International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 17 (2025) pp. 730-745

https://doi.org/10.70917/ijcisim-2025-0044
Article

INSIPR: Image-Text Synthesis Pipeline for Intelligent Retrieval and Generation

Viomesh Singh*, Prithviraj Jadhav, Hritesh Maikap, Sarvesh Jadhav, Chinmay Ingale and Sahil Jadhav

Vishwakarma Institute of Technology; Pune- 411037, India; viomesh.singh@vit.edu (V.S.); prithviraj.jadhav22@vit.edu (P.J.); hritesh.maikap22@vit.edu (H.M.); sarvesh.jadhav22@vit.edu (S.J.); chinmay.ingale22@vit.edu (C.I.); sahil.jadhav221 @vit.edu (S.J.)
* Correspondence author: viomesh.singh@vit.edu

Received date: 13 April 2025; Accepted date: 10 October 2025; Published online: 31 December 2025

Abstract: The INSPIR (Image-Text Synthesis Pipeline for Intelligent Retrieval and Generation) framework introduces an innovative approach for image captioning and image retrieval by leveraging an ensemble of state-of-the-art models. This research proposes a method that generates descriptive captions from images using an ensemble of BLIP (Bootstrapping Language-Image Pre-training), ViT-GPT2 (Vision Transformer combined with GPT-2), and GIT (Generative Image Text) and employing CLIP (Contrastive Language-Image Pre-training) for ranking the generated captions based on their relevance. The impact of temperature scaling and ensemble weights on the generated caption ranking was analyzed to evaluate the system, revealing insights regarding the balance of relevance and diversity. Testing on the Flickr8k dataset demonstrated the model's effectiveness, achieving cosine similarity, BLEU, and METEOR scores on randomly selected photos. The top-ranked captions are utilized by Llama3.1 to produce creative outputs tailored for various applications, including social media captions and image notes. By integrating multiple modalities within a unified semantic space through contrastive learning, this work aims to advance the field of image captioning beyond conventional classification tasks, offering a generalized model performance that addresses the complexities of language and vision. One of the major applications of the INSPIR model is image retrieval, where the system enhances capabilities by annotating uploaded images and enabling users to conduct text-based searches, facilitating efficient access to relevant visual content.

Keywords: multimodal; BLIP; CLIP; image captioning; retrieval; contrastive learning

1. Introduction

Generating the most lucid and contextually meaningful sentences from an image involves the intersection of computer vision and natural language processing. This task, broadly known as image captioning, describes the translation of visual content to a descriptive text, which significantly enhances accessibility and user engagement across various applications, including social media, digital asset management, and assistive technologies for visually impaired individuals. Despite significant progress, challenges persist in this field due to the inherent complexities of visual semantics, variability in visual content, ambiguity of natural language and the necessity for models to generalize across diverse datasets while maintaining the high precision and contextual relevance of captions generated from the images.

Recent advancements in neural architectures and pretraining paradigms have seen significant progress in this image captioning domain, with state-of-the-art models demonstrating impressive capabilities. For instance, "Show and Tell: A Neural Image Caption Generator" introduced a progressive approach that utilized convolutional neural networks (CNNs) in combination with recurrent neural networks (RNNs) to generate captions from images effectively [1]. Building on this foundation, researchers have explored the integration of CNNs with Long Short-Term Memory (LSTM) Networks, which enhances the capability of the model to capture the temporal dependencies in sequential data [2]. The combination CNN-LSTMs architecture significantly improved the image caption generation

capability by leveraging the strengths of both models. Moreover, researchers found that incorporating self-attention mechanisms into CNN-LSTM architectures allowed for focusing on relevant parts of the image and thereby, improving the contextual understanding [3]. However, these models often struggle with overfitting and may not generalise on diverse datasets, which limits their applications.

Notably, the introduction of transformer architectures, as highlighted in "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," has further revolutionised image processing by allowing models to capture long-range dependencies in visual data [4]. Additionally, the Contrastive Language-Image Pre-training (CLIP) model has redefined image-text matching by learning to align images and text in a shared semantic space, thus enabling more accurate retrieval and captioning tasks [5]. The recent work "ClipCap: CLIP Prefix for Image Captioning" has also shown promise by enhancing caption generation through prefix tuning with CLIP [6]. Models such as Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation (BLIP) have emerged as powerful tools for image captioning, utilizing a combination of vision and language pre-training to enhance caption generation capabilities [7].

Building upon this research, prefix-tuning and prompting strategies have come up as lightweight and efficient alternatives for fine-tuning traditional image captioning techniques. Prefix-tuning optimizes small, continuous task-specific vectors while keeping the pre-trained model frozen, allowing competitive performance with fewer parameters [8]. This approach can be effective to utilize with models like CLIP, which can provide richer and semantic features for visual data. Furthermore, prompt-based learning highlights the potential of pre-trained models for few-shot or zero-shot learning capabilities [9]. This capability can be crucial where multimodal data and coherent text are integrated for text generation, which can significantly improve the quality and relevance of generated captions.

In this paper, we present INSPIR (Image-Text Synthesis Pipeline for Intelligent Retrieval and Generation), a next-generation framework that intends to build on traditional state-of-the-art methodologies by incorporating advanced ensemble modelling, adaptive scoring techniques and context-aware synthesis into an advanced pipeline. Our approach takes into account an ensemble of state-of-the-art image captioning models like BLIP, Generative Image-to-text Transformer (GIT), and VIT-GPT2 to generate diverse and contextually rich captions. These captions are then evaluated and ranked through a novel entropy-based dynamic temperature scaling mechanism combined with multi-view scoring, taking into account factors such as relevance, diversity, and model-specific confidence. Furthermore, to introduce context-awareness prompting with Large Language Model (LLM) to refine the captions to application-specific output or to generate creative captions to adjust to the needs of the user.

The proposed method works systematically: images are processed through ensemble models to generate caption text. Each caption is evaluated using CLIP's comparative learning engine with our enhanced ranking system. It combines dynamic temperature measurement based on entropy measurement and an adaptive combination of multiple scoring criteria. This thorough evaluation facilitates the selection of the most appropriate name and increases retrieval ability. It allows users to upload images and perform text-based searches against a database of annotated images. Finally, context-aware LLM integration enables style-specific prompting and synthesis of captions that align with user expectations. The result of the system is capable of generating captions that are not only semantically accurate but also adjust to the needs of specific applications, such as social media content or semantic image retrieval. By introducing this framework, INSPIR targets the limitations of existing systems, including a lack of diversity and semantics in captions, contextual irrelevance and limited creative adaptability.

