dataset bias. Future work will extend evaluation to additional Bangla or multilingual datasets and explore transfer learning or synthetic data augmentation to further enhance robustness and domain generalization.

5.1 Quantitative Results

To further assess the effectiveness of BiCap, we compare its performance against several existing Bangla image captioning models: (Masud et al., 2025; Hossain et al., 2024; Das et al., 2023; Humaira et al., 2021). Evaluation is conducted using standard captioning metrics: BLEU-1 to BLEU-4 (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015). Table 2 summarizes the results.

BiCap achieves state-of-the-art results across all evaluation metrics. Specifically, it records a BLEU-4 score of 32.73, which is over 10 points higher than the strongest baseline (Masud et al., 2025). Improvements are also consistent in METEOR (+0.03) and ROUGE (+0.06), indicating not only higher n-gram overlap but also better recall-oriented performance and semantic adequacy. The CIDEr score of 0.29 further demonstrates that BiCap produces captions more closely aligned with human references compared to earlier works.

The gains in BLEU-1 and BLEU-2 suggest that BiCap excels at generating accurate local word choices, while improvements in BLEU-3 and BLEU-4 highlight its ability to maintain coherence over longer n-grams. This reflects the effectiveness of the Bahdanau attention mechanism, which dynamically focuses on different visual regions during decoding, thereby avoiding the information bottleneck present in earlier CNN–RNN systems.

Another factor behind BiCap's superior performance is the use of a pretrained ResNet-50 encoder, which provides rich, high-level semantic embeddings that generalize well to diverse image domains in the Chitron dataset. In contrast, earlier works often relied on shallower CNNs or training from scratch, which limited their ability to capture complex visual concepts.

Finally, the integration of attention with an LSTM decoder enables BiCap to handle the morphological richness and free word order of Bangla more effectively than prior models. This leads to higher lexical diversity, fewer incomplete sentences, and improved alignment between visual objects and their textual descriptions.

Overall, the quantitative results confirm that BiCap sets a new benchmark for Bangla image captioning, outperforming methods developed between 2021 and 2025 due to its stronger visual representations, attention-driven alignment, and language-aware decoding strategy.

5.2 Human Evaluation

To complement automatic scores, 20 randomly sampled image–caption pairs from the test set were rated by 83 native Bangla speakers on three dimensions: *fluency* (grammatical correctness), *adequacy* (coverage of salient objects/events), and *relevance* (semantic alignment with the image). Ratings were given on a 5-point Likert scale.

Table 3 summarizes the results. BiCap surpasses the baseline across all three criteria, with particularly strong gains in adequacy and relevance, confirming that attention improves semantic grounding. This clarification emphasizes that the evaluation focused on the quality of the generated captions in context of their corresponding images, not on standalone image selection. Such human ratings complement the automatic metrics and provide a more holistic measure of caption quality.

5.3 Performance Analysis

The enhanced performance of the BiCap model is attributed to three primary architectural features:

1. **Rich Feature Extraction:** The use of a pretrained ResNet-50 encoder allows BiCap to extract rich, hierarchical visual features, enabling the capture of finer details (e.g., attributes, relations) and the mention of multiple objects in complex scenes.

2. **Dynamic attention:** By attending to different image regions at each decoding step, BiCap avoids the information bottleneck of global feature vectors.

3. **Better sequence modeling:** The LSTM decoder, conditioned on context vectors, produces more coherent sequences compared to the baseline RNN.

5.4 Limitations and Future Works

Despite the aforementioned advancements, the current BiCap implementation faces several limitations that highlight clear paths for future research to further enhance its accuracy, diversity, and robustness:

1. As the encoder relies on ResNet-50, BiCap primarily captures local rather than fully global visual dependencies. Future extensions will explore hybrid CNN–ViT architectures for improved scene understanding.

Models	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	RUOUGE	CIDEr
BiCap (Proposed)	**71.42**	**59.31**	**46.40**	**32.73**	**0.37**	**0.42**	**0.29**
Masud et al., 2025	62.74	51.63	39.81	22.28	0.34	0.36	0.23
Hossain et al., 2024	59.37	47.26	31.74	20.39	0.29	0.31	0.21
Das et al., 2023	53.46	42.83	27.63	20.79	0.27	0.34	0.21
Humaira et al., 2021	51.37	41.94	26.07	19.45	0.24	0.33	0.20

Table 2: Quantitative evaluation using BLEU, METEOR, ROUGE and CIDEr scores on the Chitron dataset.

