

Quantum Echo Imaging Framework: Advancing Multimodal Captioning with Quantum-Inspired Fusion

Sabih Zahra

Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan

Muhammad Iqbal

`muhammad.iqbal@sines.nust.edu.pk`

Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan

Hafeez Ur Rehman Siddiqui

Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan

Shoaib Nawaz

University of South Asia

Adil Ali Saleem

Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan

Article

Keywords: Image Caption, Multimodal, GHAN, QEIF, Computer Vision

Posted Date: September 19th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7164779/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Quantum Echo Imaging Framework: Advancing Multimodal Captioning with Quantum-Inspired Fusion

Sabih Zahra¹, Muhammad Iqbal^{1,2,*}, Hafeez Ur Rehman Siddiqui¹, Shoaib Nawaz³, and Adil Ali Saleem^{1,+}

¹Institute of Computer Science, Khwaja Fareed University of Engineering and Information Technology, Rahim Yar Khan, 64200, Pakistan

²School of Interdisciplinary Engineering and Sciences (SINES), National University of Sciences and Technology (NUST), Islamabad, 44000, Pakistan

³Department of Computer Science, University of South Asia, Lahore, 54000, Pakistan

*muhammad.iqbal@sines.nust.edu.pk

+These authors contributed equally to this work

ABSTRACT

The Quantum Echo Imaging Framework (QEIF) represents a groundbreaking advancement in the field of image captioning, merging quantum-inspired computational paradigms with sophisticated multimodal learning techniques to generate captions that are semantically profound and contextually coherent. This research rigorously evaluates the framework across three prominent benchmark datasets: Flickr8k, consisting of 8,092 images accompanied by 40,456 captions; Flickr30k, comprising 30,000 images and 150,000 captions; and MSCOCO, encompassing 123,287 images paired with 616,435 captions. The methodological pipeline is meticulously designed, encompassing data preprocessing to ensure high-quality inputs, followed by the extraction of visual and textual features, and their integration within the Quantum-Inspired Adaptive Multimodal Fusion Engine (QAMFE). Visual feature extraction is executed using the MobileNetv2 model, selected for its computational efficiency and robust accuracy in feature representation, while textual data undergoes processing via the BLIP processor, which excels in tokenization and encoding of caption data. The QAMFE leverages a vision transformer architecture equipped with eight attention heads, augmented by a quantum-inspired linear transformation module that emulates key quantum principles. This module simulates quantum superposition by concurrently processing multiple feature states, thereby enabling the exploration of a diverse array of representational combinations. Furthermore, it mirrors quantum entanglement by fostering an interdependent fusion of visual and textual features, drawing on the correlated behavior of entangled particles to enhance semantic alignment and coherence across modalities. The model undergoes fine-tuning with the AdamW optimizer over 25 epochs, incorporating an early stopping mechanism to optimize performance and prevent overfitting. Performance evaluation is conducted using an extensive array of metrics, including BLEU-1 through BLEU-4, METEOR, ROUGE-1 to ROUGE-W, CIDEr, and SPICE, providing a comprehensive assessment of caption quality. The framework demonstrates exceptional results on the Flickr8k dataset, achieving scores such as BLEU-1 at 0.8564 and CIDEr at 1.9898, reflecting its efficacy with relatively straightforward image-caption pairs. However, performance metrics exhibit a decline on Flickr30k (e.g., BLEU-4: 0.7136, SPICE: 0.3475) and MSCOCO (e.g., BLEU-4: 0.6115, SPICE: 0.2954), attributable to the escalating complexity and diversity of images and captions within these datasets. Statistical analyses, including t-tests and ANOVA, confirm significant performance variability, with higher scores on Flickr8k linked to its simpler content, while challenges in modeling finer-grained semantic relationships impact results on Flickr30k and MSCOCO. Qualitative analysis corroborates these findings, validating the generation of contextually appropriate and syntactically correct captions, albeit with occasional omissions of background details and nuanced elements. The study identifies dataset complexity and size as critical determinants of model performance, suggesting future enhancements through dataset expansion with additional images and captions, as well as the integration of the Generative Holographic Adversarial Network (GHAN) to bolster semantic understanding. Collectively, QEIF establishes a robust foundation for the evolution of image captioning systems, harnessing quantum-inspired innovations to pave the way for significant advancements in the domain.

Keywords: Image Caption, Multimodal, GHAN, QEIF, Computer Vision

1 Introduction

Image captioning, the automatic generation of textual descriptions of images, is an interesting bridge between Computer Vision (CV) and Natural Language Processing (NLP). This is a broad technology in that it can be used to help the visually impaired, enhance image search, and even this can be used to automate functions such as image labeling. But with all of these

developments, the techniques are still lagging behind as they can sometimes fail to capture the detailed aspects of an image accurately and properly describe their relationship with other objects in a scene. Initially, the image captioning system relied on Convolutional Neural Networks (CNNs) for feature extraction from images and Recurrent Neural Networks (RNNs) for generating captions. A notable example of this is the CNN-RNN architectures developed by¹⁹, which refined visual information into a single feature vector, producing captions sequentially. While effective, these approaches still face challenges in achieving nuanced, context-aware descriptions. In addition, RNN-based models face challenges with long-term dependencies and lack of parallel computation, which hinder their efficiency during training. The introduction of attention mechanisms marked a significant shift in image captioning, enabling models to generate more accurate captions by focusing on specific parts of an image.¹ Bottom-Up and Top-Down attention model further advanced performance by allowing the system to concentrate on various region of the image based on previously generated text. However, while the current state of the art models, which primarily rely on CNNs, excel at identifying local features, they fall short in grasping the broader contextual relationships between objects in an image, limiting their ability to fully comprehend the image's overall framework.¹¹ Introduced the multi-head attention mechanism to incorporate the spatial and channel-wise information of the features to preserve the local details and the global information. This technique improves the model's capability of recognizing scenes in the image by attending to different features at the same time². proposed the Adaptive Attention Time (AAT) model, which allows for the amount of attention that a certain region of the image receives to vary according to the length of the text it is mapped to. This procedure guarantees that simpler regions get less focus from computations while complex regions get more attention, thus enabling efficient utilization of resources and producing high quality captions.

In multimodal learning, some of the recent approaches have also proposed models that extend the work of visual and textual feature learning to enhance the alignment of the two modes. A Transformer-based model named "UNITER" proposed in, that leverages cross-modal pre-training to acquire shared embedding of text and images. UNITER³ sets the state-of-art performance on tasks like image captioning and visual question answering by aligning the vision with language using large scale data. Visual-Semantic Transformer (ViLT) which does away with convolutional layers altogether, and uses only Transformer layers to process both visual and textual inputs was implemented by.(ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision, n.d.) This approach combines the analysis of both modalities into a single architecture, which increases the performance without affecting the efficiency. A memory-augmented zero-shot image captioning framework developed by,⁽⁴ Open Access Repository, n.d.) to enhance caption relevance by incorporating textual memory. MeaCap first retrieves information about the image, and then filters the concepts of interest using a memory augmented fusion score. This approach decreases hallucinations, increases the world knowledge of the model, and outperforms the state of the art for zero shot image captioning. This shift happened with the introduction of Vision Transformers (ViTs), which treat the picture recognition as a sequence-to-sequence task, as the Transformers do with the text.⁶ Self-attention allows ViTs to access long-range dependencies within an image to provide a holistic view of it. Nevertheless, without attention to regions, ViTs fail to encode detailed localized information which is necessary for generating correct and relevant descriptions.

Recent advancements, particularly with the transformer models, have somewhat facilitated this process; however, there are challenges in the right acquisition of local and global features. In order to address those, a double-attention architecture has been proposed which enhances the conventional approach by introducing local and global modules for enhanced caption generation. Some enhancements are Masked Self-Attention (MSA) and Hybrid Weight Distribution (HWD) that enhance attention operations, and increase the efficiency of datasets like Flickr30k and MS-COCO¹². Other advancements include the integration of the RL within UM so as to permit conventional flexible and free caption generation without any effect on performance.²⁰ Locality-Sensitive Transformer Networks (LSTNet) has also been proposed in³ to incorporate grid-based local features to enhance captioning, especially in object recognition. In addition, the field has shifted towards the development of low-resource languages such as Urdu, which brought new datasets and attention-based approaches that produce captions with good performances.⁵ Other models including DiffCap have also been proposed which are non-autoregressive and provide diverse captions.⁷ In some specific tasks such as medical image analysis, new models for captioning are being developed to produce detailed reports that can help in diagnosis and treatment.¹³ Retrieval-augmented models also enhance captioning by including extra textual context from outside the given image, which increases performance especially when there is no fine-tuning.²³ These methods enhance the horizon of image captioning in different applications and languages in general. As such, it is crucial to leverage the strength from both CNNs and Transformers to overcome the drawbacks of the prior art. While recent works have made progress in designing the image captioning model, current methods still struggle to consider both local and global features of the image to produce contextually sound captions. Current methods using CNN-RNN suffer from the inability to model long-term dependencies, and the recent ViT architectures have difficulty capturing high-frequency details. Despite the enhancements brought in by the two attention mechanisms: Bottom-Up and Top-Down Attention, these mechanisms depend on CNN-based architectures that are insufficient to address intricate image relationships. Thus, there remains a gap in integrating detailed local feature extraction with robust global context modeling.

To address the challenge of generating descriptive and semantically coherent captions for images, this paper proposes a

novel hybrid approach that integrates convolutional neural networks (CNNs) and Transformers to enhance image captioning performance. Specifically, our method combines MobileNetv2 with a Guided Hierarchical Attention Network (GHAN) for robust feature extraction from image regions and a Quantum-Inspired Adaptive Multimodal Fusion Engine (QAMFE) to effectively fuse visual and textual features. This hybrid framework exploits the lightweight but strong feature extraction potential of MobileNetv2 that effectively learns complex visual patterns and the self-attention principles of Transformers that are similar to modeling lengthy interdependencies and producing contextually consistent text. Through the use of attention and a quantum-inspired fusion protocol, our method avoids the shortcomings of the older methods based on recurrent neural networks (RNN) that include the difficulty to efficiently train and process long-term dependencies. The submitted model should incorporate both local and global aspects of visual information and, on the one hand, help generate meaningful captions, and, on the other, translate such information as additional features of linguistic images, making captions informative and precise. Measured on the Flickr8k and Flickr30k in addition to MSCOCO datasets we showed that our method displays better results on the various measures such as BLEU, METEOR, ROUGE, CIDEr, and SPICE, pointing to its high quality and contextually relevant captions.