The following sections explain the works related to image captioning, the detailed architecture and working of the model, with detailed discussions and effectiveness.

2. Related Work

This section reviews prior work on image captioning, grouped by surveys and trends, large-scale pretraining/zero-shot methods, encoder–decoder and attention models, Transformer-based captioners, structured/graph methods, and memory/long-context mechanisms. As an advanced review paper on captioning techniques highlights [10], the field sits at the intersection of computer vision and NLP, which demonstrates tremendous advances in the field by deep learning models in the generation of descriptive text. They establish that several issues are present here, but it also raises major concerns, as capturing how objects of an image connect is essential and required. Their paper includes all sorts of various deep learning models, datasets and evaluation metrics where they observe this gap of generalized knowledge of doing things differently than across various other domains.

Motivated by these survey findings (gaps in relation modelling and reliance on annotations), recent work has turned to large-scale pretraining; for example, Ref. [11] proposed a strong base model for image

tagging called the Recognize Anything Model, which shows its zero-shot capability to recognize any common category more accurately. Removing the reliance on manual annotations from the system is achieved through authors' utilization of large-scale and annotation-free images-text pairs, along with using very popular open-source datasets as their source for pretraining. It also develops automatic text semantic parsing and model training, unifying captioning and tagging tasks along with a data engine that refines annotations. In a word, summarizing all the above facts, RAM works much better as compared to state-of-the-art models CLIP and BLIP for multi-label classification, detection, and segmentation.

Meanwhile, extended surveys on the improvement of the caption generation on images are done in [12]. It divides the existing approaches into two broad categories: handcrafted features combined with statistical models and deep learning models based on neural networks. Handcrafted approaches rely on object detection and statistical language models but do not scale well and fail to model complex semantics. Unlike such models, the better ability of deep learning models and convolutional neural networks (CNNs) and recurrent neural networks (RNNs), respectively, makes them more effective for learning image representations as well as creating more contextually pertinent captions. Some focus here goes on attention mechanisms that enable a model to concentrate selectively on key regions of interest while generating the caption.

The paper also points out several directions towards future research, such as generation of diverse descriptions, multiple languages, improvement in the evaluation metric and efficiency improvement to address real-time applications.

Further Extension of these techniques [13] introduces MSCI, which is a novel approach to the image captioning technique. It translates the features extracted from an image into words and typically gives a suboptimal result. Such limitations are addressed in this research by integrating various levels of semantic information within an MSCI network. The proposed architecture uses a Gated Graph Neural Network (GGNN) to refine and propagate features across different semantic layers, supporting accurate representations of object relationships and overall scene context. An attention mechanism subsequently selects relevant information to enhance caption quality. It is shown through experimental analysis of captioning tasks on images in [14] that CLIP features provide a transformational ability. Multi-modal, large-scale training equips this feature of CLIP to transform the capability to enhance model performance much better than what can be provided by visual encoders.

The authors also suggest modification and simplification of the structure of the network for the perception of more objects and fusion of more sensor data for information retrieval. The study finds that even the simplest Transformer-based captioners operating on CLIP features can outperform much more complex captioners trained on reduced datasets. Over and above benchmarking improvements on caption quality as well as zero-shot transfer, this paper places CLIP as a robust benchmark for reporting image captioning performance. As having provided effective baselines for follow-up research work, it may be quite evident that one needs large-scale, multi-modal features so that meaningful textual description generation can take place from inputs in the form of images.

Further supporting the deep learning methodology, [15] presents an image model combined with NLP using a Convolutional Neural Network for feature extraction in images as well as an RNN designed for captioning. Tested by the Flickr8K dataset, the model got competitive BLEU scores that the enlargement of a training dataset brings down the mistake of captioning and general performances.

Continuing with this trend, Ref. [16] describes how image captioning has evolved from retrieval- and template-based approaches to sophisticated neural network-based models. Retrieval-based methods rely on matching an input image with similar images in a database; these often fail for novel images. Template-based methods fill pre-defined sentence structures with detected elements of an image but produce very limited sentence diversity. Such developments in CNN-RNN encoder-decoder frameworks such as NIC and m-RNN, which map directly from images to text, set a new paradigm. However, RNN is sequential in processing data, which makes the training slower. In response to this issue, Transformer models have exploited the self-attention approach to process the data in parallel, henceforth increasing the speed and efficiency. Soft-attention, as proposed by Xu et al., dynamically highlights salient image regions, whereas adaptive attention, as proposed by Lu et al., selectively uses image features only when needed. Building on these ideas, the Adaptive-Trans model combines spatial and adaptive attention within a Transformer-based decoder, selectively integrating image features to optimize accuracy. Experiments on the Flickr30k dataset show that this approach outperforms LSTM-based techniques in both accuracy and training speed.

Finally, in [17], it was proposed with Wavelet Transform-based CNN together with Visual Attention Prediction Network and Contextual Spatial Relation Extractor. In this way, WCNN considers spatial, spectral, as well as semantic details. The attention maps that VAPN produced were useful to better extract the fine-grained semantics, while the contextual spatial relationship extractor also captured the objects' relationship and led to a better understanding in the context. Here, the goal is to enhance the captioning

process within the context of feature representation refinement.

To tackle the problem of retaining long-range contextual information, [18] suggests a memory-enhanced model where words generated earlier are stored in an external memory. Traditional RNNs cannot retain earlier information in a sequence, which means partial context in captions. This selective reading mechanism fetches relevant past knowledge to be used in informing word predictions and improves coherence and contextual relevance. On the MS COCO dataset, experiments show a 3.5% improvement in CIDEr scores over other state-of-the-art models. This underlines the importance of memory modules in maximizing contextual richness in image captioning.

Another detailed overview is available in [19], which expands more on progress in methodologies in image captioning that range from basic machine learning to contemporary deep learning frameworks. It is a paper describing the evolution from rule-based systems to highly complex neural architecture that uses CNNs for feature extraction and RNNs for sequence generation. Ongoing challenges include: good capture of visual context, relevance, and output diversity. Further, the authors also discussed how large-scale datasets such as MS COCO are used during model development and testing and hinted at potential future work on attention mechanisms, reinforcement learning, and multi-modal approaches.

Finally, [20] proposed a double LSTM method whose scene factors are involved in the course of caption generation. The authors observe that many of these models disregard the contextual hints encoded in an image, meaning that their generated descriptions are often less informative. Within the implementation herein, the designed dual-LSTM architecture introduces two aspects of scene relationships in object and layout. Their respective visual features are handled through CNN before entering two streams of LSTMs for encoding the content as well as the scenes' information separately. This design produces more accurate captions based on the visual elements along with their contextual relationships. Empirical evidence confirms that such a design outperforms single-LSTM architectures by offering richer, context-aware captions that point to the increased value of scene understanding in image captioning frameworks.

