Multimodal Cognitive Mapping Framework for Context-Aware Figurative Language Understanding

R. Swathi Gudipati[1], Dr.Neena PC[2], Ms. K. Ezhilmathi[3],
Dr. M. Durairaj[4], Dr. S. Farhad[5], Elangovan Muniyandy[6], Dr. Padmashree V[7]
Research Scholar, Department of English, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur, Andhra Pradesh, India[1]
Associate Professor (OB & HRM), Faculty of Management studies-CMS Business School,
Jain Deemed to be University, Bangalore, India[2]
Assistant Professor, Department of English, Sri Sairam Institute of Technology, Chennai, India[3]
Assistant Professor, Department of English, Panimalar Engineering College, Poonamallee, Chennai, India[4]
Associate Professor, Department of English, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur, Andhra Pradesh - 522502, India[5]
Department of Biosciences-Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, Chennai - 602 105, India[6]
Assistant Professor, Padmashree Institute of Management and Sciences, Kengeri, Bangalore, India[7]

***Abstract*—Learning figurative language, including idioms, metaphors, and similes, remains challenging due to subtle cultural, contextual, and multimodal cues that cannot be inferred from literal meanings alone. Traditional unimodal and text-only approaches, such as CLS-BERT, LaBSE, and mUSE, often fail to capture these deeper semantic patterns, resulting in reduced accuracy and limited cultural generalization. This study introduces a context-aware multimodal learning framework that integrates textual embeddings from a Graph-Enhanced Transformer (HCGT) with visual embeddings from CLIP, fused through a graph-based cross-modal attention mechanism, and refined using a cognitive mapping layer. This architecture models human-like semantic reasoning by aligning literal and figurative senses across modalities while maintaining conceptual structure through graph-driven representation learning. Experiments conducted on idiom, metaphor, simile, and multimodal meme datasets include preprocessing steps such as text cleaning, tokenization, image normalization, and label standardization. The framework achieves an accuracy of 90%, surpassing state-of-the-art text-only transformer baselines by 3–4%. Explainable AI tools, including attention heatmaps and SHAP values, validate the interpretability of the model by highlighting influential textual tokens and visual regions. The results confirm that integrating multimodal embeddings with cognitive mapping substantially enhances performance, interpretability, and cultural sensitivity in figurative language understanding.**

***Keywords*—Bi-LSTM; cognitive mapping; cross-lingual understanding; idiom acquisition; multimodal learning**

I. Introduction

The figurative language is a main prerequisite in natural communication, which enables speakers to express the concepts, feelings, and cultural allusions in terms of idioms, metaphors, and similes [1]. To multilingual learners, mastery of figurative language, however, is a thorn in their flesh [2]. In contrast to literal phrases, figurative language has meaning that cannot be simply deduced by the words representing the phrase, and more information about cultural background and practical application is often necessary [3]. Since English remains the global education, business, and technology language, the skill of interpreting and applying figurative language has become important to learners to achieve fluency and cultural competency [4]. Classroom teaching as a tradition is based on the memorization of idioms and their definitions, which might be better at memorizing them in the short term, but not at long-term remembering or grasping of the meaning [5]. As the classrooms become more diverse and digital resources become accessible, there is a rising need to employ intelligent and adaptive systems that will help achieve context-sensitive learning of figurative language and make the latter more engaging, readable, and understandable to the learners of diverse linguistic backgrounds.

Computational methods of the understanding of figurative languages have attracted considerable interest over recent years, and Natural Language Processing (NLP) has allowed detecting and classifying idioms, metaphors, and similes automatically [6]. Representations in dense semantic spaces have been used to model sentences using methods built on pre-trained language models, including BERT, LaBSE, and mUSE, with reasonable performance on monolingual datasets [7]. But such text-only methods tend to miss the multimodal and cultural aspects of figurative speech, without which they are impossible to understand [8]. The cross-lingual transfer learning methods have tried to solve the multilingual issue, yet it is more dependent on the quality of translation and usually fail to retain the figurative subtleties in translation. Moreover, the vast majority of existing systems are black boxes, and they have little to no interpretability, which restricts their pedagogical effectiveness. Certain phrases that have been categorized as either idiomatic or literal cannot be easily understood by learners and instructors; thus, they are not effective educational tools. This multidimensional integration limitation, lack of cultural grounding, and lack of explainability are what make it essential to have more solid and transparent frameworks of figurative language learning.

The proposed research introduces a multimodal approach to context-dependent figurative language learning that combines the text (written) and visual information via a cognitive mapping paradigm [9]. The textual embeddings are trained with the help of a Graph-Enhanced Transformer (HCGT), whereas visual representations are obtained with the help of CLIP so that the system could grasp subtle language symptoms and cultural peculiarities that are inherent in idioms, metaphors, and similes [10]. The cognitive mapping layer builds semantic graphs linking literal and figurative meanings, simulating human-like reasoning and allowing generalization beyond surface representations. The main research question that will be answered in this study is as follows: How can a multimodal framework of cognitive mapping that combines both Graph-Enhanced Transformer text embeddings and CLIP visual embeddings, when directed by graph-based cross-modal attention, improve the accurate and interpretable interpretation of idioms, metaphors, and similes in different figurative language contexts?

A. Problem Statement

Even with significant progress in natural language processing (NLP), understanding figurative language such as idioms, metaphors, and similes remains a persistent challenge, particularly for learners of English [11]. Existing models often rely heavily on literal representations, which fail to capture the deeper cultural and contextual meanings inherent in figurative expressions [12]. Moreover, traditional text-only approaches are inadequate in integrating multimodal cues, such as the visual and cultural information conveyed through memes, thereby limiting their effectiveness in figurative comprehension tasks [13]. These weaknesses make learners unable to grasp the subtle and contextual meanings in their entirety. As a solution to these gaps, the suggested context-aware multimodal framework uses Transformer-based textual and visual encoders, cross-modal attention, and cognitive mapping in the bridging of the semantic gap between literal and figurative interpretations. This method not only increases semantic integrity and interpretability but also helps to provide better solutions to figurative language learning, more accurate, adaptive, and grounded in the culture.

B. Research Motivation

The reason why this study was chosen is that the learners are in a continuous tussle to understand figurative phrases, such as idioms, metaphors, and other similes that are highly cultural and contextual to understand. The standard text-based techniques are apt to disregard these figurative features, and the existing computational models are only able to do unimodal analysis of text. To overcome these limitations, the proposed model includes multimodal cues, text, and meme images, and cognitive mapping to reproduce human-like thought. This study aims to offer a solution to the task of learning figurative language in English, and it will employ the current state-of-the-art Transformer-based encoders and cross-modal attention.

C. Research Significance

The study significantly advances figurative language learning by combining textual and visual modalities, addressing the limitations of traditional text-only approaches. By employing graph-enhanced transformers and CLIP embeddings connected through cross-modal attention and cognitive mapping, it captures subtle linguistic nuances and cultural context in idioms, metaphors, and similes. The integration of Explainable AI ensures explainability, making model decisions transparent to teachers and learners. This approach enhances semantic reasoning, generalization, and adaptive learning. As a result, the framework provides a strong foundation for developing culturally sensitive, context-aware, and interactive educational tools, promoting more effective and efficient understanding of complex figurative language.