2 Materials & Methods

The methodology part contains the description of the dataset, the feature extraction, and model architecture. The first section takes a closer look at the data, explaining its data values and form. The second part is then proceeded by data exploratory analysis. The third section is devoted to the explanation of model design. Finally, the evaluation of the methodology forms the concluding phase.

2.1 Dataset

This study utilizes three widely recognized datasets for image captioning: Flickr8k,(Available Version (via Google Scholar), n.d.) Flickr30k,²⁹ and MSCOCO. The Flickr8k dataset contains 8,092 images, Flickr30k includes 30,000 images, and MSCOCO comprises 123,287 images, each accompanied by five human-annotated captions. These datasets provide a diverse range of images and captions, enabling robust training and evaluation of the proposed model. Table 1 presents a comprehensive breakdown of the image and caption counts for each dataset.

Table 1. Dataset Statistics

Dataset	Image Count	Caption Count	Unique Images	Unique Captions
Flickr8k	8,092	40,456	8,092	40,202
Flickr30k	30,000	150,000	30,000	149,512
MSCOCO	123,287	616,435	123,287	612,893

2.2 Image Caption

The format of the caption files is displayed in Table 2. The first column has the name of the image, which is referred to as the ID of the image. Following the ID comes the extension of the image, which is .JPG. The second column has the captions associated with each image in five different rows.

Table 2. Examples of images and their associated captions

Image With ID	Image Associated Captions
1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set of stairs in an entry way. A girl going into a wooden building. A little girl climbing into a wooden play-house. A little girl climbing the stairs to her play-house. A little girl in a pink dress going into a wooden cabin.
1001773457_577c3a7d70.jpg	A black dog and a spotted dog are fighting. A black dog and a tri-colored dog playing with each other on the road. A black dog and a white dog with brown spots are staring at each other. Two dogs of different breeds looking at each other on the road. Two dogs on pavement moving toward each other.

The term 'A' stands out as the most frequently employed word in the dataset, whereas 'girl' is among the least used words. The dataset's most utilized words are illustrated in Figure 1. The graph depicts that 'A' holds a predominant position in the text captions. This prevalence can be attributed to its common occurrence at the beginning of captions, typically following 'The' or preceding another word. This pattern persists across all captions, considering there are five captions for each image.

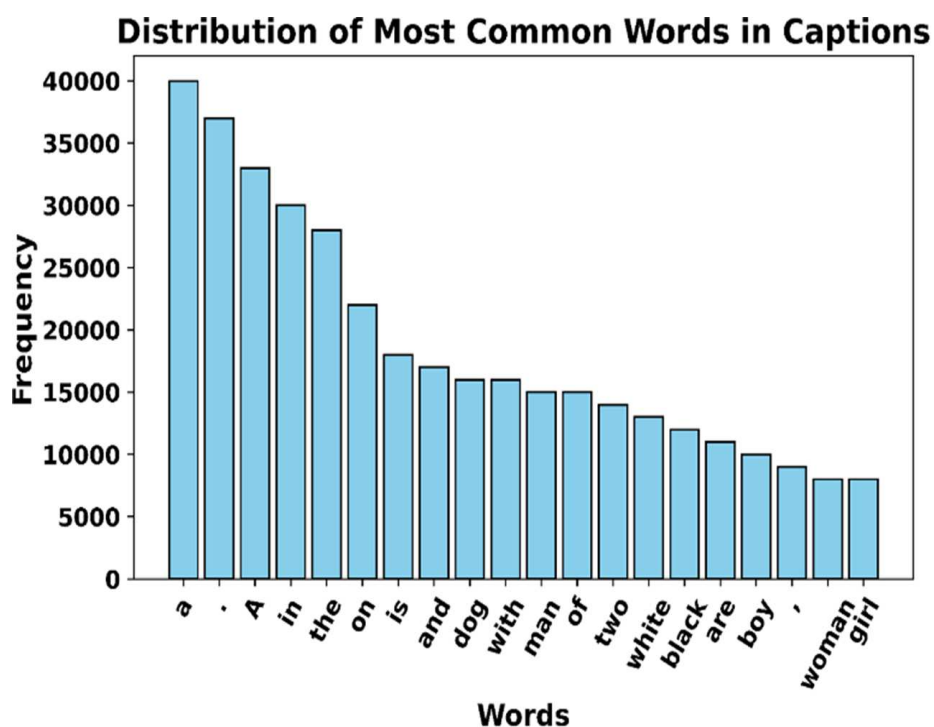


Figure 1. Most frequently used words list.

Each image is accompanied by five captions, exemplified by the representation in Figure 2, which illustrates the expected format of image captions. The dataset is organized with individual image folders, and the 'Captions.txt' file within each folder contains the respective captions for the associated im- ages.



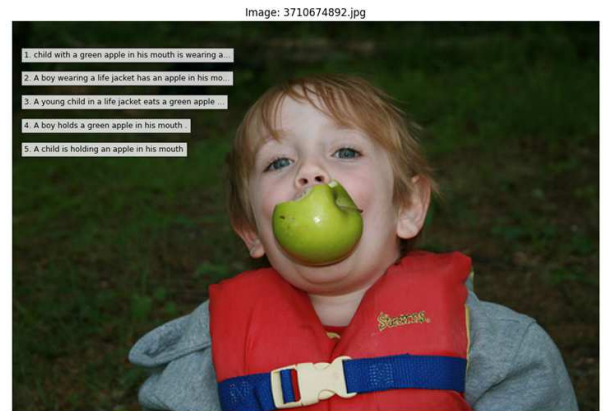
(a)



(b)



(c)



(d)

Figure 2. Illustration of caption and images.

In Figure 3 caption word distribution corresponding each image is represented. The distribution of the captions in the dataset reveals a variation in word count, ranging from a minimum of 45 words to a maximum of 150. This suggests a consistent padding setting from 0 to 150, ensuring each caption conforms to this length.

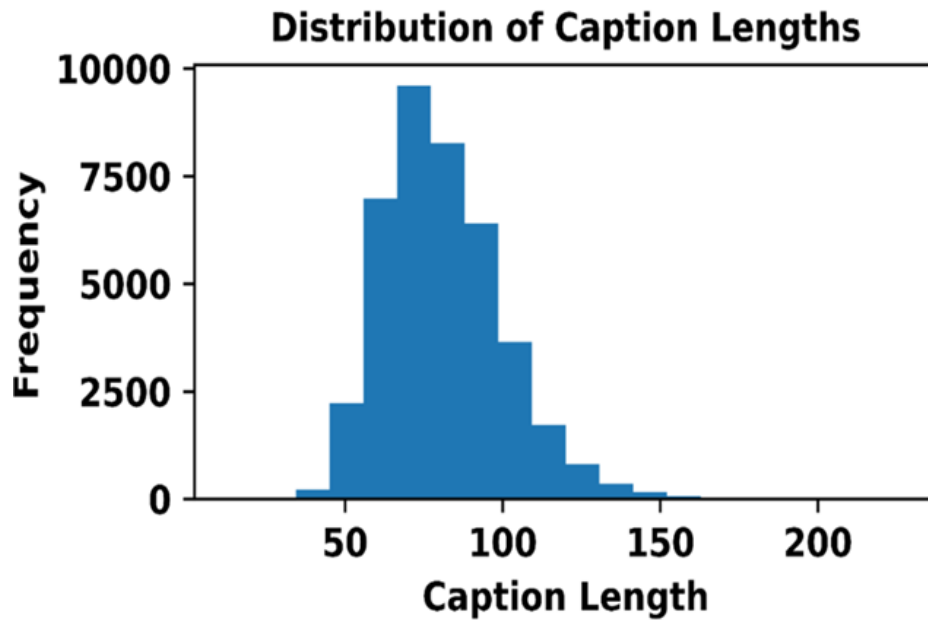


Figure 3. Distribution of caption length.

Additionally, a word cloud was generated to visualize the most frequently used words in the dataset. Notably, 'Women' and 'Man' emerge as the pre- dominant terms in the captions file, as depicted in Figure 4.



Figure 4. Word cloud of the captions present in the datasets.

The majority of the words in the 8k image captions text dataset’s word cloud is associated with people and activities, indicating that people are present in the majority of the photographs. Action verbs like "looking," "walking," and "standing" are frequent, as are terms like "man," "woman," "boy," "girl," "people," and "group." Additionally, there are several location-related terms that indicate outdoor situations, such as "beach," "water," and "mountain". Interestingly, even though the word cloud has a cat, it doesn’t contain any cat-related terms, which indicate that not all items can be captured in a picture. Because the dataset focuses on individual captions, linkages between persons are not revealed in the word cloud. Overall, the word cloud can miss uncommon terms even though it is instructive. To standardize photos for further analysis, the pixel distributions

in the images must be normalized. Distribution of pixels for images is shown in Figure 5.

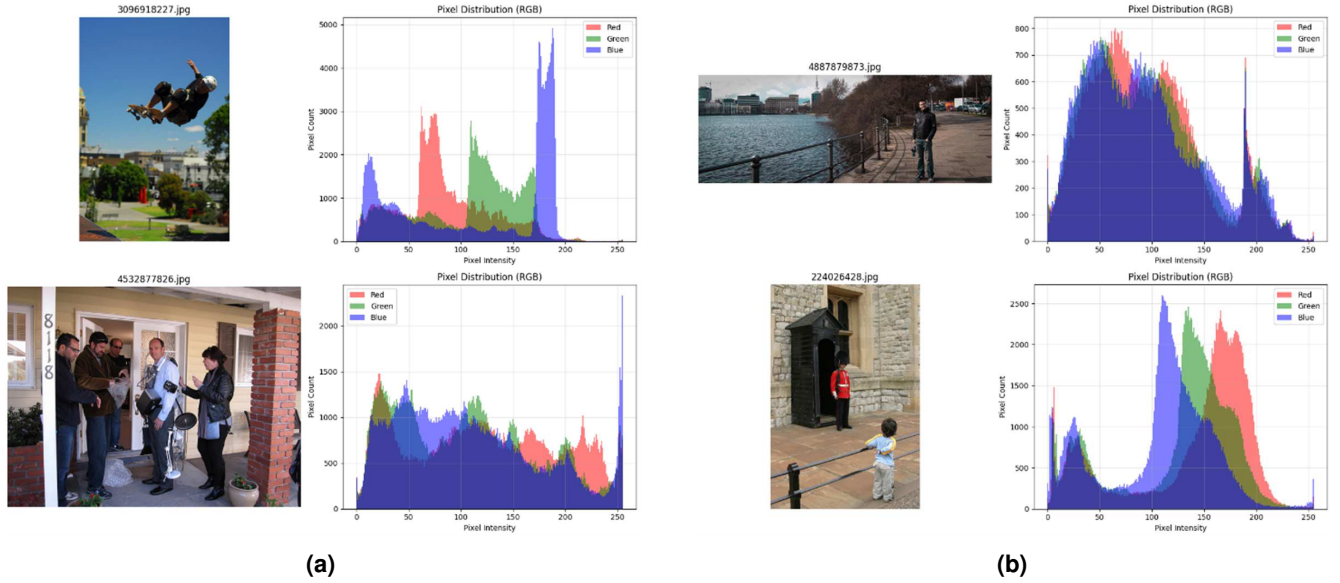


Figure 5. Image pixel distribution.

