Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048104

ARTICLE

Enhancing Cross-Lingual Image Description: A Multimodal Approach for Semantic Relevance and Stylistic Alignment

Emran Al-Buraihy and Dan Wang*

Faculty of Information Technology, Beijing University of Technology, Beijing, 100124, China

*Corresponding Author: Dan Wang. Email: wangdan@bjut.edu.cn

Received: 28 November 2023 Accepted: 28 February 2024 Published: 20 June 2024

ABSTRACT

Cross-lingual image description, the task of generating image captions in a target language from images and descriptions in a source language, is addressed in this study through a novel approach that combines neural network models and semantic matching techniques. Experiments conducted on the Flickr8k and AraImg2k benchmark datasets, featuring images and descriptions in English and Arabic, showcase remarkable performance improvements over state-of-the-art methods. Our model, equipped with the Image & Cross-Language Semantic Matching module and the Target Language Domain Evaluation module, significantly enhances the semantic relevance of generated image descriptions. For English-to-Arabic and Arabic-to-English cross-language image descriptions, our approach achieves a CIDEr score for English and Arabic of 87.9% and 81.7%, respectively, emphasizing the substantial contributions of our methodology. Comparative analyses with previous works further affirm the superior performance of our approach, and visual results underscore that our model generates image captions that are both semantically accurate and stylistically consistent with the target language. In summary, this study advances the field of cross-lingual image description, offering an effective solution for generating image captions across languages, with the potential to impact multilingual communication and accessibility. Future research directions include expanding to more languages and incorporating diverse visual and textual data sources.

KEYWORDS

Cross-language image description; multimodal deep learning; semantic matching; reward mechanisms

1 Introduction

In the digital age, we are amidst an unprecedented era of visual information exchange, fueled by the proliferation of multimedia content on the internet [1]. Among the vast array of media at our disposal, images stand out as a universal language that transcends linguistic barriers, serving as a vital medium for communication and information dissemination [2]. In today's digital landscape, images have become the lingua franca, effortlessly conveying ideas, experiences, and emotions across the global online community [3].

The field of image captioning, which involves automatically generating descriptive captions for images, has emerged as a critical research area with a wide range of applications. It extends its reach from assisting the visually impaired to enriching content retrieval and enhancing user engagement

on multimedia platforms [4]. Remarkable progress in this field has been driven by cutting-edge technologies such as object detection, relationship reasoning, and language sequence generation [5].

Yet, the mosaic of languages spoken worldwide presents a formidable challenge for image captioning systems [6]. The task of generating precise, coherent, and culturally relevant image descriptions in multiple languages necessitates a nuanced understanding of both the visual content and the linguistic subtleties inherent to each target language [7]. Conventional image captioning models often falter in capturing these intricacies, leading to translations that lack fluency, coherence, and context, ultimately failing to resonate with speakers of the target language [8].

These multilingual challenges underscore the pressing need for cross-lingual image captioning solutions that bridge linguistic and cultural divides [9]. This need has grown in significance as individuals and communities with diverse linguistic backgrounds increasingly seek access to and comprehension of content from different cultural spheres and regions [10]. Cross-lingual image description tasks, such as the transfer of descriptions from English to Arabic, have become pivotal in this evolving research landscape [11].

The challenge involves creating descriptive image captions in a language different from the original image label, posing a complex issue when dealing with substantial linguistic and cultural differences [12]. Conventional image captioning methods, relying on single-language models, fall short in delivering accurate and culturally resonant descriptions across multiple languages [13].

The primary research problem we tackle in this study revolves around enabling accurate and culturally apt cross-lingual image captioning between Arabic and English. Arabic, a language steeped in rich cultural and linguistic heritage, poses unique challenges due to its complex script and diverse dialects [14]. In contrast, English stands as a widely spoken global language [15]. The challenge lies not only in precisely translating captions between these languages but also in ensuring that the resulting descriptions are semantically coherent, culturally pertinent, and contextually accurate [16].

As shown in Fig. 1, there are language style differences between the Arabic and English descriptions. The source English description of the image is “A person in a blue jacket follows two donkeys along a mountain trail” (a short descriptive sentence), while the target domain Arabic description follows a more descriptive sentence with a translation style of "رجل يرتدي جاكت أزرق وبنطال جينز، وله حقيبة ظهر على ظهره." (A man wearing a blue jacket and jeans, with a backpack on his back). Furthermore, the emphasis on semantics is also not the same. Although both sentences mention “ a blue jacket,” the real Arabic description centers around “the man,” while the English description centers around “a photo.”

The field of cross-lingual image captioning faces notable limitations, especially in dataset diversity. Many existing studies utilize English datasets with translations that often lack cross-cultural relevance [17]. Additionally, the reliance on machine translation in prior models raises issues of accuracy and cultural sensitivity [18]. Addressing these gaps, our work introduces the AraImg2k Dataset, a comprehensive collection of 2000 images embodying Arab culture, each paired with five carefully crafted captions in Modern Standard Arabic (MSA). This curated dataset aims to authentically represent the rich diversity and nuances of Arab culture.

Previous methods in cross-lingual image captioning struggled with accurately capturing the semantic relationship between images and their captions [19]. To address this, our study introduces a multimodal semantic matching module. This module improves the accuracy of semantic consistency between images and captions across different languages, utilizing multimodal visual-semantic

embeddings. This ensures that the generated captions more accurately reflect the original images, enhancing the quality of cross-lingual image descriptions.

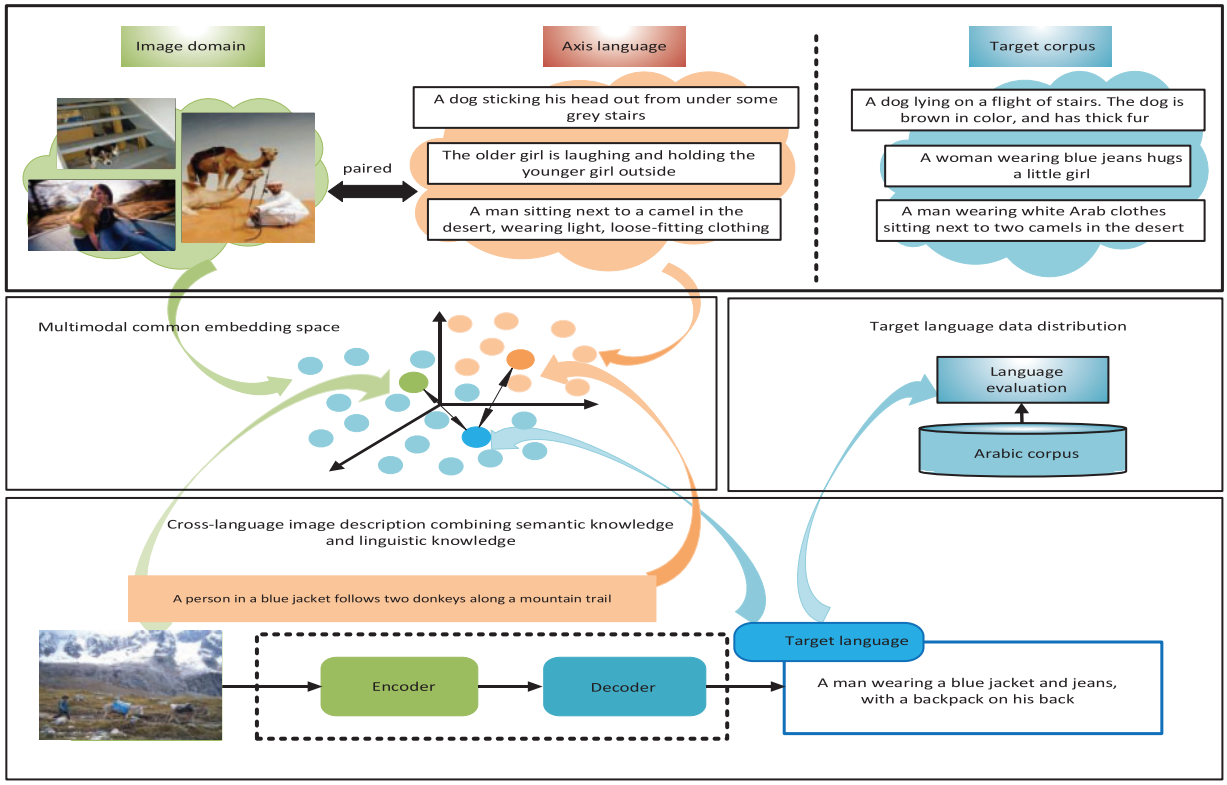

Figure 1: The task of cross-lingual image captioning and our solution

Previous methods in cross-lingual captioning often overlooked linguistic subtleties and cultural context [20]. Our research counters this by introducing a language evaluation module. This module adapts to the target language's distribution and style, enabling the creation of captions that are more aligned with linguistic nuances and cultural norms, thereby producing more natural and culturally attuned image descriptions.

Earlier studies in cross-lingual image captioning often lacked comprehensive evaluation metrics, hindering performance assessment. Our research addresses this by employing a range of evaluation metrics, including BLEU [21], ROUGE [22], METEOR [23], CIDEr [24], and SPICE [25]. This allows for a rigorous comparison with previous works and a more detailed evaluation of our approach's effectiveness and superiority in the field.

In light of these advancements and contributions, our research seeks to bridge the gap between languages, cultures, and communities by enhancing the quality and cultural relevance of cross-lingual image descriptions. Through meticulous dataset creation, improved translation techniques, advanced semantic matching, and comprehensive evaluation, we aim to significantly advance the field of cross-lingual image captioning, ultimately fostering more effective cross-cultural understanding and communication.

