Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Intelligent Cross-modal Alignment With Cataphora Dependency Modeling-based Text-to-image Synthesis and Captioning Using Gclan and Gq2phpt

Yamini Chouhan
Guru Ghasidas Vishwavidyalaya
Rohit Raja
Guru Ghasidas Vishwavidyalaya
Shilpa Choudhary
Neil Gogte Institute of Technology
Rohit Miri
Chhattisgarh Swami Vivekanand Technical University
Parul Dubey
parul.dubey@sitnagpur.siu.edu.in

Symbiosis International (Deemed University)

Article

Keywords: Cross-Modal Embeddings (CME), Image Captioning (IC), Cross-Modal Retrieval (CMR), Multimodal Learning (ML), Visual Data Integration (VDI), Cataphora Dependency (CD), and Text-to-Image Alignment (TIA)

Posted Date: August 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10163067/v1

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

Additional Declarations: No competing interests reported.

INTELLIGENT CROSS-MODAL ALIGNMENT WITH CATAPHORA DEPENDENCY MODELING-BASED TEXT-TO-IMAGE SYNTHESIS AND CAPTIONING USING GCLAN AND GQ2PHPT

[1]Yamini Chouhan [2]Rohit Raja, [3]Shilpa Choudhary, [4]Rohit Miri, Parul Dubey[5]

[1]Research Scholar, Department of Information Technology, School of Studies (Engineering and Technology) Guru Ghasidas Vishwavidyalaya, Bilaspur (CG), India, yaminichouhancs@gmail.com

[2]Department of Information Technology, School of Studies (Engineering and Technology) Guru Ghasidas Vishwavidyalaya, Bilaspur (CG), India, drrohitraja1982@gmail.com

[3]Neil Gogte Institute of Technology, Hyderabad, India shilpachoudhary1987@gmail.com

[4]Chhattisgarh Swami Vivekanand Technical University,Bhilai , Chhattisgarh, rohitmiri@csvtu.ac.in

[5]Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, India parul.dubey@sitnagpur.siu.edu.in

Corresponding Author:- Parul Dubey (parul.dubey@sitnagpur.siu.edu.in)

Funding Declaration:- This research was supported and Funded by Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune

Abstract: To enhance the semantic understanding of heterogeneous data, cross-modal alignment-based text-to-image generation and captioning is essential in recent times. Nevertheless, the existing works failed to track Cataphora Dependencies (CDs) in which pronouns appeared before their actual entities, thereby leading to poor semantic coherence and generation efficiency. Hence, in this article, an intelligent cross-modal alignment with CD modeling-based content retrieval and captioning utilizing Generative Collapsing Linear Adversarial Network (GCLAN) and Generative Quadruple Pre-trained Probabilist's Hermite Polynomials Transformer (GQ2PHPT) is proposed. Initially, the Text-to-Image Synthesis (TIS) module is designed. Here, the images are gathered, followed by pre-processing and fine-grained visual characteristics learning. Similarly, the text is collected and further fed into Natural Language Processing (NLP)-based pre-processing. Afterward, contextual word embedding, CD analysis, syntactic structure modeling, and feature extraction are carried out from pre-processed text. Then, by utilizing GCLAN, cross-modal alignment is performed. In real-time, an image is generated by the GCLAN regarding the user's query. Next, the generated image is fed into the Image Captioning (IC) module, in which a suitable caption is efficiently generated by the proposed GQ2PHPT. Thereafter, concerning the caption and generated images, cross-modal alignment is carried out, followed by score computation. Lastly, by considering the score value, the content is retrieved with a high success rate of 0.9422. When contrasted with the prevailing works, the proposed work effectively performs TIS and IC with an accuracy of 98.9734%.

Keywords: *Cross-Modal Embeddings (CME), Image Captioning (IC), Cross-Modal Retrieval (CMR), Multimodal Learning (ML), Visual Data Integration (VDI), Cataphora Dependency (CD), and Text-to-Image Alignment (TIA).*

1. INTRODUCTION

A significant demand for intelligent systems capable of effectively retrieving and understanding multimodal information has been created due to the rapid evolution of multimedia content generation via social media platforms, web repositories, and digital systems (Wang et al., 2024). Among numerous multimedia applications, IC and text-to-image generation have gained immense attention owing to their effective information access and automatic visual understanding (Rahman & Sami, 2024). By understanding contextual information present in images, IC automatically generates meaningful textual descriptions from visual content (Han et al., 2026). Furthermore, owing to TIS's potential to generate visual representations from user text queries, it plays an important role in Multimodal Learning (ML) systems (Alhabeeb & Al-shargabi, 2024). To support a content retrieval application, the TIS and IC tasks have been combined in recent times. Retrieving relevant multimedia information by examining semantic similarity between user queries and stored content is the aim of content retrieval (Sogi et al., 2024) (Zhou et al., 2023). However, the traditional Cross-Modal Retrieval (CMR) and generation studies mainly depended on keyword-based search mechanisms or visual feature matching, thus struggling to capture deeper semantic relationships between textual descriptions and images (Li et al., 2024).

Therefore, efficient content retrieval requires intelligent multimodal representation learning and effective semantic alignment betwixt visual and textual information (Li et al., 2025). To address ML challenges, Generative Adversarial Network (GAN) has evolved as a promising solution by generating semantically consistent representations and capturing complicated feature distributions (Tan et al., 2023). Moreover, for generating synthetic images from textual descriptions, the GAN has been extensively adopted (Habib, et al., 2024). Then, in numerous applications like content-based image retrieval, IC, multimodal search engine systems, and visual question answering, GAN-centric models improve TIS, cross-modal representation learning, and feature fusion (Lerner et al., 2021). The datasets, such as the MicroSoft Common Objects in COntext (MS COCO) and Flickr-30k, are widely adopted in the context of IC and cross-modal content retrieval tasks due to their availability of paired image-text annotations and rich semantic information (Han et al., 2024) (Zhang et al., 2025).

Furthermore, to perform caption generation and cross-modal content retrieval, several Artificial Intelligence (AI) techniques, such as Vision Transformer (ViT), Generative Pretrained Transformer (GPT), and Long Short-Term Memory (LSTM), were employed in the prevailing works (Alnashwan et al., 2023) (Yadav et al., 2026). Nevertheless, the traditional cross-modal content retrieval studies primarily focused on sequential text generation and local word prediction. Also, the prevailing works didn't focus on tracking CDs in which pronouns appeared before their corresponding entities, thus influencing contextual understanding and semantic coherence during content retrieval and generation. Hence, this research is motivated by the necessity to design an enhanced cross-modal alignment with a CD modeling-centric content retrieval and caption generation framework utilizing GCLAN and GQ2PHPT.

1.1 Research overview

The potential research gap and drawbacks of the prevailing works are discussed in this section. Moreover, the major goals of the proposed work are illustrated.

Research gap: The existing cross-modal content retrieval studies mainly concentrated on sequential text generation and local word prediction. Yet, the conventional works failed to focus on tracking CDs, thus mitigating contextual understanding and degrading logical coherence during content retrieval and generation.

Problem statement: The drawbacks of the traditional works are provided as follows,

- The prevailing cross-modal content retrieval system (Wang et al., 2024a) failed to address the semantic ambiguity in which the text had multiple possible meanings, thereby leading to an uncertain interpretation.
- The existing system (Zeng et al., 2025) lacked fine-grained visual detail analysis, thus reducing caption quality and weakening the alignment between visual content and textual queries.
- The traditional ML system (Pang et al., 2023) mainly depended on feature averaging or global feature concatenation. Hence, the model's potential to learn deep contextual interaction between textual characteristics and visual regions was affected.
- Some of the conventional works failed to handle the syntactic structure in the sentences, thereby degrading sentence coherence and leading to poor structural interpretation.
- Numerous prior works had reduced retrieval performance owing to the poor resolution and noisy characteristics of the images.

Significance and scope of the study: The proposed work recognizes forward references and logical relationships among entities by tracking CD, thus enhancing contextual understanding during multimodal retrieval. The existence of cross-modal alignment assists in ensuring deeper semantic understanding, thereby leading to accurate caption generation. Overall, in the proposed work, the cross-modal learning is improved during content retrieval and generation tasks.

Research aim and objectives: Designing an enhanced cross-modal alignment with CD modeling-based content retrieval and captioning is the key aim of the proposed work. The main objectives of the proposed methodology are given below,

- The proposed work uses the Coreference Jaro Similarity Resolution (CJSR) to capture CD between the sentences, thereby leading to efficient sentence understanding.
- In the proposed work, a novel Weierstrass Sentence Bidirectional Encoder Representations from Transformers (WSBERT) is introduced for reducing the semantic ambiguity.
- The proposed approach utilizes the Path Aggregation Quotient Network (PAQN) to capture the fine-grained visual characteristics.
- By employing GCLAN, cross-modal alignment is performed for ensuring effective feature fusion.

- Dependency Parsing (DP) is utilized to handle the syntactic structure of phrases.
- Effective pre-processing steps, namely noise removal and contrast enhancement, are carried out for improving the visual quality.

The paper is structured as: In Section 2, the literature review is presented, in Section 3, the mathematical modeling of the proposed work is depicted, in Section 4, the empirical assessment is illustrated, and finally, in Section 5, the paper is concluded with future enhancements.

2. LITERATURE SURVEY

(Zeng et al., 2025) introduced a multi-modal embedding transfer scheme for consistent and selective learning processes in CMR. Here, by using the embedding transfer strategy, multi-modal embeddings from the source model were transferred to the target model. This approach had high effectiveness in CMR. However, the subtle visual characteristics weren't captured by the model, thereby leading to poor contextual understanding.