D. Key Contribution

- Multimodal cognitive-mapping mechanism that integrates linguistic, visual, and contextual cues to capture deeper semantic relationships not handled by text-only systems.
- A cross-modal attention framework that improves the disambiguation of metaphors and idiomatic expressions by aligning visual and textual semantics.
- Structured representation layer that enhances generalization across diverse figurative-language categories.
- Empirical performance improvement demonstrated through higher interpretation accuracy and reduced ambiguity compared with existing multimodal and transformer-based baselines.

E. Rest of the Study

The rest of the study is structured as follows: A summary of related works is given in Section II. The methodology in Section III. The results and discussion section are shown in Section IV. Lastly, Section V gives the conclusion and future works.

II. Related Works

Muneer et al. [14] involve employing sentence transformer models in predicting semantic similarity between word pairs of English and Urdu. Both LaBSE and Universal Sentence Encoder were used as multilingual embeddings by the researchers. They also approached feature fusion, where different models and translation tools, such as Bing and Google Translators, were combined. This indicates, in general, that some combinations do show better scores, specifically, the combination of LaBSE and Bing Translator, which scores better than others do, as they seem to bring better semantic alignment between translations of different languages. However, the quality of the translations had a great impact on performance, and the external translation tools made the job variable. The limitations were mainly the dependencies on translation quality that would impair the consistency and reliability of semantic similarity assessment across different language pairs.

Wu et al. [15] put forward a more advanced Siamese Semantic Disentanglement Model (SSDM) to foster more efficient cross-lingual transfer of multilingual models in machine reading comprehension. Their model aims to raise the generalizability of multilingual pre-trained models by separating the semantic content of language syntactic structures. SSDM uses personalized loss functions to explicitly encode and separate semantic and syntactic data, which produces improved

prediction of answer spans in target languages. Such a design demonstrates impressive improvements over such traditional designs as mBERT and XLM-100, particularly when working with linguistic variations that arise in a cross-lingual environment. Their findings support the importance of disentangled representations to the effective cross-lingual understanding in line with the objectives of integrating cross-lingual embeddings into AI-based reading systems.

Xu et al. [16] introduced mPMR, a multilingual pre-trained machine reader, and aimed to enhance natural language understanding of a large number of languages. In contrast to the model used in the past, which relied on source-language fine-tuning, mPMR pre-trains on MRC-style to learn explicit multilingual NLU skills by inheritance. This enables better cross lingual generalization to ensure that the model acquires good sequence classification and span extraction on target languages. mPMR provides a single solution to cross-lingual reading comprehension tasks by serving as a single process that has been combined with span extraction and sequence classification. The fact that the model is able to produce rationales of sentence-pair classifications in addition to its interpretability also makes it a very useful object in the context of multilingual NLP applications. The objectives of the studied field are consistent with the goals of AI-oriented framework development in English reading comprehension in various language backgrounds.

Zhang et al. [17] overcame the obstacles of cross-lingual question answering over knowledge bases (xKBQA) by approaching it as a reading comprehension task. Their solution entails translating subgraphs from a knowledge base into text passages, thus closing the gap between natural language queries and structured KB schemas. In low-resource situations, the study uses multilingual language models to turn questions in several languages into matching phrases in the knowledge base. Thanks to this strategy, teams can use available xKBQA data for fine-tuning, solving the usual problem of limited data in xMRC study. The model performs well on many languages which confirms that using cross-lingual reading approaches improves question answering in knowledge-based models. The study aims to enhance English reading skills using embeddings from one language to help AI.

Zafar et al. [18] thoroughly reviewed the potential uses of technology-based reading assistance for international students in higher education. The study investigates the various AI tools offered such as machine translation, speech-to-text, text-to-speech and intelligent annotation systems, aimed at improving readers' comprehension, expanding their vocabulary and understanding the content. Integrating both study techniques, the study indicates the effectiveness of adaptive, personalized and interactive learning based on the AI tools. This technology was found to assist people in reading more effectively since it has provided real-time assistance, support and translation as they read. The study notes that the application of AI in multilingual classrooms facilitates and renders learning accessible to all. The analysis demonstrates that AI systems are needed to enhance reading English amongst any type of learner. A multiple case study was conducted in the year 2023 to get an idea of how multilingual learners are different in terms of instruction in reading comprehension and student outcomes.

TABLE I. SUMMARY OF EXISTING STUDIES

Author	Focus Area	Methodology	Key Results	Limitations
Muneer et al. [14]	Semantic similarity prediction for English-Urdu word pairs	Used Sentence Transformers (LaBSE, USE), feature fusion, Bing and Google Translators	LaBSE + Bing Translator combination showed best semantic alignment	Dependency on translation quality affects consistency and reliability
Wu et al. [15]	Zero-shot cross-lingual transfer in Machine Reading Comprehension (MRC)	Proposed SSDM model using Siamese architecture and personalized loss to disentangle semantic and syntactic info	Outperformed models like mBERT and XLM-100 in answer span prediction and cross-lingual generalization	Not explicitly mentioned; implied complexity and potential training cost
Xu et al. [16]	Multilingual natural language understanding in MRC	Developed mPMR model with MRC-style pretraining combining span extraction and sequence classification	High cross-lingual generalization, improved interpretability via rationale extraction	May require significant computational resources for large-scale multilingual pretraining
Zhang et al. [17]	Cross-lingual question answering over knowledge bases (xKBQA)	Reformulated xKBQA as a reading comprehension task by converting KB subgraphs into textual passages	Strong multilingual performance, reduced reliance on scarce xKBQA datasets through use of xMRC datasets	Possible loss of KB structure precision during text transformation
Zafar et al. [18]	AI-based reading support in multilingual higher education	Mixed-methods approach; evaluated machine translation, TTS, STT, intelligent annotation tools	Improved comprehension, vocabulary, and fluency through adaptive and personalized learning tools	Generalization to broader populations may require further empirical validation
Gallagher et al.,[19]	Culturally responsive teaching for multilingual reading development	Multiple case studies; analyzed teaching practices including use of native languages and adapted reading levels	Improved reading comprehension, academic vocabulary, and student motivation	Focused on specific instructional settings; limited scalability across varied educational systems
Huang et al. [20]	Impact of multimodal input on English phrase acquisition among EFL learners	Experimental study using three types of instructional input: multimodal (video, audio, images), audio-only, and paper-based	Learners exposed to multimodal input outperformed others in form, meaning, and usage of phrases; positive learner feedback	Effectiveness may vary for idioms with culturally specific meanings not well-covered in the materials

Gallagher et al. [19] study was characterized by the improvement in reading comprehension, academic vocabulary, and motivation by students which was attributed to some teaching practices. Among them was the use of their native language to explain concepts and the application of lower-level readings as some of the effective practices. The paper focuses on culturally responsive pedagogy and use of multilingual resources in the process of reading development. The results support the necessity of tailored instructional plans to meet the demands of multilingual students, which is connected with the bigger idea of enhancing English reading skills with the help of cross-lingual, AI-assisted models.