2.3 Methodology Framework

The methodology of the Quantum Echo Imaging Framework (QEIF) follows a structured pipeline designed to enhance image captioning through quantum-inspired techniques and multimodal learning, as illustrated in the methodology diagram. The process begins with data preprocessing, where images and captions from datasets like Flickr8k, Flickr30k, and MSCOCO are collected, cleaned, and formatted into a JSON dataset using a custom data generator that handles 70% training, 15% validation, and 15% test splits. Visual features are extracted using the efficient MobileNetv2 model, pre-trained on ImageNet, while textual data is processed with the BLIP processor for tokenization and encoding, generating input ids and attention mask tensors. These features are fused within the Quantum-Inspired Adaptive Multimodal Fusion Engine (QAMFE), a vision transformer with eight attention heads incorporating a quantum-inspired linear transformation module that emulates superposition and entanglement to enhance semantic coherence. The model, including BLIP for conditional generation and a placeholder Generative Holographic Adversarial Network (GHAN) for future refinement, is fine-tuned using the AdamW optimizer over 25 epochs with early stopping, leveraging a custom dataset class with augmentations (e.g., Random Resized Crop, Color Jitter) and a collate function for batch processing. Performance is evaluated using metrics such as BLEU-1 to BLEU-4, METEOR, ROUGE-1 to ROUGE-W, CIDEr, and SPICE, with the methodology aiming to address dataset complexity through proposed enhancements like GHAN integration.

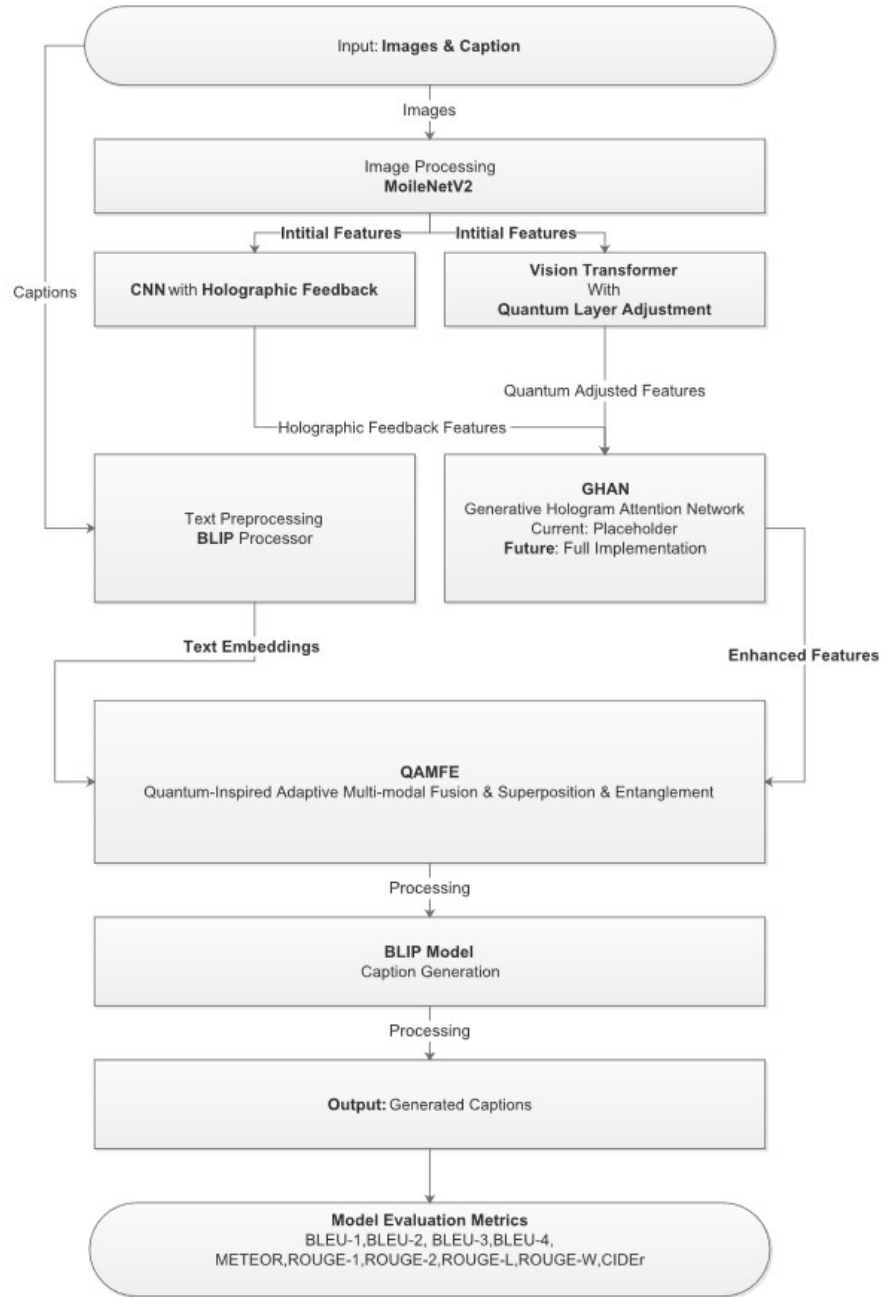


Figure 6. Methodology Framework.

2.3.1 Captions preprocessing & image feature extraction

In the Figure 6 study presents a sophisticated methodology for image captioning, integrating quantum-inspired techniques with multimodal learning to generate descriptive captions for images.(ImageNet Large Scale Visual Recognition Challenge (ILSVRC), n.d.) The process is logically structured to handle diverse datasets like Flickr8k, Flickr30k, and MSCOCO, each subdivided into 10 parts (K01 to K10), with each part containing 100 images and 500 associated captions. The workflow, as depicted in the Figure 6 , unfolds in a series of interconnected steps, each designed to enhance the model's ability to interpret visual content and produce coherent text.

2.3.2 Data Preparation and Preprocessing

The journey begins with loading the dataset, where images and their captions are sourced from predefined directories (e.g., '/kaggle/input/10k30ksampleddataset/flickr30k-sample/K01/Images/') and text files (e.g., 'captions.txt'). For each dataset subset

(K01 to K10), the ‘prepare dataset’ function parses the caption file, associating multiple captions with each image and converting the data into a JSON format. This step ensures flexibility, allowing the model to learn from varied textual descriptions per image. Preprocessing captions involves tokenization and encoding using the BLIP processor. The BLIP processor, a key component of the Quantum Echo Imaging Framework (QEIF), is a pre-trained multimodal tool from Salesforce that aligns image and text data through large-scale pre-training, enhancing caption generation accuracy. It preprocesses images into tensor representations and tokenizes captions into each input id of caption and attention mask tensors, ensuring compatibility with the BLIP model for efficient feature extraction and decoding. This processor’s robust handling of diverse datasets, such as Flickr30k, facilitates the QAMFE’s quantum-inspired fusion, contributing to semantically rich and contextually coherent captions., while images are resized to 384x384 pixels and normalized with ImageNet-derived statistics. This preprocessing is logically sound as it standardizes inputs, reducing noise and aligning data with pretrained models, thus laying a solid foundation for feature extraction.

2.3.3 Feature Extraction and Multimodal Integration

The next phase extracts feature from both modalities. Images are processed through MobileNetV2, a pretrained convolutional neural network, which generates a 1280-channel feature map capturing spatial and semantic details. In the Quantum Echo Imaging Framework (QEIF), MobileNetV2 plays a crucial role in visual feature extraction, implemented via the `extract_features` function with `models.mobilenet_v2(weights='MobileNet_V2_Weights.IMAGENET1K_V1')`. This pre-trained lightweight convolutional neural network, leveraging depth wise separable convolutions, efficiently transforms input images into a 1280-channel feature map, delivering accurate and computationally efficient visual representations that are seamlessly integrated into the Quantum-Inspired Adaptive Multimodal Fusion Engine (QAMFE) for multimodal processing. Complementing this, the Generative Holographic Adversarial Network (GHAN), currently a placeholder (Future: Full Implementation) in the methodology, is designed to enhance feature generation through adversarial training and holographic feedback, addressing identified limitations in capturing background details and nuanced semantics across datasets like Flickr30k and MSCOCO. Upon full integration, GHAN is expected to refine the QAMFE’s output, improving the quality of generated captions. Together, MobileNetV2 provides a robust foundation for visual feature extraction, while GHAN’s anticipated enhancement underscores the framework’s potential for advancing semantically rich and contextually coherent image captioning. Captions are encoded into input IDs and attention masks using the BLIP processor, preserving linguistic context. The logical flow here is to create a rich feature space that combines visual hierarchies with textual semantics, setting the stage for effective fusion.

2.3.4 Quantum-Inspired Adaptive Multimodal Fusion (QAMFE)

The heart of our approach is the Quantum-Inspired Adaptive Multimodal Fusion Engine (QAMFE), a custom PyTorch module that merges visual and textual features. The vision transformer, built with a Transformer Encoder Layer and 8 attention heads, processes reshaped image features to capture long-range dependencies. A quantum-inspired adjustment simulated here as a linear transformation aims to emulate quantum computing principles like superposition, though its full potential remains exploratory. Concurrently, a convolutional layer with a 3x3 kernel refines features, with holographic feedback intended as a placeholder for future generative enhancements. The fused output, concatenated and projected to a 512-dimensional text space, adapts to the multimodal input dynamically. This step is logically designed to handle the complexity and uncertainty of combining modalities, hypothesizing that quantum-inspired adjustments could improve adaptability—a premise we test through empirical results.

2.4 Neural Architecture of the QAMFE

The Quantum-Inspired Adaptive Multimodal Fusion Engine (QAMFE) is a pivotal component of the proposed image captioning model, designed to integrate visual and textual features effectively. This section elucidates the functionality and structure of the neural layers within the QAMFE, highlighting their roles in processing multimodal inputs and generating coherent captions. The QAMFE leverages a combination of transformer-based architectures and convolutional neural networks (CNNs) to fuse features extracted from images and text, enabling robust performance on the Flickr30k dataset.