(Wang et al., 2024a) suggested an effective IC model regarding a self-supervised model optimization transformer. Primarily, the textual features and image features were fed into the cross-modal optimization. The image-text features were transformed into a shared learning space by cross-modal optimization. Centered on the image features and paired captions, the framework significantly optimized the distribution of semantic information in images. This approach had high generalizability in IC. However, this model failed to handle the sentence polysemy problem, thus resulting in inaccurate interpretation.

(Pang et al., 2023) presented a cross-modal augmented space for composed image retrieval utilizing heterogeneous feature alignment and fusion. Here, the relative caption-aware consistency approach was used to align textual descriptions and images. After that, for capturing multiple interactions of the composed query, the cross-modal guided pooling and multimodal complementary fusion were used, thereby enhancing the content retrieval relevance. Nevertheless, this model faced improper situational understanding due to the inefficient feature fusion.

(Habib, et al., 2024a) established a TIS framework grounded on generative models utilizing attention mechanisms with contrastive learning. Here, for performing reliable TIS and IC, the conditional GAN, contrastive learning, and attention mechanism were combined. As per the experimental results, an R-Precision of 89.14% was attained by the suggested model, depicting superior efficiency. Yet, owing to the blurriness in images, the approach struggled to capture fine-scale structural details.

(Cai, et al., 2022) developed a Dual-Attention GAN (DA-GAN) for effective CMR. In this framework, intra-modal attention as well as inter-modal attention were included. To capture essential semantic features within a modality, intra-modal attention was employed. Furthermore, for identifying the semantic interaction betwixt various modalities, inter-model attention was used. The cross-modal heterogeneity was efficiently reduced by this model. However, this model had lower precision values and unstable learning performance owing to the improper correlation analysis during multi-modal learning, thereby leading to poor retrieval.

(Haoran et al., 2023) propounded a fine-grained cross-modal fusion-centric refinement approach for TIS. Here, to perform TIS, the fine-grained GAN was used. This approach attained impressive performance in Text-to-Image Alignment (TIA) owing to the utilization of a fine-grained text-image fusion block and global semantic refinement. Nevertheless, due to complex feature-level alignment, the framework incurred significant computation overhead.

(Ku & Lee, 2023) described a controllable GAN-centric TIS framework. Primarily, the image features and textual descriptions were fed into the controllable GAN. The learned cross-modal aligned features were generated by the controllable GAN. Also, concerning the user prompt query, the images were generated by considering the learned representations, thus depicting better significance. Yet, this model wasn't adequate for capturing the deeper semantic information from the textual queries, thus degrading the prominence of the model.

(Li et al., 2024a) established a deep feature-aware neighbor similarity hashing with adversarial learning for CMR. Here, to perform CMR, a contrastive language-image pre-training model was utilized. This model ensured high semantic consistency owing to the existence of a feature-based neighbor similarity matrix. Nevertheless, due to the limited sample sets, the contextual relevance of the data content was mainly affected.

(Zhang et al., 2023) presented a cross-attention-aware deep fusion transformer framework for reliable IC. In this model, a deep fusion transformer was employed for offering a deep multi-feature and multimodal information fusion strategy throughout the encoding and decoding processes. The IC reliability was effectively improved by this model. However, owing to ineffective cross-level interaction modeling, the model struggled to understand the deeper dependency analysis.

(Li et al., 2023) introduced a CMR approach for adaptive adversarial learning. Initially, from publicly available sources, the images and textual descriptions were gathered. Next, the image-text pairs were fed into the feature enhancement and adaptive balancing strategy. Overall, the semantically similar contents were retrieved by the suggested model with maximum relevance. However, this model had limited performance owing to the dimension redundancy.

3. PROPOSED METHODOLOGY FOR CROSS-MODAL CONTENT RETRIEVAL AND CAPTION GENERATION

An intelligent cross-modal alignment with a CD modeling-based content retrieval and captioning framework utilizing GCLAN and GQ2PHPT is proposed in this section. Figure 1 illustrates the conceptual design of the proposed work.

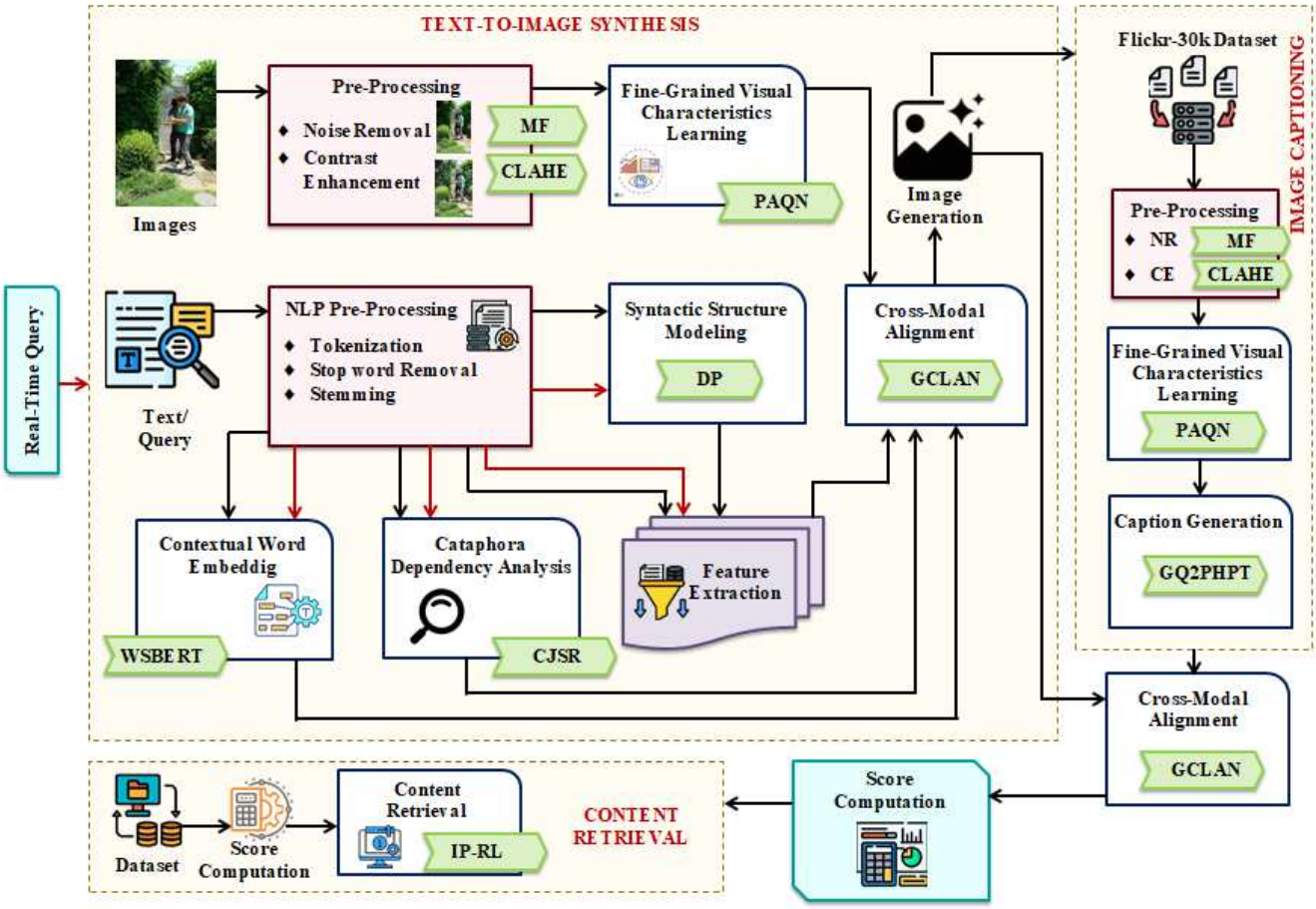


Figure 1: Architecture of the proposed work

In the proposed work, the relevance of multimodal retrieval is enhanced via TIS and image caption generation. The detailed overview of the research methodology is depicted as follows,

3.1 Text-to-image synthesis framework

In this section, the TIS framework is designed to generate an image concerning a user text prompt by learning a cross-modal relationship for facilitating relevant content retrieval.

3.1.1 Image collection

Initially, from publicly available sources, the images are collected to learn the semantic relationship betwixt image-text pairs for TIS.

$$I_x = (I_1, I_2, \ldots\ldots I_X) \text{ Here, } x = 1 \text{ to } X \tag{1}$$

Where, the collected images I_x are notated as X .

3.1.2 Pre-processing

After that, for visual quality enhancement, the raw I_x is subjected to pre-processing. Here, noise removal and contrast enhancement are included in the pre-processing. Primarily, the noise (salt-and-pepper noise) in I_x is removed by the Median Filter (MF) while preserving the crucial image properties. Moreover, the MF removes noise by replacing every single pixel with the median value of its neighboring pixels within the window.

$$N(\ddot{a},\ddot{b}) = \underset{w}{\text{med}}(I_x(a,b)) \tag{2}$$

Here, the noise-removed image with pixel coordinates $(\ddot{a},\ddot{b})$ is implied as $N(\ddot{a},\ddot{b})$, the median value is denoted as med , the pixel coordinates of the input image are exemplified as (a,b) , and the neighboring window is signified as w .

By using Contrastive Limited Adaptive Histogram Equalization (CLAHE), the resolution of $N(\ddot{a},\ddot{b})$ is improved in contrast enhancement. Furthermore, the local contrast is improved by CLAHE while reducing noise amplification, thereby enhancing visibility in low-contrast images.