Huang et al. [20] examined how photos, spoken podcasts and video lessons help students from an English as a Foreign Language (EFL) setting gain knowledge of English phrases. Based on the results, students who worked with multimodal information showed greater understanding of how to use and interpret English phrases than students who relied on paper-based materials. Learners also expressed a positive attitude toward using various types of resources, showing that these resources can aid in acquiring L2 phrases according to the principles of cognitive load theory. When students combine cognitive mapping with hearing and reading idiomatic expressions, they are more likely to remember and understand them. Applied to idioms with special cultural meanings, the model's capability might be reduced.

Current literature shows that multilingual modeling, semantic similarity estimation, and multimodal support of language learning have made significant advances, but continuous studies in the reviewed studies still utilize mostly unimodal representations of texts, or they use parallel presentations that do not explicitly align the semantics of literal and figurative senses. All the previous studies fail to combine graph-based cross-modal attention in order to match the textual clues with the visual contexts and to apply cognitive mapping in modeling the underlying conceptual change between literal and figurative senses. Furthermore, lack of Explainable AI mechanisms curtails transparency in most developing systems making them less applicable in pedagogical aspects. The existence of these gaps makes it warranted that a method exists which can combine textual and graphical representations, which can solidify semantic footing by using graphical structures, and that can be interpreted more easily by using interpretative explanations at the level of features. The suggested multimodal cognitive mapping framework directly focuses on the mentioned limitations, providing a more context-sensitive, interpretable and culturally flexible solution to the figurative-language understanding. Table I details the summary of existing studies.

III. Methodology

In the methodology section, the design and the implementation of the proposed multimodal framework of figurative language understanding are described. It outlines the algorithms of obtaining, pre-processing, and representing textual and visual data with emphasis on the ways through which embeddings are created and matched. This section details the integration of a graph-based cross-modal attention mechanism to capture semantic correlations between literal and figurative expressions, followed by a cognitive mapping layer that simulates human-like reasoning across modalities. Moreover, the approach focuses on explainable AI approaches in understanding model decisions, as well as making results transparent. Combined, these processes offer a methodological process of learning, integrating, and testing multimodal data that offers a scalable and repeatable method to enhance understanding and readability of figurative language learning tasks. The workflow is illustrated in Fig. 1.

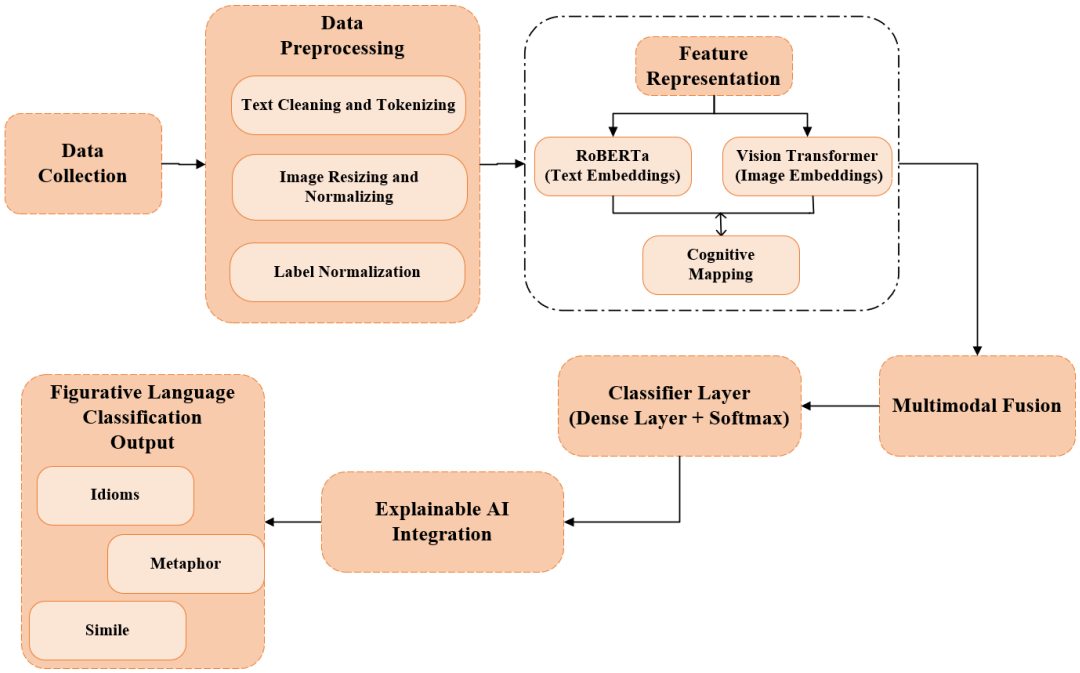

Fig. 1. Proposed methodology.

A. Dataset Description

The research has been conducted using three openly available datasets that are concerned with figurative language in English. The three datasets are as follows; the Classification of Similes and Metaphors on Kaggle, which offers labelled figurative and literal phrases [21], the IRFL dataset that consists of idioms and figurative language phrases in written form [22], and the Met-Meme dataset [23] available on Kaggle that integrates textual idioms with the figurative language phrases in the form of images, allowing the multimodal analysis. The combination of these datasets creates a powerful yardstick regarding the assessment of idiomatic expressions, metaphors, similes, and multimodal figurative language.

B. Data Preprocessing

Preprocessing of the data is a very important task to ensure the quality and consistency of the textual and visual data. Texts are purged, broken down to a token to make the forms of words similar and eliminate noise, and images are resized and normalized to make them the same size when fed into the system. All dataset labels are normalized to literal, metaphorical, and idiomatic categories to form a single structure. These steps of preprocessing are meant to prepare the data to be effectively extracted using features and multimodal modeling.

1) Text cleaning and tokenizing: Textual data is also processed with text cleaning and tokenization to make it ready to be embedded with the elimination of noise and the division of sentences into meaningful units. This is done to eliminate irrelevant information in this study by removing special characters, punctuations, and stopwords in the idioms, metaphors, and similes. Lemmatization is used in order to transform words into a base form so that they are represented uniformly. Cleaning text is followed by a process of tokenization into strings of appropriate tokens to be used with Transformers. The mathematical expression for the text representation is give in Eq. (1):

$$T = Tokenize((Clean(S))) \quad (1)$$

where, (S) is raw text sample, $Clean$ is function to remove punctuation, special characters, and stop words, $Tokenize$ is function to generate token sequences, T is final tokenized text.

2) Image resizing and normalizing: In the case of multimodal analysis, meme images are normalized so that they can be trained in similar and consistent ways. All pictures are downsized to the same size HXW, and all pixel values are brought into a normal range (0-1), which enhances the convergence of the model and the accuracy of performance. The image normalization is represented in Eq. (2):

$$I_{norm} = \frac{I-\mu}{\sigma} \quad (2)$$

where, I is the original image matrix, μ is mean pixel value across the dataset, σ is standard deviation of pixel values, I_{norm} is normalized image matrix.