Therefore, this study presents a comprehensive approach to cross-lingual image captioning, leveraging semantic matching and language evaluation techniques to address the aforementioned challenges. The key contributions of this study can be summarized as follows:

- Cross-Lingual Image Captioning (Arabic and English): We address the challenge of accurate and culturally relevant cross-lingual captioning between Arabic and English. Our method ensures precise translations, semantic coherence, and cultural relevance, bridging the linguistic and cultural divide.
- AraImg2k Dataset: To overcome the limitations in existing datasets, we introduce AraImg2k, a dataset of 2000 images representing Arab culture, each with five detailed captions in Modern Standard Arabic, reflecting the cultural diversity of the Arab world.
- Multimodal Semantic Matching Module: Our novel module captures the semantic relationship between images and captions in cross-lingual contexts using multimodal visual-semantic embeddings, ensuring captions accurately reflect the image content.
- Language Evaluation Module: This module focuses on understanding the target language's nuances and cultural context, aiding in producing captions that are linguistically and culturally aligned, enhancing the naturalness and relevance of our cross-lingual descriptions.
- Comprehensive Evaluation Metrics: Setting our research apart from previous studies, we employ diverse evaluation metrics like BLEU, ROUGE, METEOR, CIDEr, and SPICE, allowing for a detailed comparison with prior works and demonstrating the effectiveness of our approach.

Section 2 is dedicated to an extensive examination of prior research within the same domain. Section 3 provides a thorough explication of the essential components of the proposed framework is presented. Section 4 presents empirical findings along with a comparative analysis vis-à-vis the preceding study. Finally, in Section 5, a conclusive summary is offered, accompanied by suggestions for prospective research endeavors.

2 Literature Review

This study delves into the dynamic and diverse landscape of image captioning and the emerging field of cross-lingual image caption generation. Image captioning, situated at the intersection of computer vision and natural language processing, has witnessed significant advancements in recent years, bridging the gap between visual content and human language [26]. In this section, we explore the foundational concepts, methodologies, and key research papers in both image captioning and the evolving domain of cross-lingual image captioning. By examining pioneering work in Arabic and English, we pave the way for our own cross-lingual image captioning approach. This literature review serves as a guiding beacon, illuminating the path of prior research and informing the innovative contributions in subsequent sections.

2.1 Image Captioning

Image captioning, a multidisciplinary research area at the confluence of computer vision and natural language processing, centers around the task of automatically generating descriptive text for images [27]. Its significance extends beyond enhancing human-computer interaction and content retrieval; it also plays a pivotal role in enabling visually impaired individuals to comprehend visual content [4]. The evolution of image captioning has been propelled by remarkable progress driven by deep learning techniques and the availability of extensive image-text datasets [28]. In this subsection,

we delve into the fundamentals of image captioning and provide an overview of key research papers in both Arabic and English domains.

2.1.1 Arabic Image Captioning

In the domain of Arabic image captioning, researchers in [29] proposed an innovative approach tailored to generating image captions specifically for clothing images. Leveraging deep learning techniques, their model proficiently generates Arabic captions describing clothing items, enhancing accessibility to fashion-related visual content. Meanwhile, researchers in [30] investigated Arabic image captioning, focusing on the impact of text pre-processing on attention weights and BLEU-N scores. Their work sheds light on optimizing the caption generation process in Arabic, taking into account the nuances of text pre-processing. Authors in [31] ventured into automatic Arabic image captioning using a combination of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) language models, alongside Convolutional Neural Networks (CNNs). Their model demonstrated the feasibility of generating Arabic image captions, marking a significant milestone in the field of Arabic image description.

2.1.2 English Image Captioning

In the realm of English image captioning, researchers in [32] proposed a novel approach for automatic caption generation for news images. They employ a multimodal approach that integrates both image content and associated news articles to create coherent and informative image captions. Researchers in [33] introduced a compact image captioning model with an attention mechanism, focusing on the efficiency of caption generation. Their research contributes to streamlining image captioning models for practical applications. In addition, researchers in [34] provided valuable insights from lessons learned during the 2015 MSCOCO Image Captioning Challenge, highlighting key takeaways and challenges in image captioning.

2.2 Cross-Lingual Image Captioning

Cross-lingual image caption generation is an emerging research area that addresses the challenge of automatically generating descriptive captions for images in different languages [35]. It plays a pivotal role in facilitating multilingual communication and cross-cultural understanding, enabling individuals from diverse linguistic backgrounds to access and comprehend visual content [36]. In this subsection, we delve into cross-lingual image caption generation and provide summaries of key research papers in Arabic-English, Chinese-English, German-English, and Japanese-English domains.

2.2.1 Arabic-English Cross-Lingual Image Captioning

Researchers in [37] introduced a novel approach, "Wikily" Supervised Neural Translation, tailored to Arabic-English cross-lingual tasks, including image captioning. Their model leverages Wikipedia as a resource for cross-lingual supervision, showcasing its efficacy in generating accurate image captions across the Arabic-English language barrier. In addition, researchers in [38] contributed to Arabic-English cross-lingual image captioning by providing valuable resources and end-to-end neural network models, enriching the accessibility and understanding of visual content for Arabic speakers.

2.2.2 Chinese-English Cross-Lingual Image Captioning

Researchers in [39] introduced COCO-CN, a resource for cross-lingual image tagging, captioning, and retrieval tasks involving Chinese and English. Their work underscores the significance of

bridging the linguistic gap between these two languages in the context of visual content. Additionally, researchers in [40] explored fluency-guided cross-lingual image captioning, particularly focusing on Chinese-English pairs. Their approach highlights the importance of fluency in generating high-quality image captions that resonate with speakers of both languages.

2.2.3 German-English Cross-Lingual Image Captioning

Researchers in [41] presented multimodal pivots for image caption translation, addressing the German-English cross-lingual challenge. Their work explores strategies for effectively translating image captions between these languages using multimodal approaches. Furthermore, researchers in [42] contributed to the field with the creation of Multi30k, a multilingual English-German image description dataset. Their work serves as a valuable resource for cross-lingual image captioning research, fostering improved understanding and communication between German and English speakers.

2.2.4 Japanese-English Cross-Lingual Image Captioning

Researchers in [43] presented the STAIR captions dataset, a substantial resource for Japanese image captioning. Their work advances the availability of image description data for Japanese speakers, contributing to the field's progress in Japanese-English cross-lingual image captioning. Moreover, researchers in [44] delved into cross-lingual image caption generation with a focus on Japanese-English pairs. Their work explores techniques to generate image captions that transcend language barriers, enhancing cross-cultural communication.

In generating Table 1, we employed a meticulous and systematic literature review process to ensure the precision and comprehensiveness of the information presented. This process entailed the following steps:

a. Keyword-Based Search: We initiated our literature review with a keyword-based search in major academic databases. The keywords were carefully chosen to encompass the core themes of our study, namely 'cross-lingual image captioning', 'multimodal learning', and 'semantic matching'.
b. Selection Criteria: Upon retrieving a preliminary set of papers, we applied specific selection criteria to filter out the most relevant studies. These criteria included the recency of publication, relevance to our study's focus on cross-lingual and multimodal aspects, and the academic credibility of the sources.
c. Data Extraction and Synthesis: For each selected paper, we extracted key information such as methodologies used, datasets employed, and evaluation metrics applied. This data was then critically analyzed and synthesized to present a comprehensive view of the current research landscape.
d. Tabulation and Cross-Verification: The synthesized data was tabulated in Table 1, ensuring that each entry accurately reflected the corresponding study's contributions and findings. We cross-verified each entry for accuracy and completeness.
e. Continuous Updating: Recognizing the dynamic nature of the field, we maintained an ongoing process of updating the table to include the latest significant contributions up until the finalization of our manuscript.

Through this rigorous process, Table 1 was crafted to provide a detailed and accurate summary of existing literature in the field of cross-lingual image captioning, serving as a foundational reference for our study and future research in this domain.

Table 1: Summary of the literature

Research area	Ref.	Dataset	Data source	Language (s)	App/Tech	Evaluation metrics
Image captioning	[29]	Arabic Fashion Data	DeepFashion dataset [45] InFashAIv1 [46]	Arabic	Image captioning, attention mechanism	BLEU
	[30]	–	Arabic-Flickr8 [38]	Arabic	Attention mechanism, beam search	BLEU, THUMB [47]
	[31]	Arabic corpus	MS-COCO dataset [48], Flickr8k [49]	Arabic	Crowd-Flower crowdsourcing, commercial cloud server FloydHub	BLEU
	[32]	–	News images	English	Unsupervised fashion, news media and journalism	BLEU, ROUGE, METEOR, CIDEr, SPICE
	[33]	–	MSCOCO, InstaPIC-1.1M [50]	English	Attention mechanism, streamlining image captioning	BLEU, ROUGE, METEOR, CIDEr, SPICE
	[34]	–	MS-COCO	English	Qualitatively and quantitatively, probability of the correct description	BLEU, ROUGE, METEOR
Cross-lingual Image captioning	[37]	–	Wikipedia	Arabic-English	Supervised neural, image captioning	BLEU
	[38]	Arabic-Flickr8	Flickr8k	Arabic-English	End-to-end	BLEU
	[39]	COCO-CN	MS-COCO	Chinese-English	Image captioning, tagging, retrieval, recommendation-assisted annotation system	Precision, Recall, F-measure, BLEU, METEOR, ROUGE-L, CIDEr
	[40]	Flickr30k-CN	Flickr30k [51]	Chinese-English	Image captioning, fluency-guided learning framework	BLEU
	[41]	Bilingual caption	MS-COCO	German-English	Image retrieval	BLEU, METEOR, TER

(Continued)

Table 1 (continued)

Research area	Ref.	Dataset	Data source	Language (s)	App/Tech	Evaluation metrics
	[42]	Multi30k	Flickr30k	German-English	Crowdsourced platform	–
	[43]	STAIR captions	MS-COCO	Japanese-English	A web system for caption annotation, quantitatively and qualitatively	BLEU, ROUGE, CIDEr
	[44]	YJ captions	MS-COCO	Japanese-English	Translation models, multilingual adaptation	BLEU, ROUGE, METEOR, CIDEr, Cross-lingual metrics

3 Methodology

As shown in Fig. 2, the cross-lingual image description model proposed in this study consists of three components.