Initially, to improve the resolution of the local regions, $N(\ddot{a},\ddot{b})$ is divided into several non-overlapping tiles. Next, the histogram (H_{gram}) is computed for each tile (T_y) to illustrate the distribution of the local intensity values.

$$H_{\text{gram}} = \sum_{v=1}^{V} T_y; \;\; v \in (0, V-1) \tag{3}$$

Here, the number of gray levels in T_y is represented as V . Thereafter, to reduce the noise amplification, the clip limit $(L_{\otimes})$ is applied to the histogram.

$$H^{\text{clip}} = \min(H_{\text{gram}}, L_{\otimes}) \tag{4}$$

Where, the clipped histogram is indicated as H^{clip} . Further, the cumulative distribution function $(\text{Cum}(v))$ is computed for mapping the original pixel values to uniformly distributed intensity values.

$$\text{Cum}(v) = \sum_{t=1}^{T} \frac{H^{\text{clip}}(v)}{T} \tag{5}$$

Here, the total number of pixels in the tile is exemplified as $t = 1 \text{ to } T$. Also, to retransform the intensity of each pixel in the tile, the new pixel values (P_{new}) are computed by considering $\text{Cum}(v)$.

$$P_{\text{new}} = M_{\text{int}} + (\tilde{M} - M_{\text{int}}) \times \text{Cum}(v) \tag{6}$$

Where, the minimum intensity range and maximum intensity range are implied as M_{int} and $\widetilde{M}$, correspondingly. Finally, to diminish the edge artifacts, the bilinear interpolation is applied. Here, $\left(P_{\text{img}}\right)$ represents the pre-processed images (contrast-enhanced images).

3.1.3 Fine-grained visual characteristics learning

Then, by employing PAQN, the fine-grained visual characteristics are learned from P_{img} for better image-text alignment. To perform fine-grained visual characteristics learning, the prevailing Path Aggregation Network (PAN) is introduced. The enhanced low-level feature propagation is ensured by PAN, thus capturing the crucial information about small and fine objects. Moreover, bottom-up path augmentation and stronger bidirectional feature fusion are facilitated by the PAN, ensuring better visual feature learning. However, an inappropriate backbone depth selection results in weak hierarchical feature extraction, thereby mitigating poor semantic understanding. Hence, in the proposed work, the Quotient Rule (QR) is utilized to effectively determine the backbone depth by formulating depth selection as a ratio optimization problem between computational cost and representation gain.

Firstly, the input image is subjected to the backbone network that extracts hierarchical features at different scales. Here, the backbone network assists in extracting the high-level semantic features. The proposed work utilizes the QR to determine the backbone depth $\left(D_{\infty}\right)$.

$$F_{\text{map}} = D_{\infty} \cdot P_{img}\left\{f_1, f_u, \ldots f_U\right\} \tag{7}$$

$$D_{\infty} = \frac{P_{\text{img}}\, i - i P_{\text{img}}}{i^2} \tag{8}$$

Here, the number of features f_u is implied as $u = 1 \text{to} U$ and the constant value is notated as i. Subsequently, the hierarchical feature map $\left(F_{\text{map}}\right)$is subjected to the top-down feature propagation stage that examines semantic information from upper layers to lower layers.

$$\hat{U}_u = \text{Conv}\left(f_u + q\left(\hat{U}_{u+1}\right)\right) \tag{9}$$

Where, the feature map at the u^{th} pyramid level is exemplified as $\hat{U}_u$, the feature map from the higher pyramid level is indicated as $\hat{U}_{u+1}$, the upsampling operation is notated as q, and the convolution operation is denoted as Conv. Further, bottom-up path augmentation (downsampling) is done to improve the localization information from lower features.

$$\vec{A}_u = \text{Conv}\left(f_u + \widetilde{d}\left(\hat{U}_u\right)\right) \tag{10}$$

Here, the downsampled features at the u^{th} pyramid level are implied as $\vec{A}$ and the downsampling operation is represented as $\widetilde{d}$. Finally, to generate a fused feature map $\left(L_{\text{xy}}\right)$, the features from all the pyramid levels are aggregated.

3.1.4 Text collection

Next, for accurate TIS, the textual data $\left(T_{\text{data}}\right)$ are collected to perform cross-modal alignment.

3.1.5 NLP pre-processing

During pre-processing, the NLP tasks are done to enhance the data consistency and format. The NLP pre-processing steps, namely tokenization, stop-word removal, and stemming, are carried out. For efficient textual description analysis, T_{data} is divided into numerous tokens in tokenization. Also, the stop words, such as "for", "an", and "the", are removed from the tokenized data to concentrate on the most informative textual features. To enhance the semantic understanding, the word's root form is obtained through stemming. Here, $\left(T_{\text{NLP}}\right)$ signifies the pre-processed text.

3.1.6 Syntactic structure modeling

Thereafter, DP is utilized to examine the syntactic structure (grammatical form) of T_{NLP}. Moreover, by using grammatical relations between words, DP transforms sentences into a directed syntactic graph, thus facilitating structured linguistic and semantic representation. Initially, to capture important syntactic information, Part-Of-Speech (POS) tags $\left(\text{PoS}_{\text{tag}}\right)$, namely verb, adjective, and noun, are assigned to T_{NLP}, thereby depicting dependency relationships.

$$\text{Tg} \xrightarrow{\text{PoS}_{\text{tag}}} \left(T_{\text{NLP}}\right) \tag{11}$$

Where, the POS tag-assigned sentences are notated as Tg. Then, to form a dependency relationship graph, the major verb or root word identified acts as the anchor point. Next, based on syntactic roles, the relationship between words is investigated. By connecting each word to a head word, the directed connections are formed. Furthermore, a grammatical function is assigned to the dependency link. The relationship between connected words is defined by a grammatical function. Lastly, by considering the root as the node $\left(\text{n}\right)$ of the graph and edges $\left(\text{e}\right)$ as the dependencies, the tree structure $\left(\ddot{\ddot{G}}\right)$ (complete grammatical structure) is generated.

$$\ddot{\ddot{G}} = \left(\text{n},\text{e}\right) \tag{12}$$

In addition, to identify the most optimal structure, multiple parses are examined, thus improving the significance of the semantic analysis.

3.1.7 Feature extraction

Afterward, from both T_{NLP} and $\ddot{\ddot{G}}$, the important features are extracted for upgrading the learning potential of the TIS framework. The features like word length, n-grams, sentence length, and Term Frequency-Inverse Document Frequency (TF-IDF) are extracted from T_{NLP}. Similarly, from $\ddot{\ddot{G}}$, the syntactic features, namely POS tags, edge weights, node degree, parent–child relations, centrality from dependency graphs, dependency paths, node connectivity, and tree depth, are extracted.

$$O_z = (O_1, O_2, \ldots\ldots O_Z) \tag{13}$$

Here, the extracted features O_z are implied as $z = 1,2,\ldots Z$.

3.1.8 Cataphora dependency analysis

Next, the CD in T_{NLP} is tracked using the proposed CJSR to ensure logical coherence. The Coreference Resolution (CR) algorithm efficiently recognizes the correct antecedent even when it appears later in the text, thus handling CD. Furthermore, the CR has improved global text understanding and superior contextual consistency. Nevertheless, entity linking is incorrect due to the poor similarity calculation, thereby leading to inaccurate CD analysis. Hence, to enhance the contextual understanding's efficiency, the proposed methodology introduces the Jaro Similarity (JS). The character-level alignment between strings is effectively measured by JS, depicting accurate entity linking.

Initially, to identify potential mentions (referring expressions) like pronouns, named entities, and noun phrases, the T_{NLP} is scanned.

$$T_{\text{NLP}} \xrightarrow[\text{identify}]{} M_{\infty} \tag{14}$$

The features, namely semantic type (gender, number, and entity type consistency), syntactic role (grammatical role or dependency relations), and contextual features (position information and sentence distance), are extracted from the identified mentions (M_{∞}). To track coreference relationships (i.e., CD) using JS (S_{mat}), the potential mention pairs are assessed by considering the extracted features.

$$S_{\text{mat}} = \frac{1}{3}\left(\frac{X''}{M_{\infty}\left(\hat{R}_1\right)} + \frac{X''}{M_{\infty}\left|\hat{R}_2\right|} + \frac{X'' - \tilde{j}}{X''} \right) \tag{15}$$

Here, the extracted features of M_{∞} are represented as $\hat{R}_1$ and $\hat{R}_2$, the number of matching references is indicated as X'', and the number of transpositions is notated as $\tilde{j}$. Afterward, the references mentioning the same entity are grouped regarding S_{mat} to form the coreference chains (C_{chain}), including CD. For instance: "Before he entered the room, Ravi knocked on the door". In this, the He-Ravi refers to the CD because "he" appears first, while "Ravi" is introduced later.

3.1.9 Contextual semantic encoding

Additionally, by employing WSBERT, contextual semantic vectors are obtained from T_{NLP} to reduce the semantic ambiguity. The Sentence Bidirectional Encoder Representations from Transformers (SBERT) produces meaningful vector representations for complex sentences. Also, fast similarity computation is facilitated by SBERT. Yet, feature learning is insufficient owing to the improper initialization of the feed-forward dimension, affecting the embedding capability. Thus, to determine the feed-forward dimension, the proposed framework uses the Weierstrass

Function (WF) by adaptively modeling feature variations, thereby ensuring enhanced embedding efficiency and feature learning.