Table 1 provides a comparative analysis of advancements in image captioning technologies, detailing their contributions, limitations, and how INSPIR tackles these challenges.

Table 1: A timeline of technological advancements in image captioning, highlighting key limitations and how INSPIR addresses them.

Paper Title	Technique Used	Metrics Reported	Key Contributions	Limitations	How INSPIR Addresses Limitations
Show and Tell: A Neural Image Caption Generator	CNN+RNN	BLEU, METEOR	Introduced CNN for image feature extraction and RNN for sequential text generation.	Struggled with generalization across diverse datasets due to overfitting issues.	Combines multiple advanced models (BLIP, GIT, ViT-GPT2) to generalize across diverse datasets and reduce overfitting.
Image Captioning using CNN and LSTM	CNN + LSTM	BLEU, CIDEr	Enhanced temporal dependencies, allowing better sequence learning.	Limited scalability for large datasets, making it challenging to apply in real-world scenarios.	Uses scalable transformer-based models and ensemble techniques to handle large datasets efficiently while maintaining accuracy.

Table 1: A timeline of technological advancements in image captioning, highlighting key limitations and how INSPIR addresses them.

Paper Title	Technique Used	Metrics Reported	Key Contributions	Limitations	How INSPIR Addresses Limitations
Improve Image Captioning by Self-attention	CNN + LSTM with Self-Attention	BLEU, CIDEr	Introduced self-attention, enabling models to focus on relevant image regions.	Overfitting issues due to the complexity of attention mechanisms, hindering performance on smaller datasets.	Utilizes entropy-based dynamic temperature scaling and adaptive scoring to prevent overfitting and improve performance on small datasets.
An Image is Worth 16x16 Words	Vision Transformers (ViT)	BLEU, CIDEr	Revolutionized image captioning by capturing long-range dependencies in images.	Required extensive computational resources for training, limiting accessibility for smaller research teams.	INSPIR utilizes pre-trained transformer models (BLIP, GIT, ViT-GPT2) to minimize computational demands.
Learning Transferable Visual Models From Natural Language Supervision (CLIP)	CLIP	F1 Score, Precision, Recall	Established a shared semantic space for aligning image and text embeddings.	Struggled with fine-grained semantic understanding for detailed captions.	Combines CLIP with adaptive scoring and dynamic temperature scaling to enhance fine-grained semantic understanding.
ClipCap: CLIP Prefix for Image Captioning	Prefix Tuning + CLIP	CIDEr, BLEU	Enhanced CLIP features for captioning by optimizing small task-specific vectors.	Limited diversity in generated captions due to reliance on prefix tuning.	Employs ensemble models and context-aware LLM prompting to generate diverse, creative, and application-specific captions.
BLIP: Bootstrapping Language-Image Pre-training	Unified Vision-Language Pretraining	BLEU, CIDEr, METEOR	Unified framework for vision-language understanding and caption generation.	Required large-scale datasets for pretraining, which may not be feasible for all researchers.	Uses multi-modal ensemble learning and adaptive scoring to optimize results even on smaller datasets like Flickr8K.

3. Proposed Model

The INSPIR framework represents a sophisticated system for image captioning and retrieval, designed to seamlessly integrate multiple state-of-the-art (SOTA) models for enhanced image and text

interactions as illustrated in Figure 1. This section details the model's architecture, the individual components, and the processes involved in synthesizing captions, ranking their relevance, and facilitating image retrieval. A combination of image captioning, semantic ranking, and retrieval capabilities establishes a robust and generalized model.

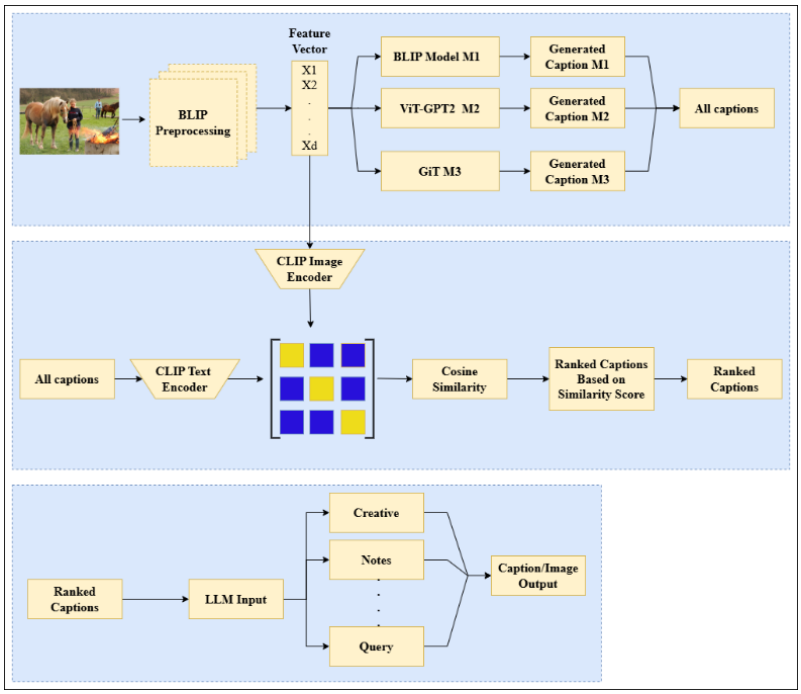

Figure 1. Model Architecture.

3.1. Image Processing

The preprocessing step before generating any caption is involves series of operations and is extremely necessary because this is the format that the trained model can understand to generate relevant captions. This step is crucial for ensuring the input format is compatible with the trained models.

1. Image Conversion: Image are first converted to RGB (Red Green Blue) scale to ensure consistency across various input files.
2. Resizing: The images are resized to 384 × 384 pixels using bicubic interpolation before it is given as input to the BLIP model. The input dimensions for all images are standardized, which is critical in deep learning models that require fixed-size inputs.
3. Normalization: The images are then converted to PyTorch Tensors, and their pixel values are normalized using fixed value mean and standard deviation.

Formally:

$$Image_{BLIP} = Normalize\left(Resize\left(Image, (384{,}384)\right)\right) \quad (1)$$

The GIT and ViT-GPT2 models apply similar preprocessing using their internal feature extractors.

4. Batch Dimension: The processed image tensors are unsqueezed to add a batch dimension (making it suitable for model inference) and then repeated to match the number of captions desired per image, allowing multiple captions to be generated from a single image input.

