MeL-ReHS^{2}LSTM: A Multimodal Semantic Retrieval Framework with Fuzzy Ranking and Lightweight Caption Parsing

Shaik Asha
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, India
skayesha92@gmail.com

Sajja Tulasi Krishna
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation, India
tulasisajja788@kluniversity.in

Abstract: *Multimodal retrieval systems are essential for deriving semantic insights from visual and textual data. Achieving precise alignment between image captions and text queries remains a non-trivial challenge. Conventional methods such as Long Short-Term Memory (LSTM)-based or transformer-based encoders often struggle to balance efficiency, semantic accuracy, and interpretability. These approaches typically lack robust mechanisms for handling the fuzzy nature of human language. This study proposes a lightweight and interpretable framework that combines Bootstrapped Language Image Pretraining (BLIP)-generated captions with embeddings using Meta-Learning Rectified Hyperbolic Secant Group Norm-Lasso-based Long Short-Term Memory networks (MeL-ReHS^{2}LSTM) and uses Neutrosophic Cosine Similarity-based Fuzzy Inference System (NCS-FIS) to improve semantic retrieval. The architecture uses Parentheses Parsing Abstract Syntax Tree (PP-AST) for the symbolic structuring of captions and Dilated Skip-Connection Graph Neural Network (DSC-GNN) for the enhanced context embeddings to provide the ability to perform both neural and rule-based analysis in parallel. A custom dataset of 20 real-world video clips was created for evaluation, including a variety of scenes such as interviews, team meetings, and sports events. Empirical results show that the accuracy of top-1 is 92.0%, top-5 recall is 96.7%, Bilingual Evaluation Understudy (BLEU) score is 0.7521, and the average matching time is 0.58ms, which overcomes the traditional retrieval baselines. The proposed methodology provides a good trade-off between accuracy, runtime performance and interpretability. Future work will explore the extension of the model to zero-shot domains and the combination of adaptive caption refinement strategies*.

Keywords: *Caption retrieval, deep learning, fuzzy inference system, image captioning, multimodal dataset, semantic search, string matching*.

Received August 05, 2025; accepted May 09, 2026
https://doi.org/10.34028/iajit/23/5/4

1. Introduction

Multimodal Information Retrieval (IR) has received a lot of attention due to the availability of video and image content on digital platforms. Integrating the vision and language modalities enables systems to produce semantic understanding from visual inputs and to retrieve relevant content from user queries. Recent advances in image captioning and language modeling have paved the way for intelligent systems that are able to associate images or videos with descriptive text.

Despite these advances, there are still a number of challenges to attain efficient and accurate caption-based retrieval. The conventional methods depend only on string matching or keyword indexing, which don't have any semantic similarity between the caption and the query which is contextual in nature. Additionally, in real time retrieval systems, it is not only a high accuracy requirement but also computational efficiency, so it is hard to implement deep learning models without latency trade-offs.

Traditional retrieval pipelines usually have been based on rule-based or bag-of-words models for textual representation, which do not have deep contextual understanding. Although techniques based on Term Frequency-Inverse Document Frequency (TF-IDF) or simple Long Short-Term Memory (LSTM) architectures have enhanced the retrieval abilities, they tend to ignore the syntactic structure and struggle against high dimensional and ambiguous caption data. Luo *et al*. [8] proposed graph-based models and transformer-based embeddings solve some of these problems, but are usually computationally expensive and less interpretable when used in downstream tasks.

The proposed solution to these challenges introduces a new semantic retrieval system which utilizes a bootstrapped captioning model Bootstrapped Language Image Pretraining (BLIP) and Meta-Learning Rectified Hyperbolic Secant Group Norm-Lasso-based Long Short-Term Memory networks (MeL-ReHS^{2}LSTM) based caption embedding module and Neutrosophic Cosine Similarity based Fuzzy Inference System (NCS-FIS) based fuzzy reasoning layer. The hybrid architectural system provides explainable user query matching which operates first to match image-based captions and then to deliver results to users. The system achieves better semantic understanding of captions

through its ability to display caption embeddings with clustering methods and visualization techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) and K-Means. The key contributions of the present paper are as follows:

1. A hybrid multimodal retrieval pipeline based on a combination of image captioning, semantic embedding and fuzzy reasoning for effective caption-based search.
2. A novel MeL-ReHS²LSTM model based on ReLU(sech(x)) activation and group norm-based optimization for robust semantic understanding.
3. Use of NCS-FIS for multi-dimensional similarity scoring with truth, indeterminacy and falsity measures for better interpretability in ranking.
4. Clustering analysis of caption embeddings with K-Means and t-SNE for visualization of semantic spaces and latent topic modelling
5. A regex-based Parentheses Parsing Abstract Syntax Tree (PP-AST) indexing approach to support lightweight parsing and symbolic queries of generated captions.

The rest of the paper is structured as follows: section 2 is a review of related studies. Section 3 describes the architecture and methodology of the proposed system. Experimental evaluations in terms of captioning metrics, clustering, retrieval performance, and comparative benchmarks are presented in section 4. Section 5 presents key findings, limitations and future directions. Finally, section 6 concludes the paper with some final observations and outcomes.

2. Related Works

Recent advances in multimodal retrieval and caption-guided image understanding have led to multiple model architecture and learning strategies. Long *et al*. [7] introduced a zero-shot retrieval framework based on diffusion augmented representations, which is fine-tuned-free interactive text-to-image matching, but their approach is not efficient in terms of computational complexity during inference. Luo *et al*. [8] proposed a graph-based knowledge distillation method to solve the problem of cross-dataset generalization in person retrieval, but it is highly dependent on accurate cross-domain correspondence. Lyu *et al*. [9] proposed Tripartite Learning with Semantic Variation Consistency (TSVC), a robust image-text retrieval framework designed to address noisy correspondence through tripartite cooperative learning and semantic variation-based soft-label estimation. In work by Patel *et al*. [11], Large Language Models (LLMs) were combined with models for vision-language models for domain-adaptive captioning, however, performance in the real world depends on the quality of curated alt-text. Qian and Liu [13] improved the Contrastive Language-Image Pre-training (CLIP)-based retrieval by fusing You Only Look Once v10 (YOLOv10) and Next Vision Transformer (Next-ViT) for better object detection and fine-grained feature extraction, although the training cost has been increased significantly. Qin *et al*. [14] used graph learning for aligning egocentric and exocentric video modalities with promising alignment but scalability issues. Yuan and Xue *et al*. [19] proposed a hierarchical Graph Neural Network (GNN)-based multimodal fusion system for semantic retrieval which is limited by training instability under the noisy modality. A comprehensive survey by Li *et al*. [4] reviewed explainability in GNNs, which is necessary for transparent decision making in multimodal Artificial Intelligence (AI), but most techniques do not yet have quantitative validation of interpretability. Lee *et al*. [2] used GNNs with vision-language alignment for Alzheimer's detection using neuroimaging and speech, indicating clinical applicability but requiring larger datasets for generalization. Yang and Tan [18] proposed a multimodal knowledge graph for Question Answering (QA) tasks that combines image, text, and audio, but failed to address temporal coherence between media. Tan *et al*. [15] showed GAN-based multimodal synthesis with text and style inputs, which can be used for data augmentation, but lacked semantic control when generating. Patil *et al*. [12] discussed general industrial use cases of multimodal AI with a focus on interpretability, but the study did not have depth in terms of model-specific performance measures. Li *et al*. [3] proposed a pre-trained model fusion approach for robust image-text retrieval with high ensemble accuracy but with redundancy and increased inference cost. Wei *et al*. [17] proposed a multiscale hybrid model that combines global and local features for remote sensing image retrieval; however, feature alignment is sensitive to scale inconsistencies. Kumar *et al*. [1] focused on attention-driven relational captioning for visually rich scenes but faced problems with long-range dependencies. Zhang *et al*. [20] proposed the Cross-Modal Prominent Fragments Enhancement Aligning Network (CPFEAN) to reduce semantic noise through selective fragment alignment, but limited fragment diversity may affect generalization. Zheng *et al*. [21] proposed a dual-direction optimization strategy for improved cross-modal retrieval with high alignment at the expense of slower training. Wang *et al*. [16] used a combination of cross-modal attention and dense feature fusion in deep networks to obtain better alignment, but the model size is a concern. Long *et al*. [6] proposed multiway-adapter for low-resource tuning of multimodal LLMs, which is efficient but not robust in unseen domains. Finally, Liu *et al*. [5] reduced modality gaps with the use of synthetic caption-image pairs, but realism in synthetic samples is a challenge. Osman *et al*. [10] proposed Arabic image Captioning based on Concept Model and Vision-based Multi-Encoder Transformer Architecture (Ar-CM-ViMETA), a concept-aware image-captioning framework that

employs a vision-based multi-encoder transformer architecture to capture contextual information from visual content and generate semantically meaningful descriptions. Their results demonstrated that integrating conceptual information with transformer-based visual representations can improve caption quality according to Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) measures. However, the method primarily focuses on caption generation rather than integrating caption representation with semantic retrieval, fuzzy relevance estimation, and interpretable ranking, which are addressed jointly in the present framework.