Primarily, to track the order of the tokens and sentence boundaries, the special tokens like classification token $\left(c^{\mathrm{ls}}\right)$(CLS) and separation token $\left(s^{\mathrm{ep}}\right)$(SEP) are included at the start and end of the T_{NLP}.

$$\mathrm{T}'' = c^{\mathrm{ls}} \oplus T_{\mathrm{NLP}} \oplus s^{\mathrm{ep}} \tag{16}$$

Then, the embedded tokens $\left(\mathrm{T}''\right)$are subjected to the multiple layers of the encoder. Further, to understand the features in several dimensions, the multi-head self-attention mechanism $\left(A_{\mathrm{tt}}\right)$is applied.

$$A_{\mathrm{tt}}\left(\mathrm{T}''\right) = S^{\max}\left(\frac{\vec{w}\vec{k}^{T}}{\sqrt{d_{\mathrm{mn}}}}\right)\vec{v} \tag{17}$$

Where, the softmax activation function is represented as $S^{\max}$, the dimensions of the key vectors are notated as d_{mn}, the query (current token) is exemplified as $\vec{w}$, the key (available description) is indicated as $\vec{k}$, and the value (actual content) is signified as $\vec{v}$. Moreover, a feed-forward network $\left(N_{\mathrm{feed}}\right)$is utilized to apply the non-linear transformations for stable learning. Thereafter, by using WF, the feed-forward dimension $\left(F_{\mathrm{dim}}\right)$is determined.

$$N_{\mathrm{feed}}\left(F_{\mathrm{dim}}\right) = S^{\max} \times \left(\left(A_{\mathrm{tt}}, \mathrm{T}''\right) \cdot w_{\mathrm{gt}}\right) + b_{\mathrm{as}} \tag{18}$$

$$F_{\mathrm{dim}}\left(A_{\mathrm{tt}}\right) = \sum_{h'=0}^{\infty} a_1^{\,h'} \cos\left(b_1^{\,h'} \pi A_{\mathrm{tt}}\right) \tag{19}$$

Here, the weight and bias of the feed-forward network are signified as w_{gt} and b_{as}, correspondingly, the Weierstrass parameter is exemplified as h', and the integer values are indicated as $\left(a_1, b_1\right)$. The semantic relationships between embeddings are captured in the pre-training phase. Once training is completed, the embeddings are passed through the encoder for obtaining contextual vectors $\left(V_{\mathrm{enc}}\right)$.

3.1.10 Cross-modal alignment

After that, the O_z, C_{chain}, V_{enc}, and L_{xy} are fed into the cross-modal alignment in which the features from text and images are aligned into a shared representation space. Here, by utilizing GCLAN, cross-modal alignment is done. The strong feature distribution alignment and enhanced cross-modal representation learning are ensured by the GAN. Nevertheless, the existence of a conventional activation function results in gradient instability and degraded feature learning. Thus, to upgrade the feature learning stability, the proposed approach uses the Collapsing Linear Unit

(CLU) activation function by maintaining a controlled linear response while suppressing negative activations.

Primarily, from O_z, C_{chain}, V_{enc}, and L_{xy}, the modality-specific features are extracted for cross-modal feature representation. In the proposed work, the CLU activation function $\left(\mathrm{C}_{\text{act}}\right)$ is introduced to enhance the feature learning.

$$T_{\text{fx}} = \left(O_z, C_{chain}, V_{\text{enc}}\right) \tag{20}$$

$$\left(I^{\text{ed}}, T^{\text{ed}}\right) = \mathrm{C}_{\text{act}} \times \left(T_{\text{fx}}, L_{\text{xy}}\right) \tag{21}$$

$$\mathrm{C}_{\text{act}}\left(T_{\text{fx}}, L_{\text{xy}}\right) = \left(T_{\text{fx}}, L_{\text{xy}}\right) \times \frac{1}{1 - \left(T_{\text{fx}}, L_{\text{xy}}\right)\exp\left| - \left(\left(T_{\text{fx}}, L_{\text{xy}}\right) + \exp\left(\left(T_{\text{fx}}, L_{\text{xy}}\right)\right)\right)\right|} \tag{22}$$

Where, the textual features are exemplified as T_{fx}, the encoded vectors of the image are implied as I^{ed}, and the encoded vectors of the text are notated as T^{ed}. Thereafter, both modalities are projected into a common embedding space.

$$\left(L^I, L^T\right) \rightarrow P_{\text{mx}}\left(I^{\text{ed}}, T^{\text{ed}}\right) \tag{23}$$

Here, the projection matrix is represented as P_{mx} and the latent space projected image and text are denoted as $\left(L^I, L^T\right)$, correspondingly. Afterward, the modality transformations are learned by the generator network $\left(G_{\text{tor}}\right)$, thus producing aligned cross-modal representations.

$$\hat{E}_T = G_{\text{tor}}\left(L^I\right) \tag{24}$$

$$\hat{E}_I = G_{\text{tor}}\left(L^T\right) \tag{25}$$

Where, the generated text embedding from image embedding and the image embedding produced from text embedding are indicated as $\hat{E}_T$ and $\hat{E}_I$, respectively. Similarly, the discriminator $\left(D^{\text{tor}}\right)$ checks whether aligned features are a real pair or a generated pair.

$$D^{\text{tor}} \rightarrow \left|0,1\right| \tag{26}$$

Subsequently, to enhance the feature alignment reliability, the generator loss $\left(G_{\text{loss}}\right)$ and discriminator loss $\left(D_{\text{loss}}\right)$ are identified.

$$G_{\text{loss}} = -\log\left(D^{\text{tor}}\left(G_{\text{tor}}\left(\left(\hat{E}_T, \hat{E}_I\right)\right)\right)\right) \tag{27}$$

$$D_{\text{loss}} = -\left[\log D^{\text{tor}}\left(1\right) + \log\left(1 - D^{\text{tor}}\left(0\right)\right)\right] \tag{28}$$

In this, the discriminator maximizes discrimination accuracy between real and generated samples, whereas the generator tries to reduce loss. Here, $\left(C_{\text{align}}\right)$ signifies the cross-modal aligned feature representations. Finally, the generator synthesizes an image corresponding to the meaning of the text regarding the learned relationships.

The pseudo-code of the proposed GCLAN is provided as follows,

---

Input: O_z, C_{chain}, V_{enc} and L_{xy}

Output: Cross-modal aligned feature representations C_{align}

Begin

Initialize $O_z, C_{\text{chain}}, V_{\text{enc}}, L_{\text{xy}}$ and T_{fx}

For 1 to each input do,

#encoder

Extract text and image embedding vectors

$$\left(I^{\text{ed}}, T^{\text{ed}}\right) = \text{Encoder} \times \text{C}_{\text{act}}\left(T_{\text{fx}}, L_{\text{xy}}\right)$$

Activate CLU function $\text{C}_{\text{act}}\left(T_{\text{fx}}, L_{\text{xy}}\right)$

Evaluate latent space representation

$$\left(L^I, L^T\right) \rightarrow P_{\text{mx}}\left(I^{\text{ed}}, T^{\text{ed}}\right)$$

#generator

Learn modality transformation

$$\hat{E}_T = G_{\text{tor}}\left(L^I\right)$$

$$\hat{E}_I = G_{\text{tor}}\left(L^T\right)$$

Execute discriminator $D^{\text{tor}} \rightarrow |0,1|$

Compute generator and discriminator loss

Repeat until converge

End For

Return C_{align}

End

The user enters the text prompt in real-time for TIS. Here, the text query is fed into the NLP pre-processing, syntactic structure modeling, CD analysis, contextual word embedding, and feature extraction. After that, the images are generated by the proposed GCLAN concerning the learned cross-modal representation. Then, for reliable contextual content interpretation, the generated image $\left(I_{\text{gen}}\right)$is subjected to the image caption generation module.

3.2 Image caption generation module

An image caption generation module that involves data collection, pre-processing, and fine-grained visual characteristics learning is designed in the proposed work.

3.2.1 Data collection

Next, from publicly available sources, the images (U_{img}) are collected to train the image caption generation module. During testing, I_{gen} is given as the input for caption generation.

3.2.2 Pre-processing

Once the data are gathered, U_{img} is pre-processed for noise removal and contrast enhancement utilizing MF and CLAHE, correspondingly. Therefore, data quality is improved. In Section 3.1.2, the step-by-step process of the image pre-processing is explained.

3.2.3 Fine-grained visual characteristics learning

By using PAQN, the fine-grained visual characteristics (H_{gl}) are learned from pre-processed images. The step-by-step procedure of PAQN is explained in Section 3.1.3.

3.2.4 Caption generation

Then, to perform caption generation, H_{gl} is subjected to the proposed GQ2PHPT. The prevailing GPT captures contextual relationships between content and queries rather than depending only on keyword matching. Then, the relationships between words are maintained by GPT, thereby enhancing caption generation efficiency. Yet, owing to the insufficient attention mechanisms, the existing GPT may overlook local contextual dependencies and fine-grained information. Thus, in the proposed work, the Quadruple Attention (QA) mechanism is introduced for capturing the fine-grained contextual information. The multiple feature interactions and local-global dependencies are simultaneously captured by QA, ensuring improved multimodal alignment. Here, the QA aggregates contextual information across multiple attention branches, thereby facilitating a better understanding of long-range feature relationships. Furthermore, slow convergence and training instability are caused by the poor initialization of the learning rate. Hence, to determine the suitable learning rate, the proposed work employs the Probabilist's Hermite Polynomials (PHP) by adaptively modeling the optimization behavior and reducing convergence instability during training. Figure 2 depicts the architecture of the proposed GQ2PHPT.