3.2. Caption generation

The INSPIR framework take into account three SOTA models—BLIP, GIT, and VIT-GPT2— to generate captions for given input image. Each model follows a probabilistic sequence-generation framework, where captions are produced token by token:

$$P(Image) = \prod_{t=1}^{T} P(w_{<t}, Image) \quad (2)$$

where w_t is the word at time step t.

1. BLIP Architecture: BLIP leverages a transformer-based architecture (Figure 2) to generate captions by integrating the visual features with language modelling. This process involves encoding of image, generating a sequence of tokens, and decoding these tokens into natural language caption.

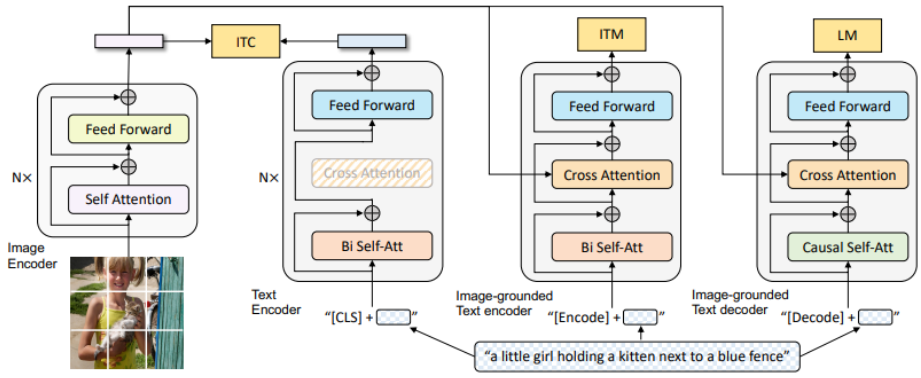

Figure 2. BLIP Architecture.

2. GIT Architecture: GIT focuses on generating captions by directly modeling the relationship between images and their corresponding textual descriptions through a generative approach (Figure 3). It follows a similar probabilistic framework which is shown by equation 2.

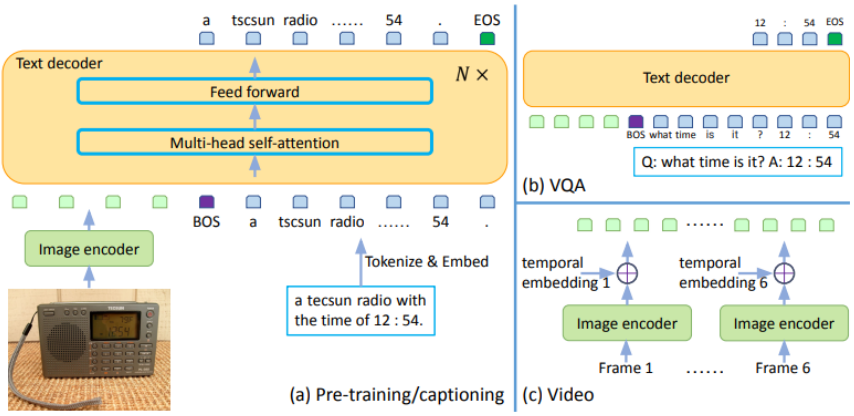

Figure 3. GIT Architecture.

3. ViT-GPT2 Architecture: ViT-GPT2 combines visual processing capabilities of Vision Transformers with the language generation capabilities of GPT-2 (Figure 4). It also follows Eq. (2), with self-attention playing a key role in linking image regions to text tokens.

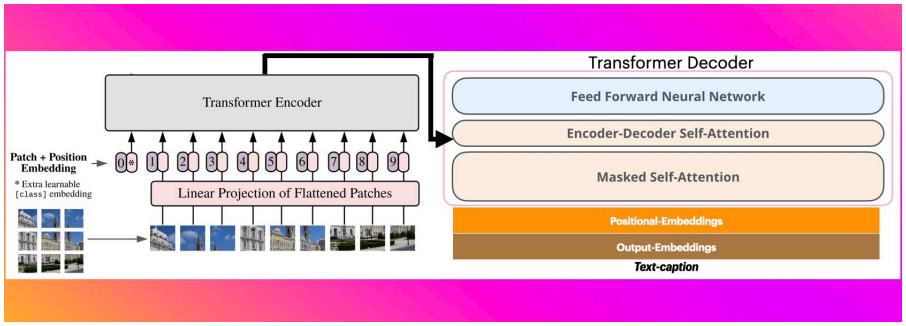

Figure 4. ViT-GPT2 Architecture.

The attention mechanism given by:

$$A = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where, Q, K, V are query, key, and value matrices derived from the input embeddings, and d_k is the dimensionality of the keys.
Across all three models, logits for each token are transformed into probabilities via SoftMax.

$$P(C_{<t}, I) = \frac{exp\ exp\ (z_t)}{\sum_{j=1}^{|V|} \quad exp\ exp\ (z_j)} \tag{4}$$

Here $P(C_{<t}, I)$ is the probability of generating token C_t given previous tokens $C_{<t}$ and visual features V, z_t is the logit for token t and $|V|$ is the size of the vocabulary.

3.3. Caption Ranking

After generating captions from multiple models, the next step is to rank these captions based on their relevance to the corresponding images. This ranking process is critical so as to ensure that the final output aligns closely with the visual content. This is achieved using CLIP (Contrastive Language-Image Pre-training), which aligns visual data with textual descriptions in a shared semantic space. CLIP ranks captions by computing similarity scores between image embeddings (p) and text embeddings (q) within a shared semantic space given by equation 5.

$$S(p, q) = \frac{\langle p, q \rangle}{||p||.||q||} \tag{5}$$

Here, $\langle p, q \rangle$ is the dot product, $||p||$ and $||q||$ are the magnitudes of the embeddings. A soft-max transformation on the scores produces probabilities, which aid in ranking given by equation 6.

$$P(Caption_i) = \frac{exp\ exp\ \left(S(p, q_i)\right)}{\sum_j \quad exp\ exp\ \left(S(p, q_j)\right)} \tag{6}$$

3.4. Creative Generation

In this final step, the ranked captions are applied creatively to social media posts or notes about the images. Llama3.1 receives the best captions to be further processed. The user then chooses a kind of caption to be generated. A prompt is built that includes instructions to Llama3.1 to come up with the specific type of caption based on the options produced in earlier steps. The best caption is refined by Llama3.1 through prompting, enabling user-preferred variations.