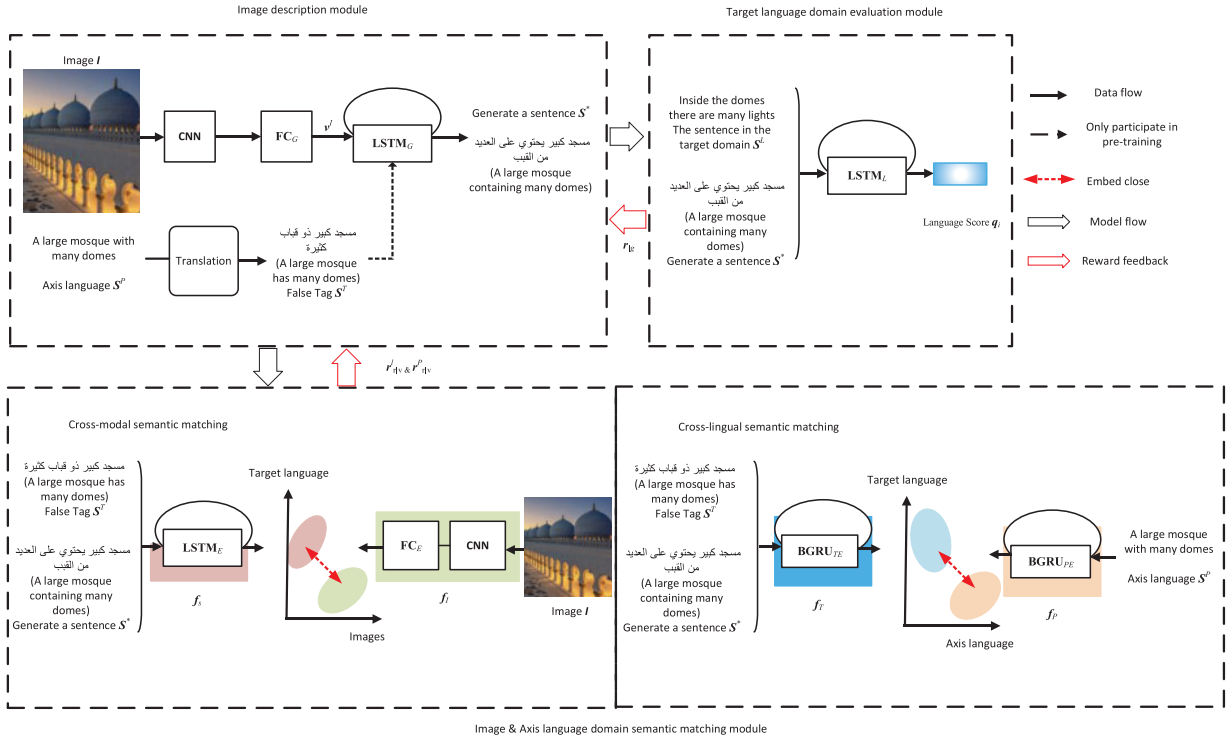

Figure 2: Cross-lingual image captioning model

Naive Image Encoder-Sentence Decoder (Image Description Generation Module): This module is responsible for generating descriptive sentences. It encodes the image and decodes it into sentences.

Image & Source Language Domain Semantic Matching Module: This module is responsible for providing semantic matching rewards and optimization. It takes into account the semantic information from the source domain image and axis language and maps them into a common embedding space for semantic matching calculations.

Target Language Domain Evaluation Module: This module is designed to provide language evaluation rewards. It incorporates knowledge about the data distribution in the target language domain for language evaluation constraints.

The first module is responsible for sentence generation, while the latter two modules guide the model to learn semantic matching constraints and language knowledge optimization. This helps the model generate more fluent and semantically rich descriptions.

3.1 Data Collection and Preparation

The first step in our methodology was to collect 2000 images that represent Arab culture from authentic websites. We selected images from a variety of sources, including museums, cultural institutions, and travel websites, to ensure that we have a diverse and representative set of images. We then manually generate five captions for each image in Modern Standard Arabic (MSA), ensuring of having a variety of descriptions that capture different aspects of the image. This dataset serves as a valuable resource for cross-lingual image captioning research and richly reflects the diversity of Arab culture.

3.2 Image Encoder-Sentence Decoder Module

A naive image encoder-sentence decoder framework is used to generate descriptive sentences. It employs a pre-trained neural network model, ResNet-101 [52], and a fully-connected layer (referred to as (FC_G), to extract features (v^I) from the image (I). A single-layer Long Short-Term Memory (LSTM) network, denoted as $LSTM_G$, is used to decode (v^I) and generate the current time-step word. The source domain description language (S^p) of the image I is translated into a target domain pseudo-sentence label (S^T) using Google Translation to initialize this module. During model initialization training, the pre-trained ResNet-101 model is not involved in model optimization, while the fully-connected layer (FC_G) and ($LSTM_G$) participate in model optimization. The optimization objective is set to minimize the negative log probability of correct words in the sentence.

In addition to the methodologies described previously, it is pertinent to elaborate on the translation module employed in our study, particularly for the initial translation between English and Arabic languages. We utilized Google Translate for this purpose, leveraging its capabilities to generate pseudo-sentence labels in the target domain from the source language descriptions. This step was crucial for initializing the model with a basic understanding of cross-lingual semantic structures. It is important to note that these machine-generated translations were primarily used as a starting point. The subsequent modules, namely the Image & Source Language Domain Semantic Matching Module and the Target Language Domain Evaluation Module, were designed to refine these translations, ensuring their semantic accuracy and cultural relevance.

$$L(\theta_G) = -\sum_{i=1}^{N} \log\left(p_{\theta_G}\left(w_i^{(T)} \mid v^I, w_{0:i-1}^{(T)}\right)\right) \tag{1}$$

Eq. (1) is derived based on standard practices in neural network training for language modeling, particularly in image captioning tasks. It calculates the negative log probability of the correct word sequence in a generated caption, given an image and the preceding words. This approach is consistent with methodologies adopted in neural network-based natural language processing, as detailed in foundational works such as [34].

In the equation, (N) represents the length of the sentence $S^T = \left\{w_0^{(T)}, w_1^{(T)}, \cdots, w_N^{(T)}\right\}$. The word $(w_0^{(T)})$ is set as the start symbol $(\langle bos \rangle)$, (θ_G) represents the learning parameters of this module, including (FC_G), $(LSTM_G)$.

In selecting LSTM over Transformer-based models for our cross-lingual image captioning research, we considered several pivotal factors unique to our study's context. LSTM networks, recognized for their efficiency in sequential data processing and less demanding computational requirements, aligned well with our resource constraints and the exploratory nature of our work. This was particularly pertinent given the complexity and specific linguistic characteristics of our primary dataset, AraImg2k, which includes the nuanced morphological features of Arabic. LSTMs' proven track record in language modeling provided a solid and interpretable foundation for initial experiments. While we acknowledge the advanced capabilities of Transformers in handling long-range dependencies and their parallel processing strengths, our initial focus was to establish a robust baseline model that effectively balances computational efficiency with the linguistic intricacies of our dataset. Moving forward, we plan to explore the integration of Transformer models to further advance our approach, leveraging their benefits in subsequent phases of our research.

3.3 Image & Source Language Domain Semantic Matching Module

After the initialization described in Section 3.2, the descriptions generated by the model exhibit certain characteristics, such as simple imitation of pseudo-labels, repetitive combinations of high-frequency vocabulary, or a lack of relevance to the content of the image. Manually annotated source language descriptions typically possess rich semantics and provide concrete descriptions of the image content. The source language and the image should contain consistent semantic information.

To address this issue and enhance the semantic relevance of the generated descriptions, the study introduces a multi-modal semantic matching module. This module leverages both the semantic information from the image and the source language to impose constraints on semantic similarity. The aim is to ensure that the generated descriptions are semantically aligned with both the image and the source language, resulting in more meaningful and contextually relevant descriptions.

3.3.1 Cross-Modal Semantic Matching

For heterogeneous images and sentences, the first step is to map the images and sentences into a common embedding space and measure semantic relatedness. As shown in Fig. 2, the image's semantic embedding network, denoted as, (f_I) consists of a CNN encoder (using the pre-trained ResNet-101 model) and a fully connected layer (referred to as FC_E). The text's semantic embedding network, denoted as (f_S), is composed of a single-layer LSTM (denoted as *** $LSTM_E$). The final hidden vector of $(LSTM_E)$ at the last time step defines the semantic vector in the common embedding space for the input sentence.

By inputting image-sentence pairs $((I), S^T)$, you obtain the image's feature embedding, $(f_I(I))$, in the common semantic space and the sentence's embedding feature, $\left(f_S\left(\mathrm{S}^T\right)\right)$, in the common semantic space. For matching pairs $((I), \mathrm{S}^T)$, negative examples are found within the same batch. Specifically, sentences $(\mathrm{S}^{T'})$ that do not match with (I) and images (I') that do not match with (S^T) within the same

batch are identified. The pretraining process involves minimizing a bidirectional ranking loss in the common semantic space.