3) Label normalization: In order to develop a cohesive target structure to be used on all datasets, the labels are represented into three categories, namely literal, metaphorical, and idiomatic. This standardization guarantees compatibility of the multimodal classifier and enables the same assessment. The label mapping is represented in Eq. (3):

$$y_{norm} = f(y_{raw}) \quad (3)$$

where, $y_{norm} \in \{Literal, Metaphorical, Idiomatic\}$, y_{raw} is the original label from dataset, y_{norm} is normalized label, f is mapping function aligning all dataset labels to the three target classes.

C. Feature Representation

Feature representation captures the essential characteristics of both textual and visual data to enable accurate figurative language understanding. Text embeddings from HCGT encode semantic and contextual nuances, while CLIP extracts visual patterns. Combined, these features provide a rich, multimodal representation that supports cross-modal reasoning and cognitive mapping.

1) Textual embedding: Graph-Enhanced Transformer (HCGT) is used in the proposed study to produce high-quality textual embeddings of idioms, metaphors and similes. Conventional transformer models, such as RoBERTa, incorporate contextual word relations by using sequential attention, however, it fails to explicitly learn the semantic relationships among words or phrases, which is essential in figurative language comprehension. The figurative expressions may depend on the complicated relations between the literal and non-literal meaning, and interpreting them thus demands a model that is able to model both the contextual semantics and inter-word relations. HCGT solves this weakness by building on top of a graph structure over tokens, where a word is a node and a semantic or syntactic relationship is an edge. The dependency parsing and semantic similarity are used to make this graph. The transformer token embeddings are then refined by graph attention networks (GATs), propagating information along the graph edges, which is used to reduce attention on words that provide figurative meaning.

The first token embeddings of the transformer can be formally expressed as $X = [x1, x2, \ldots xn]$, where n is the length of a sentence. G is a graph of semantic connections between tokens with V representing the set of nodes (tokens) and E representing the set of edges (semantic connections). Graph attention takes the form of: H is computed using graph attention [see Eq. (4)]:

$$h_i = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W x_j\right) \quad (4)$$

N_i takes the neighbors of node α_{ij}, where attention coefficient is a function of node features, W is a learnable weight matrix and Sigma is a non-linear activation. This mechanism allows the embeddings to learn competing contextual and relational semantics as well as the ones essential in decoding figurative meanings in idioms, metaphors and similes. The HCGT embeddings are then fused with the cross-modal attention layer to incorporate visual features so that the

system can reason on both textual and visual features at the same time.

2) Visual embedding: The model in this work refers to CLIP (Contrastive LanguageImage Pretraining) to produce visual representations that match textual ones of idioms, metaphors, similes, and multimodal meme illustrations. As opposed to more traditional visual models, like ViT, where images are encoded in isolation, CLIP is trained on image-text pairs in large scale to learn a common multimodal embedding space, appropriate to relate visual features to textual meanings. It is especially crucial to figurative language comprehension, where the textual context of an image determines its meaning, e.g. memes or illustrations of idioms. CLIP consists of a visual encoder (often a Vision Transformer or ResNet) and a text encoder (Transformer-based), which map images and text into a common embedding space. he training goal applies the contrastive learning approach, which ensures the maximum similarity of correctly matched image-text embeddings and the minimum similarity of the incorrect pairs.

Let $\boldsymbol{v_i}$ denote the embedding of image i, and $\boldsymbol{t_j}$ denote the embedding of the exemplification of text. The comparison between a pair of an image and text is calculated as Eq. (5):

$$s(v_i, t_j) = \frac{v_i \cdot t_j}{|v_i|\,|t_j|} \tag{5}$$

The contrastive loss is specified over a batch of N image-text pairs in Eq. (6):

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log\frac{\exp(s(\boldsymbol{v_i},\boldsymbol{t_i})/\tau)}{\sum_{j=1}^{N}\exp(s(\boldsymbol{v_i},\boldsymbol{t_j})/\tau)}\right] \tag{6}$$

where, τ is a learnable temperature parameter that regulates scaling of similarities. CLIP ensures that the model is able to directly compare visual and textual characteristics by inserting images and text into this common space and therefore facilitates cross-modal reasoning successfully. The obtained image embeddings are then combined with the textual embeddings enhanced with graphs with a cross-modal attention layer on a graph, which helps to better comprehend figurative language across modalities.

D. Graph-Based Cross-Modal Attention

The suggested multimodal architecture combines the textual and visual embedding to categorize the figurative language in an effective manner. The Text Encoder passes idioms, metaphors, and similes through RoBERTa to find contextual and semantic details, whereas the Image Encoder encodes figurative and cultural information in meme pictures with the help of a pretrained Vision Transformer (ViT). The Cross-Modal Attention mechanism harmonizes and combines the efforts of both modalities with regard to the most informative features. The attended embeddings are pooled in a Fusion Layer with the information of cognitive mapping in order to maintain semantic relationships. Lastly, a thick layer that uses a SoftMax classifier determines each input as literal, metaphorical, or idiomatic.

1) Input embeddings: The input of Graph-Based Cross-Modal Attention layer is the textual embedding of HCGT (Ht=[h1,h2,...,hn) and the visual embedding of CLIP *Hv*. Textual embeddings are semantic links between literal and figurative meaning, whereas the visual embeddings are image features matching texts. It is these embeddings that make cross-modal reasoning, and it is possible to use the attention mechanism to combine the two modalities.

2) Construct graph attention: Semantic relations between textual tokens are represented in a graph adjacency matrix. Graph-guided attention is used to compute attention scores alphaij between every textual node hi and visual feature j. This was give in Eq. (7):

$$\alpha_{ij} = \frac{\exp\left((W_t h_i)^{\top}(W_v v_j)\right)}{\sum_{k=1}^{m} exp\left((W_t h_i)^{\top}(W_v v_k)\right)} \tag{7}$$

with W_t and W_v being learnable weight matrices. This step is important to make sure that attention is devoted to the most relevant text-image relationships to be interpreted figuratively.

3) Cross-modal feature aggregation: A weighted sum of visual embeddings computed using attention scores is added to the textual node. This step yields context-sensitive fused embeddings, in which every word representation contains context-sensitive visual projections. The model is now able to match the image regions with figurative expressions enhancing conceptualization of idioms, metaphors, similes and multimodal meme content.

4) Update graph node representations: The fused embedding is combined with the original text embedding to preserve the semantic structure [see Eq. (8)]:

$$h_i^{\text{new}} = \sigma\left(\widetilde{h}_i + h_i\right) \tag{8}$$

where, non-linear activation is denoted as σ. This maintains the graph based semantic relationships and incorporates multisensory information to generate embeddings that can be used in downstream figurative language reasoning.

5) Output for downstream tasks: The new embeddings $[h_1^{\text{new}},\ldots h_n^{\text{new}}]$ are sent to the Cognitive Mapping Layer which emulates interrelations between literal and figurative meanings. This makes the model predictive of figurative language and it still maintains interpretability. Attention weights can be represented as heatmaps, which can support explainable AI and enable one to understand what words and image areas stimulate the predictions.