3.5. Image Retrieval

Beyond image captioning, INSPIR enables text-to-image retrieval. This image-retrieval pipeline preprocesses dataset images (convert to RGB, resize to 384×384, normalize) and stores them in a database. Captioning models (BLIP, GIT, ViT-GPT2) generate candidate captions; CLIP scores and ranks them and the top caption is saved with each image. CLIP also encodes images and captions into fixed-size multimodal embeddings. A user query (e.g., "A dog playing on the beach") is encoded by CLIP, images are ranked by cosine similarity to the query embedding, and the top matches are returned with their stored captions.

Figure 5 illustrates this entire retrieval process, providing a visual representation of how user queries are processed within the INSPIR framework.

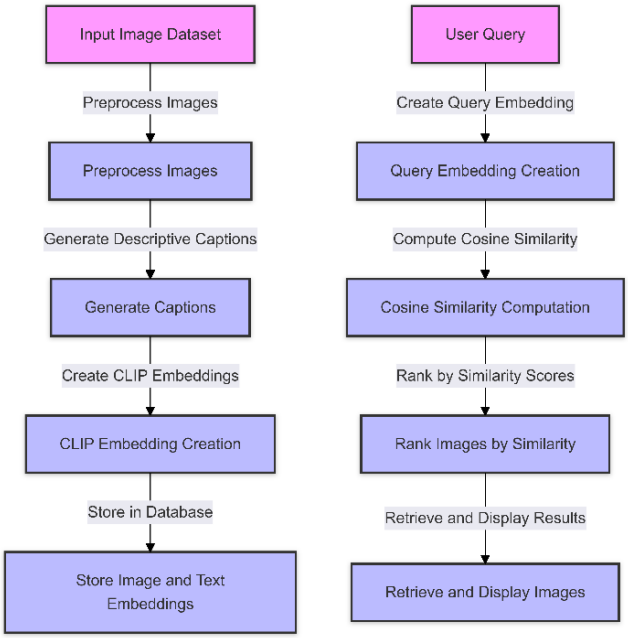

Figure 5. Image Retrieval Process.

4. Testing Methodology

Following Section 3, we evaluate INSPIR on the Flickr8k dataset. This section outlines the evaluation focusing on assessing the effectiveness of the image captioning system through quantitative metrics such as cosine similarity, BLEU, METEOR scores, as well as qualitative assessment.

4.1. Performance of Individual Models

Utilizing the Flickr8k dataset, we evaluated three advanced captioning models: BLIP, GIT, and ViT-GPT2. The performance of each model was assessed by measuring cosine similarity, BLEU, and METEOR scores between the generated captions and the ground truth. The results obtained are summarized in Table 2.

Table 2. Evaluation Metrics for Individual Models

Model	Average Cosine Similarity Score	BLEU Score	METEOR Score
BLIP	0.1832	0.1804	0.3921
GIT	0.2041	0.2076	0.4425
ViT-GPT2	0.1508	0.1250	0.3653
Combination	0.2119	0.2192	0.4663

Also, to give a holistic view of performance metrics across all the models we are evaluating, we plot the BLEU, METEOR, and cosine similarity scores of BLIPS, GIT, and ViT-GPT2 side-by-side in the following Figure 6 for an easy comparison of these models with respect to all these crucial evaluation metrics.

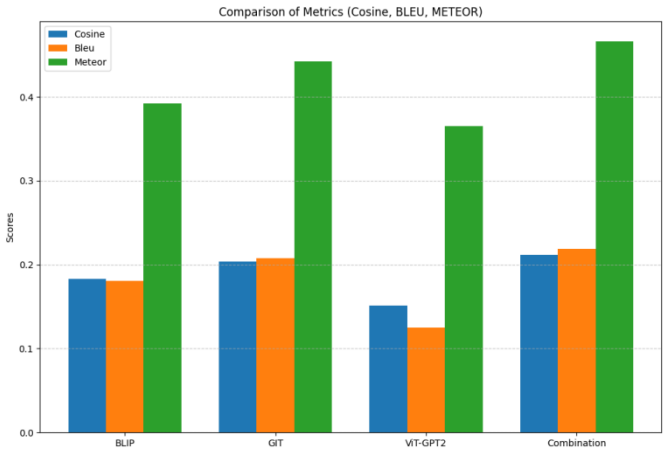

Figure 6. Comparison of BLEU, METEOR, and Cosine Similarity Scores Across Models.

Conclusions drawn from the results are:

- The Combination model (CLIP-ranked selection among BLIP, GIT and ViT-GPT2 captions) achieved the highest mean Cosine of 0.211920, indicating the greatest lexical overlap with the reference captions on this sample.
- **GIT** obtained the next highest mean Cosine of 0.204082.
- **BLIP** performed slightly worse than GIT but better than ViT-GPT2, with a score of 0.1832.
- **ViT-GPT2** had the lowest score (0.1508), suggesting that it struggles to generate captions that align well with the ground truth.

The cosine similarity was calculated using equation 7

$$Cosine\ Similarity\ (A,B)\ =\ \frac{A\ .\ B}{||A||\ ||B||} \tag{7}$$

where A and B are the TF-IDF vectors of the generated caption and ground truth captions, respectively.

The BLEU score measures n-gram overlap between generated captions and ground truth, which was calculated by using equation 8

$$BLEU\ =\ BP\ .exp\ (\sum_{n\ =1}^{N}\ \ w_n\ log\ p_n) \tag{8}$$

where BP is brevity penalty, p_n is the precision of grams and w_n is the weight of an n-gram

And finally, the METEOR score that evaluated alignment based on synonym matching and stemming was calculated using equation 9.

$$METEOR\ =\ (1\ -\ Penalty)\ .\frac{Precision\ .\ Recall}{\alpha\ .\ Precision\ +\ (1\ -\ \alpha)\ .\ Recall} \tag{9}$$

where, Penalty accounts for word order mismatches and α balances precision and recall.

4.2. Combination Model with Ranking

To enhance performance further, we combined the captions generated by BLIP, GIT, and ViT-GPT2 and used CLIP to rank them. The Combination model achieved a mean Cosine (CLIP) = 0.2119, exceeding each model (GIT = 0.2041, BLIP = 0.1832, ViT-GPT2 = 0.1508). This indicates that combining multiple models and using CLIP for ranking can improve performance over individual models like BLIP, ViT-GPT2 and GIT's performance.

This ranking process uses CLIP to compute similarity scores between image embeddings (p) and text embeddings (q), and the similarity equation can be represented using equation 10.

$$Similarity(Image, Caption)\ =\ p\ .q \tag{10}$$

And the probabilities for the ranking can be computed by using equation 11

$$P(Caption_i|Image)\ =\ \frac{exp(Similarity(Image, Caption_i))}{\sum_{j=1}^{N}\ \ exp(Similarity(Image, Caption_i))} \tag{11}$$

4.3. Impact of Weighted Combination

We further experimented with assigning random weights to the captions generated by BLIP, GIT, and ViT-GPT2 and normalizing these weights to create a weighted combination model. The results are summarized in Table 3.