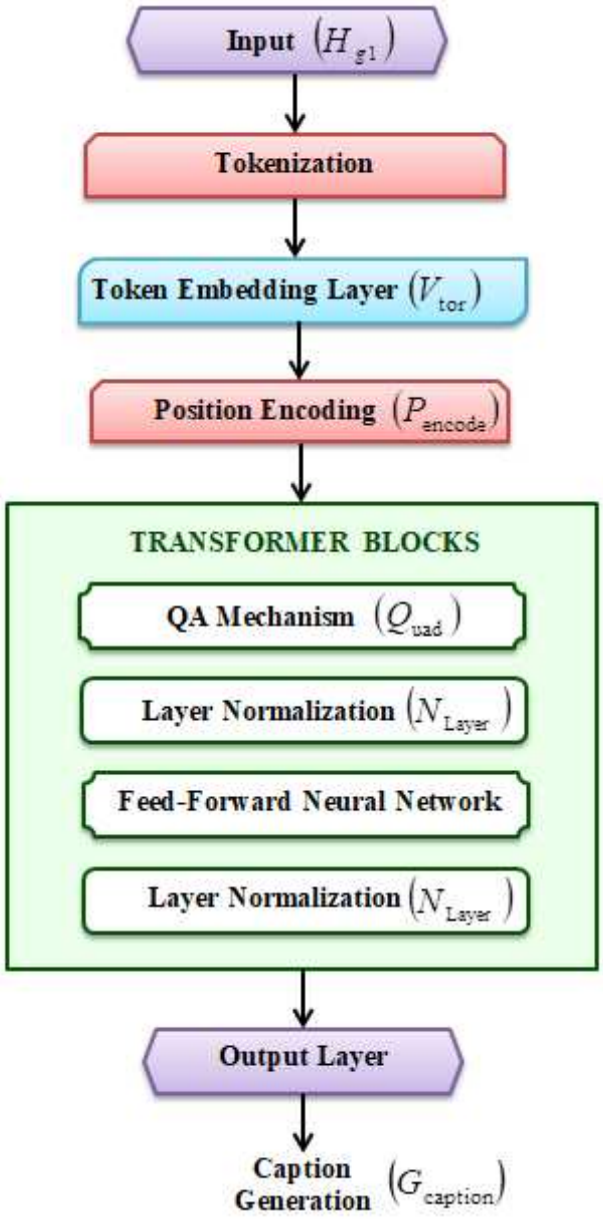


Figure 2: Design of the proposed GQ2PHPT

Initially, the H_{gl} is fed into the tokenization phase, where H_{gl} is split into several tokens. Further, every single token is transformed into a dense numerical vector (V_{tor}) that captures semantic meaning.

$$H_{gl} \xrightarrow{\text{convert}} V_{tor} \tag{29}$$

Subsequently, to include position information (I_{pos}) in the embeddings, the positional encoding is done, thus maintaining the order of tokens.

$$P_{\text{encode}} = \left(V_{\text{tor}} \oplus I_{\text{pos}}\right) \tag{30}$$

Where, the positional encoded outcome is exemplified as P_{encode} .

After that, the P_{encode} is inputted to the transformer block. A QA mechanism, layer normalization, and a feed-forward neural network are included in the transformer block. Also, the relationships between tokens are learned by the QA mechanism $\left(Q_{\text{uad}}\right)$, thus concentrating on the important features.

$$Q_{\text{uad}} = \frac{1}{4}\left(S_{\text{ig}}\left(\vec{H}\left(P_{\text{encode}}\right)\right) + S_{\text{ig}}\left(\vec{W}\left(P_{\text{encode}}\right)\right) + S_{\text{ig}}\left(\vec{J}\left(P_{\text{encode}}\right)\right) + S_{\text{ig}}\left(\vec{M}\left(P_{\text{encode}}\right)\right)\right) * P_{\text{encode}} \tag{31}$$

Here, the sigmoid activation function is notated as S_{ig}, the height-channel attention is indicated as $\vec{H}$, the width-channel attention is represented as $\vec{W}$, the height-width attention is denoted as $\vec{J}$, and the spatial attention is implied as $\vec{M}$.

Moreover, to improve the consistency of the model's training, layer normalization $\left(N_{\text{layer}}\right)$is applied.

$$N_{\text{layer}} = S_{\text{pt}} \frac{P_{\text{encode}}\left(Q_{\text{uad}}\right) - m_{\text{ean}}}{S_{\text{dev}}} + L^{\text{shift}} \tag{32}$$

Where, the learnable scaling parameter is signified as S_{pt}, the learnable shift parameter is notated as L^{shift} , the mean value of P_{encode} is indicated as m_{ean}, and the standard deviation of P_{encode} is exemplified as S_{dev}. Furthermore, to learn complex patterns, the feed-forward neural network processes each token independently. Then, for determining the learning rate $\left(L^{\text{rate}}\right)$, PHP is utilized.

$$L^{\text{rate}} = (-1)^{X_{\text{ab}}} \exp\left(\frac{{N_{\text{layer}}}^2}{2}\right) \frac{d''^{X_{\text{ab}}}}{d''{N_{\text{layer}}}^{X_{\text{ab}}}} \exp\left(-\frac{{N_{\text{layer}}}^2}{2}\right) \tag{33}$$

Here, the derivative factor is represented as d'' and the probabilistic hermite factor is denoted as X_{ab}. Finally, in the output layer, processed representations are converted into predicted tokens, thus generating corresponding captions $\left(G_{\text{caption}}\right)$. The pseudocode of the proposed GQ2PHPT is given as follows,

---

Input: Fine-grained visual characteristics H_{gl}

Output: Image captions G_{caption}

Begin

Initialize H_{gl} , I_{pos} and N_{layer}

For 1 to each input do,

Perform tokenization and vector embedding

$$H_{\text{gl}} \xrightarrow{\text{convert}} V_{\text{tor}}$$

Include positional information

$$P_{\text{encode}} = \left(V_{\text{tor}} \oplus I_{\text{pos}}\right)$$

#Transformer block

Apply QA mechanism Q_{uad}

Include layer normalization

$$N_{\text{layer}} = S_{\text{pt}} \frac{P_{\text{encode}}\left(Q_{\text{uad}}\right) - m_{\text{ean}}}{S_{\text{dev}}} + L^{\text{shift}}$$

Evaluate feed-forward network

Initialize learning rate using PHP L^{rate}

Implement output layer

Generate captions G_{caption}

End For

Return G_{caption}

End

3.2.5 Cross-modal feature alignment

Next, by employing the proposed GCLAN, the cross-modal feature alignment is carried out between I_{gen} and G_{caption} for enhancing the relevance of the content retrieval. In Section 3.1.10, the step-by-step process of the proposed GCLAN is described. Here, the aligned generated captions and generated images assist in preserving semantic correspondence betwixt textual and visual modalities. Furthermore, to obtain cross-modal aligned features $\left(A_{\text{feature}}\right)$, the features from the generated images and generated captions are projected into the shared space model representation.

3.3 Score computation

In this section, the alignment score value $\left(S_{\text{align}}\right)$is calculated for A_{feature} to measure the semantic similarity between cross-modal features and recognize the most relevant images from the database during retrieval. Here, the shared feature space representation of the image-text pairs is signified as A_{feature} . To compute the alignment score, the image embedding vector and text embedding vector are considered from the aligned feature space.

$$S_{\text{align}} = A_{\text{feature}} \cdot \frac{\hat{O}^{I} \times \hat{O}^{T}}{\left\|\hat{O}^{I}\right\| \left\|\hat{O}^{T}\right\|} \tag{34}$$

Where, the image embedding vector is implied as $\hat{O}^{I}$ and the text embedding vector is notated as $\hat{O}^{T}$.

After that, the S_{align} is fed into the content retrieval phase.

3.4 Content retrieval module

Thereafter, the content retrieval module is designed to retrieve relevant content regarding S_{align}, thus ensuring efficient semantic understanding and ML.

The database contents (image-text pairs) undergo cross-modal alignment and score computation during content retrieval training for establishing aligned multimodal representations. During retrieval, the system observes the current query score S_{align} and analogizes it with the learned database scores for taking retrieval actions.

Content retrieval – In this, to perform accurate content retrieval, the proposed Incomplete Polylogarithm-Reinforcement Learning (IP-RL) is utilized. By dynamically optimizing retrieval decisions concerning user feedback and relevance rewards, Reinforcement Learning (RL) improves content retrieval, thus enhancing retrieval accuracy. However, owing to the improper learning rate selection, the RL has unstable training outcomes. If the learning rate is too high, then exploding gradients take place. If the learning rate is too low, then convergence is very slow. In the proposed work, Incomplete Polylogarithm (IP) is used to determine the suitable learning rate. The IP evaluates the convergence behavior of the reward sequence. Also, the learning rate that reduces instability during policy updates is chosen by IP.

In the RL environment, the action indicates the retrieval tasks and the state signifies the query score. The RL agent continuously observes the environment and takes retrieval actions to choose semantically similar images. Then, for actions, the agent receives a reward $\left(r_{\text{ward}}\right)$. To obtain relevant content, the agent updates its learning policy concerning the reward. Moreover, for determining the learning rate $\left(B_{\text{rate}}\right)$, the IP is utilized.

$$\breve{K}\left(\hat{s},\hat{a}\right)=\dot{K}\left(\dot{s},\dot{a}\right)+B_{\text{rate}}\left[r_{\text{ward}}+D^{\text{fac}}\max_{\dot{a}}\dot{K}\left(\dot{s},\dot{a}\right)-\dot{K}\left(\dot{s},\dot{a}\right)\right] \tag{35}$$

$$B_{\text{rate}}=\frac{1}{I_{\text{ptr}}\left(x_{\text{va}}\right)}\int_{x_{\text{va}}}^{\infty}\frac{\dot{s}^{x_{\text{va}}-1}}{\exp\left(\dot{s}\right)/\dot{a}-1}d''\dot{s} \tag{36}$$

Here, the updated q-value with new state $\hat{s}$ and action $\hat{a}$ is implied as $\breve{K}\left(\hat{s},\hat{a}\right)$, the previous q-value with state $\dot{s}$ and action $\dot{a}$ is notated as $\dot{K}\left(\dot{s},\dot{a}\right)$, the polylogarithm parameter is exemplified as I_{ptr}, the constant value is denoted as x_{va}, and the discount factor is represented as D^{fac}. Overall, in the proposed work, adaptive and relevance-aware content retrieval is facilitated.