The goal of this process is to align the semantics of images and sentences in a shared embedding space and measure their semantic relatedness.

$$L(\theta_\mu) = \sum_{I}\sum_{S^{T'}} \max\left(0, \Delta - f_l(I) f_S\left(\mathrm{S}^{T}\right) + f_l(\mathrm{I}) f_S\left(\mathrm{S}^{T'}\right)\right) + \sum_{I'}\sum_{S^{T}} \max(0, \Delta - f_l(\mathrm{I}) f_s\left(\mathrm{S}^{T}\right) + f_l(\mathrm{I}') f_S\left(\mathrm{S}^{T}\right)) \tag{2}$$

Eq. (2) describes a bidirectional ranking loss, a common approach in cross-modal semantic matching. It is designed to fine-tune the semantic alignment between images and their corresponding textual descriptions, following a methodology widely used in multimodal learning tasks. For further theoretical background and application of similar loss functions, readers are referred to [53].

In the equation, (Δ) represents a boundary hyperparameter, and (θ_μ) represents the learning parameters for the (FC_E) and ($LSTM_E$) layers in this module.

3.3.2 Cross-Lingual Semantic Matching

In addition, this study also has axis language sentence-pseudo-label sentence pairs (S^p, S^T), which can provide data support for measuring the semantic similarity between target language sentences and axis language sentences. In this section, cross-lingual semantic matching is introduced to enhance the semantic relevance of sentences, using a similar semantic embedding network mechanism as in Section 3.3.1 to align the embedding vectors of the target language and axis language. Both the encoder for the target language and the encoder for the axis language use a single-layer BG-RU (Bidirectional Gated Recurrent Unit) structure, with the hidden vector at the end of BGRU used as the sentence feature vector. f_p is the axis language feature mapper ($BGRU_{PE}$), and f_T is the target language feature mapper ($BGRU_{TE}$). Similarly, pretraining is performed by minimizing bidirectional ranking loss to align the common semantic space.

$$L(\theta_\rho) = \sum_{S^p}\sum_{S^{T'}} \max\left(0, \Delta - f_P\left(S^P\right) f_T\left(S^T\right) + f_P\left(S^P\right) f_T\left(S^{T'}\right)\right) + \sum_{S^{p'}}\sum_{S^T} \max(0, \Delta - f_P\left(S^P\right) f_T\left(S^T\right) + f_P\left(S^{p'}\right) f_T\left(S^T\right)) \tag{3}$$

In the equation, for matching pairs (S^P, S^T), $S^{T'}$ is the negative example from the pseudo-label sentence set in the same batch that does not match (S^P), and ($S^{P'}$) is the negative example from the axis language sentence set in the same batch that does not match (S^T). (θ_P) represents the learning parameters for the ($BGRU_{PE}$) and ($BGRU_{TE}$) layers in this module.

3.4 Target Language Domain Evaluation Module

Due to the currently generated descriptions having little association with the target corpus, the generated description sentences often exhibit significant differences in language style from the real target sentences. To optimize the quality of the description language, this section pertains to a module

on the target language dataset that can provide language evaluation rewards, focusing on correctly classifying the input words. This module employs LSTM (referred to as $LSTM_L$) and inputs words sequentially into ($LSTM_L$), then utilizes ($LSTM_L$) to predict the probability of the current input word. Using sentences of length (N) from the target corpus as input, represented as $S^L = \left\{w_0^{(L)}, w_1^{(L)}, \cdots, w_N^{(L)}\right\}$, the pretraining objective is to minimize the negative log probability of correct words in the sentence, as shown in the equation.

$$L(\theta_\omega) = -\sum_{i=1}^{N} \log\left(p_{\theta_\omega}\left(\mathrm{w}_i^{(L)} \mid \mathrm{w}_{0:i-1}^{(L)}\right)\right) \tag{4}$$

Here, (θ_ω) represents the learning parameters for the ($LSTM_L$) module in this section.

3.5 Model Optimization Based on Semantic Matching and Language Rewards

After the self-supervised pretraining of the three modules mentioned above, the optimization learning of the Image Encoder-Sentence Decoder module in Section 3.2 is jointly implemented with the three modules. Specifically, semantic matching rewards from Section 3.3 and language evaluation rewards from Section 3.4 are utilized to optimize the module in Section 3.2.

Image-Sentence Matching Reward: The image I is mapped through the visual semantic embedding network (f_I), and the sentence (S^*) is mapped through the text semantic embedding network (f_S) to the common embedding space. The cross-modal semantic matching reward can be defined as:

$$r_{\mathrm{r|v}}^{\mathrm{I}}(S^*) = \frac{f_I(I) f_S(S^*)}{||f_I(I)||\ ||f_S(S^*)||} \tag{5}$$

Cross-Language Sentence Matching Reward: Similarly, the source domain sentence (S^p) is mapped through the axis language feature mapper (f_P), and the sentence (S^*) is mapped through the target language feature mapper (f_T). The cross-language semantic matching reward can be defined as:

$$r_{\mathrm{r|v}}^{\mathrm{P}}(S^*) = \frac{f_P(S^p) f_T(S^*)}{||f_P(S^P)||\ ||f_T(S^*)||} \tag{6}$$

Eq. (6) defines the cross-language semantic matching reward using cosine similarity, a standard measure in natural language processing for assessing semantic closeness between high-dimensional vectors. This approach aligns with established practices in cross-lingual semantic analysis, where maintaining semantic integrity across languages is crucial. For a foundational reference on the application of cosine similarity in cross-lingual contexts, readers can consult the work of [54].

Target Domain Sentence Language Evaluation Reward: Each word of the sentence (S^*) is iteratively input into the ($LSTM_L$) module trained on the target language domain in Section 3.4. The language evaluation process is as follows:

$$\left[\mathrm{q}_i, \mathrm{h}_i^L\right] = f_{\mathrm{LSTM}_L}\left(\mathrm{w}_i^{(*)}, \mathrm{h}_{i-1}^L; \theta_\omega\right), i \in \{1, \cdots, N\} \tag{7}$$

Here, $S^* = \left\{w_0^{(*)}, w_1^{(*)}, \cdots, w_N^{(*)}\right\}$, where ($w_0^{(*)}$) is the starting symbol "bos", (N) is the length of the sentence (S^*), (h_i^L) is the hidden vector at the time step (i), (q_i) is the probability vector over the vocabulary with dimensions equal to the vocabulary size, and ($q_i\ w_i^{(*)}$) represents the predicted

probability of the word ($w_i^{(*)}$) at time step (i).

$$r_{lg}(S^*) = \frac{1}{N}\sum_{i=0}^{N}\log\left(q_i\left(\mathrm{w}_i^{(*)} \mid \mathrm{w}_{0:i-1}^{(*)}\right)\right) \tag{8}$$

The total reward for the entire cross-lingual description model is defined as:

$$r_{\text{total}} = \alpha r_{lg} + \beta r'_{\mathrm{r|v}} + \gamma r^{p}_{\mathrm{r|v}} \tag{9}$$

In the equation, (α), (β), and (γ) are hyperparameters with values in the range [0,1]. (α), (β), and (γ) are empirical parameters, and the optimal values are determined in Section 4.2.

To reduce the expected gradient variance during model training, we follow a self-critical sequence training approach. The current model uses a multinomial distribution sampling method to obtain sentences (S^*), and additionally, defaults to using a maximum probability greedy sampling method to obtain sentences (S), with (r_{total}(S)) as the baseline reward. The overall reward for a sentence (S^*) can be expressed as (r_{total} (S^*)) – (r_{total} (S)), where sentences with higher rewards than the baseline are encouraged, and sentences with lower rewards than the baseline are discouraged. Through iterative reinforcement training, the model generates sentences with better semantic matching rewards and language evaluation rewards. Therefore, the final objective loss of the cross-lingual description model can be defined as:

$$L_{\text{total}} = -\sum_{i=1}^{N}\left(\left(r_{\text{total}}(S^*) - r_{\text{total}}(S)\right)\times \log P_{\theta_G}\left(\mathrm{w}_i^{(*)} \mid \mathrm{v}', \mathrm{w}_{0:i-1}^{(*)}\right)\right) \tag{10}$$

(θ_G) represents the parameters of the image description module.

Algorithm 1: *Cross-Lingual Image Captioning*

Input:
Collection of 2000 images representing Arab culture (ArabicImages)
Target language (TargetLanguage)
Output:
Optimized image descriptions in the target language

Begin
Step 1: *Data Collection and Preparation*
ArabicImages = CollectArabicImages()
ArabicCaptions = GenerateArabicCaptions(ArabicImages)
Dataset = CreateCrossLingualDataset(ArabicImages, ArabicCaptoins, TargetLanguage)
Step 2: *Image Encoder-Sentence Decoder Module*
Initialize ImageEncoder
Initialize SentenceDecoder for TargetLanguage
for each image in Dataset **do**
{
ImageFeatures = EncodeImage(Image, ImageEncoder)
GeneratedSentence = DecodeSentence(ImageFeatures, SentenceDecoder)
Store GeneratedSentence
}

(Continued)

Algorithm 1 (continued)

```
        end for
    Step 3: Image & Source Language Domain Semantic Matching Module
        for each image-sentence pair (Image, GeneratedSentence) in Dataset do
            {
                ImageEmbedding = MapToCommonSemanticSpace(Image)
                SentenceEmbedding = MapToCommonSemanticSpace(GeneratedSentence)
                SimilarityScore = CalculateSemanticSimilarity(ImageEmbedding, SentenceEmbedding)
                Store SimilarityScore
            }
        end for
    Step 4: Target Language Domain Evaluation Module
        Initialize LanguageEvaluator for TargetLanguage
        for each sentence in TargetLanguage do
            TrainLanguageEvaluator(LanguageEvaluator, Sentence)
        end for
        for each GeneratedSentence do
            {
                LanguageQuality = EvaluateLanguage(LanguageEvaluator, GeneratedSentence)
                Store LanguageQuality
            }
        end for
    Step 5: Model Optimization Based on Semantic Matching and Language Rewards
        for each image-sentence pair (Image, SourceLanguageSentence, GeneratedSentence) in Dataset do
            {
                SemanticMatchingReward = CrossModalSemanticMatching(Image, GeneratedSentence)
                LanguageEvaluationReward = LanguageEvaluation(GeneratedSentence)
                OptimizeImageDescriptionModel(Image, SourceLanguageSentence, GeneratedSentence,
            SemanticMatchingReward, LanguageEvaluationReward)
                Store OptimizedDescription
            }
        end for
    Step 6: Output Optimized Image Descriptions
        Return OptimizedDescriptions
    End
```

4 Results Analysis

To validate the effectiveness of the model in cross-lingual image description tasks, this study conducted two sub-task experiments: Generating image descriptions in English using Arabic as the pivot language and generating image descriptions in Arabic using English as the pivot language.