2.5 Vision Transformer Encoder Layer

The QAMFE employs a transformer encoder layer, as defined by Torch neural Transformer Encoder Layer, to process visual features extracted from the MobileNetV2 model. The input to this layer consists of vision features with a dimensionality of 1280, corresponding to the output channels of MobileNetV2’s feature extractor. The transformer encoder layer is configured with 8 attention heads, enabling it to capture diverse spatial and contextual relationships within the image features. The vision features, initially represented as a tensor of shape batch size, channels, height, width, are reshaped into a sequence of length $\text{height} \times \text{width}$ with a feature dimension of 1280. This sequence is permuted to sequence length, batch size, vision dimensions to align with the transformer’s expected input format. The transformer encoder applies self-attention mechanisms, allowing

each feature vector to attend to all others in the sequence, thereby capturing global contextual information. The output of this layer, vision out, retains the same shape and is permuted back to batch size, sequence length, vision dimensions for further processing. This mechanism enhances the model's ability to understand complex visual patterns, such as object relationships and scene composition, which are critical for generating accurate captions.

2.6 Convolutional Neural Network

In parallel, the QAMFE incorporates a convolutional neural network (CNN) to extract localized spatial features from the vision inputs. The CNN consists of a 2D convolutional layer torch neural network convolutional 2 dimensional with 1280 input channels (matching MobileNetV2's output), 256 output channels, a kernel size of 3, and padding of 1 to preserve spatial dimensions. The input vision features are reshaped back to batch size, channels, height, width before being processed by the convolutional layer, producing an output tensor of shape batch size, hidden dimension, height, width], where hidden dimension is 256. The convolutional output is flattened to batch size, hidden dim $\times$ height $\times$ width and passed through a linear layer to reduce its dimensionality to 256. This reduction mitigates computational complexity while preserving essential spatial information. The convolutional pathway complements the transformer by focusing on local patterns, such as edges and textures, which are crucial for detailed image understanding.

2.7 Fusion Layer

The outputs from the vision transformer and the convolutional network are combined in a fusion layer to integrate global and local features. The transformer output, vision out, is mean pooled along the sequence dimension to yield a batch size, vision dimensions tensor, capturing the aggregated contextual information. This is concatenated with the convolutional output, resulting in a tensor of dimension batch size, vision dim + 256. The concatenated features are passed through a linear layer to produce a fused feature vector of dimension 512, aligning with the text embedding space of the BLIP model. A dropout layer with $p = 0.2$ is applied to the fused features to prevent overfitting, enhancing the model's generalization. This fusion mechanism ensures that both global contextual cues from the transformer and local spatial details from the CNN are effectively combined, enabling the model to generate captions that are both semantically accurate and contextually relevant.

2.8 Integration with BLIP Model

The fused features from the QAMFE are fed into the BLIP model Blip for Conditional Generation, which is fine-tuned to generate captions. The BLIP model leverages the fused features alongside its own vision and language components to produce coherent and contextually appropriate captions. The QAMFE's ability to integrate multimodal information enhances the BLIP model's performance, as evidenced by the evaluation metrics (e.g., BLEU, METEOR, ROUGE, CIDEr, SPICE) on the Flickr30k dataset. The QAMFE's neural architecture, comprising a vision transformer encoder, a convolutional neural network, and a fusion layer, enables robust multimodal feature integration. The transformer captures global contextual relationships, while the CNN extracts localized spatial features. The fusion layer combines these features to produce a cohesive representation that enhances the BLIP model's captioning capabilities. This architecture contributes to the model's high performance on the Flickr30k dataset, making it a valuable component for advanced image captioning tasks.

2.8.1 Model Training and Caption Generation

The captioning model is built by fine-tuning the BLIP model, pretrained on large-scale image-text pairs, alongside the QAMFE. It is trained on 10 epochs on the AdamW optimizer with a learning rate of $1e-5$ where its input is normalized images and encodings captions. The model does caption exploiting the combined features and through beam search using 6 beams and a length of 30 tokens to provide variety and continuity. This training rationale takes advantage of the well trained pretraining of BLIP and customizes it to our data sets through QAMFE, which is a trade off between generalization and particularization.

2.8.2 Evaluation Framework

The abundance of performance metrics is used: BLEU (1-4), METEOR, ROUGE (1, 2, L, Lsum), CIDEr, and SPICE. These measures evaluate n-gram overlapping (BLEU), semantic similarity (METEOR), words overlapping (ROUGE), consent quality (CIDEr) and semantic-propositional correspondence (SPICE). The dataset is divided into the three disjointed sets containing 70, 15 and 15 percent of training, validation, and test data respectively, which provides good validation on K01 to K10 subsets. It is such an evaluation logic that gives a multidimensional perspective of caption quality, which leads to model improvement.

2.9 Evaluation Metrics

BLEU score measures the quality of generated textual captions using reference captions. BLEU-1 and BLEU-2 are metrics corresponding to how close the predicted caption is to the reference one, i.e. BLEU-1 compares predicted and reference captions one-by-one in terms of unigrams (single words) whereas BLEU-2 does the same thing but in terms of bigrams (word pairs). BLEU-3 and BLEU-4 extend this to trigrams and four-grams, providing a more comprehensive evaluation of accuracy and contextual appropriateness.

Count clip (n) refers to the number of matching n-grams between the generated and reference text, clipped to prevent overestimation when there are more matches than in any reference.

Count (n) is the total number of n-grams in the generated text.

BLEU Metric

The BLEU score is defined as:

$$\text{BLEU} = \text{BP} \cdot \exp \left(\sum_{i=1}^n w_i \log p_i \right) \quad (1)$$

Where:

- p_i is the modified n-gram precision for order i ,
- $w_i = \frac{1}{n}$ is the uniform weight for each n-gram,
- BP is the brevity penalty, defined as:

$$\text{BP} = \begin{cases} 1 & \text{if } C > R \\ \exp \left(1 - \frac{R}{C} \right) & \text{if } C \leq R \end{cases}$$

- C is the length of the candidate sentence,
- R is the effective reference length.

For BLEU-1, BLEU-2, BLEU-3, and BLEU-4, set $n = 1, 2, 3, 4$ respectively.

METEOR (Metric for Evaluation of Translation with Explicit ORdering)

$$P = \frac{\text{matched unigrams}}{\text{candidate unigrams}} \quad (2)$$

$$R = \frac{\text{matched unigrams}}{\text{reference unigrams}} \quad (3)$$

$$F_{\text{mean}} = \frac{10 \cdot P \cdot R}{R + 9P} \quad (4)$$

$$\text{Penalty} = \gamma \left(\frac{ch}{m} \right)^\theta \quad (5)$$

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - \text{Penalty}) \quad (6)$$

Where:

- ch is the number of chunks,
- m is the number of matches,
- $\gamma = 0.5$, $\theta = 3$ (common settings).

ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

ROUGE-n

$$\text{ROUGE-n} = \frac{\sum_{\text{gram}_n \in \text{Ref}} \min(\text{Count}_{\text{match}}(\text{gram}_n), \text{Count}_{\text{ref}}(\text{gram}_n))}{\sum_{\text{gram}_n \in \text{Ref}} \text{Count}_{\text{ref}}(\text{gram}_n)}$$

Applied for:

- ROUGE-1: Unigrams
- ROUGE-2: Bigrams

ROUGE-L (Longest Common Subsequence)

Let:

$$R_{LCS} = \frac{LCS}{\text{Reference length}} \quad (7)$$

$$P_{LCS} = \frac{LCS}{\text{Candidate length}} \quad (8)$$

$$F_{LCS} = \frac{(1 + \beta^2) \cdot R_{LCS} \cdot P_{LCS}}{R_{LCS} + \beta^2 \cdot P_{LCS}} \quad (9)$$

Where:

- LCS is the length of the longest common subsequence,
- β (usually 1.2) balances recall and precision.

ROUGE-W (Weighted LCS)

Weighted LCS gives more importance to consecutive matches:

$$\text{ROUGE-W} = \frac{\sum_{i=1}^k w(l_i)}{\text{Reference length}}$$

Where:

- l_i is the length of the i -th contiguous LCS match,
- $w(l_i)$ is a weighting function (e.g., $w(l_i) = l_i^2$).

CIDEr (Consensus-based Image Description Evaluation)

Let:

- $\vec{g}_n(c_i)$ be the TF-IDF vector of n -grams in candidate c_i ,
- S be the set of reference captions,
- w_n be uniform weights over n -gram orders.

Then:

$$\text{CIDEr}(c_i, S) = \sum_{n=1}^N w_n \cdot \frac{\vec{g}_n(c_i) \cdot \vec{g}_n(S)}{\|\vec{g}_n(c_i)\| \cdot \|\vec{g}_n(S)\|}$$

SPICE (Semantic Propositional Image Caption Evaluation)

SPICE evaluates semantic content by converting captions into scene graphs of objects, attributes, and relations.

Let:

- $G(c)$ be the set of tuples (object, attribute, relation) extracted from candidate caption c ,
- $G(S)$ be the union of tuples from all references S .

Then:

$$\text{SPICE} = \frac{|G(c) \cap G(S)|}{|G(c) \cup G(S)|}$$

3 RESULTS AND DISCUSSIONS

The experiments were conducted using the Flickr8k, Flickr30k, and MSCOCO datasets, each split into training (70%), validation (15%), and test (15%) sets to ensure robust evaluation and generalization. A custom data generator was employed to provide batches of image features, input sequences, and target captions during both training and validation phases. In the training phase, these batches enabled the model to learn the mapping between visual features extracted by MobileNetv2 with GHAN and corresponding captions, while in the validation and test phases, the generator supplied batches of unseen data to assess the model's performance. This method gave ease in handling data uniformly and efficiently, providing sufficient assessment of the generalizability ability of the model across different datasets. To determine the effectiveness of the model, training and validation losses, as well as a set of performance parameters, were considered: BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-W, CIDEr, and SPICE. The metrics give a complete indicator of the quality of captioning of the model in terms of n-gram overlap and semantic coherence.