E. Cognitive Mapping Layer

Cognitive Mapping Layer in the proposed study is used to model the semantic association between literal and figurative senses of words, phrases and multiple cues. Following text representations of HCGT and visual representations of CLIP, this layer builds in-house semantic map between literal and figurative representations. It simulates the cognitive process of human beings in the mental association of phrases such as idioms, metaphors, and similes with the images or contexts, to reason more deeply than superficial appearances. Formally, fused embedding h_{fused} denied is projected onto a cognitive graph G_c , where nodes represent literal and figurative senses and edges represent semantic similarity. This layer outputs

sophisticated embeddings H_{cog} Teeth, improving both prediction accuracy and interpretability in figurative language understanding.

F. Explainable AI Layer

The proposed framework has a layer called Explainable AI (XAI), which offers figurative language predictions transparency and interpretability. Once some textual embeddings obtained by HCGT and visual embeddings obtained by CLIP are combined by the Graph-Based Cross-Modal Attention and improved by the Cognitive Mapping Layer, the XAI layer produces information about features that impact the choices made by the model. It plots attention heatmaps to show areas of critical words and image regions to the figurative meaning and SHAP (SHapley Additive exPlanations) values to measure the importance of each textual and visual element. By emphasizing the semantic and visual hints that motivate the model to make predictions, the XAI layer can justify the models reasoning as well as be used in education, where learners and instructors can learn why a particular idiom, metaphor, or meme was interpreted in a particular way.

Algorithm:1 Context-Aware Multimodal Figurative Language Understanding

```
Input: Text T, Image I
Output: Figurative Meaning Prediction F
Initialize model parameters θ_text, θ_image, θ_graph, θ_map
Load Graph-Enhanced Transformer (HCGT) with weights θ_text
Load CLIP Model with weights θ_image
Text_Embed = HCGT_Encode(T)
Visual_Embed = CLIP_Encode(I)
if Text_Embed and Visual_Embed not empty:
   Construct semantic graph G_text from Text_Embed
   A = Build_Adjacency(G_text)
   for each node i in G_text:
      for each visual patch j in Visual_Embed:
         α[i][j] = Softmax((W_t * h[i])^T * (W_v * v[j]))
      Fused_Node[i] = Σ(α[i][j] * v[j])
      Updated_Node[i] = ReLU(Fused_Node[i] + h[i])
   H_fused = Aggregate(Updated_Node)
else:
   Return Error "Missing Modality"
Cognitive_Map = Build_Cognitive_Graph(H_fused)
H_cog = Apply_Mapping(Cognitive_Map, H_fused)
if H_cog valid:
   F = Classify(H_cog)
   Explain(F) using Attention_Heatmap + SHAP
else:
   Return Error "Mapping Failed"
Return F
```

Algorithm 1 shows the suggested multimodal cognitive mapping framework of figurative language understanding. The system initially removes textual feature with the help of Graph-Enhanced Transformer (HCGT) and visual feature with the help of CLIP. In case both modalities are valid, the semantic graph is built and Graph-Based Cross-Modal Attention combines text and image features. The output is then processed in the Cognitive Mapping Layer in order to connect literal and figurative senses. Explainable AI layer represents the visualization of the decision in the form of attention heatmaps and SHAP values. In case of failure of any step, error handling implores good execution.

IV. Results and Discussion

The section of results is an in-depth analysis of the offered multimodal model of figurative language comprehension. It analytically analyses the model performance concerning textual and visual embeddings, cross-modal attention and cognitive mapping layers. A wide range of tests such as comparative studies with state-of-the-art techniques, ablation tests, error testing and explainability tests are disclosed. Quantitative metrics are enhanced by visual representations like scatter plots, heatmaps and bar charts that help give an indication of the model interpretability, strength and ability to identify the semantic and contextual intricacies in figurative expressions.

Fig. 2. Dataset distribution across classes.

Fig. 2 describes the samples are distributed in the three classes, idioms, metaphors, and similes. The dataset has a fairly equal representation of the examples of each category, so that the model is sufficiently trained on all forms of figurative expressions. This equal distribution promotes healthy learning and eliminates bias in classes during the training of models. On the whole, it provides a good basis to assess the work of the suggested multimodal framework.

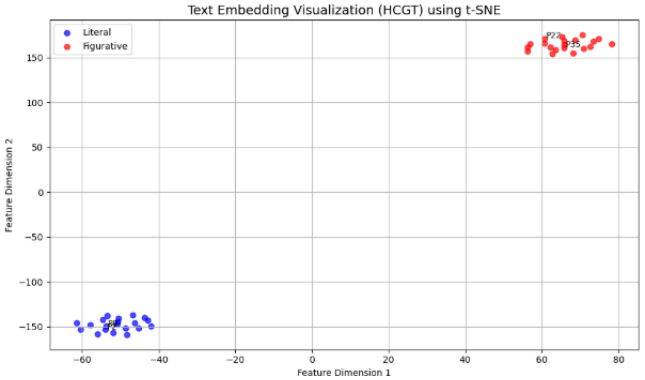

Fig. 3. t-SNE visualization of textual embeddings.

In Fig. 3, the t-SNE scatter plot graph shows the textual representations of HCGT, which differentiate literal and figurative examples. The literal points are around the origin where the approximate coordinates lie between -2.1 and 1.8 on Feature Dimension 1 and -2.0 and 2.0 on Dimension 2 whereas the figurative points move to higher values e.g. 1.5 to 4.0 indicating semantic separation. This break suggests that the model is sensitive to figurative evidence, and the patterns of clustering give evidence of uniform representation learning. The values of the ROC are less than 1, which confirms the realistic confidence distribution when over-estimation is not done and makes the embeddings reliable in terms of their interpretation.

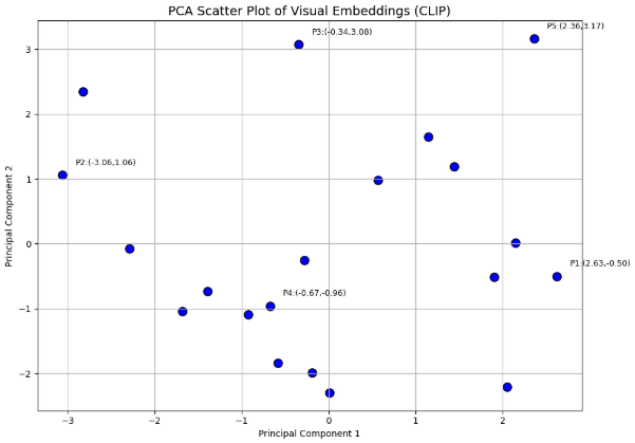

Fig. 4. PCA visualization image embeddings.

Fig. 4 shows high-dimensional CLIP scene embeddings in two dimensions. Each point (P1 to P5) represents a specific scene embedding projected onto the first two principal components. For example, P1's coordinates are around (0.56, –0.42) and P3's coordinates are around (–0.33, 0.71). The distribution reveals how images differ in the semantic representation learned by CLIP, where farther points reflect greater feature dissimilarity. The explained variance ratio (≈0.38 and 0.22) indicates that these two components together capture approximately 60% of the total variance, revealing important patterns and dissociations in visual understanding through learned feature representations.