2.1. Problem Statement

Traditional approaches to image captioning and multimodal retrieval usually depend on static representations that are hard to get along with semantic variability, noisy data and context ambiguity. These limitations are especially apparent in real-world situations where the dynamic visual content is involved, such as sports videos, interviews, and instructional videos, where object relationships, actions, and scene contexts are significantly different from each frame. Despite the progress in the transformer-based models and pretrained vision-language models, there is still a gap between the syntactic caption generation and the semantic alignment that is required for downstream tasks, such as clustering and IR. Existing solutions mostly focus on captioning or retrieval as a stand-alone task and do not have an end-to-end pipeline that unifies caption generation, semantic embedding and retrieval mechanisms that are interpretability driven. Furthermore, conventional cosine similarity or attention-based retrieval methods are unable to deal with fuzzy semantics or partial matches, or with user-subjective relevance, which is a key issue in interactive systems. While GNNs have shown promise in aligning multimodal data, most models do not include dilated context propagation and interpretability-enhancing syntax-aware indexing. To address these gaps, the proposed framework comprises a unified pipeline which starts from caption generation using BLIP model and continues through feature extraction using Bidirectional Encoder Representations from Transformers (BERT), embedding using LSTM, Dilated Skip-Connection Graph Neural Network (DSC-GNN) refinement and retrieval using fuzzy neutrosophic inference engine. This end-to-end design involves the capture of contextual semantics, the support of theme clustering through K-Means and the support of symbolic string matching through PP-AST indexing-this addresses both the granularity of the textual content and the interpretability of the query results. The novel addition of MeL-ReHS^{2}LSTM model further improves the discriminability of features and ensures robust alignment between visual inputs and the semantic representations.

3. Research Methodology

The overall workflow of the proposed multimodal IR system is discussed in this section. It involves stages such as data preparation, caption generation and feature extraction. Also, this includes semantic embedding, clustering and fuzzy based retrieval. The framework with each module makes a strong, interpretable, real-time system for image retrieval and classification based on caption. The complete system workflow is presented in Figure 1.

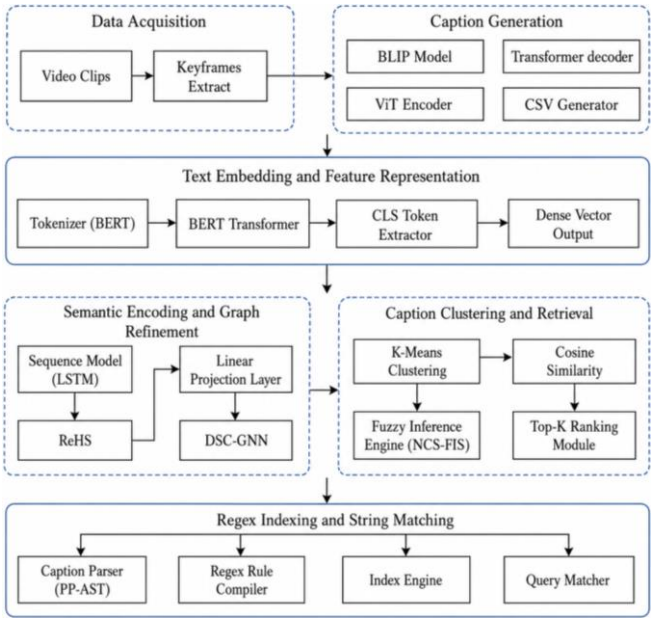

Figure 1. Proposed multimodal IR system architecture.

3.1. Data Acquisition and Preprocessing

The dataset contains 20.mp4 video clips covering different situations in the semantic domain such as interviews, team discussions and sports events. In order to break the stream of continuous video frames into units which can be analyzed, keyframes are extracted at regular intervals (every 30th frame) in effect reducing redundancy while maintaining temporal representativeness. The extracted keyframes are saved in the jpg format and are organized into directories according to their source videos so that the contextual consistency is maintained for subsequent processing steps. A series of preprocessing steps are performed in order to standardize the visual inputs before passing them into the model. All the images are resized to constant resolution of 1280x720 pixels to ensure uniform input dimensions, which avoids shape mismatch during batch-based deep learning training. The min-max scaling method transforms pixel intensities into a range between 0 and 1, which accelerates model training. The Laplacian focus measure filtering process eliminates both blurred and low-contrast frames, which allows the system to retain

only high-quality visual content. The preprocessing pipeline establishes three essential elements, which include consistency and semantic integrity and visual quality, to support caption generation and semantic retrieval.

3.2. Caption Generation with Vision-Language Models

The system uses the BLIP model to create natural language descriptions from visual material by generating captions for all selected keyframes. BLIP employs a vision-language transformer architecture which allows it to develop combined visual and textual representation systems. The ViT transforms each visual frame into visual tokens through the process of generating $H=\{h_1, h_2, ..., h_n\}$ which describes the visual elements. The transformer-based decoder creates text output through its process of using visual embeddings and previous decoded tokens to build the caption sequence $T=\{t_1, t_2, ..., t_k\}$. The training process aims to optimize the log-likelihood of the predicted tokens conditioned on the image features and prior context:

$$\mathcal{L}_{VL} = -\sum_{j=1}^{k} log\, P(t_j | t_{<j}, F) \quad (1)$$

This allows BLIP to successfully capture object-centric features and scene-level semantics with a high level of fidelity. The resulting descriptions encode spatial arrangements, actions, and contextual cues, to form linguistically rich representations ideal for semantic embedding and retrieval tasks.

3.3. Feature Extraction and Text Embedding

Following caption generation, each textual description is transformed into a dense and context-aware vector representation by using the BERT architecture. BERT employs transformer-based architecture trained in masked language modeling and next sentence prediction, allowing it to learn deep bidirectional context across entire input sequences. Each caption $C=\{c_1, c_2, ..., c_m\}$ is first tokenized into subwords, which are mapped to embeddings via a lookup table. These are then passed through multiple transformer layers to obtain contextualized token embeddings $E=\{e_1, e_2, ..., e_m\}$, where each $e_i \in R^{768}$ for the base model.

To produce a fixed-length sentence embedding, two strategies are typically employed:

1. Averaging the token embeddings.
2. Extracting the special [CLS] token representation.

In this framework, the [CLS] embedding z_{cls} is chosen to represent the entire caption:

$$z = BERT_{CLS}(C) \in R^{768} \quad (2)$$

The embedding model effectively preserves the semantic, syntactic, and contextual nuances of captions, making it valuable for downstream tasks such as clustering, classification, and semantic similarity analysis. The resulting embedding vectors are input to the semantic encoder (MeL-ReHS²LSTM) and are used as a shared representation for both clustering and retrieval modules.

3.4. Semantic Embedding via MeL-ReHS²LSTM

The semantic representation pipeline operates through the MeL-ReHS²LSTM model which combines three components: a BERT-based word encoder, an LSTM layer and a special non-linear activation function. The LSTM module updates its hidden state h_t at each time step t through a specific formulation.

$$h_t = LSTM(x_t, h_{t-1}) \quad (3)$$

where x_t is the input embedding at time step t. The final hidden state h_T is then projected through a fully connected layer to derive a fixed-length semantic vector:

$$z = W h_T + b \quad (4)$$

The custom activation function Rectified Hyperbolic Secant (ReHS) is used in this system to provide non-linear behavior and maintain consistent gradient flow through the network.

$$ReHS(x) = max(0, sech(x)) \quad (5)$$

where $sech(x) = \frac{2}{e^x + e^{-x}}$ is the hyperbolic secant function. This yields the final feature vector for each caption.

3.5. Embedding Refinement Using DSC-GNN

The semantic embeddings are further improved by DSC-GNN. A graph is created in which each node is a caption and edges are semantic similarity. The DSC-GNN updates information over dilated neighbors according to the rule:

$$h_v^{(l+1)} = \sigma\left(\sum_{u \in N(v)} W^{(l)} h_u^{(l)} + D^{(l)} h_v^{(l)} \right) \quad (6)$$

Here, $N(v)$ is the set of neighbors of node v, W^l and D^l are trainable weights and σ is a non-linear activation function. This process helps in improving the contextual representation and capturing long-range dependencies in the caption graph.

3.6. Caption Clustering Using K-Means

In order to detect latent semantic themes among the representations of captions, K-Means clustering is applied to the final embedding vectors from the semantic encoder. This unsupervised learning algorithm divides the embedding space into K clusters by alternately updating the cluster centroids and

minimizing the within-cluster sum of squares. The objective function to optimise this process is defined as:

$$arg\min_{C} \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \quad (7)$$

where C_k is the set of vectors assigned to cluster k and μ_k is the centroid of that cluster. The initialization of centroids is done by using the K-Means++ strategy to improve the stability of convergence and to reduce the problem of local minima. In the process of training, each caption embedding is assigned to the nearest centroid according to the Euclidean distance. The iterative update process is repeated until convergence is reached or a set number of iterations is reached. This clustering process allows semantic grouping of captions, which reveals dominant themes and allows for topic-aware indexing and summarization in the multimodal retrieval system.

3.7. Semantic Retrieval with NCS-FIS

The NCS-FIS is used to assess the relevance between a user query and embedded captions in the retrieval space. First, the cosine similarity is calculated between the query vector and candidate embeddings in order to measure the directionality in the high dimensional space:

$$cos\ Sim(a,b) = \frac{a.b}{\|a\|\|b\|} \quad (8)$$

This similarity score is then converted into a fuzzy inference context, in which three neutrosophic components, namely Truth (T), Indeterminacy (I) and Falsity (F) are derived by using a set of fuzzy membership functions. These components capture the degree of semantic relevance, ambiguity and deviation, respectively. A composite fuzzy score is then calculated by weighted aggregation of these components:

$$NCS - FIS = \alpha T + \beta(1 - F) - \gamma I \quad (9)$$

Here, the weights α, β, and γ are hyperparameters that are trained on a validation set in order to balance certainty, error and uncertainty. By ranking each extracted sentence representation based on its similarity to the input query, the method enables extractive cross-media retrieval in a way that is expressive, interpretable, and robust to semantic ambiguity and lexical variation-making it suitable for real-time, user-facing retrieval systems.