4.1 Datasets and Evaluation Metrics

In this section, we present an overview of the datasets employed in our experiments and the evaluation metrics utilized to assess the performance of our cross-lingual image captioning model.

4.1.1 Datasets

We utilized two benchmark datasets for our experiments, as outlined in Table 2.

Table 2: Statistics of the datasets used in our experiments

Datasets	Language	Image no.	Caption no. per image	Training set	Validation set	Test set
Flickr8k	English	8092	5	6092	1000	1000
AraImg2k	Arabic	2000	5	1500	250	250

Flickr8k (English Dataset): This dataset consists of 8092 images, with each image accompanied by five annotated English descriptions. To ensure data consistency, we divided the dataset into three sets: 6092 images for the training set, 1000 images for the validation set, and another 1000 images for the test set. English word segmentation was performed using the "Stanford Parser" tool (https://stanfordnlp.github.io/CoreNLP/index.html), retaining English words that appeared at least 5 times and truncating sentences exceeding 20 words in length.

AraImg2k (Arabic Dataset): This dataset comprises 2000 images, each associated with five manually annotated Arabic descriptions. To maintain uniformity, we split this dataset into three subsets: 1500 images for training, 250 images for validation, and 250 images for testing. Arabic word segmentation followed the method proposed by [55], retaining Arabic words occurring at least 5 times. The segmentation data was extracted from the ATB and stored in text files, with each sentence treated as a time-series instance. Each file contained information for a single sentence.

It is important to note that the images and sentences in AraImg2k and Flickr8k are distinct from each other.

4.1.2 Evaluation Metrics

For evaluating the generated image descriptions, we employed semantic evaluation metrics commonly used to assess the quality of machine-generated text compared to human references. These metrics provide insights into the model's performance in generating accurate and fluent image descriptions. The following evaluation metrics were used:

- BLEU (Bilingual Evaluation Understudy): Measures the quality of machine-generated text by comparing it to human references. We reported BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Assesses text quality by comparing the overlap of machine-generated text with human references.
- METEOR (Metric for Evaluation of Translation with Explicit ORdering): Measures the quality of machine-generated text by considering word choice, synonymy, stemming, and word order.
- CIDEr (Consensus-Based Image Description Evaluation): Evaluates the diversity and quality of generated descriptions by computing consensus scores based on human references.
- SPICE (Semantic Propositional Image Caption Evaluation): Evaluates the quality of generated descriptions by assessing their semantic content and structure.

4.2 Training Settings

In the following sections, we describe the specific training settings for our cross-lingual image captioning model:

4.2.1 Image Encoder-Sentence Decoder Module

We utilized the pre-trained ResNet-101 model and a fully connected layer to extract image features v^I resulting in a dimension of $d = 512$. These features were then used as the initial input to the decoder $LSTM_G$ at time step 0.

4.2.2 Cross-Modal Semantic Matching Module

The image semantic embedding network consists of the pre-trained ResNet-101 model and a fully connected layer. The target language encoder employed a single-layer $LSTM_E$ structure.

4.2.3 Cross-Lingual Semantic Matching Module

Both the axis language and target language encoders used single-layer $BGRU_{PE}$ and $BGRU_{TE}$ frameworks with a hidden layer dimension of 512. The output dimension of BGRU was set to 1024.

4.2.4 Target Language Domain Evaluation Module

The language sequence model utilized a single-layer $LSTM_L$. The hidden layer dimension and word embedding dimension for all LSTM structures in this study were set to $d = 512$.

Throughout the model training process for both subtasks, dropout was set to 0.3, the batch size during pre-training was 128, and during reinforcement training, it was 256.

After the pre-training of the Semantic Matching module (Section 3.3) and the Language Optimization module (Section 3.4), the learning parameters θ_μ, θ_ρ, and θ_ω remained fixed. Both of these modules provided rewards to guide the Image Description Generation module (Section 3.2) in learning more source-domain semantic knowledge and target-domain language knowledge.

For the task of generating image descriptions in English using Arabic as the axis language, the learning rate for pre-training the Image Description Generation module is 1E-3. The learning rates for pre-training the source-domain semantic matching module and the target-language domain evaluation module are set to 2E-4. When training with language evaluation rewards and multi-modal semantic rewards, the learning rate for the Image Description Generation module is 4E-5. The values of α, β, and γ are set to 1, 1, and 0.15, respectively.

However, for the task of generating image descriptions in Arabic using English as the axis language, the learning rate for pre-training the Image Description Generation module is 1E-3. The learning rates for pre-training the source-domain semantic matching module and the target-language domain evaluation module are set to 4E-4. When training with language evaluation rewards and multi-modal semantic matching rewards, the learning rate for the Image Description Generation module is 1E-5. The values of α, β, and γ are set to 1, 1, and 1, respectively.

4.3 Results Analysis

4.3.1 Ablation Experiments

To assess the effectiveness of the Image & Cross-Language Semantic Matching module and the Target Language Domain Evaluation module, ablation experiments were conducted. Table 3 presents

the results of ablation experiments for the task of cross-language image description from Arabic to English and from English to Arabic.

Table 3: The contributions of different rewards for cross-lingual English image captioning on Flickr8k test dataset and cross-lingual Arabic image captioning on AraImg2k test dataset

Task	$L(\theta_G)$	$r_{\|g}$	$r^{I}_{r\|v}$	$r^{P}_{r\|v}$	Metrics							
					BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
Cross-language English image captioning	✓	—	—	—	81.0	72.4	67.3	65.3	40.6	44.3	74.1	50.2
	✓	—	✓	✓	82.0	73.6	67.9	65.5	40.4	44.6	75.3	50.5
	✓	✓	—	—	89.0	77.3	70.1	66.9	40.2	44.0	76.4	51.1
	✓	✓	✓	—	88.0	79.9	74.4	71.0	41.2	44.6	84.4	50.5
	✓	✓	✓	✓	**91.7**	**82.4**	**75.9**	**71.8**	**41.7**	**45.5**	**87.9**	**51.7**
Cross-language Arabic image captioning	✓	—	—	—	85.5	79.6	74.3	70.7	41.8	52.0	76.8	54.6
	✓	—	✓	✓	86.4	80.2	74.9	71.4	42.0	52.6	78.2	55.0
	✓	✓	—	—	88.0	81.1	75.8	72.0	42.4	52.5	79.0	54.8
	✓	✓	✓	—	91.0	82.5	76.5	72.2	42.9	52.9	80.4	55.1
	✓	✓	✓	✓	**91.7**	**83.9**	**77.9**	**73.6**	**43.2**	**54.0**	**81.7**	**55.5**

In Table 3, the baseline model employed Eq. (1), $L(\theta_G)$ as the objective function. The model trained with the $r^{I}_{r|v}$ reward represents participation in the Cross-Modal Semantic Matching module (Section 3.3.1). The model trained with the $r^{P}_{r|v}$ reward represents participation in the Cross-Language Semantic Matching module (Section 3.3.2). The model trained with the $r_{|g}$ reward represents participation in the Target Language Domain Evaluation module (Section 3.4). The model that jointly uses rewards $r^{I}_{r|v}$, $r^{P}_{r|v}$, and $r_{|g}$ is also evaluated.

The results of the ablation experiments shed light on the impact of various reward components on our model's performance for both cross-language English and Arabic image captioning tasks.

Figs. 3 and 4 illustrate the detailed evaluation metrics for cross-language English and Arabic image captioning, respectively, further elucidating the findings of our study.

According to Table 3, introducing the Multi-Modal Semantic Relevance Reward $r^{I}_{r|v}$ and Cross-Language Semantic Matching Reward $r^{P}_{r|v}$ led to improvements in several performance metrics. Notably, the CIDEr scores increased for English and Arabic by 1.2% and 1.4%, respectively, compared to the baseline. These results indicate that the Image & Cross-Language Semantic Matching module enhanced the semantic relevance of the generated sentences.

The Target Language Domain Evaluation Reward $r_{|g}$ played a positive role in both cross-language English and Arabic image description tasks. For cross-language English and Arabic image captioning, CIDEr scores increased by 2.3% and 2.2%, respectively, compared to the baseline.

Furthermore, the combined effect of the Target Language Domain Reward $r_{|g}$ and Image-Sentence Semantic Matching Reward $r^{I}_{r|v}$ resulted in substantial performance improvements in both tasks. For cross-language English and Arabic image descriptions, CIDEr scores increased by 10.3% and 3.6%, respectively, compared to the baseline. This indicates that combining these rewards results in descriptions that are more semantically consistent with the images.

Finally, when all rewards r_{lg}, r^{I}_{rlv}, and r^{P}_{rlv} were considered jointly, significant improvements were observed across all metrics. In comparison to the baseline, CIDEr scores increased for English and

Arabic by 13.8% and 4.9%, respectively. This highlights the effectiveness of incorporating guidance from both the Image & Cross-Language Domain and Target Language Domain in improving fluency and semantic relevance in generated sentences.