Table 3. Evaluation of Ensemble Weights on Average Cosine Similarity Score.

Weights (BLIP, GIT, ViT-GPT2)	Average Cosine Similarity Score
(0.232, 0.376, 0.392)	0.4125
(0.123, 0.021, 0.856)	0.4125
(0.435, 0.094, 0.472)	0.4125
(0.109, 0.355, 0.536)	0.4125
(0.291, 0.430, 0.279)	0.3472
(0.199, 0.498, 0.303)	0.3472
(0.136, 0.841, 0.023)	0.3472
(0.199, 0.498, 0.303)	0.3282
(0.159, 0.458, 0.383)	0.3472
(0.291, 0.430, 0.279)	0.3282

The best performance (0.4125) was achieved when the weights were BLIP=0.232, GIT=0.376, and ViT-GPT2=0.392. Such combinations had the highest similarity score. Indicating a combination of these weights gives a balance between the three models. The weights are normalized to sum to one by using the formula in 12.

$$Normalized\ Weights\ = \frac{Weights}{\sum_{i=1}^{3} \quad Weights_i} \tag{12}$$

4.4. Impact of Entropy-Based Temperature Scaling

We investigated the impact of entropy-based temperature scaling on the CLIP ranking process This approach dynamically adjusts the temperature parameter (T) to refine the probability distribution and improve ranking accuracy. The scaled logits are computed as using the equation 13.

$$Scaled\ Logits\ = \frac{Logits}{T} \tag{13}$$

where T>0 is the temperature parameter. Lower values of T result in sharper probability distributions, and higher values result in smoother distributions.

SoftMax Probabilities are recalculated using the equation 4. Entropy (H) was calculated to measure the uncertainty in the probability distribution which is given in equation 14.

$$H\ = -\sum_{i=1}^{N} \quad P_i\ .\ log(P_i) \tag{14}$$

Based on the entropy value, T was dynamically adjusted using the scaling function shown in equation 15.

$$T_{new}\ =\ T.\ f(H) \tag{15}$$

where *f(H)* is a linear function defined using equation 16.

$$f(H)\ = \propto . H\ +\ \beta \tag{16}$$

with constants α and β controlling the sensitivity of scaling. The adjusted T_{new} was then used to recompute probabilities and rank captions.

Entropy-based temperature scaling adjusts T based on the entropy H of the SoftMax distribution; this reduces overconfidence for high-entropy caption sets and improves ranking stability. Figure 7 outlines

the entropy distribution across the images. The dynamic adjustment was very effective in reducing uncertainty for the high-entropy captions. Dynamic scaling improved the ranking stability. Dynamic scaling can rank captions that semantically differ subtly. Figure 8 details the adjusted temperature values for the given images.

Key takeaways were that temperature adjustment enhanced the diversity of the captions without a loss in relevance. And, ranking variability was decreased, thus giving consistent high-quality outputs.

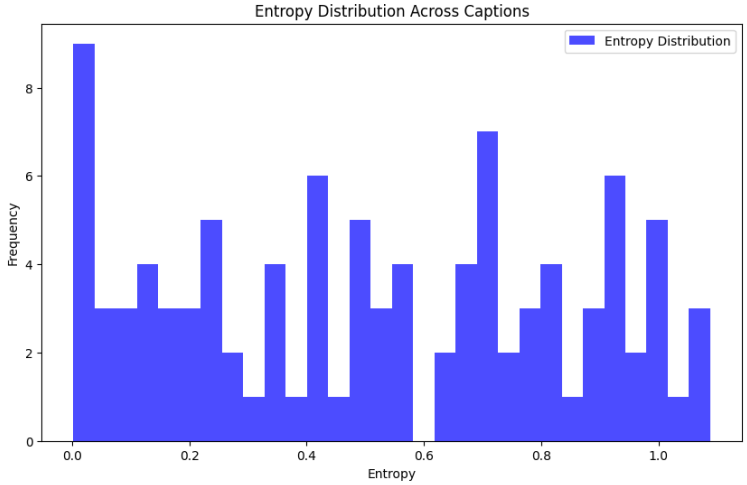

Figure 7. Entropy Distribution Across Images.

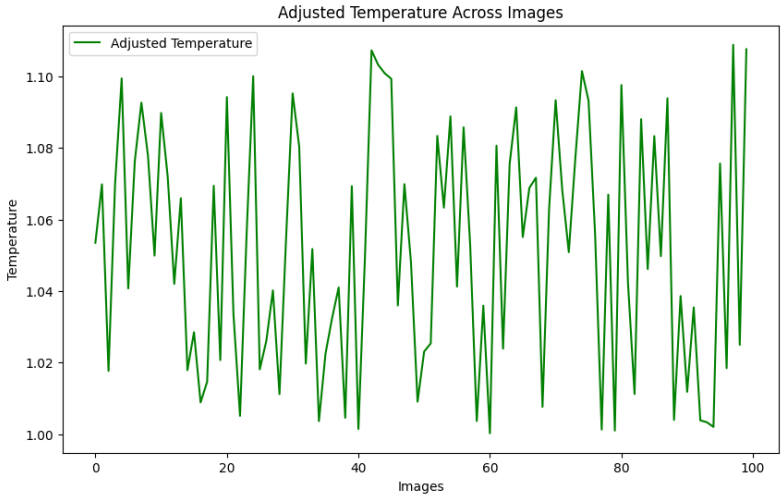

Figure *8.* Adjusted Temperature Across Images.

4.5. Analysis of Logits and Probabilities

The updated implementation addressed earlier concerns regarding deterministic logits. By dynamically scaling temperatures, the ranking probabilities became more flexible, reflecting the model's confidence more accurately. For example,

- **Raw Logits**: tensor ([[24.5699, 29.0210, 30.3251]])
- **Scaled Logits (Temperature=0.5)**: tensor ([[49.1399, 58.0421, 60.6501]])
- **Probabilities (Temperature=0.5)**: tensor ([[9.3383e-06, 6.8620e-02, 9.3137e-01]])

Some of the key takeaways were

- Combination as best overall: The CLIP-ranked Combination model achieved the highest mean score (Cosine = 0.211920), outperforming the best single model GIT (Cosine = 0.204082), BLIP (0.183212) and ViT-GPT2 (0.150848) on samples of images. This indicates that selecting among multiple generators yields measurable gains in average alignment to the reference captions.