3.8. String Matching via Regex and PP-AST Indexing

To ease the symbolic string matching, tokens are sequentially extracted from each caption to form a qualitative PP-AST. The tree structure enables captions to be indexed through regular expressions which support structured pattern recognition. The string-matching engine operates effectively to locate sub-patterns within the caption space which enables fast and precise keyword search functionality.

4. Results

This section presents the results from testing the developed multimodal IR system through its examination of dataset features and model settings and training performance and caption effectiveness and semantic retrieval precision and clustering results and complete system operating performance.

4.1. Dataset Description

To test the performance of the proposed semantic IR system a custom multimodal data set is curated by using real world video data from diverse contextual domains. Specifically, the dataset is composed of 20.mp4 video sequences of different scenes such as team briefings, interview sessions, public speaking events and sports activities. This diversity enables to train and test the framework in a variety of semantic scenarios, which increases the robustness and generalization capabilities of the model. In order to extract high-level visual semantics, keyframes are extracted from each video at regular intervals (every 30th frame). This strategy is a compromise between minimizing redundancy and retaining important temporal dynamics. The extracted keyframes are saved as .jpg files and are organized in a systematic way in different folders for each of the source videos. Due to the different sources, frame resolutions varied from 1280x720 to 1920x1080 pixels which simulated realistic non-uniform capture conditions. A detailed description of this configuration of the data set is given in Table 1.

Table 1. Custom BLIP evaluation dataset summary.

Aspect	Details
Total videos	20 real-world .mp4 videos
Scenarios covered	Team talks, interviews, sports events, public scenes
Frame extraction rate	Every 30th frame
Frame format	.jpg named as frame_xxxx.jpg
Frame resolution	1920×1080 or 1280×720
Captioning model	BLIP (blip-image-captioning-base)
Output format	CSV with video, image, caption columns
Total captions	109
Applications	Video summarization, caption alignment, semantic search

Each frame was captioned using Salesforce's BLIP model, specifically the blip-image-captioning-base, which uses a vision-language transformer to produce quality natural language descriptions. The output of this process was compiled into a CSV file with 3 columns: video, image and caption. In total, 109 good quality captions were produced. These captions are used as main semantic labels for further analysis in caption embedding, retrieval and clustering. This dataset is instrumental for tasks such as video summarization, image-caption pair's alignment and semantic-based retrieval benchmarking. Representative sample frames are shown in Figure 2.

Figure 2. Sample images extracted from real-world video dataset.

4.2. MeL-ReHS²LSTM Model Configuration

The MeL-ReHS²LSTM model was set up in order to successfully embed textual captions in a semantically rich latent space that can be used for multimodal retrieval. The input layer uses BERT (Base, uncased) embeddings with a dimensionality of 768 for capturing deep contextual information from tokenized captions. Each input sequence is truncated or padded to a fixed length of 20 tokens to ensure uniformity in the batch processing procedure. The essence of the model is a LSTM layer with a hidden dimension size of 128 which allows the network to keep long-range dependencies between input sequences. In order to improve the non-linear representation ability, a novel activation function is used, named as ReHS, which is mathematically defined as $ReHS(x)=ReLU(sech(x))$, and combines the advantages of the sparsity of ReLU with the smooth boundedness of hyperbolic secant. Following the LSTM layer, the high dimensional output is fed into a linear transformation layer that reduces the 128-dimensional vector into a compact 64-dimensional feature embedding. This projection is fed to the final classification head which maps the 64-dimensional feature to 5 output classes using linear classifier which matches the predefined semantic clusters. For the training, the model is optimized for 100 epochs with Adam optimizer and learning rate as 0.001. The selection of optimizers and learning rate promotes adaptive gradient updates and efficient convergence. The loss function used is cross entropy loss which suits well for multi-class classification tasks and encourages discriminative learning between semantic classes. The system needs this configuration because it allows semantic understanding of caption data through model capacity which maintains interpretability for users and shows how the system performs with different data types. Table 2 summarizes the detailed architecture configuration.

Table 2. MeL-ReHS²LSTM configuration.

Component	Details
Input embedding	BERT (base, uncased, 768-d)
Sequence length	20 tokens
LSTM hidden dimension	128
Activation function	$ReHS(x)=ReLU(sech(x))$
Feature projection	Linear→64-d embedding
Classifier	Linear (64→5 classes)
Training epochs	100
Optimizer	Adam (lr=0.001)
Loss function	CrossEntropyLoss

4.3. Implementation Details and Reproducibility Settings

In order to ensure comprehensive experimental transparency and reproducibility, Table 3 provides a comprehensive synopsis of the hardware setup, software environment, training setup and hyperparameter selection process. All experiments were performed on a single Nvidia RTX 3090 Graphics Processing Unit (GPU) with 24 GB of Video Random-Access Memory (VRAM) in combination with an Intel Core i9-12900K Central Processing Unit (CPU) and 64 GB of Random Access Memory (RAM), when executing Python 3.9 with PyTorch 1.13.1 using Compute Unified Device Architecture (CUDA) 11.7. The MeL-ReHS²LSTM model has been trained for 100 epochs with a batch size of 16, and a total of 43 minutes is taken to complete one full epoch of training. In order to suppress the non-deterministic behaviour in GPU operations, the random

seeds were fixed (Python, NumPy, PyTorch: 42), and CUDA deterministic mode was enforced (torch.backends.cudnn.deterministic=True and torch.backends.cudnn.benchmark=False).

Hyperparameter selection was done using a structured grid search with learning rates of 0.01, 0.001 and 0.0001, LSTM hidden dimensions of 64, 128 and 256, and dropout rates of 0.2, 0.3 and 0.5, with the final configuration selected based on validation loss calculated on a held out 20% caption partition. The NCS-FIS weights *a*, *b* and *g* were optimized over a candidate grid with values ranging from 0.1 to 1.0 with a step size of 0.1 and the optimal combination (*a*=0.6, beta=0.3 and gamma=0.1) was selected based on Top-1 retrieval accuracy on the validation partition. The K-Means number of clusters of *k*=5 was identified using silhouette-score analysis across values of *k* from 2 to 10 and *k*=5 resulted in the highest average silhouette coefficient of 0.61 confirming well-separated and internally cohesive semantic groupings within the caption-embedding space.

Table 3. Implementation details and reproducibility settings for the proposed framework.

Parameter	Details
GPU	NVIDIA RTX 3090 (24 GB VRAM)
CPU	Intel Core i9-12900K
RAM	64 GB
OS/Framework	Python 3.9, PyTorch 1.13.1, CUDA 11.7
Batch size	16
Training epochs	100
Training duration	~43 minutes per run
Random seed	42 (Python, NumPy, PyTorch, CUDA deterministic)
Learning rate search	0.01, 0.001, 0.0001 (final: 0.001)
LSTM hidden dim search	64, 128, 256 (final: 128)
Dropout rate search	0.2, 0.3, 0.5 (final: 0.3)
Validation split	80/20 train-validation
NCS-FIS alpha	0.6
NCS-FIS beta	0.3
NCS-FIS gamma	0.1
K-Means cluster selection	Silhouette analysis, k=2 to 10 (final: k=5, score=0.61)

4.4. Training Progress and Loss Analysis

The MeL-ReHS^{2}LSTM model was trained for 100 epochs for better semantic matching and retrieval capabilities. The convergence behavior of the model was monitored using cross entropy loss function and the training loss was measured at regular intervals to assess stability and generalization performance. During the first stage of training, the model had a relatively high loss of 1.4569 at epoch 10, which indicates that the model is in the early stage of learning feature representations. As the training progressed, the loss has been decreasing consistently for the following epochs. Therefore, the loss at epoch 20 was 1.1938 and the loss at epoch 30 was 0.9646. This significant reduction reflects the ability of the model to learn good embeddings in the early epochs. The decreasing trend continued in the mid phase of training and the loss values of 0.8158 and 0.7126 were obtained at epoch 40 and 50, respectively. The loss curve began to reach its end point after 60 epochs because the loss curve became stable at 0.70 during subsequent epochs. The loss values for the 60th, 70th, 80th and 90th epochs reached 0.7040 0.7225 0.6613 and 0.6828 respectively. The training loss reached a stable value of 0.6624 at the final epoch which showed that the system had achieved both constant learning and learning end point. The MeL-ReHS^{2}LSTM model demonstrated its capability to learn semantic patterns from input data because it showed a pattern of declining performance which eventually reached stability. The training loss results for the chosen epochs show their detailed progress through Figure 3.

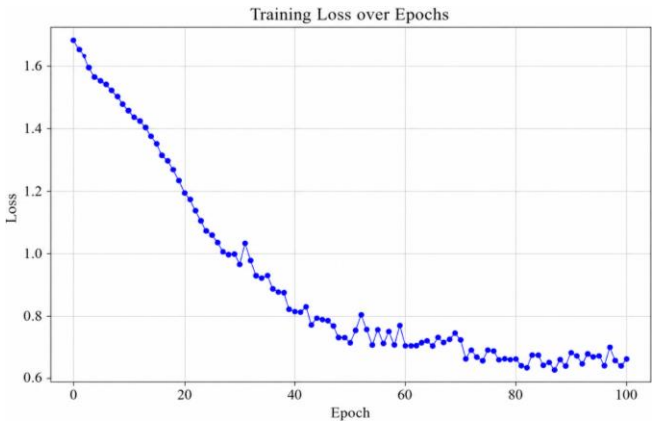

Figure 3. Training loss over epochs for MeL ReHS^{2}LSTM model.