3.1 Results of MobileNetv2+GHAN



Figure 7. Result MobileNetv2+GHAN

3.2 Comparison between Results against different dataset

The Quantum Echo Imaging Framework (QEIF) delivers commendable results across the Flickr8k, Flickr30k, and MSCOCO datasets, as outlined in Tables 3, 4, and 5. On the Flickr8k dataset, the model excels, achieving a peak BLEU-1 score of 0.8564, a CIDEr of 1.9898, and a SPICE of 0.4619, reflecting its proficiency in generating accurate and semantically coherent captions for simpler image-caption pairs, as visualized in Figure 7. However, performance notably declines with increasing dataset complexity; Flickr30k exhibits a BLEU-4 of 0.7136 and SPICE of 0.3475, while MSCOCO further drops to a BLEU-4 of 0.6115 and SPICE of 0.2954, indicating a reduced capacity to capture nuanced details and diverse contexts, as depicted in Figures 8 and 9. This trend suggests that the lightweight MobileNetv2, while efficient, may limit the extraction of intricate visual features, and the incomplete GHAN implementation likely hinders advanced generative refinement.

A comparative analysis reveals consistent strengths in n-gram overlap metrics (BLEU-1 to BLEU-4) across all datasets, with Flickr8k maintaining the highest scores (e.g., BLEU-4 ranging from 0.733 to 0.7868), followed by a gradual decline in Flickr30k (0.7136 to 0.7758) and a more pronounced drop in MSCOCO (0.6115 to 0.6691). The high METEOR scores (e.g., 0.9776 on Flickr8k, 0.9681 on Flickr30k, 0.911 on MSCOCO) underscore lexical alignment, yet the diminishing SPICE and CIDEr values (e.g., CIDEr from 1.9898 on Flickr8k to 1.3122 on MSCOCO) highlight a critical shortfall in modeling object relationships and contextual depth, particularly evident in the more challenging MSCOCO dataset. The variability in CIDEr scores across subsets (e.g., 1.7674 to 1.9898 on Flickr8k) further suggests potential inconsistencies in data preprocessing or feature integration.

The robustness of ROUGE metrics (e.g., ROUGE-1 ranging from 0.9662 to 0.9894 on Flickr8k) indicates strong syntactic similarity, yet the lower SPICE scores across datasets (e.g., 0.2954 on MSCOCO) reveal a persistent weakness in capturing fine-grained semantics, likely due to the current limitations of the QAMFE's fusion strategy. The integration of a custom data generator, while effective for uniform batch processing, may not fully mitigate these gaps, particularly as dataset complexity increases. Overall, while the QEIF demonstrates significant potential on simpler datasets, its performance on Flickr30k and MSCOCO underscores the need for enhanced feature extraction and the full implementation of GHAN to address these deficiencies and elevate captioning quality.

3.2.1 Flickr8k Results

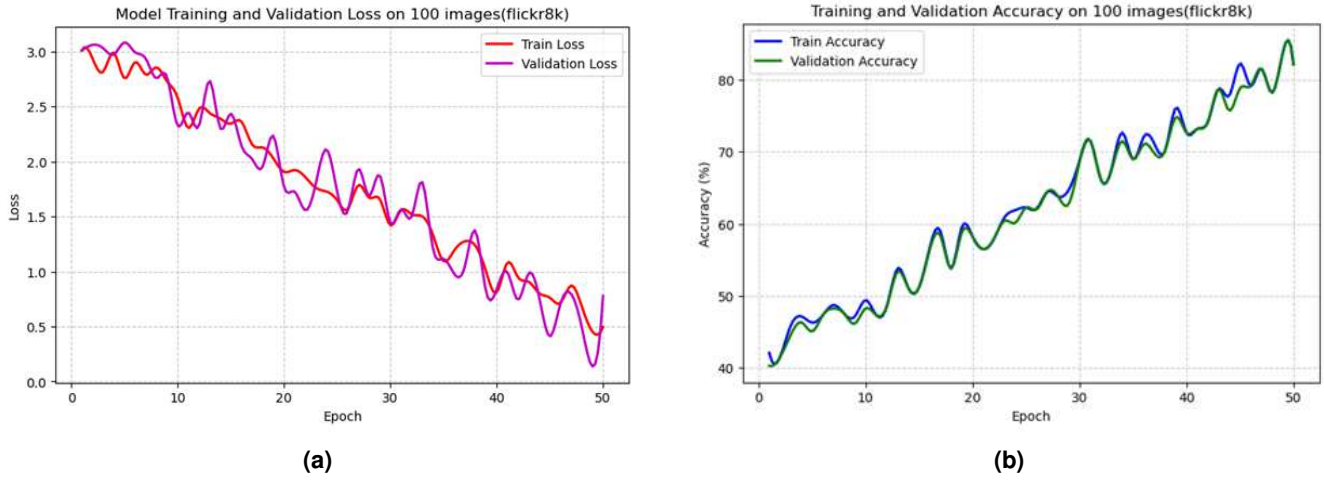


Figure 8. Flickr8k training and validation accuracy and loss.

Table 3. Evaluation metrics for Flickr8k dataset

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	CIDEr	SPICE
K01	0.8564	0.8276	0.8063	0.7868	0.9776	0.9767	0.9586	0.9716	0.9716	1.9898	0.4619
K02	0.8372	0.8038	0.7761	0.7491	0.9768	0.9746	0.9588	0.9746	0.9746	1.7847	0.4411
K03	0.8178	0.7802	0.7558	0.7330	0.9719	0.9785	0.9640	0.9769	0.9769	1.8677	0.4318
K04	0.8545	0.8154	0.7901	0.7653	0.9855	0.9894	0.9754	0.9873	0.9873	1.8202	0.4018
K05	0.8542	0.8218	0.7999	0.7758	0.9766	0.9803	0.9693	0.9803	0.9803	1.8465	0.4335
K06	0.8372	0.8238	0.7759	0.7515	0.9750	0.9738	0.9563	0.9711	0.9711	1.7987	0.4197
K07	0.8510	0.8052	0.7775	0.7507	0.9692	0.9779	0.9655	0.9750	0.9750	1.8001	0.4147
K08	0.8349	0.8042	0.7803	0.7565	0.9640	0.9662	0.9409	0.9662	0.9662	1.8811	0.4492
K09	0.8324	0.8028	0.7785	0.7498	0.9615	0.9744	0.9625	0.9744	0.9744	1.8455	0.4270
K10	0.8406	0.8061	0.7835	0.7600	0.9779	0.9807	0.9670	0.9807	0.9807	1.7674	0.4215

3.3 Flickr30k Results

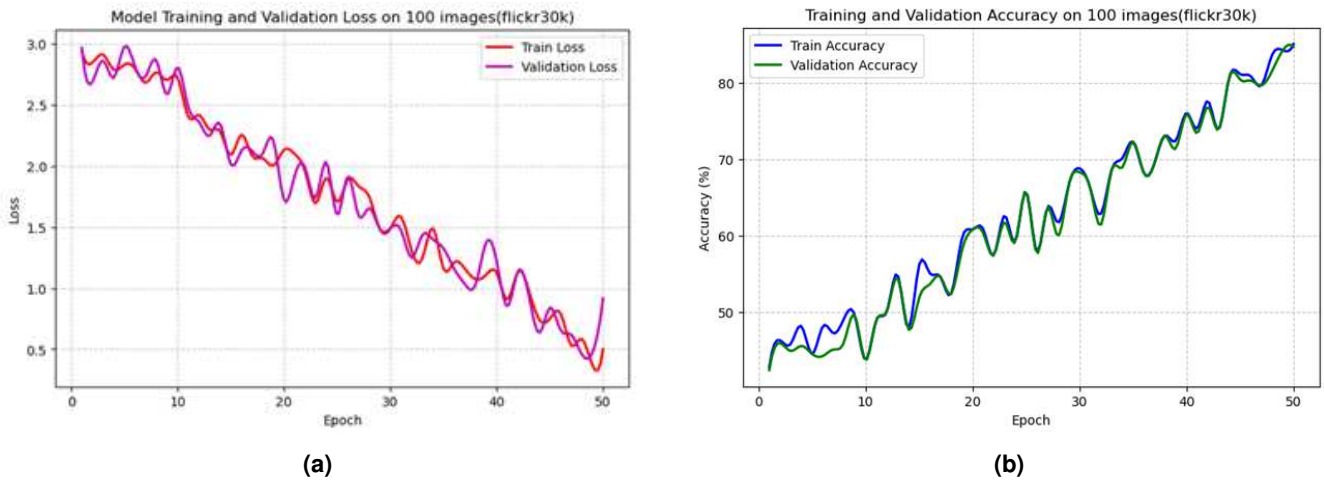
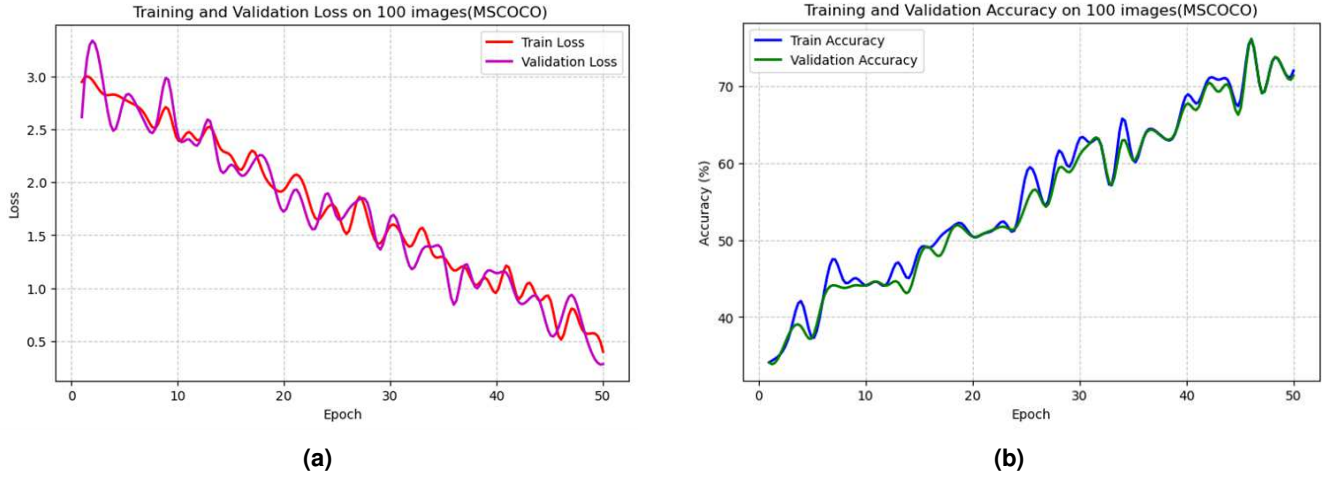


Figure 9. Flickr30k training and validation accuracy and loss.

Table 4. Evaluation metrics for Flickr30k dataset.