4. RESULTS AND DISCUSSION

In this phase, the proposed methodology's empirical validation is carried out for demonstrating its significance in cross-modal learning and generation tasks. Here, the proposed work is implemented on the PYTHON platform.

4.1 Dataset description

By using the publicly available dataset named the Flick 30k dataset (Adityajn, 2020), the proposed model's performance is assessed. The Flick 30k dataset encompasses 30,000 images paired with

five different captions. Significant descriptions of the entities and events are provided by the captions. The images included in the Flick 30k dataset cover a variety of scenes and situations, thereby ensuring the model's generalizability. Hence, the proposed modules, namely TIS, IC, and content retrieval, are efficiently assessed by the Flick 30k dataset. For every single phase, 80% (24000) and 20% (6000) of the data are used for training and testing, correspondingly.

Table 1: Sample image outcomes of the proposed framework

Phases	**TIS**		**IC**	
Samples	**Sample 1**	**Sample 2**	**Sample 1**	**Sample 2**
Input				
Noise removal				
Contrast enhancement				

The sample image outcomes of the proposed TIS module and IC phase concerning noise removal and contrast enhancement are presented in Table 1.

4.2 Performance validation for CD analysis

Here, a performance evaluation is performed for depicting the proposed CJSR's efficiency in CD analysis.

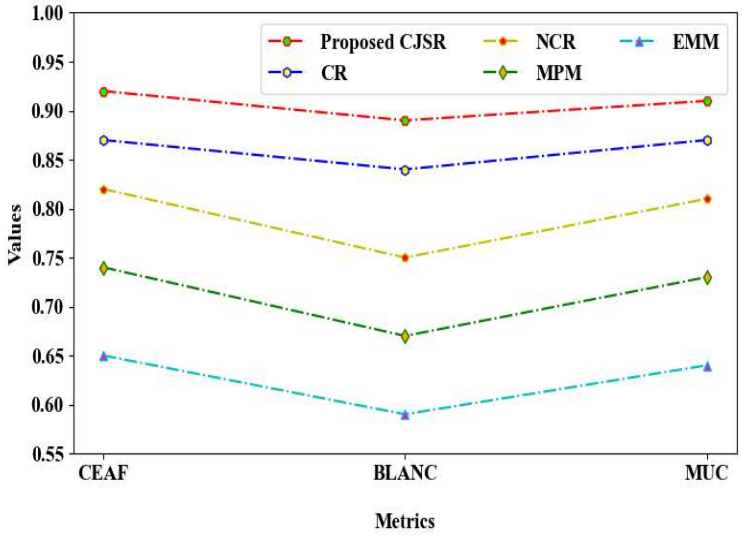


Figure 3: Robustness evaluation of the proposed CJSR and existing algorithms

In Figure 3, the proposed CJSR's performance is analogized with the prevailing algorithms, namely CR, Mention-Pair Model (MPM), Neural Coreference Resolution (NCR), and Entity-Mention Model (EMM), concerning CD analysis. To calculate the similarity between mentions, the proposed approach introduced JS, thus reducing poor entity linkage. Nevertheless, due to improper similarity matching, the prevailing CR struggled to effectively identify entity linkages. Here, Constrained Entity Alignment F-measure (CEAF) of 0.92, Mention Understanding Coreference (MUC) of 0.91, and BiLateral Assessment of Noun Coreference (BLANC) of 0.89 were attained by the proposed CJSR. However, the prevailing CR obtained minimum CEAF, BLANC, and MUC values of 0.87, 0.84, and 0.87, respectively. According to the investigation outcomes, the proposed algorithm had more robustness in CD analysis.

4.3 Experimental analysis for cross-modal alignment

In this phase, the performance evaluation of the proposed GCLAN and conventional techniques, such as GAN, DA-GAN, Deep Visual-Semantic Embedding (DeViSE) model, and Deep Canonical Correlation Analysis (DCCA), is done.

Table 2: Numerical analysis for cross-modal alignment

Techniques	Peak Signal-to-Noise Ratio (PSNR)	Maximum Mean	Structured Similarity Index Measure (SSIM)

		Discrepancy (MMD)	
Proposed GCLAN	45.78	0.042	0.964
GAN	38.22	0.095	0.925
DA-GAN	34.53	0.158	0.864
DCCA	28.49	0.274	0.814
DeViSE	26.34	0.357	0.754

Table 2 depicts the performance of the proposed GCLAN and conventional models for the cross-modal alignment concerning the PSNR, MMD, and SSIM. The proposed GCLAN achieved a PSNR of 45.78, an MMD of 0.042, and an SSIM of 0.964 by significantly enhancing the gradient stability utilizing the CLU activation function. However, the traditional GAN faced a gradient instability issue owing to the improper activation function, thereby leading to poor performance. Here, the PSNR of 28.49, MMD of 0.274, and SSIM of 0.814 were obtained by the prevailing DCCA. Moreover, due to the improper similarity calculation, the conventional DCCA attained limited outcomes in cross-modal alignment. Therefore, the effectiveness of the proposed model in cross-modal alignment was proved.

4.4 Numerical investigation for caption generation

Here, regarding numerous quantitative metrics like accuracy, Metric for Evaluation of Translation with Explicit ORdering (METEOR), BiLingual Evaluation Understudy (BLEU), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE), the performance evaluation for caption generation is carried out.

Table 3: Experimental analysis of the proposed GQ2PHPT

Techniques	ROUGE	BLEU	METEOR	Accuracy (%)	Precision (%)	Recall (%)
Proposed GQ2PHPT	0.934	0.922	0.912	98.9734	98.9564	98.9622
GPT	0.917	0.878	0.882	96.2892	96.8675	96.3943
Text-To-Text Transfer Transformer (T5)	0.864	0.784	0.724	93.5755	94.5655	94.2732

Retrieval-Augmented Generation (RAG)	0.828	0.728	0.703	90.4644	92.6565	92.5893
Retrieval-Enhanced Transformer (RET)	0.792	0.694	0.682	87.4644	89.5654	90.9433

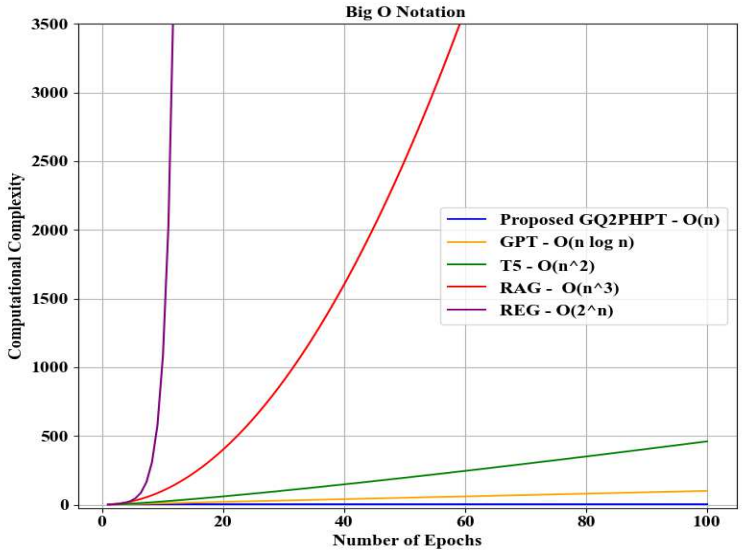


Figure 4: Complexity evaluation

Table 3 and Figure 4 validate the efficiency of the proposed GQ2PHPT by analogizing it with various prevailing algorithms, namely GPT, T5, RAG, and RET, concerning caption generation. Owing to the utilization of an efficient attention mechanism and learning rate initialization, the performance of the proposed GQ2PHPT was enhanced. Here, maximum ROUGE (0.934), BLEU (0.922), METEOR (0.912), accuracy (98.9734%), precision (98.9564%), and recall (98.9622%) were achieved by the proposed GQ2PHPT. Nevertheless, due to a lack of a sufficient attention mechanism, the traditional RET obtained limited performance. Also, an average ROUGE, BLEU, METEOR, accuracy, precision, and recall of 0.8502, 0.771, 0.7477, 91.9483%, 93.4137%, and 93.5500%, respectively, were attained by the conventional algorithms. As the number of epochs increased, the complexity of the proposed work was gradually increased. However, the traditional algorithms experienced sudden fluctuations in computational complexity. Furthermore, the prevailing algorithms lacked appropriate learning rate initialization, thereby leading to poor performance. Overall, the analysis results depicted that when contrasted with the prevailing methods, the proposed model had superior prominence in caption generation.

4.5 Efficiency assessment for content retrieval

Here, for evaluating the proposed IP-RL's performance in the content retrieval task, the performance metrics, such as cumulative reward, learning efficiency, and success rate, are taken into account.