- GIT as strongest standalone: GIT remains the top single model, consistently outperforming BLIP and ViT-GPT2 across the evaluated metrics.
- Weighted ensembles: Appropriately weighted combinations of the generators can further leverage GIT's strengths while incorporating complementary outputs from BLIP and ViT-GPT2; in practice such weighted schemes can match or exceed the performance of individual models depending on the chosen weights.
- Temperature scaling benefits: Dynamic temperature adjustment reduced the dominance of a single caption in many cases, produced smoother and more informative probability distributions for ranking, and improved ranking stability—particularly for candidate sets with subtle semantic differences.

In summary, while GIT is the strongest standalone captioner in our experiments, the Combination strategy (with CLIP ranking and optional weighted aggregation) yields the best average performance on the evaluated Flickr8k subset; dynamic temperature scaling further improves ranking robustness and fairness.

5. Results and Discussions

To assess the performance of the INSPIR model and its subcomponents, we utilize three widely recognized metrics for evaluating image captioning systems: BLEU, METEOR, and CIDEr. Additionally, the practical utility of the model in image retrieval was tested, showcasing its application-oriented capabilities.

4.1. Quantitative results

4.1.1. Evaluation metrics for Individual Models

The performance of individual models, namely, BLIP, GIT, and ViT-GPT2, was evaluated using BLEU, METEOR, and CIDEr scores. This framework for evaluating caption quality encompasses all facets.BLEU (Bilingual Evaluation Understudy) measures n-gram overlap between generated and reference captions by focusing more on precision. It is most helpful in judging fluency and adequacy of the generated text. The BLEU-4 score measures overlap for up to 4-grams, thus balancing the short and long-term contextual accuracy. METEOR (Metric for Evaluation of Translation with Explicit Ordering) incorporates synonyms, stemming, and word order, which better correlate with human judgment than BLEU. This measure accounts for semantic meaning and syntactic structure. In this regard, it is considered to be a more robust measure for evaluating generated captions. CIDEr (Consensus-based Image Description Evaluation) focuses on consensus between generated captions and multiple references based on a TF-IDF weighting scheme. By evaluating term importance, the CIDEr highlights contextually rich and meaningful descriptions. Table 4 summarizes results from key research studies that also justify the performance of the INSPIR framework.

Table 4. Metrics of SOTA models**.**

Model	**Evaluation Metric**		
	BLEU-4	***METEOR***	***CIDEr***
CNN+LSTM	0.27	0.24	0.85
CNN+LSTM (Self Attention)	0.30	0.25	0.94
ViT-GPT-2	0.34	0.28	1.02
GIT	0.35	0.29	1.05
CLIPCap	0.36	0.30	1.13
BLIP	0.37	0.30	1.16

Evaluation metrics of standard models that have been utilized for image captioing, trained, and tested on MS COCO Dataset [1,2,3,4,5]

Transformer models use multi-headed self-attention to capture global dependencies in images and text. ViT-GPT-2 achieves substantial gains across BLEU-4 (34.4%), METEOR (28.7%), and CIDEr (102.3%), whereas GIT scores marginally higher than it on all metrics. These results validate the capacity of transformers to generalize across diverse visual inputs and generate semantically rich captions. These essentially form the backbone of the ensemble model. BLIP excels across all metrics, setting a new benchmark for image captioning tasks.

4.1.2. Performance of INSPIR's Combination Model

The INSPIR model offers significant qualitative advantages, particularly in creative tasks and efficient image retrieval. The INSPIR framework enhances creative caption generation and simplifies the process of finding images based on textual queries, addressing key challenges in multimodal AI applications. An advantage of INSPIR is its ability to generate creative captions that engage users effectively. Traditional models based on simple text performed poorly in real-time interaction regarding visual content, as they cannot gain much utility or value in this regard in creativity-based applications. However, by employing an ensemble of advanced models—BLIP, GIT, and ViT-GPT2—INSPIR manages these limitations. The model generates dynamic and engaging captions tailored for social media platforms. By utilising Llama3.1, captions can adapt tone and style, creating humorous, poetic, or professional descriptions, and enhancing user engagement.

The combination model, which integrates captions generated by BLIP, GIT, and ViT-GPT2 ranked using CLIP, exhibited substantial improvements over individual models in terms of diversity and contextual relevance. Per-model BLEU-4, METEOR and CIDEr scores and the per-model cosine values are reported in Table 2; a compact summary of combination performance and top ensemble weights appears in Table 3.

4.1.3. Impact of Weighted Combinations

The evaluation of the top weight combinations for the ensemble was performed to analyze their effect on cosine similarity scores which are given in Table 3.

4.1.4. Impact of Entropy-Based Temperature Scaling

Entropy-based dynamic temperature scaling was analyzed to evaluate its impact on ranking stability. By dynamically adjusting temperature (T) based on entropy (H), the model refined probability distributions, leading to improved ranking stability.

4.1.5. Explanation of Determinism:

The logits produced by CLIP exhibited a clear hierarchy where one caption consistently received a higher score than others:

- **Raw Logits:** tensor ([[24.5699,29.0210,30.3251]]) tensor ([[24.5699,29.0210,30.3251]])
- **Scaled Logits (T=0.5):** tensor ([[49.1399,58.0421,60.6501]]) tensor ([[49.1399,58.0421,60.6501]])
- **Probabilities:** `tensor([[9.34e-06,6.86e-02,9.31e-01]]) tensor([[9.34e-06,6.86e-02,9.31e-01]])`

4.2. Qualitative Results: Image Retrieval

INSPIR can perform semantic image retrieval from natural-language queries. Given a query, the system ranks stored images by CLIP similarity and returns the top matches. Figure 9 shows all four images, which are used to upload within the system so that all four images set different scenarios. When a user inputs the query "A man sliding on the water," then the system activates the retrieval mechanism, as shown in Figure 5. This figure depicts the workflow of the retrieval process. It demonstrates how user queries are passed through the algorithm to identify and rank potential matches within the uploaded images. The model processes the semantic content of the query and compares it to the features of each image it has stored in the database. The system fetched an image captioned "A man laying on top of a black tarp in the middle of water," which it identified as the best match for this particular query. Figure 10 depicts a visual example of the outcome of retrieval. It shows an image retrieved for the given input query along with its caption. The outcome shows the ability of the model to classify contextual details and subtleties in user queries even with such a small amount of data. The image retrieved not only aligns very much with the keywords of the query but also captures something about the action and setting that reflects what was asked for. This example underlines how effectively the INSPIR framework connects the user inputs with adequate visual content. This improves both accessibility and searching efficiency across different datasets. Images can be found based on finer queries, an example of the opportunity for practical utility in various contexts, such as digital asset management, e-commerce, and the creation of digital content, where immediate and accurate access to visual information becomes necessary.