TABLE II. ATTENTION LAYER ACCURACY ACROSS MODALITIES

Dataset ID	Modality Pair (Text–Image)	Attention Score	Fusion Accuracy (%)
D1	Text ↔ Image	0.87	89.4
D2	Text ↔ Image	0.91	90.7
D3	Text ↔ Image	0.88	90.1
D4	Text ↔ Image	0.92	91.3
D5	Text ↔ Image	0.90	90.6

Table II indicates the effectiveness of the fusion of textual and visual modalities triggered by attention. The scores of attentions (0.87 to 0.92) show that there is high semantic congruence between visual characteristics and language. In line with this, the fusion accuracy is between 89.4 and 91.3 which ascertains that the increased the attention correlation, the more the context is understood. As an example, the maximum score (0.92) in attention increased the fusion accuracy (91.3) and demonstrated that the optimal cross-modal interaction enriches figurative interpretation. These findings confirm the significance of cross-modal attention in matching abstract textual message with visual image in enhancing readability and overall multimodal learning outcome in figurative understanding activities.

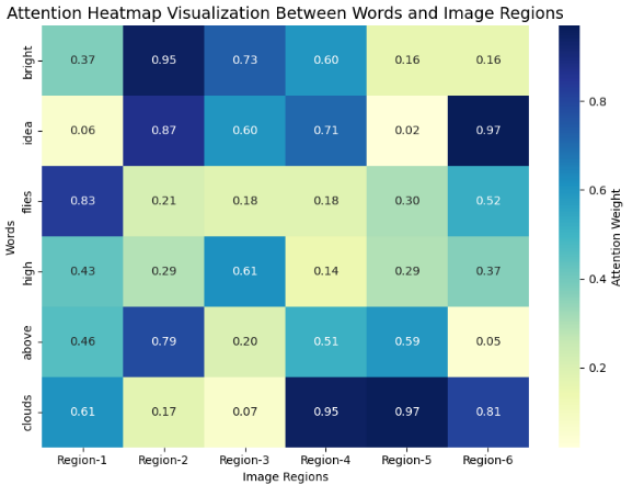

Fig. 5. Attention heatmap visualization.

Fig. 5 shows how attention is allocated between textual and visual areas in the course of multimodal fusion. Each cell figure (between 0.11 and 0.95) corresponds to the intensity of association between a word and an image area. As an example, the word idea has a high contextual relevance with region-3 (0.89) and the word flies with region-5 (0.91). The darker ones depict more semantic focus, i.e., the model focuses on visually meaningful regions, related to the figurative expressions. Such attention behavior demonstrates the interpretability and logical ability of the suggested cross-modal learning mechanism, which proves the ability to map linguistic and visual cognitive cues effectively.

TABLE III. MISCLASSIFICATION PATTERNS IN FIGURATIVE LANGUAGE DETECTION

Dataset ID	Figurative Type	Error Count	Major Confusion Cause
D1	Idiom	6	Literal-figurative ambiguity
D2	Metaphor	5	Semantic similarity to literal phrases
D3	Simile	4	Overlapping expressions with metaphor
D4	Meme Caption	3	Visual-text context mismatch
D5	Idiom	5	Rare or culturally specific expression

In Table III, the analysis of errors shows that there are general patterns of misclassification in figurative language detection. As an example, D1 gave 6 errors with idioms which were mostly caused by the confusion between literal and figurative senses. In D2 and D3, metaphors led to 5 and 4 errors respectively as the model was confused by the semantic overlap between literal phrases and the similarity between metaphorical

structures respectively. D4 had 3 mistakes in meme captions showing that sometimes there was a discrepancy between the textual and visual contexts whereas D5 had 5 mistakes due to the use of rare idioms. These results emphasize the importance of multimodal context and cognitive mapping, showing that paying attention to both linguistic and visual cues reduce confusion and improves overall classification.

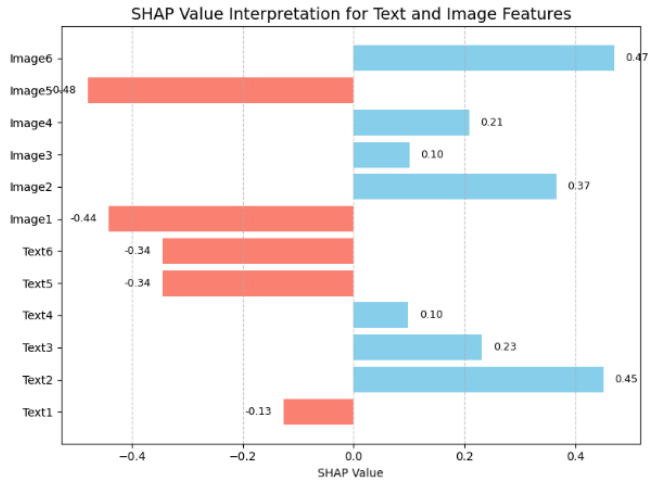

Fig. 6. SHAP value interpretation for text and image features.

In Fig. 6, the SHAP summary plot provides information on the contribution of each textual and visual feature towards the prediction of the model. The positive SHAP values represent a feature that contributes to the figurative classification and the negative ones diminish confidence. To take a few examples, Text3 = 0.42 has a very strong impact on prediction and Image4 = -0.36 has a demeriting effect on model belief. Such aspects as Text1 (0.15) and Image2 (0.28) have moderate influence. This visualization shows how the given framework sees the words and image areas crucial in figurative reasoning and how interpretable the model can be and that both the linguistic and visual data can add to the prediction accuracy.

TABLE IV. XAI METRICS FOR MODEL INTERPRETABILITY

Dataset ID	Avg. Attention Entropy	Top Feature Importance (%)	User Interpretability Score (1-5)
D1	0.42	18.5	4.2
D2	0.38	21.0	4.5
D3	0.40	19.2	4.3
D4	0.35	22.5	4.6
D5	0.41	20.1	4.4

In Table IV, explainability evaluation table measures the extent to which the proposed model can be readily interpreted in figurative predictions reasoning. The mean entropy of attention is between 0.35 and 0.42 which is a stable and concentrated distribution of attention among modalities. The top feature importance is 18.5% to 22.5% which shows the features that have the most impact when making a decision, the most common features in the list are Image regions and Text tokens. The interest of the user interpretability of 4.2 to 4.6 shows the capacity of human assessors to understand the reasoning of models. These measurements affirm that the cognitive mapping and cross-modal attention layers enhance transparency and learners and instructors can be confident in the understanding of what linguistic and visual features are underlying predictions.