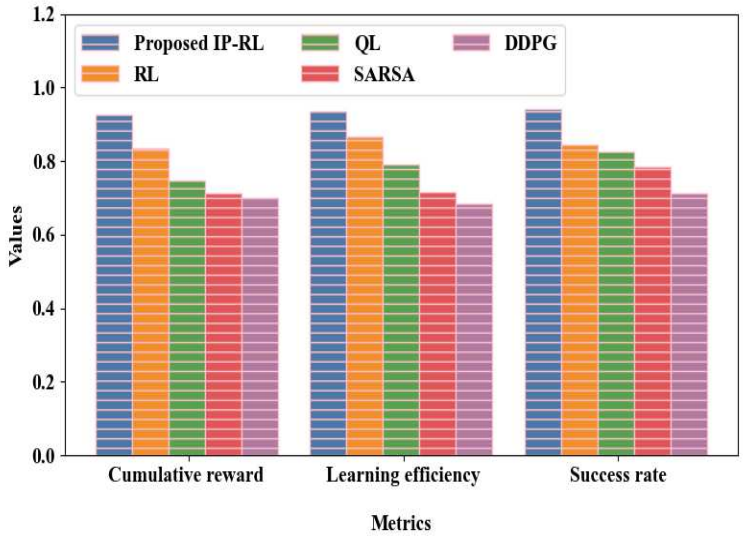


Figure 5: Performance validation for content retrieval

Figure 5 illustrates the performance validation of the proposed IP-RL and traditional models like RL, State-Action-Reward-State-Action (SARSA), Q-Learning (QL), and Deep Deterministic Policy Gradient (DDPG). In the proposed work, the IP-based learning rate selection was used for enhancing the significance of the proposed IP-RL in content retrieval. Here, the proposed IP-RL obtained maximum cumulative reward, learning efficiency, and success rate of 0.9262, 0.9345, and 0.9422, respectively. Nevertheless, cumulative reward of 0.7454, learning efficiency of 0.7895, and success rate of 0.8245 were attained by the traditional QL. The conventional RL and QL achieved poor robustness due to the improper learning rate and discount factor selection. Hence, the importance of the proposed model was proved.

4.6 Robustness evaluation for fine-grained visual characteristics learning

The performance analysis of the proposed PAQN in fine-grained visual characteristics learning is done in this section.

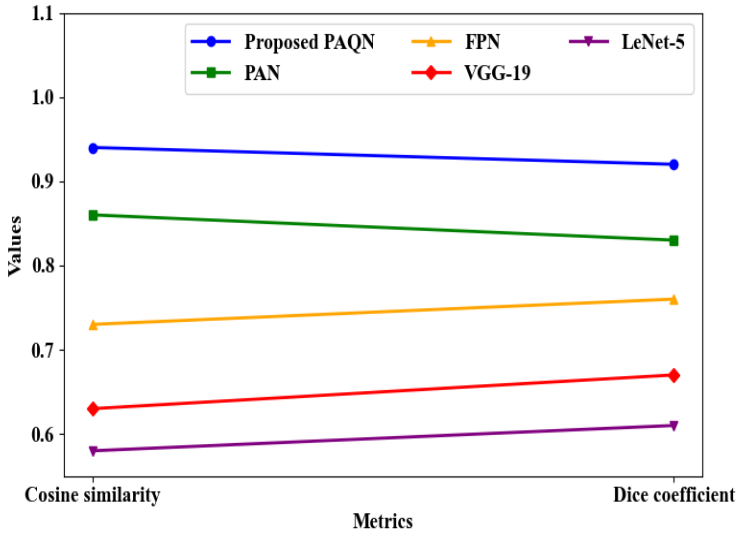


Figure 6: Experimental validation of the proposed PAQN

In Figure 6, the empirical evaluation of the proposed PAQN and conventional algorithms, namely PAN, Visual Geometry Group-19 (VGG-19), Feature Pyramid Network (FPN), and LeNet-5, is displayed. Here, a cosine similarity of 0.94 and a dice coefficient of 0.92 were obtained by the proposed PAQN. Nevertheless, the traditional methods, such as PAN, FPN, VGG-19, and LeNet-5, achieved cosine similarities of 0.86, 0.73, 0.63, and 0.58, correspondingly. Furthermore, an average dice coefficient of 0.7175 was attained by the conventional algorithms. The prevailing techniques had poor efficiency owing to the inappropriate layer parameter initialization. Therefore, due to the utilization of QR-based backbone depth selection, the proposed model had superior performance.

4.7 Experimental validation for contextual semantic encoding

In this section, an empirical assessment is carried out for evaluating the efficiency of the proposed WSBERT and conventional algorithms in contextual semantic encoding.

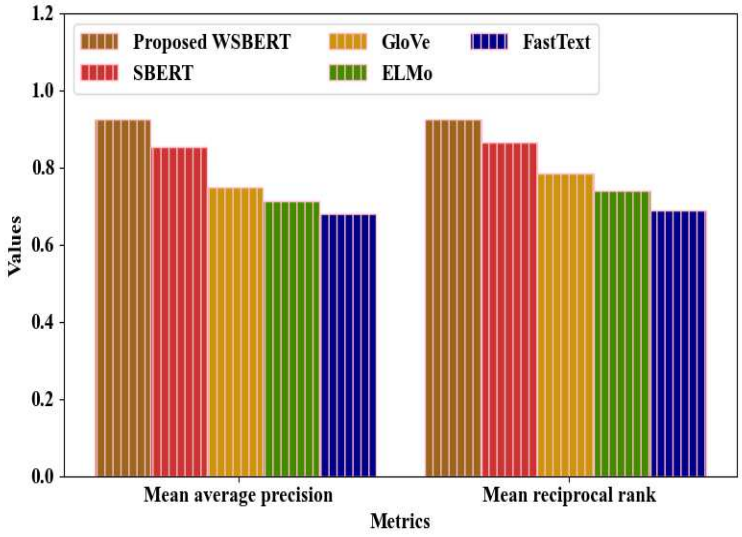


Figure 7: Performance validation of the proposed WSBERT and traditional algorithms

Figure 7 depicts the empirical evaluation of the proposed WSBERT and prevailing techniques like SBERT, FastText, Global Vectors (GloVe), and Embeddings from Language Models (ELMo). The proposed work used the WF to determine the feed-forward dimension for enhancing the performance of the proposed WSBERT in contextual semantic encoding. Yet, owing to the inappropriate feed-forward dimension selection, the prevailing SBERT had poor performance. The traditional FastText attained a mean average precision of 0.6783 and a mean reciprocal rank of 0.6893 due to the poor attention mechanism. In contrast, a mean average precision of 0.9244 and a mean reciprocal rank of 0.9247 were obtained by the proposed WSBERT. When contrasted with the existing methods, the proposed model effectively performed contextual semantic encoding.

4.8 Comparative validation of the proposed work

Here, to illustrate the proposed work's robustness in the cross-modal alignment-based IC task, a comparative analysis of the proposed framework is done.

Table 4: Comparative assessment of the proposed work

Authors' name	Algorithms	ROUGE	METEOR	BLEU
Proposed work	GQ2PHPT	0.93	0.91	0.92
(Abinaya et al., 2024)	VisualCaptionNet (VCN)	0.51	0.20	0.73
(Badarneh et al., 2025)	Ensemble-deep learning approach	0.43	0.23	0.72
(Arasi et al., 2023)	Improved Mobile Network Version 2	-	0.29	0.72

	(MobileNetv2) model			
(Xiao et al., 2022)	Attentional LSTM	0.58	0.28	0.78
(Pourkeshavarz et al., 2023)	Stacked cross-modal feature consolidation attention-LSTM	0.58	0.29	0.80

In Table 4, the comparative evaluation of the proposed work and related studies is provided. To capture the subtle visual properties, fine-grained visual characteristics learning was performed in the proposed work, thus offering deeper and more accurate captions. Here, maximum ROUGE (0.93), METEOR (0.91), and BLEU (0.92) were attained by the proposed GQ2PHPT. Likewise, in the prevailing works, numerous techniques, namely an improved MobileNetv2 model, attentional LSTM, ensemble-deep learning approach, and VCN, were employed for performing IC. The traditional VCN (Abinaya et al., 2024) achieved ROUGE of 0.51, METEOR of 0.20, and BLEU of 0.73 due to inadequate deeper visual feature understanding. Moreover, the prevailing attentional LSTM (Xiao et al., 2022) obtained minimum ROUGE, METEOR, and BLEU of 0.58, 0.28, and 0.78, correspondingly. This prevailing work exhibited poor performance in caption generation owing to the lack of a traditional activation function. According to the evaluation outcomes, the proposed work had superior efficiency in captioning.

5. CONCLUSION

An intelligent cross-modal alignment with CD modeling-centric TIS and IC framework utilizing GCLAN and GQ2PHPT was proposed in this article. The TIS module, an IC module, and a content retrieval phase were included in the proposed work. Here, by using the Flick–30k dataset, the performance of the proposed work was assessed. To perform cross-modal alignment, the proposed GCLAN was utilized, thus improving the efficiency of the ML. Also, the proposed GCLAN attained the maximum PSNR (45.78) for cross-modal alignment. In the proposed work, the relevance of the cross-model retrieval and generation tasks was enhanced by tracking the CD in the sentences. Here, the proposed GQ2PHPT was generalized well enough to produce appropriate captions for the corresponding images, thereby ensuring superior ML and visual understanding. Moreover, in caption generation, the proposed GQ2PHPT achieved an accuracy of 98.9734% and a BLEU of 0.922. The proposed IP-RL was utilized for retrieving the relevant content. Next, by employing the proposed IP-RL, semantically similar images were significantly retrieved regarding the learned alignment scores with a maximum success rate of 0.9422. Also, to further improve the significance of the cross-modal learning and retrieval, the proposed methodology included fine-grained visual characteristics learning, syntactic structure modeling, and contextual semantic encoding. Overall, in content retrieval and caption generation, the proposed work performed better than the prevailing works.