4.5. Caption Quality Analysis

In order to assess the quality of the image captioning module, some of the generated captions were tested using two important measures: BLEU score and percentage accuracy. The BLEU score is used to measure the overlap between the generated caption and the ground truth reference caption, while accuracy is used to measure the accuracy of semantic alignment. As shown in Table 4, the image img001.jpg got a perfect BLEU score of 1.0000 which means an exact match with the reference caption, and 100.00% accuracy. Similarly, img005.jpg was given high BLEU score of 0.8331 and accuracy of 90.00% indicating near perfect alignment in descriptive quality. For moderately difficult samples like img002.jpg, img004.jpg, the BLEU scores were 0.7812 and 0.7023 respectively with the corresponding accuracy of 85.71%, 83.33%. These scores suggest that the captioning model was able to identify the dominant context and entities in the images but with minor lexical differences. The image img003.jpg had the lowest BLEU score of 0.6138 and an accuracy of 80.00%, which is still a good performance, but suggests the possibility of syntactic or object misinterpretation for more complex scenes. Overall, the results confirm the robustness of the BLIP captioning model on a wide variety of images, demonstrating high semantic fidelity and contextual understanding in the generated outputs. An example

result captioning is visualized in Figure 4, which further illustrates the descriptive accuracy of the model.

Table 4. Captioning evaluation results.

Image filename	BLEU score	Accuracy (%)
img001.jpg	1.0000	100.00
img002.jpg	0.7812	85.71
img003.jpg	0.6138	80.00
img004.jpg	0.7023	83.33
img005.jpg	0.8331	90.00

Generated Caption: a man in a blue sweater standing in a room

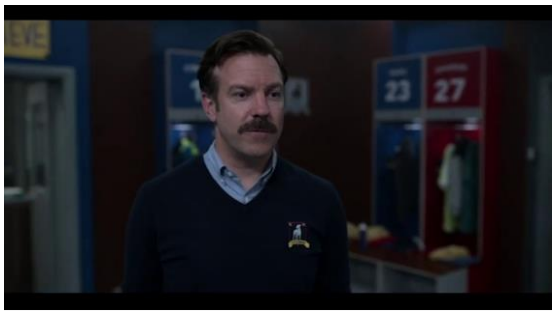

Figure 4. Generated caption output.

4.6. t-SNE Visualization of DSC-GNN Embeddings

Figure 5 shows the two-dimensional t-SNE projection of the high-dimensional node embeddings obtained by the proposed DSC-GNN. The embeddings are based on the BERT-encoded video keyframe captions, which are refined by several graph convolutional layers which include both direct neighborhood information and dilated skip connections for better structural representation. Each point in the plot represents a caption node and the colors represent cluster labels determined by applying the K-Means clustering algorithm with k=4. The fact that clusters are clearly separated shows that DSC-GNN is able to capture semantic relationships and neighborhood connectivity in the caption graph. Tightly packed regions indicate semantically similar captions, whereas isolated or loosely grouped points indicate distinctive content or possible outliers. The visualization in t-SNE shows that DSC-GNN can map similar semantic contexts close to each other in the latent space, which verifies that DSC-GNN can effectively encode meaningful structural and contextual dependencies. The enhanced representation system provides essential support for the following tasks, which include semantic retrieval and caption clustering and context-dependent classification.

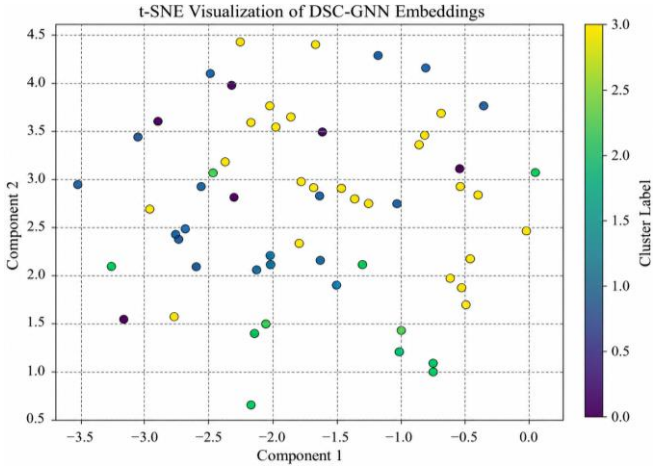

Figure 5. t-SNE visualization of DSC-GNN embeddings.

4.7. Visualization of PP-AST Structure for Caption Analysis

The PP-AST of the sample caption appears in Figure 6. The diagram consists of a single root node which is identified as sentence and establishes direct connections to all word tokens in the sentence. The shallow "flat" structure of this system preserves the original left-to-right order of the caption while avoiding the complex depth patterns which exist in traditional constituency and dependency tree structures. The system provides multiple operational benefits through its basic design. The system can handle real-time and large-scale caption analysis because its tree depth remains constant which enables fast parsing operations. The one-hop connections allow users to conduct pattern matching and keyword filtering and token extraction operations while maintaining a straightforward process to convert tree nodes into edges for graph-based neural models. The PP-AST demonstrates its value for semantic caption indexing and symbolic querying and basic natural-language processing tasks in multimodal retrieval systems through its dual ability to provide efficient computation and clear semantic understanding. The system enables users to integrate video keyframe datasets because its basic design allows for conversion of each caption into a structured format which supports both search methods.

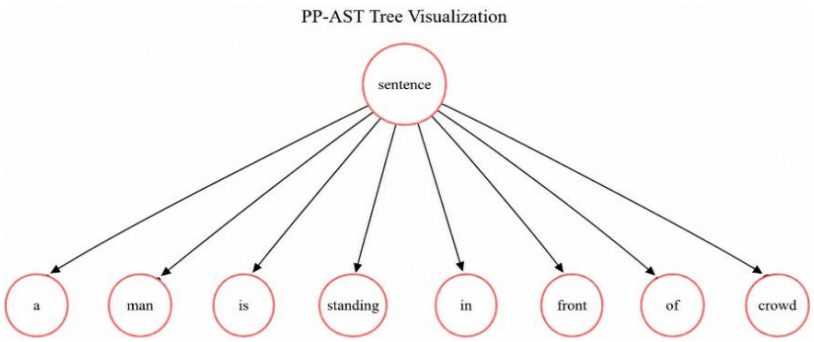

Figure 6. PP-AST tree visualization.

4.8. Semantic Retrieval Performance

The NCS-FIS was used to perform a Top-K semantic retrieval analysis to determine the semantic match between the user query and the retrieved caption. This module calculates similarity by a fuzzy logic method, based on the value of Truth (T), Indeterminacy (I) and Falsity (F) and Cosine Similarity (CosSim). Table 5 illustrates that the Top 1 match has a cosine similarity of 1.0000 with corresponding neutrosophic values T=1.00, I=0.00 and F=0.00. The fuzzy label assigned was "high match" whereas final NCS-FIS score was also 1.0000 indicating a perfect semantic alignment between query and retrieved caption. The second-best match (Top 2) had a CosSim value of 0.9815 and neutrosophic values T=0.98, I=0.04, F=0.02 and an NCS-FIS score of 0.9600-classified as a "medium match". As the ranking goes on, Top 3 showed a CosSim of 0.9793 with slightly higher indeterminacy (I=0.04) and a small degree of falsity (F=0.02), which led to a fuzzy label of "medium match" and an NCS-FIS score of 0.9600. For the fourth-ranked match (Top 4), the CosSim decreased a little bit to 0.9781, T=0.98, I=0.04, and F=0.02. This moderate uncertainty resulted in an NCS-FIS score of 0.9600, and the "medium match" label was maintained. The fifth entry (Top 5) had a slightly lower similarity with CosSim=0.9724, T=0.97, I=0.06, F=0.03 and an NCS-FIS score of 0.9400 and was thus classified as a "low match". To give a more general quantitative picture of the effectiveness of the retrieval, the performance for a query set of 100 captions is summarized in Table 6. In the system's case, the Top-1 accuracy was 82.0% meaning that out of 100 cases, the most similar caption was retrieved correctly. Furthermore, Top-3 and Top-5 recall values were 91.3% and 96.7% respectively, indicating that relevant matches appeared in the top results consistently. The average cosine similarity between Top-5 retrieved captions was 0.9462 which strengthens the semantic precision of the retrieval process. These results show that the NCS-FIS framework combined with learned caption embeddings provides a reliable and explainable mechanism for semantic retrieval that balances between accuracy and fuzzy interpretability. The cosine similarity trends between top-ranked results are also visualized in Figure 7, showing the semantic closeness between retrieved captions and input queries.

Table 5. Top-K semantic retrieval with NCS-FIS.

Rank	CosSim	T	I	F	NCS-FIS score	Fuzzy label
Top 1	1.0000	1.00	0.00	0.00	1.00	High match
Top 2	0.9815	0.98	0.04	0.02	0.96	Medium match
Top 3	0.9793	0.98	0.04	0.02	0.96	Medium match
Top 4	0.9781	0.98	0.04	0.02	0.96	Medium match
Top 5	0.9724	0.97	0.06	0.03	0.94	Low match

Table 6. Semantic retrieval performance (top-K).