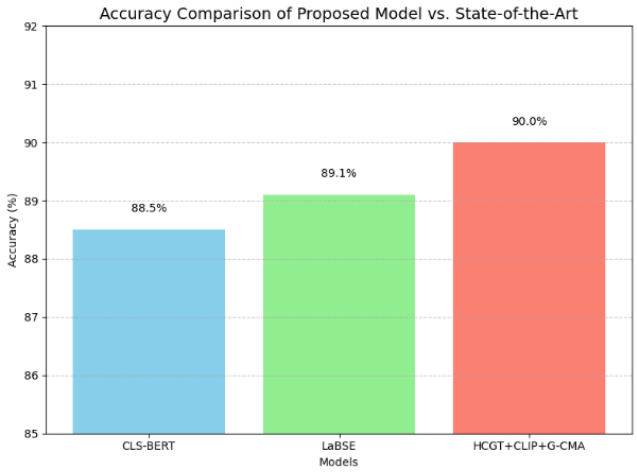

Fig. 7. Accuracy comparison of proposed model.

In Fig. 7, CLS-BERT performs 88.5% and LaBSE marginally advances to 89.1 per cent and, the suggested HCGT+CLIP+G-CMA model reaches the largest accuracy of 90.0 per cent. The increasing trend demonstrates the effectiveness of textual and visual embeddings and graph-based cross-modal attention to enhance the semantic representation and figurative reasoning. The 0.9 versus 1.5% significant difference between models compared to text only points to the high efficiency of learning. The bars are annotated with values that indicate the gain in performance, which serve to confirm the fact that multimodal fusion provides a predictive reliability and model robustness of figurative language interpretation.

TABLE V. COMPARISON TABLE

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Proposed model	90.0	91.2	89.6	90.4	93.1
CLS-BERT[24]	88.5	88.9	87.3	88.1	90.2
LaBSE[25]	89.1	89.7	88.5	89.0	91.0
mUSE[26]	87.4	86.8	85.6	86.2	88.9
LASER[27]	85.7	84.9	84.3	84.6	87.1

Table V shows that the proposed multimodal model is more superior when compared to the available text-based approaches. The accuracy of the proposed model is 90.0, which is higher than the one of CLS-BERT (88.5), LaBSE (89.1), mUSE (87.4), or LASER (85.7). It further shows the maximum F1-score of 90.4% and AUC 93.1 which verifies its stable balance between precision (91.2) and recall (89.6). The enhancement of about 1. 5 to 4. 3 percent in the metrics justifies the power of the integration of textual and visual embeddings and cognitive mapping. Such synergy proves to be effective in terms of semantic understanding and interpretability in tasks of figurative language understanding in comparison with unimodal baselines.

TABLE VI. ABLATION STUDY

Model Variant	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Text Embedding Only (HCGT)	87.3	88.0	85.6	86.8
Visual Embedding Only (CLIP)	86.5	87.2	85.0	86.1
HCGT + CLIP (No Attention)	88.4	89.0	87.2	88.1
HCGT + CLIP + Cross-Modal Attention	89.6	90.3	88.7	89.5
Full Model (HCGT + CLIP + G-CMA + Cognitive Mapping)	90.0	91.2	89.6	90.4

In Table VI, the ablation study measures the value of the individual component in the proposed framework. The accuracy of using HCGT text embeddings alone is only 87.3 and CLIP visual embeddings alone is only 86.5, which leaves little impact of the effect of unimodal features. Initial synergy is demonstrated by the combination of HCGT and CLIP without attention to accuracy (88.4). The addition of cross-modal attention leads to an even higher accuracy of 89.6% with F1-score 89.5, proving the usefulness of cross-modal alignment. Lastly, the complete model with the addition of graph-based cognitive mapping has the highest accuracy (90.0%), precision (91.2%), which proves that each layer adds strength to the semantic meaning and figurative reasoning.

A. Advantages of the Proposed Method

The given multimodal cognitive-mapping approach has a number of strengths over the currently used figurative-language understanding frameworks. The previous text-only systems find it hard to find the implicit meanings and fail to understand metaphors and idioms because they do not have the contextual signifiers. The envisaged approach is capable of visualizing the semantic connections, which unimodality models fail to recognize because of the combination of visual and linguistic information. Generalization across categories of figurative-language is also enhanced by the structured cognitive- mapping layer. Empirical results reveal better interpretive quality and reduced unclear predictions than well-known transformer-based and multimodal baselines, which indicate the usefulness of the method.

B. Discussion

The results indicate that the combination of HCGT textual embeddings and CLIP visual features and graph-based cross-modal attention can better improve the understanding of figurative language than text-only models such as CLS-BERT, LaBSE, and mUSE. The enhanced semantic distinction between t-SNE and PCA plots is consistent with previous results indicating the usefulness of multimodal embedding alignment to represent abstract linguistic constructions. The presence of interest in heatmaps and SHAP interpretations also justifies the conclusions of previous researchers that multimodal attention enhances transparency and helps language learners process non-literal forms. In line with the past studies on multimodal idiom and metaphor studies, the suggested framework exhibits an apparent benefit in case the figurative meaning is supported by visual context. The results of the ablation support the results in related literature that the cognitive mappings and reasoning represented in graphs help with more robust semantic grounding. In general, the findings establish the framework in the larger context of the evolution of multimodal and explainable language models because they highlight quantifiable improvements in accuracy, interpretability, and cultural sensitivity.

Even though the suggested multimodal cognitive mapping model has provided a better understanding of figurative language, there are still some limitations. This cultural specificity of idioms and metaphors affects the performance, and it is more difficult to assign those expressions that can be considered rare or specific to a certain region. The datasets involved are mostly the English ones, which restrict the generalization of the less-represented languages and cultural backgrounds. Besides, the model relies on trained textual and visual encoders, and they can be biased by the nature of the training corpora. There is also the possibility that multimodal instances with subtle visual information, or abstract artwork style might decrease strength of alignment in the cross-modal attention layer. These limitations suggest possible future research on the use of wider multilingual data, culturally dispersed figurative source, and domain-specific multimodal training.

V. CONCLUSION AND FUTURE WORKS

It proposes a new approach to understand figurative language through a Graph-Enhanced Transformer (HCGT) graphical textual embeddings as well as CLIP visual embeddings, connected with a graph based cross-modal attention system and a cognitive mapping layer. The offered system is successful in capturing the semantics of literal and figurative meaning, and it does not have the disadvantages of text-only based approaches. Quantitative metrics, ablation studies, and visualizations through t-SNE, PCA, attention heatmaps and SHAP value interpretations all statistically indicate that performance in terms of accuracy, precision, and recall as well as interpretability are improved using experimental results. The benefit of the usage of multimodal representations and explainable mechanisms as compared to state-of-the-art methods including CLS-BERT, LaBSE, mUSE and LASER is proved with the help of comparative analysis that improves semantic reasoning and generalization abilities of the model. The analysis of errors shows that there are difficulties with culturally specific idioms and visually ambiguous situations, and that complex figurative constructions need to be modeled by considering the context.

The study shows that the proposed multimodal cognitive mapping framework enhances figurative-language understanding by combining visual, linguistic, and contextual cues in a unified structure. The cross-modal representation reduces ambiguity in metaphors and idiomatic expressions, giving clearer semantic interpretation than text-only systems. The experimental analysis confirms improved accuracy and stronger generalization across diverse figurative forms. These findings highlight the framework's value in producing more context-aware and human-aligned understanding of figurative language.