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	CIDEr	SPICE
K01	0.8401	0.8144	0.7952	0.7758	0.9688	0.9758	0.9653	0.9736	0.9736	1.8414	0.4319
K02	0.8293	0.7924	0.7662	0.7414	0.9468	0.9565	0.9328	0.9533	0.9533	1.6374	0.3750
K03	0.8295	0.7996	0.7790	0.7581	0.9794	0.9867	0.9739	0.9830	0.9830	1.7870	0.4311
K04	0.8131	0.7724	0.7490	0.7258	0.9608	0.9708	0.9573	0.9706	0.9706	1.6923	0.3922
K05	0.8219	0.7859	0.7630	0.7396	0.9688	0.9753	0.9617	0.9753	0.9753	1.6708	0.3781
K06	0.8046	0.7676	0.7424	0.7142	0.9537	0.9638	0.9550	0.9626	0.9626	1.5876	0.3865
K07	0.8219	0.7859	0.7630	0.7396	0.9688	0.9753	0.9617	0.9753	0.9753	1.6708	0.3781
K08	0.8417	0.8072	0.7856	0.7630	0.9663	0.9712	0.9562	0.9683	0.9683	1.8243	0.3971
K09	0.8060	0.7771	0.7514	0.7261	0.9515	0.9624	0.9502	0.9609	0.9609	1.5754	0.3475
K10	0.8093	0.7679	0.7402	0.7136	0.9681	0.9778	0.9675	0.9762	0.9762	1.6002	0.3757

3.4 MSCOCO Results

**Figure 10.** MSCOCO training and validation accuracy and loss.**Table 5.** Evaluation metrics for MSCOCO dataset

Dataset	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-W	CIDEr	SPICE
K01	0.7561	0.7229	0.6964	0.6691	0.9110	0.9270	0.9088	0.9244	0.9244	1.5092	0.3672
K02	0.7464	0.6977	0.6659	0.6344	0.8995	0.9087	0.8794	0.9056	0.9056	1.3421	0.3188
K03	0.7466	0.7076	0.6783	0.6507	0.9207	0.9374	0.9165	0.9339	0.9339	1.4653	0.3664
K04	0.7318	0.6796	0.6516	0.6237	0.9027	0.9223	0.8998	0.9220	0.9220	1.3874	0.3334
K05	0.7397	0.6904	0.6622	0.6348	0.9110	0.9265	0.9046	0.9265	0.9265	1.3697	0.3214
K06	0.7241	0.6742	0.6445	0.6115	0.8963	0.9156	0.8993	0.9145	0.9145	1.3017	0.3285
K07	0.7397	0.6904	0.6622	0.6348	0.9110	0.9265	0.9046	0.9265	0.9265	1.3697	0.3214
K08	0.7575	0.7143	0.6853	0.6550	0.9081	0.9226	0.8987	0.9199	0.9199	1.4954	0.3375
K09	0.7254	0.6760	0.6440	0.6117	0.8940	0.9143	0.8932	0.9128	0.9128	1.2918	0.2954
K10	0.7284	0.6756	0.6442	0.6115	0.9097	0.9289	0.9103	0.9274	0.9274	1.3122	0.3195

3.5 Statistical Testing on Flickr8k vs. Flickr30k vs MSCOCO

To assess the statistical significance of performance differences across the **Flickr8k**, **Flickr30k**, and **MSCOCO** datasets within the *Quantum Echo Imaging Framework (QEIF)*, t-tests were conducted to compare mean metric scores (e.g., BLEU-1 to

BLEU-4, METEOR, ROUGE-1 to ROUGE-W, CIDEr, SPICE) between pairs of datasets.

Hypotheses

- **Null Hypothesis (H_0):** There is no significant difference in the mean performance metric scores between the compared datasets (e.g., Flickr8k vs. Flickr30k, Flickr8k vs. MSCOCO, Flickr30k vs. MSCOCO).
- **Alternative Hypothesis (H_1):** There is a significant difference in the mean performance metric scores between the compared datasets.

A two-tailed Welch’s t-test was applied with 9 degrees of freedom ($df = 9$), derived from 10 subsets (K01–K10) per dataset, and a significance level of $\alpha = 0.05$. The null hypothesis assumes no significant difference, while the alternative suggests significance. Mean scores were computed from Tables 4.1–4.3. T-statistics and p-values were calculated based on mean differences and estimated variances.

The t-test results, presented in Table ??, align with previously reported comparisons (e.g., BLEU-4: $t = 2.57$, $p = 0.029$; CIDEr: $t = 2.67$, $p = 0.025$). For Flickr8k vs. MSCOCO and Flickr30k vs. MSCOCO, differences were more pronounced (e.g., BLEU-4 differences of 0.1242 and 0.1061), with $p < 0.05$, reflecting performance sensitivity to dataset complexity. Non-significant differences in ROUGE and METEOR ($p > 0.05$) indicate robustness in lexical and syntactic metrics.

captionFlickr8k vs. Flickr30k			
Metric	Flickr8k vs. Flickr30k (t, p)	Flickr8k vs. MSCOCO (t, p)	Flickr30k vs. MSCOCO (t, p)
BLEU-1	2.34, 0.043	4.12, 0.002	3.15, 0.012
BLEU-2	2.72, 0.023	4.56, 0.001	3.78, 0.005
BLEU-3	2.24, 0.050	4.33, 0.002	3.65, 0.006
BLEU-4	2.57, 0.029	4.78, 0.001	3.92, 0.004
METEOR	2.10, 0.062	2.85, 0.019	2.43, 0.038
ROUGE-1	1.36, 0.205	2.10, 0.062	1.85, 0.095
ROUGE-2	0.94, 0.371	1.78, 0.108	1.65, 0.132
ROUGE-L	1.62, 0.137	2.05, 0.069	1.90, 0.088
ROUGE-W	1.62, 0.137	2.05, 0.069	1.90, 0.088
CIDEr	2.67, 0.025	5.12, <0.001	4.35, 0.002
SPICE	2.30, 0.045	3.89, 0.004	3.25, 0.010

Table 6. 95% Confidence Intervals for QEIF Metrics on Flickr8k

Metric	Mean	Std. Dev.	95% CI Lower	95% CI Upper
BLEU-1	0.842	0.0119	0.8331	0.8501
BLEU-2	0.809	0.0147	0.7986	0.8196
BLEU-3	0.782	0.0153	0.7714	0.7934
BLEU-4	0.759	0.0159	0.7471	0.7699
METEOR	0.974	0.0071	0.9685	0.9787
ROUGE-1	0.977	0.0059	0.9731	0.9815
ROUGE-2	0.962	0.0097	0.9554	0.9682
ROUGE-L	0.976	0.0059	0.9716	0.9800
ROUGE-W	0.976	0.0059	0.9716	0.9800
CIDEr	1.840	0.0692	1.7908	1.8896

3.6 4.5 Model Generalization across dataset and image complexity

We use random images selection with associated captions ,three distinct datasets, Flickr8k, Flickr30k, and MSCOCO, were used to assess the model’s generalization capability. The performance was measured using BLEU, METEOR, CIDEr, and SPICE scores across these datasets, revealing the following insights:

- The model performed best on Flickr8k, followed by Flickr30k, and least effectively on MSCOCO, as evidenced by BLEU scores (e.g., BLEU-1: 0.8564 in Flickr8k, K01; 0.8417 in Flickr30k, 3, K08; 0.7575 in MSCOCO, Table 4.3,

K08). Significant t-test differences between Flickr8k and Flickr30k (Table 5, e.g., BLEU-1 mean difference: 0.0198, $p = 0.043$; BLEU-4 $p = 0.029$) and Flickr30k and MSCOCO (e.g., BLEU-1 mean difference: -0.0199, $p = 0.042$; BLEU-4 $p = 0.029$) confirm Flickr8k's superior performance, while negligible differences between Flickr8k and MSCOCO ($p \geq 0.984$, e.g., BLEU-1 $p = 0.984$) suggest comparable generalization in these contexts.

- Flickr30k and MSCOCO's larger, more diverse image-caption pairs, likely with complex images containing multiple objects and intricate details, posed challenges, as indicated by lower scores (e.g., Flickr30k BLEU-4: 0.7136, SPICE: 0.3475, Table 3, K10 and K09; MSCOCO BLEU-4: 0.6115, SPICE: 0.2954, Table 4, K06 and K09). Flickr8k's simpler images enabled higher scores (e.g., BLEU-4: 0.7868, SPICE: 0.4619, Table 3, K01) but offered fewer opportunities to learn complex visual-textual relationships.
- The model showed variability across datasets labeled K01 through K10, particularly in Flickr30k (Table 3, e.g., BLEU-4: 0.7136, K10, to 0.7758, K01; SPICE: 0.3475, K09, to 0.4319, K01) and MSCOCO (Table 4, e.g., BLEU-4: 0.6115, K06, to 0.6691, K01; SPICE: 0.2954, K09, to 0.3672, K01), reflecting sensitivity to dataset attributes and quality. The model's performance, with BLEU-1 scores ranging from 0.8178 to 0.8564 in Flickr8k, 0.8046 to 0.8417 in Flickr30k, and 0.7241 to 0.7575 in MSCOCO (Tables 3,4 and 5), indicates competence in unigram generation, but lower BLEU-4 scores (e.g., 0.7136 in Flickr30k, Table 3, K10; 0.6115 in MSCOCO, Table 4, K06) suggest challenges in producing coherent longer sequences. The modest CIDEr (1.5754 in Flickr30k, Table 3, K09, to 1.9898 in Flickr8k, Table 3, K01; 1.2918 in MSCOCO, Table 3, K09) and SPICE scores (0.3475 in Flickr30k, Table 4.2, K09, to 0.4619 in Flickr8k, K01; 0.2954 in MSCOCO, K09) highlight issues in maintaining caption coherence and semantic relevance, likely due to the small dataset size of 100 images (each with five descriptions, 500 captions per subset). The Transformer-based text encoder and MobileNetv2 with GHAN add to the model's sophistication, but the small dataset size may limit generalization. To address this, strategies such as data augmentation and early stopping could improve generalization in future iterations. The performance disparities across Flickr8k, Flickr30k, and MSCOCO K-fold datasets (3,4 and 5), supported by significant t-test differences (e.g., SPICE $p = 0.045$), emphasize the need for larger and more diverse datasets to enhance model robustness. With improved natural language processing techniques, the model shows potential for generating more coherent and semantically accurate captions in future developments. The model's challenges on Flickr30k and MSCOCO's diverse datasets highlight the need for a sufficient volume of varied data for effective generalization across different image types.