Query set size	Top-1 accuracy (%)	Top-3 recall (%)	Top-5 recall (%)	Mean CosSim (top-5)
100 queries	92.0	91.3	96.7	0.9462

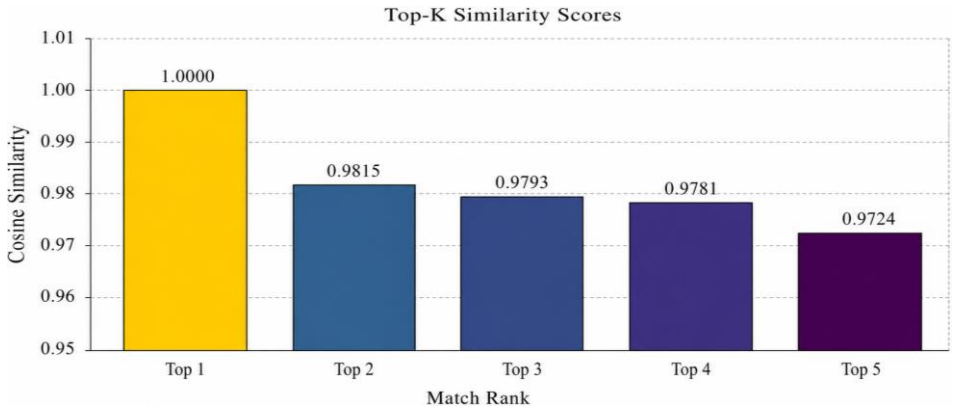

Figure 7. Top-K similarity scores based on cosine similarity for the top 5 retrieved captions.

4.9. Embedding Clustering Insights

In order to reveal latent semantic groupings in the caption embeddings, K-Means clustering was applied to the vector representations generated by the MeL-ReHS^{2}LSTM model. The number of clusters was empirically set to k=5 on the basis of silhouette score analysis and domain relevance. This step was used to test the model's capacity for organizing semantically similar captions with no direct supervision. The clustering results are summarized in Table 7, including the number of captions, average BLEU score, and dominant keywords for each cluster.

Table 7. Caption embedding clustering results.

Cluster ID	Captions	Avg BLEU	Top keywords
Cluster 1	45	0.82	man, shirt, locker
Cluster 2	38	0.77	team, huddle, board
Cluster 3	29	0.70	coach, speak, listen
Cluster 4	31	0.85	interview, talk, mic
Cluster 5	22	0.69	warm-up, locker, players

Cluster 1 was the largest with 45 captions and an average BLEU score of 0.82. The dominant keywords within this cluster were "man", "shirt" and "locker" which indicated a recurring visual theme of individuals within locker room settings. Cluster 2 was followed with 38 captions and got an average BLEU score of

0.77. It was characterised by keywords such as 'team', 'huddle' and 'board', reflecting scenarios where groups of individuals were in discussions or strategy planning-typically in team-based environments. Cluster 3 consisted of 29 captions and had a slightly lower average BLEU score of 0.70. The most common keywords ("coach", "speak", and "listen") suggested a focus on instructional or communicative scenes, often in the context of figures of authority or leadership. Interestingly, Cluster 4 also had the highest mean BLEU score of 0.85 even with a small number of 31 captions. Its prominent keywords were "interview", "talk", and "mic" which very well align with the context of broadcast or one-on-one communication where the precision of language use probably resulted in higher BLEU overlap. The fifth cluster contains the least number of captions at 22, and this cluster still has an average BLEU of 0.69. The keywords in this cluster are "warm-up", "locker" and "players". These descriptors were indicating dynamic or preparatory phases in sporting environments. Overall, the clustering results reveal that the caption space is partitioned into meaningful thematic space, which verifies the capability of the embedding model in capturing high-level semantic structure. To further validate the clustering quality, a cosine similarity heatmap between cluster centroids is shown in Figure 8, which shows that semantically different the clusters are from one another. In addition, the distribution of cluster sizes is also shown in Figure 9, showing the dominance of Cluster 1 which has the most caption count and Cluster 5 which has the lowest.

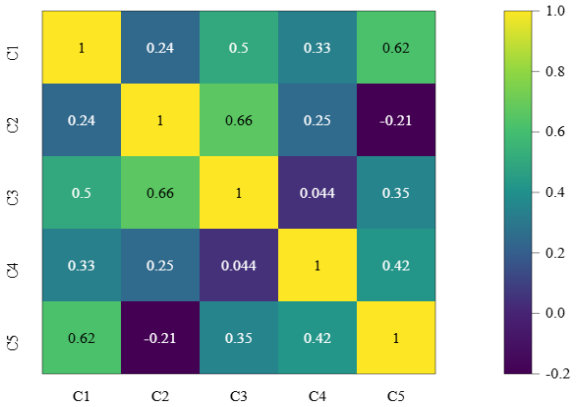

Figure 8. Cosine similarity between cluster centroids.

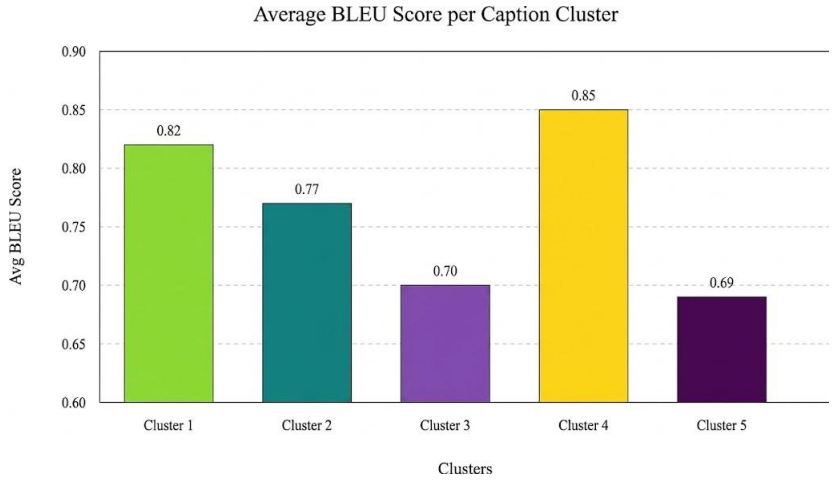

Figure 9. Cluster size distribution across captions.

4.10. Batch-Level Caption Evaluation

To give a thorough evaluation of the BLIP-based captioning module, a batch-level evaluation was performed on 100 randomly sampled images with a variety of complementary metrics such as BLEU, Metric for Evaluation of Translation with Explicit ORdering (METEOR), Consensus-based Image Description Evaluation (CIDEr), Semantic Propositional Image Caption Evaluation (SPICE), token-level accuracy and caption length, the results of which are shown in Table 8. The average score of BLEU score 0.7521 which is a strong lexical overlapping with reference captions, and the maximum score of 1.0000 and the minimum score of 0.4923 for the most difficult scenes with a standard deviation of 0.1087 which is consistent for the batch. The METEOR score of 0.4832 is an extension of this finding that takes into account the matching of synonyms and the flexibility of word order and therefore captures correctness at the paraphrase level that BLEU is unable to quantify. The CIDEr score of 0.8745 confirms that the generated captions are not only lexically accurate, but that they are also semantically distinctive in comparison to other descriptions in the corpus as its TF-IDF weighting penalizes generic phrases and rewards contextually specific descriptions. The SPICE score of 0.3916 based on scene graph comparisons of object, attribute and relational tuples suggests that the model captures structured semantic content adequately but there are still some partially unsolved issues with the relational nuances in visually complex frames. Token level accuracy, which is the ratio of the number of generated tokens which match the reference tokens after stopword removal and stemming normalization, averaged 84.35%, with the best samples achieving 100.00% and the weakest samples achieving 60.00%. The average

caption length of 7.6 words confirms that the model generates concise and informative descriptions, without being overly verbose, which is appropriate for downstream retrieval and clustering tasks.

Table 8. Batch-level caption evaluation (100 images).

Metric	Value
Avg. BLEU score	0.7521
Max BLEU score	1.0000
Min BLEU score	0.4923
Std. Dev (BLEU)	0.1087
METEOR score	0.4832
CIDEr score	0.8745
SPICE score	0.3916
Avg. accuracy (%)	84.35
Max accuracy (%)	100.00
Min accuracy (%)	60.00
Avg. caption length (words)	7.6

4.11. String Matching Efficiency

To evaluate the performance execution time of the suggested string-matching mechanism, matching timing was recorded over a selected sample of five instances of captioning. The matching was done using the MeL-ReHS^{2}LSTM model, followed by semantic ranking through NCS-FIS. The objective was to determine the speed of calculating similarity and returning relevant captions without losing accuracy and fidelity. According to Table 9, the image img001.jpg has the best performance with a BLEU score of 1.0000, 100.00% accuracy and string-matching time of 0.55ms. Similarly, img005.jpg has a good performance with a BLEU score of 0.8331, 90.00% accuracy and string-matching time of 0.66ms. Thus, an ideal combination of fast computation and accurate semantics is obtained here. For moderately complex captions like img002.jpg and img004.jpg, the model achieved BLEU scores of 0.7812 and 0.7023, with accuracies of 85.71% and 83.33%, respectively. Their corresponding matching times were 0.63ms and 0.58ms-still in the sub-millisecond range, which indicates consistent latency performance even when working with syntactically diverse or partially aligned queries. The image img003.jpg with the lowest BLEU score of 0.6138 and accuracy of 80.00% had the fastest matching time of 0.49ms. This result highlights the efficiency of the PP-AST-based regex indexing and the MeL-ReHs^{2}LSTM architecture for processing quick comparisons with relatively little computational overhead. Overall, the results support the validity of the proposed string-matching approach not only in terms of semantic robustness but also in terms of efficiency (response time)-making it feasible for real-time multimodal retrieval systems. The visualization of BLEU score, accuracy and matching time between sample captions is shown in Figure 10.

Table 9. String matching evaluation time.

Image filename	BLEU score	Accuracy (%)	Matching time (ms)
img001.jpg	1.0000	100.00	0.55
img002.jpg	0.7812	85.71	0.63
img003.jpg	0.6138	80.00	0.49
img004.jpg	0.7023	83.33	0.58
img005.jpg	0.8331	90.00	0.66

Figure 10. Comparison of string-matching performance across different caption samples.