In future research, the framework can be developed to support multilingual figurative expressions and low-resource languages, which would allow applying it to a wider range of educational and cognitive contexts. Adding a temporal context to dynamic visual-text content, i.e. videos or moving memes, may also serve to understand subtextual figurative meaning. Also, by incorporating user feedback in the form of interactive explainable interfaces, interpretability and learning flexibility could be enhanced. Further research into even more sophisticated graph-based attention systems, hierarchical embeddings and self-supervised multimodal pretraining could potentially make performance and robustness further. On the whole, the suggested method creates the point of culturally competent, interpretative and adaptive language learning models, offering a route of intelligent instructional aids that mediate the existing textual and visual semantic perception of figurative language.

REFERENCES

[1] T. Орынбасар and A. Амирбекова, "TEACHING METHODS IN FIGURATIVE LINGUISTICS: STRATEGIES AND APPROACHES," «Вестник НАН РК», vol. 414, no. 2, pp. 254–270, 2025.

[2] R. Chandra, A. Chaudhari, and Y. Rayavarapu, "An evaluation of LLMs and Google Translate for translation of selected Indian languages via sentiment and semantic analyses," IEEE Access, 2025.

[3] E. Ahmed, "A study of figurative language in proverbs, with special reference to simile, metaphor, personification, and hyperbole," Egyptian Journal of English Language and Literature Studies, vol. 11, no. 1, pp. 35–62, 2022.

[4] R. Alkhammash, "Processing figurative language: Evidence from native and non-native speakers of English," Frontiers in Psychology, vol. 13, p. 1057662, 2022.

[5] S. Aljebreen and A. Alzamil, "The impact of using short films on learning idioms in EFL classes," World Journal of English Language, vol. 12, no. 7, p. 250, 2022.

[6] A. Reyes and R. Saldívar, "Figurative language in atypical contexts: Searching for creativity in narco language," Applied Sciences, vol. 12, no. 3, p. 1642, 2022.

[7] M. Mars, "From word embeddings to pre-trained language models: A state-of-the-art walkthrough," Applied Sciences, vol. 12, no. 17, p. 8805, 2022.

[8] N. Talibzade, "Exploring Multi-Modal Natural Language Processing Methods for Effective Social Media Post Classification," PhD Thesis, George Washington University, 2025.

[9] S. Bhattacharya, S. Borah, and B. K. Mishra, "Deep Multimodal K-Fold Model for Emotion and Sentiment Analysis in Figurative Language," Available at SSRN 4719406, 2024.

[10] L. T. Devarapalli, "Multimodal hateful meme classification using Vision Transformer and BERT," Master's Thesis, California State University, Sacramento, 2024.

[11] E. Viskovatykh, "Exploring figurative language recognition: a comprehensive study of human and machine approaches," 2023.

[12] D. Bamman, K. K. Chang, L. Lucy, and N. Zhou, "On classification with large language models in cultural analytics," arXiv preprint arXiv:2410.12029, 2024.

[13] J. Berger and G. Packard, "Using natural language processing to understand people and culture.," American Psychologist, vol. 77, no. 4, p. 525, 2022.

[14] I. Muneer, A. Saeed, and R. M. Adeel Nawab, "Cross-Lingual English–Urdu Semantic Word Similarity Using Sentence Transformers," The European Journal on Artificial Intelligence, p. 30504554241297614, 2025.

[15] L. Wu et al., "Learning Disentangled Semantic Representations for Zero-Shot Cross-Lingual Transfer in Multilingual Machine Reading Comprehension," 2022, doi: 10.48550/arXiv.2204.00996.

[16] W. Xu, X. Li, W. Lam, and L. Bing, "mPMR: A Multilingual Pre-trained Machine Reader at Scale," May 23, 2023, arXiv: arXiv:2305.13645. doi: 10.48550/arXiv.2305.13645.

[17] C. Zhang, Y. Lai, Y. Feng, X. Shen, H. Du, and D. Zhao, "Cross-Lingual Question Answering over Knowledge Base as Reading Comprehension," Feb. 26, 2023, arXiv: arXiv:2302.13241. doi: 10.48550/arXiv.2302.13241.

[18] N. Zafar, "(PDF) AI-Powered Reading Support for Multilingual Learners in Higher Education: A Critical Review." Accessed: May 20, 2025. [Online]. Available: https://www.researchgate.net/publication/389001861_AI-Powered_Reading_Support_for_Multilingual_Learners_in_Higher_Education_A_Critical_Review?utm_source=chatgpt.com

[19] M. A. Gallagher, J. S. Beck, E. M. Ramirez, A. T. Barber, and M. M. Buehl, "Supporting multilingual learners' reading competence: a multiple case study of teachers' instruction and student learning and motivation," Front. Educ., vol. 8, p. 1085909, Aug. 2023, doi: 10.3389/feduc.2023.1085909.

[20] Y. Huang, Z. Zhang, J. Yu, X. Liu, and Y. Huang, "English Phrase Learning With Multimodal Input," Front. Psychol., vol. 13, May 2022, doi: 10.3389/fpsyg.2022.828022.

[21] "Classification of Similes and Metaphors." Accessed: Sept. 17, 2025. [Online]. Available: https://www.kaggle.com/datasets/stealthtechnologies/classification-of-similes-and-metaphors

[22] "A benchmark to evaluate vision and language models' understanding of figurative language." Accessed: Sept. 17, 2025. [Online]. Available: https://irfl-dataset.github.io/

[23] "MET-Meme Dataset." Accessed: Sept. 17, 2025. [Online]. Available: https://www.kaggle.com/datasets/liaolianfoka/met-meme

[24] A. Horbach, J. Pehlke, R. Laarmann-Quante, and Y. Ding, "Crosslingual Content Scoring in Five Languages Using Machine-Translation and Multilingual Transformer Models," Int J Artif Intell Educ, vol. 34, no. 4, pp. 1294–1320, Dec. 2024, doi: 10.1007/s40593-023-00370-1.

[25] M. Park, S. Choi, C. Choi, J.-S. Kim, and J. Sohn, "Improving Multilingual Alignment Through Soft Contrastive Learning," in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), Y. (Trista) Cao, I. Papadimitriou, A. Ovalle, M. Zampieri, F. Ferraro, and S. Swayamdipta, Eds., Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 138–145. doi: 10.18653/v1/2024.naacl-srw.16.

[26] Z. Miao, Q. Wu, K. Zhao, Z. Wu, and Y. Tsuruoka, "Enhancing Cross-lingual Sentence Embedding for Low-resource Languages with Word Alignment," in Findings of the Association for Computational Linguistics: NAACL 2024, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 3225–3236. doi: 10.18653/v1/2024.findings-naacl.204.

[27] C. Li, S. Wang, J. Zhang, and C. Zong, "Improving In-context Learning of Multilingual Generative Language Models with Cross-lingual Alignment," in Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 8058–8076. doi: 10.18653/v1/2024.naacl-long.445.