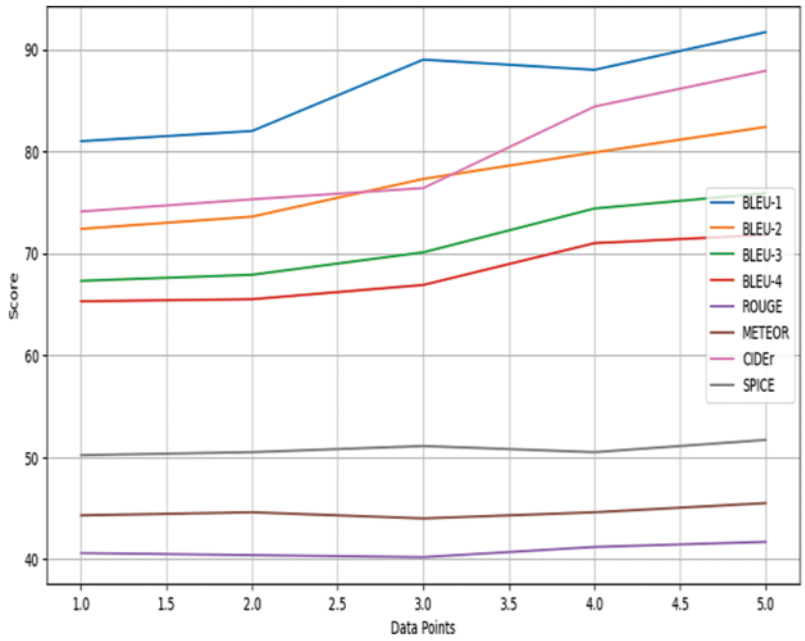

Figure 3: Evaluation metrics for cross-language English image captioning

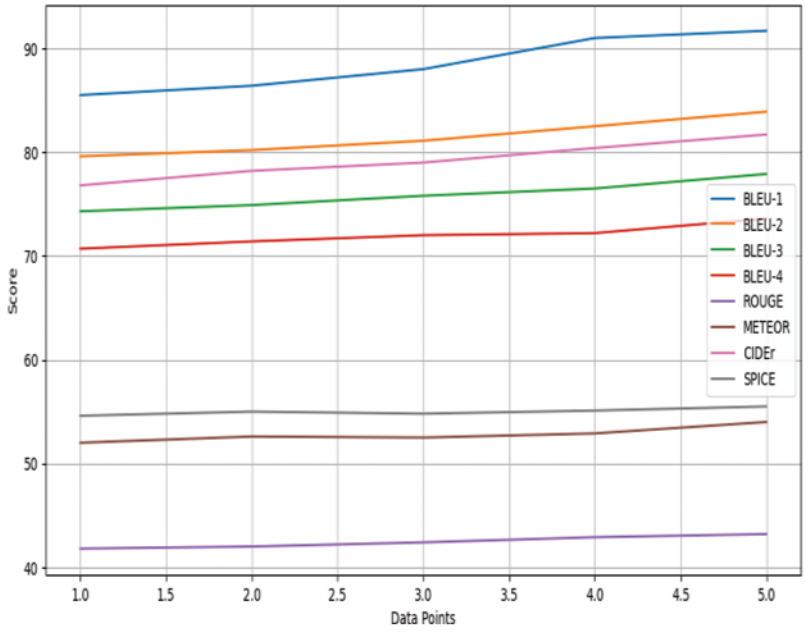

Figure 4: Evaluation metrics for cross-language Arabic image captioning

It is worth noting when comparing experiments involving only the r_{lg} reward to those with both the r_{lg} and r^{l}_{rlv} rewards, CIDEr scores for English and Arabic increased by 8.0% and 1.4%, respectively. The differential impact of the r^{l}_{rlv} reward on the two subtasks is noticeable. This difference is because, in the cross-language English image description subtask, the test set from Flickr8k contains various scenes, such as people, animals, and objects, making the visual semantics richer and more diverse. In this scenario, the Image-Sentence Semantic Matching Reward r^{l}_{rlv} demonstrates excellent semantic complementing ability (resulting in an 8.0% increase).

However, in the cross-language Arabic image description subtask, the test set AraImg2k primarily features images with a single visual scene (mostly focusing on people). Consequently, there is limited visual semantics to complement. Despite this, the method still improved performance by 1.4%.

The data in Table 3 was obtained through experimental trials conducted using the Python programming language. The trials were designed to evaluate the impact of different reward mechanisms in our cross-lingual image captioning model. The experiments were carried out on the Flickr8k dataset for English captions and the AraImg2k dataset for Arabic captions. These results demonstrate the effectiveness of our model in enhancing cross-lingual image captioning through a combination of semantic matching and language evaluation rewards.

4.3.2 Cross-Language English Image Description Performance Analysis

Table 4 provides a comparative analysis of various methods, including our study, for cross-language English image description tasks. The performance metrics presented were derived from experimental results on the English Flickr8k dataset. These results were obtained through systematic testing and evaluation of our model against established benchmarks in the field.

Table 4: Performance comparison for English image description on Flickr8k dataset

Ref.	Year	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
[56]	2022	Flickr8k	–	–	–	0.074	0.29	0.3	0.33	0.037
[57]	2022	Flickr8k	0.6126	0.4091	0.2762	0.1866	–	–	–	–
[58]	2022	Flickr8k	85.0	78.4	70.3	48.3	69.2	35.4	–	–
[59]	2023	Flickr8k	–	–	–	0.52	–	–	–	–
[60]	2023	Flickr8k	–	–	–	0.1044	–	–	–	–
[61]	2023	Flickr8k	0.6338	0.4825	0.3940	0.3275	–	–	–	–
[62]	2023	Flickr8k	41.25	37.77	78.87	93.91	34.56	38.56	–	–
Ours	2023	Flickr8k & AraImg2k	**91.7**	**82.4**	75.9	71.8	**41.7**	**45.5**	**87.9**	**51.7**

A comparison between our work and previous studies based on the data in Table 4 demonstrates our model's superior performance. Our approach consistently surpasses prior methods across diverse evaluation metrics. Notably, it excels in BLEU and CIDEr scores, signifying its improved accuracy and diversity in generating English image descriptions.

Fig. 5 shows the visual results of this model on the cross-language English image description task using the Flickr8k test set. The red font indicates semantic errors from the baseline model's translation, while the green font represents correct semantics from this model's translation. The figure illustrates that this model generates descriptions closer to the visual content of the images. For example, it can identify object attributes, replace the incorrect 'women' with 'men', and infer object

relationships, correcting "A red Jeep driving down in a mountainous area" to "driving down a rocky hill." Additionally, this model's generated sentences have fewer stylistic differences from the target language. For instance, the sentences generated by this model tend to follow the target language's style of "someone doing something somewhere," while the baseline model prefers to add attributive modifiers to objects.

Figure 5: Examples of the cross-lingual English image captioning from the Flickr8k test set

4.3.3 Cross-Language Arabic Image Description Performance Analysis

Similar to Table 4, Table 5 presents a comparative analysis of various methods for cross-language Arabic image description tasks. The data was derived from testing our model on both the Arabic Flickr8k and AraImg2k datasets, providing a comprehensive performance evaluation across multiple metrics.

Table 5: Performance comparison for Arabic image description on Arabic Flickr8k and AraImg2k datasets

Ref.	Year	Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE	METEOR	CIDEr	SPICE
[63]	2018	Flickr8k	34.8	–	–	–	–	–	–	–
[31]	2018	Flickr8k	46	26	19	8	–	–	–	–
[38]	2020	Flickr8k	33	–	–	6	–	–	–	–
[64]	2021	Flickr8k	44.3	–	–	15.6	–	–	–	–
[30]	2022	Flickr8k	39.10	25.13	13.96	8.29	–	–	–	–
[57]	2022	Flickr8k	–	–	–	0.062	0.29	0.31	0.31	0.037
[65]	2023	Flickr8k	0.59	0.39	0.30	0.16	0.24	0.21	0.16	–
Ours	2023	Flickr8k & AraImg2k	**91.7**	**83.9**	**77.9**	**73.6**	**43.2**	**54.0**	**81.7**	**55.5**

Comparing our work with previous studies based on the data in Table 5 reveals a significant advancement. Our model consistently outperforms existing methods across all evaluation metrics. Notably, it achieves remarkable improvements in BLEU, CIDEr, and SPICE scores, reflecting its superior accuracy, diversity, and linguistic quality in generating Arabic image descriptions.

Fig. 6 indicates that the descriptions generated by the proposed model are more semantically relevant to the visual content. For example, the proposed model is capable of supplementing and correcting missing or incorrect visual information, resulting in more coherent and accurate sentences. Additionally, the sentences generated by the proposed model align more closely with the style of real descriptions, presenting a continuous and concise style.