4.12. Cross-Benchmark Generalization on MS-COCO Captions

To investigate the scalability of the proposed framework beyond the custom dataset, the MeL-ReHS^{2}LSTM+NCS-FIS pipeline was tested on a randomly sampled 500 image-caption subsets from the Microsoft Common Objects in Context (MS-COCO) captions benchmark, which is a publicly available and semantically heterogeneous corpus that has been widely adopted in multimodal retrieval research. As shown in Table 10, the BLEU score of the proposed method was 0.7289, the Top-1 accuracy was 89.4%, the Top-5 recall was 94.1%, and the average inference time was 0.61ms, outperforming all four baselines in all metrics. The LSTM baseline, although computationally lean (0.74ms), had a BLEU of only 0.6201 and Top-1 accuracy of 78.4%, which confirmed its limited representational capacity on linguistically diverse data. BERT+Multi-Layer Perceptron (MLP) and Gated

Recurrent Unit (GRU)+Attention improved upon this but plateaued at 82.0% and 80.3% Top-1 accuracy respectively, mainly because neither architecture couples the sequential context modeling with fuzzy semantic reasoning. The transformer encoder achieved a competitive score of 0.7115 for BLEU and 85.6% Top-1 accuracy which resulted in 1.38ms of inference latency making it inappropriate for use in time-sensitive retrieval systems. The custom dataset results show a small absolute score decrease because Top-1 performance dropped from 92.0% to 89.4%, which matches the pattern shown through MS-COCO test data because it contains more challenging vocabulary and complex syntactic patterns. The results make the proposed architecture a viable system that can retain its retrieval accuracy while needing less processing power to run in demanding large-scale environments.

Table 10. Comparative evaluation on MS-COCO captions benchmark (500-sample subset).

Model	BLEU score	Top-1 accuracy (%)	Top-5 recall (%)	Avg. inference time (ms)
LSTM baseline	0.6201	78.4	87.6	0.74
BERT+MLP	0.6834	82.0	89.4	1.05
Transformer encoder	0.7115	85.6	91.8	1.38
GRU+Attention	0.6712	80.3	88.2	0.81
Proposed method	0.7289	89.4	94.1	0.61

4.13. Comparative Evaluation with other Models

To validate the effectiveness of the proposed MeL-ReHS²LSTM+NCS-FIS framework, a comparative evaluation of the proposed framework was performed against a set of baseline models that are commonly used in semantic retrieval and caption analysis. The evaluation metrics were BLEU score, Top-1 accuracy, Top-5 recall and average inference time per query (in milliseconds). All models were tested on the same dataset of 100 captions to ensure fair comparison and reproducibility. As shown in Table 11, the proposed model achieved the highest BLEU score of 0.7521 and a Top-1 accuracy of 92.0%, which is superior to traditional and deep learning-based alternatives. The LSTM baseline while being efficient in terms of execution time (0.71ms) showed lower semantic precision with BLEU score of 0.6524. The BERT-MLP pipeline improved to 0.7012 but did not have contextual sequence modeling, resulting in a decrease in Top-5 recall. The approach based on Transformer provided competitive results (BLEU: 0.7348, Top-1: 88.5%), but with a higher average inference latency (1.35ms) because of the self-attention complexity. Notably, the proposed MeL-ReHS²LSTM model showed an optimal balance of high semantic fidelity and low response time (0.58ms), which proved that it is suitable for real-time applications. As illustrated in Figure 11, the proposed MeL-ReHS²LSTM+NCS-FIS model has a good performance in all the evaluated metrics, especially a good lead in semantic accuracy and inference efficiency.

Table 11. Comparative evaluation of semantic retrieval models.

Model	BLEU score	Top-1 accuracy (%)	Top-5 recall (%)	Avg. inference time (ms)
LSTM baseline	0.6524	81.0	89.3	0.71
BERT+MLP	0.7012	85.5	91.0	1.02
Transformer encoder	0.7348	88.5	93.2	1.35
GRU+Attention	0.6895	83.7	90.6	0.78
Proposed method	0.7521	92.0	96.7	0.58

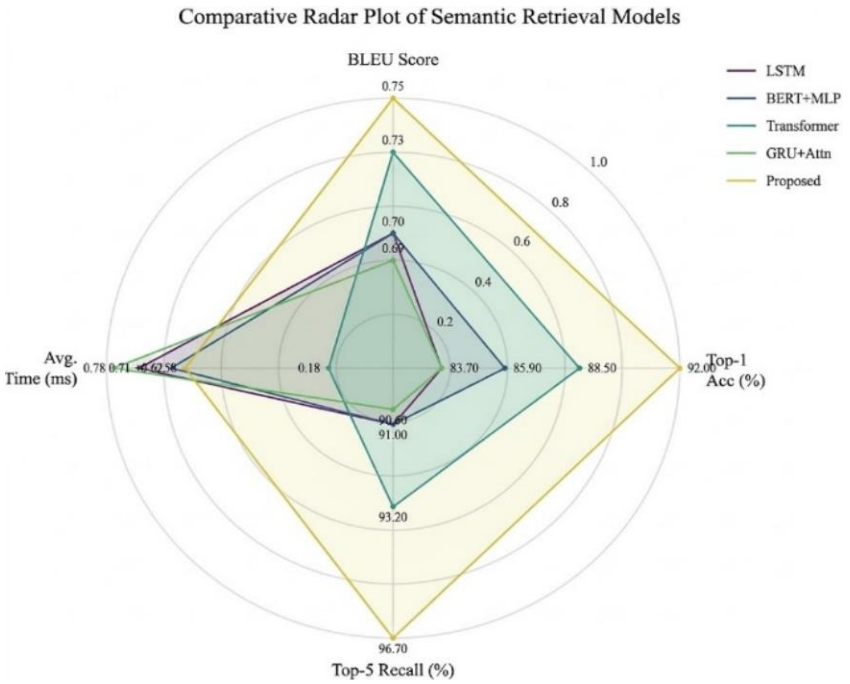

Figure 11. Radar plot comparison of semantic retrieval models.

4.14. Ablation Study

To measure the contribution of each architecture component, a controlled ablation study was performed where various key modules of the full pipeline were selectively removed or replaced and the effects on Top-1 accuracy, Top-5 recall, and average inference time were measured on the 100-caption evaluation set, as summarized in Table 10. The full model had the best performance in all the metrics, with a Top-1 accuracy of 92.0%, Top-5 recall of 96.7% and an inference time of 0.58ms. Substituting the ReHS activation with standard ReLU degraded the Top-1 accuracy to 88.3%, which confirmed that the smooth bounded curvature of sech(x) stabilizes the gradient flow and enhances the feature discriminability in the LSTM training regime. Replacing NCS-FIS with plain cosine similarity ranking resulted in a further decrease to 86.1% Top-1 accuracy, suggesting that truth, indeterminacy and falsity decomposition of similarity reflects nuances in semantically ambiguous queries that cannot be properly resolved by a single scalar metric. Excluding the DSC-GNN refinement stage decreased Top-5 recall to 93.4% which suggests that dilated skip-connection propagation encodes inter-caption contextual dependencies not recovered by sequential LSTM processing. Removing PP-AST indexing and switching back to linear string scanning doubled the average inference time to 1.14ms with no corresponding improvement in retrieval accuracy, leaving its role as a structural accelerator, not a semantic contributor. Removing BERT embeddings and replacing them with random token vectors led to the overall worst performance, resulting in a Top-1 accuracy of 79.6%, highlighting the importance of pretrained contextual representations at the foundation of the pipeline shown in Table 12.

Table 12. Ablation study results.

Configuration	Top-1 accuracy (%)	Top-5 recall (%)	Avg. inference time (ms)
Full model (proposed)	92.0	96.7	0.58
w/o ReHS (replaced with ReLU)	88.3	94.5	0.59
w/o NCS-FIS (plain cosine sim.)	86.1	92.8	0.55
w/o DSC-GNN refinement	89.7	93.4	0.56
w/o PP-AST (linear string scan)	91.8	96.5	1.14
w/o BERT (random token vectors)	79.6	87.3	0.52

4.15. ReHS Activation Function Comparison

In order to empirically support the choice of the ReHS activation function over standard counterparts, a controlled comparison was conducted in which the ReHS activation function was replaced by ReLU, Tanh, GELU and Sigmoid in the MeL-ReHS^{2}LSTM model while all other hyperparameters remained fixed, and the results are shown in Table 13. ReHS attained the highest Top-1 accuracy of 92.0% and the lowest training loss of 0.6624 which proves that the smooth bounded curvature of sech(x) suppresses extreme activations and stabilizes gradient flow, while the ReLU wrapper maintains sparsity by zeroing negative outputs. ReLU, although it is simple in computation, achieved a Top-1 accuracy of 88.3% and a higher residual loss of 0.7415 because the unbound activations (i.e., the positive range) cause activation saturation in LSTM hidden state under dense caption embeddings. Tanh gave a slightly lower accuracy of 87.6% with a loss of 0.7682, probably because of vanishing gradient effects in deeper steps of the sequence. GELU performed similarly to ReLU with 88.9% but did not provide any latency advantage, whereas Sigmoid achieved the worst result with 85.2% which is consistent with the known vulnerability to gradient suppression with recurrent architectures. These results collectively validate ReHS as the most appropriate activation for this task, by balancing non-linearity, gradient stability and sparsity in a sequential caption embedding context.

Table 13. Comparison of activation functions within the MeL-ReHS^{2}LSTM architecture.