Figure 9. Images from Dataset

Figure 10. Result of Image Retrieval Application

6. Conclusion

INSPIR model brings together the best of many state-of-the-art technologies like BLIP, ClipCap, ViT-GPT-2, and GIT, and therefore can really go further than previous state-of-the-art image captioning models. A smart ranking system with temperature scaling and multi-view scoring enables it to balance between relevance, diversity, and fluency. INSPIR combines all those technologies into this single smooth-flowing process producing captions that would not only seem more fluent but also more contextually relevant than the captions are meaningfully sharp. Ranking using CLIP ensures it's caption-precise-according to an image. This design can be beneficial in a wide range of aspects, including social media content enrichment, enhanced image search results, and accessibility options for visually impaired users. Later, INSPIR may further be extended to support audio and video content in order to be more interactive and immersive. Ability to refine the captions based on user feedback could make it even more effective. It might also help INSPIR scale heights by reaching new applications in real-time, such as live events and in e-commerce.

Author Contributions

V.S. contributed to the conceptualization, supervision, project administration, and critical review of the manuscript. C.I., H.M., and P.J. contributed to the methodology, software development, investigation, formal analysis, data curation, visualization, and preparation of the original draft. S.J. (Sarvesh Jadhav) and S.J. (Sahil Jadhav) contributed to the literature review and manuscript review and editing. All authors read and approved the final manuscript.

Funding

This research received no external funding.

Conflict of Interest Statement

The authors declare no conflicts of interest.

Data Availability Statement

The data supporting the findings of this study are based on the publicly available **Flickr8k image dataset (publicly available at https://www.kaggle.com/datasets/adityajn105/flickr8k)**, which consists of images collected from Flickr along with corresponding textual captions. The dataset is openly accessible

for research purposes and was used for training and evaluation in this study. No new datasets were generated during the current research.

Acknowledgment
We would like to thank our honorable Director Prof. (Dr.) R.M Jalnekar, Vishwakarma Institute of Technology, Pune and HOD Prof. (Dr.) Mrs. Shital Dongre for extending strong moral support and encouragement.

References

1. Vinyals et al., "Show and Tell: A Neural Image Caption Generator," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
2. G. Sairam, M. Mandha, P. Prashanth, and P. Swetha, “Image Captioning using CNN and LSTM,” IEEE Xplore, Nov. 01, 2021. https://ieeexplore.ieee.org/document/9770764
3. Z. Li, Y. Li, and H. Lu, “Improve Image Captioning by Self-attention,” Communications in computer and information science, pp. 91–98, Jan. 2019, doi: https://doi.org/10.1007/978-3-030-36802-9_11.
4. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in Proceedings of the International Conference on Learning Representations (ICLR), 2021. K. Elissa, “Title of paper if known,” unpublished.
5. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proceedings of the 38th International Conference on Machine Learning (ICML), 2021.
6. R. Mokady, A. Hertz, and A. H. Bermano, “ClipCap: CLIP Prefix for Image Captioning,” arXiv:2111.09734 [cs], Nov. 2021, Available: https://arxiv.org/abs/2111.09734
7. Li et al., "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding," arXiv preprint arXiv:2201.12086.
8. X. L. Li and P. Liang, “Prefix-Tuning: Optimizing Continuous Prompts for Generation,” arXiv:2101.00190 [cs], Jan. 2021, Available: https://arxiv.org/abs/2101.00190
9. P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, “Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing,” ACM Computing Surveys, vol. 55, no. 9, Sep. 2022, doi: https://doi.org/10.1145/3560815.
10. L. Agarwal and B. Verma, “From methods to datasets: A survey on Image-Caption Generators,” Multimedia Tools and Applications, vol. 83, no. 9, pp. 28077–28123, Aug. 2023, doi: https://doi.org/10.1007/s11042-023-16560-x.
11. Y. Zhang et al., “Recognize Anything: A Strong Image Tagging Model,” arXiv.org, 2023. https://arxiv.org/abs/2306.03514
12. H. Wang, Y. Zhang, and X. Yu, “An Overview of Image Caption Generation Methods,” Computational Intelligence and Neuroscience, vol. 2020, pp. 1–13, Jan. 2020, doi: https://doi.org/10.1155/2020/3062706.
13. P. Tian, H. Mo, and L. Jiang, “Image Caption Generation Using Multi-Level Semantic Context Information,” Symmetry, vol. 13, no. 7, p. 1184, Jun. 2021, doi: https://doi.org/10.3390/sym13071184.
14. M. Barraco, M. Cornia, S. Cascianelli, L. Baraldi, and R. Cucchiara, “The Unreasonable Effectiveness of CLIP Features for Image Captioning: An Experimental Analysis.” Available: https://openaccess.thecvf.com/content/CVPR2022W/MULA/papers/Barraco_The_Unreasonable_Effectiveness_of_CLIP_Features_for_Image_Captioning_An_CVPRW_2022_paper.pdf
15. C. Amritkar and V. Jabade, “Image Caption Generation Using Deep Learning Technique,” IEEE Xplore, Aug. 01, 2018. https://ieeexplore.ieee.org/abstract/document/8697360
16. “Image Caption Generation With Adaptive Transformer | IEEE Conference Publication | IEEE Xplore,” ieeexplore.ieee.org. https://ieeexplore.ieee.org/abstract/document/8787715
17. R. Sasibhooshan, S. Kumaraswamy, and S. Sasidharan, “Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction,” Journal of Big Data, vol. 10, no. 1, Feb. 2023, doi: https://doi.org/10.1186/s40537-023-00693-9.
18. H. Chen, G. Ding, Z. Lin, Y. Guo, C. Shan, and J. Han, “Image Captioning with Memorized Knowledge,” *Cognitive Computation*, vol. 13, no. 4, pp. 807–820, Jun. 2019, doi: https://doi.org/10.1007/s12559-019-09656-w.
19. S. Bai and S. An, “A survey on automatic image caption generation,” *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018, doi: https://doi.org/10.1016/j.neucom.2018.05.080.
20. Y. Peng, X. Liu, W. Wang, X. Zhao, and M. Wei, “Image caption model of double LSTM with scene factors,” *Image and Vision Computing*, vol. 86, pp. 38–44, Jun. 2019, doi: https://doi.org/10.1016/j.imavis.201