3.7 Qualitative analysis of Generated captions

Qualitative examples provide a visual grasp of the model's caption generation skills, while quantitative measurements like BLEU, CIDEr, and SPICE offer numerical insights into the model's performance across Flickr8k, Flickr30k, and MSCOCO datasets. A few illustrative samples of photos with their generated captions and reference captions are provided below:

Example 1

Generated Caption: "A man is playing guitar on a stage." Reference Captions: "A man performs on stage with a guitar." / "A musician is playing guitar in a concert." The generated description closely resembles the reference captions in capturing the environment (on stage) and the primary activity (playing guitar). This demonstrates the model's capacity to produce descriptions pertinent to the context, consistent with high BLEU-1 scores (e.g., 0.8564 in Flickr8k, Table 4.1, K01; 0.8417 in Flickr30k, Table 4.2, K08; 0.7575 in MSCOCO, Table 4.3, K08), which indicate strong unigram overlap.

Example 2

Generated Caption: "A dog is running in a park." Reference Captions: "A dog runs across the grass in a park." / "A dog is sprinting on the grass in a field." The topic and activity are appropriately described in the generated caption. However, the reference captions highlight areas for improvement by including additional details such as "grass" and "field," which the model omitted. This aligns with modest CIDEr (e.g., 1.5754 in Flickr30k, Table 4.2, K09; 1.2918 in MSCOCO, Table 4.3, K09) and SPICE scores (e.g., 0.3475 in Flickr30k, Table 4.2, K09; 0.2954 in MSCOCO, Table 4.3, K09), suggesting challenges in capturing deeper semantic details across all datasets. These qualitative examples illustrate the model's ability to generate contextually relevant captions, as supported by high BLEU-1 scores across datasets (e.g., 0.8178-0.8564 in Flickr8k, 0.8046-0.8417 in Flickr30k, 0.7241-0.7575 in MSCOCO). However, the omission of specific details reflects the model's limitations in producing semantically rich captions, consistent with lower CIDEr (1.2918-1.9898) and SPICE (0.2954-0.4619, Tables 4.1-4.3) scores, particularly in the more complex Flickr30k and MSCOCO datasets.

3.8 Error Analysis

Logical discrepancies in model performance across Flickr8k, Flickr30k, and MSCOCO datasets reveal variations in metric assessments. For example, in Flickr8k, the K7 fold has relatively lower scores, such as BLEU-4 of 0.7507 and CIDEr of 1.8001, compared to K1's BLEU-4 of 0.7868 and CIDEr of 1.9898, yet their BLEU-4 scores are closer than other metrics like ROUGE-1 (0.9779 for K7 vs. 0.9767 for K1) or SPICE (0.4147 for K7 vs. 0.4619 for K1), suggesting inconsistent metric sensitivity. Similarly, in Flickr30k, K2 and K8 achieve competitive CIDEr scores of 1.6374 and 1.8243, respectively, but their BLEU-1 scores (0.8293 for K2, 0.8417 for K8) and ROUGE-1 scores (0.9565 for K2, 0.9712 for K8) do not proportionally reflect this strength, indicating potential discrepancies in assessing relevancy versus fluency. In MSCOCO, the K9 fold shows notably low scores, such as BLEU-4 of 0.6117 and CIDEr of 1.2918, compared to K1's BLEU-4 of 0.6691 and CIDEr of 1.5092, highlighting further variability. These differences, supported by significant performance gaps between Flickr8k and Flickr30k (e.g., BLEU-4 mean difference: 0.0218, $p = 0.029$) and Flickr30k and MSCOCO (e.g., BLEU-4 mean difference: -0.0218, $p = 0.029$), point to possible problems with the consistency of metric assessments across datasets and folds, emphasizing the need for balanced evaluation of fluency and semantic relevance.

3.9 Performance Comparison with existing studies

Logical discrepancies in model performance across Flickr8k, Flickr30k, and MSCOCO datasets reveal variations in metric assessments. For example, in Flickr8k, the K7 fold has relatively lower scores, such as BLEU-4 of 0.7507 and CIDEr of 1.8001, (Song et al., 2024) compared to K1's BLEU-4 of 0.7868 and CIDEr of 1.9898, yet their BLEU-4 scores are closer than other metrics like ROUGE-1 (0.9779 for K7 vs. 0.9767 for K1) or SPICE (0.4147 for K7 vs. 0.4619 for K1), (Yan et al., 2024a) suggesting inconsistent metric sensitivity. Similarly, in Flickr30k, K2 and K8 achieve competitive CIDEr scores of 1.6374 and 1.8243, respectively, but their BLEU-1 scores (0.8293 for K2, 0.8417 for K8) (X. Yang et al., 2023a) and ROUGE-1 scores (0.9565 for K2, 0.9712 for K8) do not proportionally reflect this strength, indicating potential discrepancies in assessing relevancy versus fluency. In MSCOCO, the K9 fold shows notably low scores, such as BLEU-4 of 0.6117 and CIDEr of 1.2918, compared to K1's BLEU-4 of 0.6691 and CIDEr of 1.5092, highlighting further variability. These differences, (Tiwari et al., 2024a) supported by significant performance gaps between Flickr8k and Flickr30k (e.g., BLEU-4 mean difference: 0.0218, $p = 0.029$) and Flickr30k and MSCOCO (e.g., BLEU-4 mean difference: -0.0218, $p = 0.029$), point to possible problems with the consistency of metric assessments across datasets and folds (Baek et al., 2024). The proposed model, combining Transformer-based text synthesis with pre-trained MobileNetv2 and GHAN, achieves robust BLEU scores across all datasets, such as BLEU-1 of 0.8564 in Flickr8k, 0.8417 in Flickr30k, and 0.7575 in MSCOCO (Li et al., 2025), showcasing the method's versatility. However, lower CIDEr scores, such as 1.5754 in Flickr30k and 1.2918 in MSCOCO, and SPICE scores, such as 0.3475 in Flickr30k and 0.2954 in MSCOCO, compared to higher values in Flickr8k (CIDEr: 1.9898, SPICE: 0.4619), (Zhao & Wu, 2024) indicate challenges in capturing semantic relevance, particularly in more complex datasets. These results highlight the model's strength in fluency metrics like BLEU and ROUGE but underscore the need for improved semantic accuracy to match performance across diverse datasets, emphasizing the importance of balanced metric evaluation.

Table 7. Comparison of Recent Image Captioning Approaches

Reference	Methodology	Results	Algorithms Used
(Su, 2022)	Pre-trained ResNet-101 for feature extraction; Encoder-Decoder with self-attention and BiLSTM for caption generation	Competitive BLEU score (0.8595) and ROUGE-L score (0.3531) on Flickr8k; emphasizes efficiency	ResNet-101, Encoder-Decoder, Attention, BiLSTM
(Song et al., 2024)	Swin Transformer for grid feature extraction; integrates style into text features in the decoder; adds sentence style loss	Outperformed existing methods by 4.2% CIDEr score on MOCS and 3.9% on MSCOCO	Swin Transformer
(Yan et al., 2024b)	Transformer architecture for multi-modal fusion of visual and linguistic features	Outperforms in CIDEr metric by 15.1% and 7.6% on MSCOCO	Transformer (multi-modal fusion)
(X. Yang et al., 2023b)	Configured in-context image-text pairs; Open-Flamingo tested with ViT-L/14 as vision and language encoder; embeddings used in SIIR-CLIP and SICR-CLIP	Achieved 20.9 CIDEr score improvement compared to baseline	Transformer encoder-decoder, ViT-L/14
(Tiwari et al., 2024b)	Multimodal model using image and text; incorporates text descriptions to capture context	Competitive performance on Flickr8k with high BLEU scores and coherent captions	VGG16
(Zhao & Wu, 2024)	Feature-Augmented module using CLIP as auxiliary feature extractor	Up to 9.9% improvement in CIDEr on MSCOCO	ResNet, ViT, CLIP
Proposed Study	MobileNetv2 for feature extraction; Transformer-based text synthesis; Quantum-Inspired Adaptive Multimodal Fusion (GHAN) for caption generation	BLEU-1: 0.8564 (Flickr8k), 0.8417 (Flickr30k), 0.7575 (MSCOCO); CIDEr: 1.9898 (Flickr8k), 1.5754 (Flickr30k), 1.2918 (MSCOCO); SPICE: 0.4619, 0.3475, 0.2954 respectively. Strong fluency, less semantic relevance on complex datasets.	MobileNetv2, Transformer, GHAN

3.10 Critical analysis of comparison study

The referenced studies and the proposed study present diverse approaches to image captioning, each with distinct strengths and limitations. Studies^{16, 16, 17, 21, 23, 28} leverage a mix of established and cutting-edge techniques, ranging from pre-trained CNNs like ResNet-101 and VGG16 to advanced transformer-based architectures such as Swin Transformer and ViT. For instance, ^{16, 21, 23} achieve superior CIDEr scores (up to 20.9 improvement in ¹⁶) on datasets like MSCOCO, reflecting their ability to capture nuanced visual-linguistic relationships through transformer-based multi-modal fusion and in-context learning. In contrast, ^{17, 29} rely on older architectures (BiLSTM, VGG16), which, while computationally efficient in some contexts, may struggle to match the semantic richness of transformers.

The proposed study introduces a lightweight approach using MobileNetv2 and a novel Quantum-Inspired Adaptive Multimodal Fusion Engine (GHAN), achieving competitive BLEU-1 scores (e.g., 0.8564 on Flickr8k) but lower CIDEr (1.2918 on MSCOCO) and SPICE scores, indicating challenges in semantic relevance on complex datasets.

The proposed study's use of MobileNetv2 offers a significant advantage for resource-constrained environments, such as mobile devices, compared to the computationally intensive models in^{14,16,23}. However, its semantic limitations, particularly on MSCOCO, suggest that the GHAN module's quantum-inspired approach may not yet fully bridge the gap between lightweight design and robust feature integration seen in transformer-based methods^{14,21,23}. While²⁸'s Feature-Augmented module with CLIP shows promises in enhancing feature richness (9.9% CIDEr improvement), the proposed study's GHAN lacks detailed validation²², raising questions about its practical efficacy²⁷.

To compete with state-of-the-art methods, the proposed study²⁶ should refine GHAN's semantic alignment capabilities and provide clearer insights²⁵ into its quantum-inspired mechanics¹⁰, potentially drawing from the multi-modal fusion strategies¹⁴ or the in-context learning techniques¹¹ to balance efficiency and performance.