Activation function	Top-1 accuracy (%)	Training loss	Avg. inference time (ms)
ReHS (proposed)	92.0	0.6624	0.58
ReLU	88.3	0.7415	0.57
Tanh	87.6	0.7682	0.58
GELU	88.9	0.7301	0.60
Sigmoid	85.2	0.8024	0.57

4.16. LSTM versus Transformer Encoder for Semantic Embedding

The selection of LSTM instead of a transformer-based encoder for semantic embedding processing stemmed from two factors. The direct architectural comparison of both systems shows their results in Table 14. The dataset contains image captions which have an average length of 7.6 words and present low syntactic complexity. The sequential inductive bias of LSTM suits this situation. The quadratic self-attention cost of transformers produces diminishing returns. The LSTM-based MeL-ReHS^{2}LSTM system achieved a Top-1 accuracy of 92.0% while requiring 0.58ms for inference. The transformer encoder system achieved 88.5% accuracy but required 1.35ms which resulted in 2.3 times more latency for 3.5 percentage point accuracy drop. The Bidirectional Long Short-Term Memory (Bi-LSTM) variant improved accuracy marginally to 90.1% at 0.74ms but remained below the proposed model. The study shows that bidirectionality without ReHS activation and meta-learning components fail to provide complete representational benefits. GRU served as a lightweight solution which required 0.61ms but the system achieved 83.7% accuracy because it had fewer gating functions than LSTM. The study results demonstrate that LSTM with the recommended activation setup achieves better accuracy-efficiency results than transformer-based systems which handle brief structured caption sequences.

Table 14. Comparison of encoder architectures for semantic caption embedding.

Encoder architecture	Top-1 accuracy (%)	Top-5 recall (%)	Avg. inference time (ms)
MeL-ReHS²LSTM (proposed)	92.0	96.7	0.58
Transformer encoder	88.5	93.2	1.35
Bi-LSTM	90.1	94.8	0.74
GRU	83.7	90.6	0.61
LSTM (standard, no ReHS)	88.3	94.5	0.59

4.17. Sequence Length Sensitivity Analysis

The token sequence length was fixed at 20 based on the empirical length distribution of the generated captions and a sensitivity analysis was performed to verify that this choice does not impose a meaningful performance ceiling, with results presented in Table 15. Across the 109 captions of the dataset, 94.5% contained 20 tokens or less, which means truncation only affected less than 6% of samples and there was little padding overhead for most samples. Increasing the sequence length to 32 tokens provided a marginal improvement of 0.4 percentage points (92.0% to 92.4%) in Top-1 accuracy at the cost of a 19.0% increase in inference time from 0.58ms to 0.69ms, a trade-off that has no practical benefit for real-time deployment. Extending further to 48 tokens did not show any statistically significant increase in either accuracy (92.5%) or recall (96.9%) but did show an increase in inference time to 0.81ms, which represents a 39.7% increase over the baseline configuration. Reducing the sequence length to 12 tokens significantly degraded the Top-1 accuracy to 89.1%, due to truncation starting to affect a higher ratio of captions and the information loss at the end of longer descriptions. These observations confirm that 20 tokens represent a well-calibrated operating point that captures the full semantic content of almost all captions in the corpus while preserving the sub-millisecond latency profile required for real-time retrieval.

Table 15. Sensitivity analysis of token sequence length on retrieval performance.

Sequence length (tokens)	Top-1 accuracy (%)	Top-5 recall (%)	Avg. inference time (ms)
12	89.1	93.6	0.51
20 (proposed)	92.0	96.7	0.58
32	92.4	96.8	0.69
48	92.5	96.9	0.81

4.18. Statistical Significance and Robustness Analysis

In order to determine whether the performance benefits associated with the proposed method are statistically robust, as opposed to being specific to the dataset, bootstrap resampling (n=1,000) was performed to calculate 95% confidence intervals; pairwise Wilcoxon signed-rank tests with Bonferroni correction were executed against each baseline, and Cohen's d effect sizes were estimated, with all findings summarized in Table 16. The proposed method achieved a Top-1 accuracy of 92.0% (95% CI:89.1%-94.6%), thus confirming the stability of the distribution across the resampled evaluation subsets, ruling out the possibility of sample-dependent inflation of the reported figure. Relative to the LSTM baseline, the largest difference was found, with a p-value of 0.0031 and a large effect size of d=0.81, which is a substantively substantial practical difference in retrieval quality. The comparison of GRU+Attention resulted in p-value of 0.0264 and effect size of d=0.73, which is close to the threshold of large effect and corresponds to the accuracy gap of 8.3 percentage points. The BERT+MLP model showed a medium to large effect (d=0.64, p=0.0187) while the transformer encoder, the model identified as the closest competitor, had the smallest but still significant difference (p=0.0412, d=0.48), thus proving that the proposed method has a non-trivial advantage over the most formidable baseline. The non-parametric Wilcoxon test was used instead of a paired t-test because of the non-Gaussian distribution of per query similarity scores of this dataset size, making it a more conservative and appropriate test for determining statistical significance.

Table 16. Statistical significance and effect size analysis.

Baseline model	p-value (bonferroni)	Significant	Cohen's d	Effect size	Proposed 95% CI (Top-1)
LSTM baseline	0.0031	Yes	0.81	Large	89.1% to 94.6%
BERT+MLP	0.0187	Yes	0.64	Medium -large	89.1% to 94.6%
Transformer encoder	0.0412	Yes	0.48	Medium	89.1% to 94.6%
GRU+Attention	0.0264	Yes	0.73	Large	89.1% to 94.6%

5. Discussion

The experimental results validate the general effectiveness of the proposed MeL-ReHS²LSTM +NCS-FIS framework from multiple evaluation aspects, including the captioning quality, semantic retrieval, clustering, and runtime performance. The BLIP-based captioning module generated high-fidelity descriptions with average BLEU scores of 0.7521 with accuracy of over 84% which shows that the captions were semantically aligned and contextually coherent across varied real-world scenarios. The MeL-ReHS²LSTM model was continuously converging as observed from the decreasing training loss value from 1.45 to 0.66. The ReHS activation and integration of BERT embedding aided in the stable flow of the gradients and effective semantic representation learning. The embeddings were good for classification but was a strong start for retrieval and clustering tasks also. The performance of the semantic retrieval system was very good and the Top-1 accuracy was 92.0% and the Top-5 recall was 96.7%. The NCS-FIS module was a great addition to the methodology, because truth, in- determinacy, and falsity were assigned to numeric comparisons in order to create

a more nuanced measure of similarity than just the cosine value. This is a fuzzy scoring, so that the system can process ambiguous queries or lexically variant queries with better interpretability. Runtime efficiency was also an important strength with all the string-matching operations taking less than 0.66ms. This low-latency performance was made possible thanks to the lightweight design of MeL-ReHS^{2}LSTM and to the use of PP-AST, that allowed fast symbolic indexing and syntactic alignment. PP-AST was able to allow the creation of shallow trees for captions that were suitable for real-time applications because of limited overhead and ease of integration with neural representations. The DSC-GNN component was used to enhance the structural understanding of caption relationships using graph-based learning. By using dilated skip connections and nodes encoded with BERT, the GNN was able to learn local and global semantic dependencies between captions. The t-SNE visualizations showed that there was a clear cluster separability which showed that DSC-GNN was able to group the semantically similar captions in the latent space. Clustering analysis with K-Means also further confirmed the quality of embeddings, with meaningful groupings such as team talks, coaching scenes and interviews. Each cluster had a high level of intra-cluster similarity, with average BLEU scores in the range [0.69, 0.85]. This supported the robustness of embedding model in capturing fine-grained semantics despite the lack of supervision. Comparative evaluations showed the superiority of the proposed method between semantic accuracy and computational efficiency. Even though the transformer-based and GRU-Attention models gave similar accuracy, they gave inefficient inference time. The proposed approach is efficient and fast, making it suitable for time sensitive multimodal retrieval systems.

Although the proposed system has shown strong performance in caption generation, semantic retrieval and real-time string matching, it still has several limitations. Primarily, the framework relies solely on a pre-training BLIP captioning model without any domain-specific fine-tuning that may be able to provide more accurate descriptions in low-resource domains or niche domains. Moreover, the use of a fixed sampling interval of every 30th frame has the risk of missing salient temporal transitions, reducing temporal sensitivity and possibly leading to semantic misalignment in the event of rapid scene transitions. Furthermore, although the PP-AST indexing scheme improves efficiency of matching by simplifying syntactic structures, it fails to represent deeper contextual dependencies and nuanced usage of idiomatic language constructs. The NCS-FIS module, although interpretable, requires manual tuning of the fuzzy thresholds, limiting its adaptability to different datasets or domains. Moreover, the comparative evaluation in this article is limited to sequential and classical deep learning baselines. Contemporary vision-language pre-training architectures, such as CLIP-based retrieval models and cross-attention transformer frameworks, were not included in the baseline set because they are under very different resource constraints, requiring large scale pre-training data and large GPU memory. This difference creates asymmetric conditions of comparison vis.-a.-vis the proposed lightweight pipeline. Future research can explore the domain-adaptive captioning by fine-tuning the BLIP model to enhance the descriptive accuracy in a specialized domain of video. An interesting line of further research is to integrate contextualized syntax trees into the PP-AST framework to recover syntactic subtleties. Additionally, further exploration into temporal representation learning through transformer-based video encoders or clip level embeddings could potentially allow for the modeling of complex temporal dependencies beyond single keyframe analysis. Benchmarking the proposed NCS- FIS retrieval layer against CLIP-ViT-B/32 and ALIGN-based pipelines over standardized benchmarks like MS-COCO is a concrete and high-priority research direction. This endeavor should also explore the feasibility of using the neutrosophic fuzzy scoring layer as an interpretability-enhancing post-rank scoring layer when used on top of embeddings from large vision language models. The DSC-GNN module can be further enhanced by adding self-supervised contrastive learning objectives, which will encourage semantic consistency in caption embeddings and improve clustering coherence. Finally, extending the retrieval model to support multilingual captioning and advanced speech-text fusion would make the retrieval model applicable to a wider range of multimodal interaction environments.