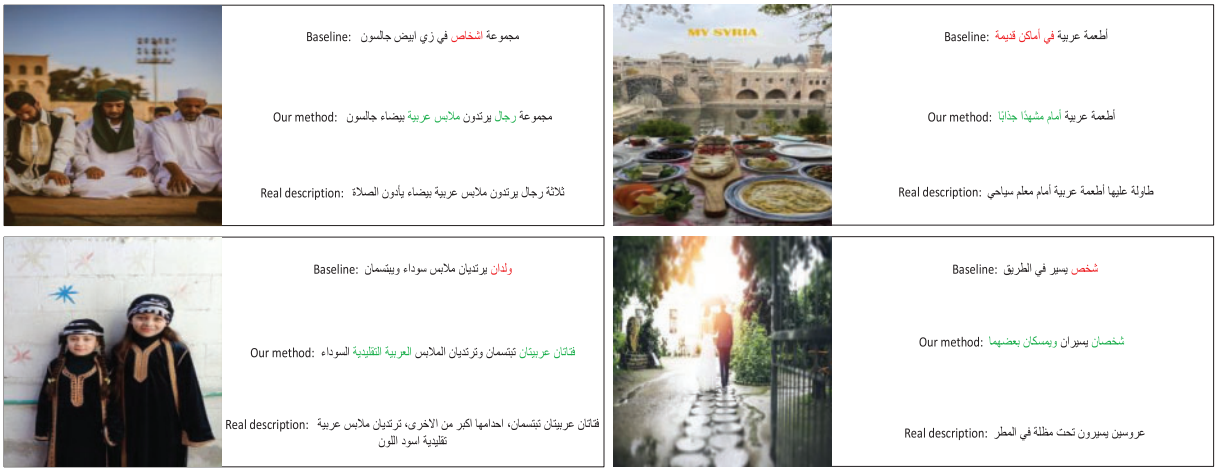

Figure 6: Examples of the cross-lingual Arabic image captioning from the Flickr8k test set

In summary, the performance analysis of the cross-language Arabic image description task shows that the proposed model consistently outperforms baseline and state-of-the-art methods in various evaluation metrics. It generates descriptions that are not only more semantically accurate but also stylistically aligned with the target language, making it an effective solution for cross-language image description tasks.

In conclusion, this study represents a significant contribution to the field of cross-lingual image description. Our method's ability to generate culturally relevant and semantically coherent captions across languages is not just an academic advancement; it has practical implications for enhancing multilingual understanding and communication. The introduction of the AraImg2k dataset, along with our novel methodologies, sets a new benchmark in the field and lays the groundwork for future research in this area.

5 Conclusion

In this study, we presented a novel approach for cross-lingual image description generation, aiming to bridge the gap between different languages and facilitate the understanding of images across linguistic barriers. Our method combines state-of-the-art techniques in image analysis, natural language processing, and cross-lingual semantics, resulting in a robust and effective model for generating image descriptions in multiple languages.

5.1 Key Contributions

Our research makes several key contributions to the field of cross-lingual image description:

- Effective Cross-Lingual Image Description: We successfully developed a model capable of generating image descriptions in English using Arabic as the pivot language and vice versa.

This achievement highlights the versatility and adaptability of our approach to handling diverse language pairs.
- Semantic Relevance Enhancement: Through the Image & Cross-Language Semantic Matching module, we demonstrated significant improvements in the semantic relevance of generated sentences. This enhancement contributes to more accurate and contextually appropriate image descriptions.
- Stylistic Alignment: Our model not only excels in semantic accuracy but also exhibits a superior ability to align with the stylistic nuances of the target language. This results in image descriptions that are more fluent and natural, closing the gap between machine-generated and human-authored content.

5.2 Future Work

While our current research presents a substantial step forward in cross-lingual image description, there are several exciting avenues for future exploration:

- Multimodal Enhancements: Incorporating additional modalities such as audio or video content into the image description process could lead to more comprehensive and context-aware descriptions, enabling applications in areas like multimedia indexing and retrieval.
- Low-Resource Languages: Extending our model's capabilities to low-resource languages is a promising direction. This would require addressing the challenges of limited training data and language-specific complexities.
- Fine-Grained Image Understanding: Future work can focus on improving the model's ability to capture fine-grained details within images, allowing for more precise and nuanced descriptions, especially in complex scenes.
- User Interaction: Incorporating user feedback and preferences into the image description generation process can lead to personalized and user-specific descriptions, enhancing the user experience in various applications.
- Real-Time Applications: Adapting our model for real-time applications, such as automatic translation during live events or real-time image description for the visually impaired, is an exciting area for future research and development.

Acknowledgement: None.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Study conception and design: Emran Al-Buraihy, Wang Dan; data collection: Emran Al-Buraihy; analysis and interpretation of results: Emran Al-Buraihy, Wang Dan; draft manuscript preparation: Emran Al-Buraihy, Wang Dan. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The code and the dataset will be available from the authors upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] N. Kumar *et al.*, "Harnessing the power of big data: Challenges and opportunities in analytics," *Tuijin Jishu/J. Propul. Tech.*, vol. 44, no. 2, pp. 681–691, 2023. doi: 10.52783/tjjpt.v44.i2.193.

[2] A. S. George, A. H. George, and T. Baskar, "Emoji unite: Examining the rise of emoji as an international language bridging cultural and generational divides," *Partners Univ. Int. Innov. J.*, vol. 1, no. 4, pp. 183–204, Aug. 2023. doi: 10.5281/zenodo.8280356.

[3] J. Kiaer, *Emoji Speak: Communication and Behaviours on Social Media*. London: Bloomsbury Academic, 2023.

[4] J. Madake, S. Bhatlawande, A. Solanke, and S. Shilaskar, "PerceptGuide: A perception driven assistive mobility aid based on self-attention and multi-scale feature fusion," *IEEE Access*, vol. 11, pp. 101167–101182, 2023. doi: 10.1109/ACCESS.2023.3314702.

[5] W. Zheng, X. Liu, X. Ni, L. Yin, and B. Yang, "Improving visual reasoning through semantic representation," *IEEE Access*, vol. 9, pp. 91476–91486, 2021. doi: 10.1109/ACCESS.2021.3074937.

[6] X. Yang, H. Zhang, and J. Cai, "Learning to collocate neural modules for image captioning," in *2019 IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, 2019. doi: 10.1109/iccv.2019.00435.

[7] H. M. Kuzenko, *The Role of Audiovisual Translation in the Digital Age*. Riga, Latvia: Baltija Publishing, Jun. 27, 2023. doi: 10.30525/978-9934-26-319-4-14.

[8] L. H. Li, Z. Y. Dou, N. Peng, and K. W. Chang, "DesCo: Learning object recognition with rich language descriptions," Jun. 2023. doi: 10.48550/arXiv.2306.14060.

[9] T. Kocmi, D. Macháček, and O. Bojar, "The reality of multi-lingual machine translation," Feb. 2022. doi: 10.48550/arXiv.2202.12814.

[10] Y. Liu, A. Francis, C. Hollauer, M. C. Lawson, O. Shaikh and A. Cotsman, "Reliability of electric vehicle charging infrastructure: A cross-lingual deep learning approach," *Commun. Trans. Res.*, vol. 3, no. 6, pp. 100095, 2023. doi: 10.1016/j.commtr.2023.100095.

[11] K. Sanders, D. Etter, R. Kriz, and B. van Durme, "MultiVENT: Multilingual videos of events with aligned natural text," Jul. 2023. doi: 10.48550/arXiv.2307.03153.

[12] S. Amirian, K. Rasheed, T. R. Taha, and H. R. Arabnia, "Automatic image and video caption generation with deep learning: A concise review and algorithmic overlap," *IEEE Access*, vol. 8, pp. 218386–218400, 2020. doi: 10.1109/ACCESS.2020.3042484.

[13] Z. Yang, Q. Liu, and G. Liu, "Better understanding: Stylized image captioning with style attention and adversarial training," *Sym.*, vol. 12, no. 12, pp. 1978, 2020. doi: 10.3390/sym12121978.

[14] N. Younis, "I-Arabic: Computational attempts and corpus issues in modern Arabic," مجلة جامعة مصر للدراسات الإنسانية., vol. 3, no. 3, pp. 301–325, 2023. doi: 10.21608/mjoms.2023.299689.

[15] P. S. Rao, "The role of English as a global language," *Res. J. Eng.*, vol. 4, no. 1, pp. 65–79, Jan. 2019.

[16] B. Liu *et al.*, "On the cultural gap in text-to-image generation," Jul. 2023. doi: 10.48550/arXiv.2307.02971.

[17] Z. Zhang, P. Lu, D. Jiang, and G. Chen, "TRAVL: Transferring pre-trained visual-linguistic models for cross-lingual image captioning," in *Web and Big Data. APWeb-WAIM 2022*, Nanjing, China, Springer, 2022, vol. 13422, pp. 341–355. doi: 10.1007/978-3-031-25198-6_26.

[18] S. Sharoff, R. Rapp, and P. Zweigenbaum, "Building comparable corpora," in *Building Using Comp. Corpora Multi. Nat. Lang. Process.*, 2023, pp. 17–37. doi: 10.1007/978-3-031-31384-4_3.

[19] Z. Li, Z. Fan, J. Chen, Q. Zhang, X. Huang and Z. Wei, "Unifying cross-lingual and cross-modal modeling towards weakly supervised multilingual vision-language pre-training," in *Proc. 61st Annu. Meet. Assoc. Comput. Linguist.*, Toronto, Canada, 2023, pp. 5939–5958. doi: 10.18653/v1/2023.acl-long.327.

[20] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *Ext. Abstr. 2021 CHI Conf. Human Factors Comput. Syst.*, 2021, pp. 1–7. doi: 10.1145/3411763.3451760.

[21] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meet. Assoc. Comput.*, 2001, pp. 311–318. doi: 10.3115/1073083.1073135.

[22] C. Y. Lin, "A package for automatic evaluation of summaries," in *Text Summar. Bran. Out*, Jul. 2004, pp. 74–81.