Practical implications: A real-time multimodal search engine, text-to-image generation framework, CMR scheme, and image-to-text captioning systems are supported by the proposed CD modeling-based cross-model retrieval and captioning framework.

Future scope: This work will concentrate on incorporating video-text ML and temporal cross-modal understanding in the future for ensuring better generalizability and scalability. Moreover, to prove the trustworthiness of the model, explainable AI will be facilitated.

Funding Declaration:- This research was supported and Funded by Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune

Data Availability Statement:- Data Availability Statement: The data supporting the findings of this study are publicly available in the Flickr30k dataset (https://www.kaggle.com/datasets/adityajn105/flickr30k). No new datasets were created during this study.

REFERENCES:

Abinaya, S., Deepak, M., & Alphonse, A. S. (2024). Enhanced Image Captioning Using Bahdanau Attention Mechanism and Heuristic Beam Search Algorithm. *IEEE Access*, *12*, 100991–101003. https://doi.org/10.1109/ACCESS.2024.3431091

Adityajn, (2020). Flick 30k Dataset. Kaggle. https://www.kaggle.com/datasets/adityajn105/flickr30k

Alhabeeb, S. K., & Al-shargabi, A. A. (2024). Text-to-Image Synthesis With Generative Models: Methods, Datasets , Performance Metrics, Challenges, and Future Direction. *IEEE Access*, *12*, 24412–24427. https://doi.org/10.1109/ACCESS.2024.3365043

Alnashwan, R. O., Chelloug, S. A., Almalki, N. S., Issaoui, I., Motwakel, A., & Sayed, A. (2023). Lighting Search Algorithm With Convolutional Neural Network-Based Image Captioning System for Natural Language Processing. *IEEE Access*, *11*, 142643–142651. https://doi.org/10.1109/ACCESS.2023.3342703

Arasi, M. A., Alshahrani, H. M., Alruwais, N., Motwakel, A., Ahmed, N. A., & Mohamed, A. (2023). Automated Image Captioning Using Sparrow Search Algorithm With Improved Deep Learning Model. *IEEE Access*, *11*, 104633–104642. https://doi.org/10.1109/ACCESS.2023.3317276

Badarneh, I. Al, Hammo, B. H., & Al-kadi, O. (2025). An Ensemble Model with Attention Based Mechanism for Image Captioning. *Computers and Electrical Engineering*, 1–35. https://arxiv.org/pdf/2501.14828

Cai, L., Zhu, L., Zhang, H., & Zhu, X. (2022). DA-GAN : Dual Attention Generative Adversarial Network for Cross-Modal Retrieval. *Future Internet*, *14*, 1–23. https://doi.org/10.3390/fi1402004

Habib, A., Wadud, A. H., Patwary, F. K., Rahman, M. M., Mridha, M. F., Okuyama, Y., & Shin, J. (2024). Exploring Progress in Text-to-Image Synthesis : An In-Depth Survey on the Evolution of Generative Adversarial Networks. *IEEE Access*, *12*, 178401–178440. https://doi.org/10.1109/ACCESS.2024.3435541

Habib, A., Wadud, A. H., Pinky, L. Y., Talukder, M. H., Rahman, M. M., Mridha, M. F., Okuyama, Y., & Shin, J. (2024a). GACnet-Text-to-Image Synthesis With Generative Models Using Attention Mechanisms With Contrastive Learning. *IEEE Access*, *12*, 9572–9585. https://doi.org/10.1109/ACCESS.2023.3342866

Han, X., Wu, Z., Li, Y., Zhang, X., Wang, G., & Hou, B. (2026). CSSA: A Cross-Modal Spatial–Semantic Alignment Framework for Remote Sensing Image Captioning. *Remote Sensing*, *18*(3), 1-22. https://doi.org/10.3390/rs18030522

Han, Z., Azman, A. Bin, Mustaffa, M. R. B., & Khalid, F. B. (2024). Cross-Modal Retrieval : A Review of Methodologies, Datasets, and Future Perspectives. *IEEE Access*, *12*, 115716–115741. https://doi.org/10.1109/ACCESS.2024.3444817

Haoran, S., Yang, W., Haipeng, L., & Biao, Q. (2023). Fine-Grained Cross-Modal Fusion Based Refinement for Text-to-Image Synthesis. *Chinese Journal of Electronics*, *32*(6), 1–12. https://doi.org/10.23919/cje.2022.00.227

Ku, H., & Lee, M. (2023). TextControlGAN : Text-to-Image Synthesis with Controllable Generative Adversarial Networks. *Applied Sciences*, *13*, 1–12. https://doi.org/10.3390/app13085098

Lerner, P., Ferret, O., & Guinaudeau, C. (2021). Cross-modal Retrieval for Knowledge-based Visual Question Answering. *European Conference on Information Retrieval*, 1–19. https://arxiv.org/pdf/2401.05736

Li, K., Zhang, Y., Wang, F., Liu, G., & Wei, X. (2024a). Deep Feature-Based Neighbor Similarity Hashing With Adversarial Learning for Cross-Modal Retrieval. *IEEE Access*, *12*, 128559–128569. https://doi.org/10.1109/ACCESS.2024.3413186

Li, Z., Lu, H., Fu, H., Meng, F., & Gu, G. (2025). Csan : cross-coupled semantic adversarial network for cross- modal retrieval. *Artificial Intelligence Review*, *58*, 1–27. https://doi.org/10.1007/s10462-025-11152-7

Li, Z., Lu, H., Fu, H., Wang, Z., & Gu, G. (2023). Adaptive Adversarial Learning based cross-modal retrieval. *Engineering Applications of Artificial Intelligence*, *123*, 1–12. https://doi.org/10.1016/j.engappai.2023.106439

Li, Z., Zhao, W., Du, X., Zhou, G., & Zhang, S. (2024). Cross-Modal Retrieval and Semantic Refinement for Remote Sensing Image Captioning. *Remote Sensing*, *16*, 1–20. https://doi.org/10.3390/rs16010196

Pang, H., Wei, S., Zhang, G., Zhang, S., Qiu, S., & Zhao, Y. (2023). Heterogeneous Feature Alignment and Fusion in Cross-Modal Augmented Space for Composed Image Retrieval. *IEEE Transactions on Multimedia*, *25*, 6446–6457. https://doi.org/10.1109/TMM.2022.3208742

Pourkeshavarz, M., Nabavi, S., Moghaddam, M. E., & Shamsfard, M. (2023). Stacked Cross-modal Feature Consolidation Attention Networks for Image Captioning. *Multimedia Tools and Applications*, *83*, 1–21. https://arxiv.org/pdf/2302.04676

Rahman, M., & Sami, S. I. (2024). A comprehensive construction of deep neural network-based encoder–decoder framework for automatic image captioning. *IET Image Processing*, *18*(14), 4778–4798. https://doi.org/10.1049/ipr2.13287

Sogi, N., Shibata, T., & Terao, M. (2024). Object-Aware Query Perturbation for Cross-Modal Image-Text Retrieval. *European Conference on Computer Vision*, 1–25. https://arxiv.org/pdf/2407.12346?

Tan, Y. X., Lee, C. P., Neo, M., Lim, K. M., Lim, J. Y., & Alqahtani, A. (2023). Recent Advances in Text-to-Image Synthesis: Approaches, Datasets and Future Research Prospects. *IEEE Access*, *11*, 88099–88115. https://doi.org/10.1109/ACCESS.2023.3306422

Wang, T., Li, F., Zhu, L., Li, J., Zhang, Z., & Shen, H. T. (2024). Cross-Modal Retrieval : A Systematic Review of Methods and Future Directions. *Proceedings of the IEEE*, *112*(11), 1–35. https://doi.org/10.1109/JPROC.2024.3525147

Wang, Y., Li, D., Liu, Q., Liu, L., & Wang, G. (2024a). Self-supervised modal optimization transformer for image captioning. *Neural Computing and Applications*, *36*(31), 19863–19878. https://doi.org/10.1007/s00521-024-10211-4

Xiao, F., Xue, W., Shen, Y., & Gao, X. (2022). A New Attention-Based LSTM for Image Captioning. *Neural Processing Letters*, *54*(4), 3157–3171. https://doi.org/10.1007/s11063-022-10759-z

Yadav, S. S., Reddy, S. S. M., Divakarachari, P. B., Penmatsa, R. K. V., & Praveena, H. D. (2026). BCMT: Bidirectional Cross-Modality Transformer for Remote Sensing Image Captioning. *IEEE Access, 14*. 12560 – 12578. https://doi.org/10.1109/ACCESS.2026.3652136

Zeng, Z., He, S., Zhang, Y., & Mao, W. (2025). A multimodal embedding transfer approach for consistent and selective learning processes in cross-modal retrieval. *Information Sciences*, *704*, 1–14. https://doi.org/10.1016/j.ins.2025.121974

Zhang, B., Zhang, Y., Li, J., Chen, J., Akutsu, T., Cheung, Y.-M., & Cai, H. (2025). Unsupervised Dual Deep Hashing With Semantic-Index and Content-Code for Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(1), 387–399. https://doi.org/10.1109/TPAMI.2024.3467130

Zhang, J., Xie, Y., Ding, W., & Wang, Z. (2023a). Cross on Cross Attention : Deep Fusion Transformer for Image Captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(8), 4257–4268. https://doi.org/10.1109/TCSVT.2023.3243725

Zhou, K. U. N., Hassan, F. H., & Hoon, G. K. (2023). The State of the Art for Cross-Modal Retrieval: A Survey. *IEEE Access*, *11*, 1–22. https://doi.org/10.1109/ACCESS.2023.3338548