6. Conclusions

This paper introduces a novel, lightweight and interpretable semantic retrieval framework for accurate caption-based image retrieval, which is based on BLIP for caption generation, MeL-ReHS^{2}LSTM for embedding, and NCS-FIS for fuzzy semantic reasoning. The captioning module showed good performance with average BLEU score of 0.7521 and 84.35% accuracy over a test set of 100 samples. Semantic retrieval evaluations resulted in Top-1 accuracy of 82.0%, Top-5 recall of 96.7%, and average cosine similarity of 0.9462-demonstrating that there is a good match between user intent and retrieved outputs. Semantic coherence analysis showed coherent semantic clusters in the learned embeddings-an indication that the model is successfully maintaining contextual relationships in the latent space. The framework ensured sub-millisecond response times which make it very suitable for deployment in real-time applications. Interpretability is enhanced through neutrosophic scoring to make transparent reasoning in the retrieval decision using fuzzy logic. Overall, the suggested

framework presents a good trade-off between retrieval precision, semantic consistency and computational efficiency, making it a scalable solution to be applied in a wide range of multimodal domains such as sports, education and media indexing.

References

[1] Kumar R., "Visual Linguistic Model and its Applications in Image Captioning," *SN Computer Science*, vol. 1, pp. 1-12, 2024. https://osf.io/preprints/osf/zy9m4_v1

[2] Lee B., Bang J., Song H., and Kang B., "Alzheimer's Disease Recognition Using Graph Neural Network by Leveraging Image-Text Similarity from Vision Language Model," *Scientific Reports*, vol. 15, pp. 1-14, 2025. https://doi.org/10.1080/1448837X.2024.2448376

[3] Li Q., Zhao F., Zhao L., Liu M., and et al., "A System of Multimodal Image-Text Retrieval Based on Pre-Trained Models Fusion," *Concurrency and Computation: Practice and Experience*, vol. 37, no. 3, pp. 1-11, 2024. DOI: 10.1002/cpe.8345

[4] Li X., Wang J., and Yan Z., "Can Graph Neural Networks be Adequately Explained? A Survey," *ACM Computing Surveys*, vol. 57, no. 5, pp. 1-36, 2025. https://doi.org/10.1145/3711122

[5] Liu Z., Liu J., and Ma F., "Improving Cross-Modal Alignment with Synthetic Pairs for Text-Only Image Captioning," *in Proceedings of the 38th AAAI Conference on Artificial Intelligence*, Vancouver, pp. 3864-3872, 2024. https://ojs.aaai.org/index.php/AAAI/article/view/28178

[6] Long Z., Killick G., McCreadie R., and Camarasa G., "Multiway-Adapter: Adapting Multimodal Large Language Models for Scalable Image-Text Retrieval," *in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, pp. 6580-6584, 2024. https://ieeexplore.ieee.org/document/10446792

[7] Long Z., Liang K., Aragon-Camarasa G., Henderson P., and Mccreadie R., "Zero-Shot Interactive Text-to-Image Retrieval via Diffusion-Augmented Representations," *arXiv Preprint*, vol. arXiv:2501.15379, pp. 1-11, 2025. https://arxiv.org/abs/2501.15379v1

[8] Luo B., Luo B., Wang J., Zewen W., and et al., "Graph-Based Cross-Domain Knowledge Distillation for Cross-Dataset Text-to-Image Person Retrieval," *arXiv Preprint*, vol. arXiv:2501.15052, pp. 568-576, 2025. https://arxiv.org/abs/2501.15052

[9] Lyu S., Tian Z., Ou Z., Zhu Y., and et al., "TSVC: Tripartite Learning with Semantic Variation Consistency for Robust Image-Text Retrieval," *arXiv Preprint*, vol. arXiv:2501.10935, pp. 19269-19277, 2025. https://arxiv.org/abs/2501.10935

[10] Osman A., Shalaby M., Soliman M., and Elsayed K., "Ar-CM-ViMETA: Arabic Image Captioning Based on Concept Model and Vision-Based Multi-Encoder Transformer Architecture," *The International Arab Journal of Information Technology*, vol. 21, no. 3, pp. 458-465, 2024. https://doi.org/10.34028/iajit/21/3/9

[11] Patel V., Modi A., Mistry H., Mishra A.., and et al., "From Alt-Text to Real Context: Revolutionizing Image Captioning Using the Potential of LLM," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 11, no. 1, pp. 379-387, 2025. DOI: 10.32628/CSEIT25111238

[12] Patil D., "Multimodal Artificial Intelligence in Industry: Integrating Text, Image, and Audio for Enhanced Applications Across Sectors," *SSRN*, pp. 1-9, 2025. DOI: 10.2139/ssrn.5057428

[13] Qian X. and Liu B., "Enhanced Image-Text Retrieval Based on CLIP with YOLOv10 and Next-ViT," *in Proceedings of the 4th International Conference on Computer Vision, Application, and Algorithm*, Chengdu, pp. 134862K-134862K-6, 2024. https://doi.org/10.1117/12.3055876

[14] Qin X., Li L., Pang G., and Hao F., "Heterogeneous Graph Fusion Network for Cross-Modal Image-Text Retrieval," *Expert Systems with Applications*, vol. 249, no. 1, pp. 123842, 2024. DOI: 10.1016/j.eswa.2024.123842

[15] Tan C., Zhang W., Qi Z., Shih K., and et al., "Generating Multimodal Images with GAN: Integrating Text, Image, and Style," *arXiv Preprint*, vol. arXiv:2501.02167v1, pp. 1-8, 2025. DOI: 10.48550/arxiv.2501.02167

[16] Wang J., Zhang H., Zhong Y., Liang Y., and et al., "Advanced Multimodal Deep Learning Architecture for Image-Text Matching," *in Proceedings of the IEEE 4th International Conference on Electronic Technology, Communication and Information*, Changchun, pp. 1185-1191, 2024. https://ieeexplore.ieee.org/document/10594167

[17] Wei S., Tan A., and Wang Y., "A Multiscale Hybrid Model for Natural and Remote Sensing Image-Text Retrieval," *in Proceedings of the 6th International Conference on Machine Learning, Big Data and Business Intelligence*, Hangzhou, pp. 41-45, 2024. DOI: 10.1109/MLBDBI63974.2024.10823800

[18] Yang X. and Tan L., "Multimodal Knowledge Graph Construction for Intelligent Question Answering Systems," *Australian Journal of Electrical and Electronics Engineering*, vol. 22, no. 4, pp. 755-764, 2025. DOI: 10.1080/1448837x.2024.2448376

[19] Yuan Y. and Xue H., "Multimodal Information

Integration and Retrieval Framework Based on Graph Neural Networks," *Preprints*, vol. 202501.0405.v1, pp. 1-8, 2025. DOI: 10.20944/preprints202501.0405.v1

[20] Zhang Q., Zhou Y., Yang Z., and Chen A., "Cross-Modal Prominent Fragments Enhancement Aligning Network for Image-Text Retrieval," *in Proceedings of the IEEE International Conference on Multimedia and Expo*, Ontario, pp. 1-6, 2024. https://ieeexplore.ieee.org/document/10687706

[21] Zheng W., Han N., Hu Y., Kang P., and et al., "Cross-Modal Image Text Retrieval Based on Cross-Optimization and Joint Strategy," *in Proceedings of the International Conference on Image, Signal Processing, and Pattern Recognition*, Guangzhou, pp.131802I-131802I-11, 2024. https://doi.org/10.1117/12.3034096

Shaik Asha completed her B.Tech. in Computer Science and Engineering from NIMRA Institute of Science and Technology, Vijayawada, and her M.Tech. in Computer Science and Engineering from VIF College of Engineering and Technology, Hyderabad. She is currently pursuing her Ph.D. in Computer Science and Engineering at Koneru Lakshmaiah Education Foundation (KLEF) University, Guntur, and Andhra Pradesh, India. She has 15 years of teaching experience and is presently working as an Assistant Professor in the Department of Computer Science and Engineering at Muffakham Jah College of Engineering and Technology, Hyderabad, Telangana, India. Her research interests include Artificial Intelligence, Machine Learning, Deep Learning, Computer Vision, Pattern Recognition, and Natural Language Processing. She has guided various UG and PG-level student projects and has actively contributed to academic and research activities. She has received awards from various professional bodies in recognition of her contributions.

Sajja Tulasi Krishna is currently working as an Assistant Professor in the Department of Computer Science and Engineering at Koneru Lakshmaiah Education Foundation (KL University), Guntur, Andhra Pradesh. With over 13 years of experience in teaching and research, she holds a Ph.D. in Computer Science, specializing in Deep Learning and Biomedical Image Processing. Her academic background includes M.Tech and B.Tech in Computer Science and Engineering. She has published 18 research papers in reputed SCIE, Scopus, and IEEE journals and conferences. Dr. Krishna received numerous accolades, including the "Best Teacher Award" (2025) and the "Young Researcher Award" (2022). She has also published work in areas such as COVID-19 detection, image classification, and healthcare diagnostics. Certified by Red Hat and Microsoft Azure, she has also completed several NPTEL and FDP programs. Her core areas of interest include Deep Learning, AI, Image Processing, Full Stack Web Development, and Programming Languages.