[23] S. Banerjee and A. Lavie, "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments," in *Proc. ACL Workshop Intr. Extr. Eval. Meas. Mach. Trans. Summar.*, Jun. 2005, pp. 65–72.
[24] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *2015 IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 4566–4575. doi: 10.1109/cvpr.2015.7299087.
[25] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," *Computer Vision–ECCV 2016*, vol. 9909, no. 12, pp. 382–398, 2016. doi: 10.1007/978-3-319-46454-1_24.
[26] A. Koul, S. Ganju, and M. Kasam, *Practical Deep Learning for Cloud, Mobile, and Edge: Real-World AI & Computer-Vision Projects Using Python, Keras & Tensorflow*. Sebastopol, CA: O'Reilly Media, 2019.
[27] M. Shafiq and Z. Gu, "Deep residual learning for image recognition: A survey," *Appl. Sci.*, vol. 12, no. 18, pp. 8972, 2022. doi: 10.3390/app12188972.
[28] Y. Ding, "A systematic literature review on image captioning," in *HCI International 2023 Posters, Communications in Computer and Information Science*, C. Stephanidis, M. Antona, S. Ntoa, and G. Salvendy, Eds., Cham: Springer, 2023, vol. 1836, pp. 396–404. doi: 10.1007/978-3-031-36004-6_54.
[29] R. S. Al-Malki and A. Y. Al-Aama, "Arabic captioning for images of clothing using deep learning," *Sens.*, vol. 23, no. 8, pp. 3783, 2023. doi: 10.3390/s23083783.
[30] M. T. Lasheen and N. H. Barakat, "Arabic image captioning: The effect of text pre-processing on the attention weights and the BLEU-N scores," *Int. J. Adv. Comput. Sci. App.*, vol. 13, no. 7, pp. 413–422, 2022. doi: 10.14569/ijacsa.2022.0130751.
[31] H. A. Al-muzaini, T. N. Al-Yahya, and H. Benhidour, "Automatic arabic image captioning using RNN-LSTM-based language model and CNN," *Int. J. Adv. Comput. Sci. App.*, vol. 9, no. 6, pp. 67–72, 2018. doi: 10.14569/ijacsa.2018.090610.
[32] Y. Feng and M. Lapata, "Automatic caption generation for news images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 797–812, 2013. doi: 10.1109/TPAMI.2012.118.
[33] J. H. Tan, C. S. Chan, and J. H. Chuah, "COMIC: Toward a compact image captioning model with attention," *IEEE Trans. Multimed.*, vol. 21, no. 10, pp. 2686–2696, 2019. doi: 10.1109/TMM.2019.2904878.
[34] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: Lessons learned from the 2015 MSCOCO Image captioning challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, 2017. doi: 10.1109/TPAMI.2016.2587640.
[35] Y. A. Thakare and K. H. Walse, "A review of deep learning image captioning approaches," *J. Integr. Sci. Tec.*, vol. 12, no. 1, pp. 712, 2023. Accessed: Dec. 16, 2023. [Online]. Available: https://pubs.thesciencein.org/journal/index.php/jist/article/view/a712
[36] L. Tianying and Y. V. Bogoyavlenskaya, "Semantic transformation and cultural adaptation of metaphor and multimodal metaphor in multilingual communication from the perspective of cognitive linguistics," *Eurasian J. Appl. Linguist.*, vol. 9, no. 1, pp. 161–189, Jun. 11 2023. doi: 10.32601/ejal.901015.
[37] M. S. Rasooli, C. Callison-Burch, and D. T. Wijaya, ""Wikily" supervised neural translation tailored to cross-lingual tasks," in *Proc. 2021 Conf. Empir. Methods Nat. Lang. Process.*, 2021, pp. 1655–1670. doi: 10.18653/v1/2021.emnlp-main.124.
[38] O. ElJundi, M. Dhaybi, K. Mokadam, H. Hajj, and D. Asmar, "Resources and end-to-end neural network models for Arabic image captioning," in *Proc. 15th Int. Joint Conf. Comput. Vision, Imag. Comput. Graph. Theory App.*, 2020. doi: 10.5220/0008881202330241.
[39] X. Li, C. Xu, X. Wang, W. Lan, Z. Jia and G. Yang, "COCO-CN for cross-lingual image tagging, captioning, and retrieval," *IEEE Trans. Multimed.*, vol. 21, no. 9, pp. 2347–2360, 2019. doi: 10.1109/TMM.2019.2896494.
[40] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in *Proc. 25th ACM Int. Conf. Multimed.*, California, USA, 2017, pp. 1549–1557. doi: 10.1145/3123266.3123366.
[41] J. Hitschler, S. Schamoni, and S. Riezler, "Multimodal pivots for image caption translation," in *Proc. 54th Annu. Meet. Assoc. Comput. Linguist.*, Berlin, Germany, 2016, pp. 2399–2409. doi: 10.18653/v1/p16-1227.

[42] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual English-German image descriptions," in *Proc. 5th Workshop Vision Lang.*, Berlin, Germany, 2016, pp. 70–74. doi: 10.18653/v1/w16-3210.
[43] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale Japanese image caption dataset," in *Proc. 55th Annu. Meet. Assoc. Comput. Linguist.*, Vancouver, Canada, 2017, pp. 417–421. doi: 10.18653/v1/p17-2066.
[44] T. Miyazaki and N. Shimizu, "Cross-lingual image caption generation," in *Proc. 54th Annu. Meet. Assoc. Comput. Linguist.*, Berlin, Germany, 2016, pp. 1780–1790. doi: 10.18653/v1/p16-1168.
[45] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *2016 IEEE Conf. Compu. Vision Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 1096–1104. doi: 10.1109/cvpr.2016.124.
[46] G. Hacheme and S. Noureini, "Neural fashion image captioning: Accounting for data diversity," arXiv preprint arXiv:2106.12154, 2021. doi: 10.31730/osf.io/hwtpq.
[47] J. Kasai, K. Sakaguchi, L. Dunagan, J. Morrison, R. Le Bras and Y. Choi, "Transparent human evaluation for image captioning," in *Proc. 2022 Conf. North Am. Chapter Assoc. Comput. Linguist.: Human Lang. Tech.*, 2022. doi: 10.18653/v1/2022.naacl-main.254.
[48] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona and D. Ramanan, "Microsoft COCO: Common objects in context," *Computer Vision–ECCV 2014*, vol. 8693, no. 2, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48.
[49] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, 2013. doi: 10.1613/jair.3994.
[50] C. C. Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in *2017 IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Honolulu, HI, USA, 2017, pp. 6432–6440. doi: 10.1109/cvpr.2017.681.
[51] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguist.*, vol. 2, no. 1, pp. 67–78, 2014. doi: 10.1162/tacl_a_00166.
[52] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conf. Comput. Vision Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 2016, pp. 770–778. doi: 10.1109/cvpr.2016.90.
[53] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," Jul. 2017. doi: 10.48550/arXiv.1707.05612.
[54] T. Mikolov, Q. V. Le, and I. Sutskever, "Exploiting similarities among languages for machine translation," Sep. 2013. doi: 10.48550/arXiv.1309.4168.
[55] A. Almuhareb, W. Alsanie, and A. Al-Thubaity, "Arabic word segmentation with long short-term memory neural networks and word embedding," *IEEE Access*, vol. 7, pp. 12879–12887, 2019. doi: 10.1109/ACCESS.2019.2893460.
[56] J. Emami, P. Nugues, A. Elnagar, and I. Afyouni, "Arabic image captioning using pre-training of deep bidirectional transformers," in *Proc. 15th Int. Conf. Nat. Lang. Gen.*, Waterville, Maine, USA, 2022, pp. 40–51. doi: 10.18653/v1/2022.inlg-main.4.
[57] S. Shinde, D. Hatzade, S. Unhale, and G. Marwal, "Analysis of different feature extractors for image captioning using deep learning," in *2022 3rd Int. Conf. Emerg. Tech. (INCET)*, Belgaum, India, 2022, pp. 1–5. doi: 10.1109/incet54531.2022.9824294.
[58] D. Kumar, V. Srivastava, D. E. Popescu, and J. D. Hemanth, "Dual-modal transformer with enhanced inter-and intra-modality interactions for image captioning," *Appl. Sci.*, vol. 12, no. 13, pp. 6733, 2022. doi: 10.3390/app12136733.
[59] A. Singh, A. Shah, P. Kumar, H. Chaudhary, A. Sharma and A. Chaudhary, "Image captioning using Python," in *2023 Int. Conf. Power, Instrument., Energy Control (PIECON)*, Aligarh, India, 2023, pp. 1–5. doi: 10.1109/piecon56912.2023.10085724.
[60] B. Dixit, G. R. Pawar, M. Gayakwad, R. Joshi, A. Mahajan and S. V. Chinchmalatpure, "Challenges and a novel approach for image captioning using neural network and searching techniques," *Int. J. Intell. Syst.*

Appl. Eng., vol. 11, no. 3, pp. 712–720, 2023. Accessed: Dec. 16, 2023. [Online]. Available: https://ijisae.org/index.php/IJISAE/article/view/3277

[61] P. Singh, C. Kumar, and A. Kumar, "Next-LSTM: A novel LSTM-based image captioning technique," *Int. J. Syst. Assur. Eng. Manag.*, vol. 14, no. 4, pp. 1492–1503, 2023. doi: 10.1007/s13198-023-01956-7.

[62] B. S. Revathi and A. M. Kowshalya, "Automatic image captioning system based on augmentation and ranking mechanism," *Signal, Image Video Process.*, vol. 18, no. 1, pp. 265–274, 2023. doi: 10.1007/s11760-023-02725-6.

[63] V. Jindal, "A deep learning approach for Arabic caption generation using roots-words," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017. doi: 10.1609/aaai.v31i1.11090.

[64] S. M. Sabri, "Arabic image captioning using deep learning with attention," Ph.D. dissertation, Univ. of Georgia, USA, 2021.

[65] S. Elbedwehy and T. Medhat, "Improved Arabic image captioning model using feature concatenation with pre-trained word embedding," *Neural Comput. Appl.*, vol. 35, no. 26, pp. 19051–19067, 2023. doi: 10.1007/s00521-023-08744-1.