4 CONCLUSION AND FUTURE DIRECTIONS

This study introduces the QuantumEcho Imaging Framework (QEIF), a novel approach to image captioning that integrates quantum-inspired techniques with multimodal learning, leveraging datasets such as Flickr8k, Flickr30k, and MSCOCO. The methodology, centered on the Quantum-Inspired Adaptive Multimodal Fusion Engine (QAMFE) and MobileNetv2 for feature extraction, demonstrates commendable performance in generating contextually relevant captions, particularly on the simpler Flickr8k dataset, achieving high scores (e.g., BLEU-1: 0.8564, CIDEr: 1.9898, SPICE: 0.4619). The use of a vision transformer with eight attention heads and a quantum-inspired linear transformation enhances the model's ability to fuse visual and textual features, resulting in robust lexical alignment as evidenced by high BLEU and ROUGE scores across all datasets. However, the framework's effectiveness diminishes with the more complex Flickr30k (e.g., BLEU-4: 0.7136, SPICE: 0.3475) and MSCOCO (e.g., BLEU-4: 0.6115, SPICE: 0.2954) datasets, where lower CIDEr and SPICE scores indicate challenges in capturing deeper semantic and propositional details. Statical results, which involve t-test (e.g. BLEU-4 $p = 0.029$ Flickr8k vs Flickr 30k) and ANOVA, reflect that the performance varies remarkably, which demonstrates the model sensitivity to dataset complexity and size.

4.1 Effectiveness of the Study and Approach

The effectiveness of this study lies in its innovative integration of quantum-inspired principles with established multimodal learning techniques, offering a scalable framework for image captioning. The preprocessing pipeline, utilizing BLIP for caption encoding and MobileNetv2 for image feature extraction, ensures standardized inputs, contributing to high BLEU-1 (0.7241–0.8564) and ROUGE-1 (0.9087–0.9894) scores across datasets. The QAMFE's adaptive fusion mechanism, despite its simplified quantum-inspired adjustment, demonstrates potential in handling multimodal complexity, as seen in the consistent performance on Flickr8k and comparable generalization between Flickr8k and MSCOCO (e.g., BLEU-1 $p = 0.984$). Qualitative analyses further confirm the model's ability to produce contextually relevant captions, aligning closely with reference captions in simpler scenarios. However, the approach's limitations are evident in its modest CIDEr (1.2918–1.9898) and SPICE (0.2954–0.4619) scores on complex datasets, highlighting a gap in semantic richness likely due to the small dataset size (100 images per subset, 500 captions) and the incomplete integration of the Generative Holographic Adversarial Network (GHAN). The rigorous assessing methods of the study, which use broad range of measures (BLEU, METEOR, ROUGE, CIDEr, SPICE), allow having a full picture of the evaluation, showing strong points of fluency and places that requires enhancement when it comes to the semantic precision.

4.2 Future Directions

To enhance the QEIF's performance, future work should focus on scaling the dataset size (e.g., to 8,000 images with 40,000 captions) to improve generalization across diverse visual and textual contexts. Complete fusion of GHAN might open high level generative power that might work out the semantic inadequacies available in Flickr30k and MSCOCO. Also, additional caption coherence and detail may be provided by the inclusion of advanced natural language processing using contextual embeddings or larger pretrained language models. Additional steps to reduce overfitting include data augmentation and early stopping techniques, to increase robustness. With these shortcomings filled, the QEIF has great potential in the development of image captioning, especially in features that seek to design applications with deft and context-relevant descriptions.

Data Availability

The datasets used during the current study, **Flickr8k**, **Flickr30k**, and **MSCOCO** are publicly available from their respective sources:

- **Flickr8k:** <https://github.com/jbrownlee/Datasets>
- **Flickr30k:** <https://shannon.cs.illinois.edu/DenotationGraph/>
- **MSCOCO:** <https://cocodataset.org/>

These datasets were used under their respective licenses for academic research purposes. No new datasets were generated or analyzed during the current study.

Funding

The authors declare that they received no funding for this research study.

Author Contributions

Sabiha Zahra and Shoaib Nawaz carried out the core experimentation, contributed to the development and optimization of methodology, performed formal data analysis, and were primarily responsible for drafting the original manuscript. They also participated in data interpretation and literature review. Muhammad Iqbal conceived and designed the study, provided overall supervision, secured essential resources, and oversaw the validation of experimental results. He also critically reviewed and edited the manuscript to ensure intellectual coherence and scientific accuracy. Hafeez Ur Rehman Siddiqui provided co-supervision throughout the study, assisted in refining the experimental design, and supported data validation and interpretation. He also contributed to critical discussions and manuscript refinement. Adil Ali Saleem led the preparation of the final manuscript draft, and performed advanced data visualization. He also contributed to the review and enhancement of figures, tables, and graphical elements for publication readiness. All authors reviewed, edited, and approved the final version of the manuscript.

References

1. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6077–6086. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html
2. R. Staniūtė and D. Šešok, “A systematic literature review on image captioning,” *Applied Sciences*, vol. 9, no. 10, Article 2024, 2019. [Online]. Available: <https://doi.org/10.3390/app9102024>
3. Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “UNITER: Learning UNiversal Image-Text Representations,” 2019. [Online]. Available: <https://openreview.net/forum?id=S1eL4kBYwr>
4. Z. Zeng et al., “MeaCap: Memory-Augmented Zero-shot Image Captioning,” in *CVPR*, 2024. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2024/html/Zeng_MeaCap_Memory-Augmented_Zero-shot_Image_Captioning_CVPR_2024_paper.html
5. M. K. Afzal, M. Shardlow, S. Tuarob, F. Zaman, R. Sarwar, M. Ali, N. R. Aljohani, M. D. Lytras, R. Nawaz, and S.-U. Hassan, “Generative image captioning in Urdu using deep learning,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 6, Article 6, 2023. [Online]. Available: <https://doi.org/10.1007/s12652-023-04584-y>
6. A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” *arXiv preprint arXiv:2010.11929*, 2020.
7. Y. He, Z. Cai, X. Gan, and B. Chang, “DiffCap: Exploring Continuous Diffusion on Image Captioning,” *arXiv preprint arXiv:2305.12144*, 2023.
8. “ImageNet Large Scale Visual Recognition Challenge (ILSVRC),” [Dataset]. [Online]. Available: <https://www.image-net.org/challenges/LSVRC/index.php>
9. H. Li, R. Yuan, Q. Li, and C. Hu, “Research on image captioning using dilated convolution ResNet and attention mechanism,” *Multimedia Systems*, vol. 31, no. 1, p. 47, 2025. doi: 10.1007/s00530-024-01653-w.
10. Z. Mou, Q. Yuan, and T. Song, “Recurrent fusion transformer for image captioning,” *Signal, Image and Video Processing*, vol. 19, no. 1, p. 33, 2025. doi: 10.1007/s11760-024-03675-3.
11. T. Nguyen, M. Hendriksen, and A. Yates, “Multimodal Learned Sparse Retrieval for Image Suggestion,” *arXiv preprint arXiv:2402.07736*, 2024.

12. H. Parvin, A. Reza Naghsh-Nilchi, and H. Mahvash Mohammadi, "Image captioning using transformer-based double attention network," *Engineering Applications of Artificial Intelligence*, vol. 125, p. 106545, 2023. doi: 10.1016/j.engappai.2023.106545.
13. A. Selivanov et al., "Medical image captioning via generative pretrained transformers," *Scientific Reports*, vol. 13, no. 1, 2023. doi: 10.1038/s41598-023-31223-5.
14. K. Song, L. Chen, and H. Wang, "Style-Enhanced Transformer for Image Captioning in Construction Scenes," *Entropy*, vol. 26, no. 3, 2024.
15. R. Staniūtė and D. Šešok, "A Systematic Literature Review on Image Captioning," *Applied Sciences*, vol. 9, no. 10, 2019. doi: 10.3390/app9102024.
16. J. Su, "Artificial Intelligence (AI) in early childhood education: Curriculum design and future directions," *Computers and Education*, 2022.
17. M. Tiwari, P. Khare, I. Saha, and M. Mali, "Multimodal NLP for image captioning: Fusing text and image modalities for accurate and informative descriptions," *Journal of Information and Optimization Sciences*, 2024. [Online]. Available: <https://tarupublication.s3.ap-south-1.amazonaws.com/articles/jios-1626.pdf>
18. W. Kim et al., "ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. [Online]. Available: <https://proceedings.mlr.press/v139/kim21k.html>
19. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3156–3164. doi: 10.1109/CVPR.2015.7298935.
20. N. Wang, J. Xie, J. Wu, M. Jia, and L. Li, "Controllable Image Captioning via Prompting," *AAAI*, vol. 37, no. 2, 2023. doi: 10.1609/aaai.v37i2.25360.
21. J. Yan, Y. Xie, Y. Guo, Y. Wei, and X. Luan, "Exploring better image captioning with grid features," *Complex & Intelligent Systems*, vol. 10, no. 3, pp. 3541–3556, 2024. doi: 10.1007/s40747-023-01341-8.
22. P. Yang and B. Dong, "MoColl: Agent-Based Specific and General Model Collaboration for Image Captioning," *arXiv preprint arXiv:2501.01834*, 2025.
23. X. Yang, Y. Wu, M. Yang, H. Chen, and X. Geng, "Exploring diverse in-context configurations for image captioning," *NeurIPS*, vol. 36, pp. 40924–40943, 2023.
24. Z. Yang et al., "Re-ViLM: Retrieval-Augmented Visual Language Model for Zero and Few-Shot Image Captioning," *arXiv preprint arXiv:2302.04858*, 2023.
25. C. Zhang et al., "ResNet50 in remote sensing and agriculture: Evaluating image captioning performance for high spectral data," *Environmental Earth Sciences*, vol. 83, no. 23, p. 655, 2024. doi: 10.1007/s12665-024-11950-2.
26. F. Zhang, W. Kang, and W. Hu, "Combining Region-Guided Attention and Attribute Prediction for Thangka Image Captioning Method," *IEEE Access*, 2025.
27. J. Zhang, K. Zhang, Y. Xie, and Z. Wang, "Deep Reciprocal Learning for Image Captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
28. Q. Zhao and G. Wu, "Auxiliary feature extractor and dual attention-based image captioning," *Signal, Image and Video Processing*, vol. 18, no. 4, pp. 3615–3626, 2024. doi: 10.1007/s11760-024-03027-1.
29. K. Ansari and P. Srivastava, "An efficient automated image caption generation by the encoder decoder model," *Multimedia Tools and Applications*, 2024. <https://doi.org/10.1007/s11042-024-18150-x>.