Models	Fluency	Adequacy	Relevance
BiCap (Proposed)	**4.3**	**4.4**	**4.6**
Masud et al., 2025	3.6	3.3	3.9
Hossain et al., 2024	3.7	2.8	2.9
Das et al., 2023	3.7	3.1	3.4
Humaira et al., 2021	3.4	2.6	2.4

Table 3: Human evaluation of generated captions (Likert scale 1-5).

2. While effective in low-resource settings, BiCap's RNN-based decoder is inherently sequential and may not fully capture long-range dependencies. Future work will explore hybrid Transformer–RNN extensions.

3. Although BiCap performs well on the 15K-image Chitron dataset, its generalizability to unseen domains remains unverified due to limited Bangla multimodal data. Future work will focus on additional Bangla or multilingual datasets, cross-domain evaluation and multilingual transfer learning to enhance robustness.

6 Conclusion

This study introduced BiCap, an attention-based encoder-decoder framework for Bangla image captioning. By combining ResNet-50 for visual feature extraction, Bahdanau attention for dynamic region selection, and an LSTM decoder for sequential text generation, BiCap addresses key challenges in Bangla captioning such as semantic grounding and syntactic fluency.

Experimental results on the Chitron dataset demonstrated that BiCap significantly outperforms the existing baselines, achieving higher BLEU, METEOR, ROUGE and CIDEr scores. Superior human evaluation ratings further highlighted its ability to generate more descriptive and contextually accurate Bangla captions.

Overall, BiCap establishes a strong foundation for research in low-resource multimodal learning, particularly for Bangla. Future extensions could integrate transformer-based decoders, leverage multilingual embeddings for rare word coverage, and explore larger-scale multimodal datasets to further enhance caption quality. This work represents a step forward in bridging the gap between computer vision and natural language processing for underrepresented languages.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, and 1 others. 2024. On scaling up a multilingual vision and language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14432–14444.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, and 1 others. Pali: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations*.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. *Advances in neural information processing systems*, 28.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N

Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Bidyut Das, Ratnabali Pal, Mukta Majumder, Santanu Phadikar, and Arif Ahmed Sekh. 2023. A visual attention-based model for bengali image captioning. *SN Computer Science*, 4(2):208.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Md Anwar Hossain, Mirza AFM Rashidul Hasan, Sajeeb Kumar Ray, and Naima Islam. 2024. Generating bangla image captions with deep learning techniques. In *2024 6th International Conference on Sustainable Technologies for Industry 5.0 (STI)*, pages 1–6. IEEE.

Mayeesha Humaira, Paul Shimul, Md Abidur Rahman Khan Jim, Amit Saha Ami, and Faisal Muhammad Shah. 2021. A hybridized deep learning method for bengali image captioning. *International Journal of Advanced Computer Science and Applications*, 12(2).

Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2891–2903.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, and 1 others. 2024. Llava-plus: Learning to use tools for creating multimodal agents. In *European conference on computer vision*, pages 126–142. Springer.

Tongwei Lu, Jiarong Wang, and Fen Min. 2023. Full-memory transformer for image captioning. *Symmetry*, 15(1):190.

Adiba Masud, Md Biplob Hosen, Md Habibullah, Mehrin Anannya, and M Shamim Kaiser. 2025. Image captioning in bengali language using visual attention. *PloS one*, 20(2):e0309364.

Faisal Muhammad Shah, Mayeesha Humaira, Md Abidur Rahman Khan Jim, Amit Saha Ami, and Shimul Paul. 2022. Bornon: Bengali image captioning with transformer-based deep learning approach. *SN Computer Science*, 3(1):90.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamila Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

A. Rahman and 1 others. 2019. Bnlit: Bangla natural language image-to-text dataset and baseline. In *Proceedings of the International Conference on Bangla Speech and Language Processing (ICBSLP)*.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024.

Sumon Sazzed. 2020. Chittron: A bangla image captioning dataset. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 751–758.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International conference on machine learning*, pages 23318–23340. PMLR.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*.