Course Project Report

MDCKE:Multimodal Deep Context Knowledge Extractor that Integrates Contextual Information

Submitted By

Gaurav Kumar (221IT027)
Vishwajeet H D (221IT084)
Krishna Tulsyan (221EE130

as part of the requirements of the course

IT416 Computer Vision [Jul-Nov 2025]

in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology in Information Technology

under the guidance of

Dr. Dinesh Naik, Dept of IT, NITK Surathkal

undergone at

DEPARTMENT OF INFORMATION TECHNOLOGY

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA, SURATHKAL

JUL-NOV 2025

DEPARTMENT OF INFORMATION TECHNOLOGY

National Institute of Technology Karnataka, Surathkal

CERTIFICATE

This is to certify that the Course project Work Report entitled **"MDCKE:Multimodal Deep Context Knowledge Extractor that Integrates Contextual Information"** is submitted by the group mentioned below -

Details of Project Group

Name of the Student	Register No.	Signature with Date
Gaurav Kumar	221IT027	
Vishwajeet H D	221IT084	
Krishna Tulsyan	221EE130	

this report is a record of the work carried out by them as part of the course **Computer Vision (IT416)** during the semester **Jul-Nov 2025**. It is accepted as the Course Project Report submission in the partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Information Technology.**

(Name and Signature of Course Instructor)
Dr. Dinesh Naik
Assistant Professors
Dept. of IT, NITK Surathkal

DECLARATION

We hereby declare that the project report entitled **"MDCKE:Multimodal Deep Context Knowledge Extractor that Integrates Contextual Information"** submitted by us for the course **Computer Vision (IT416)** during the semester **Jul-Nov 2025**, as part of the partial course requirements for the award of the degree of Bachelor of Technology in Information Technology at NITK Surathkal is our original work. We declare that the project has not formed the basis for the award of any degree, associateship, fellowship or any other similar titles elsewhere.

Details of Project Group

Name of the Student	Register No.	Signature with Date
1. Gaurav Kumar	221IT027	
2. Vishwajeet H D	221IT084	
3. Krishna Tulsyan	221EE130	

Place: NITK, Surathkal
Date: 28/11/25

MDCKE:Multimodal Deep Context Knowledge Extractor that Integrates Contextual Information

Gaurav Kumar[1], Vishwajeet H D[2],Krishna Tulsyan[3]

Abstract— **The swift transformation of multimodal information, which involves textual information, incorporation of visual information, and also the integration of information and reasoning on the heterogeneous source of information, has posed a grave need to have systems that can extract, integrate and reason about the information. The Multimodal Knowledge Graph Construction Framework proposed in the offered project presupposes the implementation of the existing models of deep learning and natural language processing to form one and understandable knowledge form. It may utilize two sources (images and text) to execute them with the help of such architects as YOLOv3 (object detection and localization) and Swin Transformer (hierarchical visual feature extraction) and textual data with the help of a Constituency Parser which identifies syntactic structures. It contains a Vision Transformer (ViT) that creates high-level semantic features of images, which is required to add additional textual information after enhanced interpretations of the visual information. These textual and visual attributes are obtained and combined with a Visual Prefix Fusion mechanism in order to get the cross-modes semantic correspondence. The fused multimodal presentation is in turn fed into BERT to generate deep contextual representations that retain association between visual and textual chunks. Then, the Named Entity Recognition (NER) and Transformer-based Relation Extraction (RE) modules detect entities and relations between entities and their meanings and create the nodes and the edges of the resulting Knowledge Graph. This semantic graph based on images and text allows acquiring more contextual knowledge and forming improved reasoning as evidenced by the fact that multimodal learning may possibly lead to the increased use of structured knowledge in an AI system.**

Keywords: **Multimodal learning, Knowledge graph, BERT,NER,Visualtext fusion, Object detection,Transformer models, Vision Transformer,Semantic representation.**

Dataset Link `https://www.kaggle.com/datasets/marquis03/more-a-multimodal-relation-extraction-dataset`

I. INTRODUCTION

This massive increase of multimodal data, the visual and textual information, has posed great difficulties in the representation of knowledge, semantic interpretation and context-based inferences. The systems of traditional knowledge extraction have concentrated on the text based inputs thus ignoring the rich semantic information woven in images, and vision systems whereas effective in object detection and scene interpretation, frequently do not provide linguistic context. Such disconnection is a hindrance to the creation of interpretable and comprehensive AI models. As the remedy to these shortcomings, the Multimodal Knowledge Graph Construction Framework is proposed which includes high-level deep learning and natural language processing models to combine the information sources of images and text. Object detection and hierarchical visual feature extraction of visual data are performed by the system on the basis of YOLOv3 and the Swin Transformer, respectively. Concurrently, a Constituency Parser is used to extract syntactic structures on textual data, and a Vision Transformer (ViT) is used to produce high-level semantic representations based on images that supplement textual interpretation without the use of BLIP. Visual prefix fusion It is a mechanism that aligns visual and textual embeddings in a common semantic space, followed by the further refinement of the fused representation by BERT. Entities and relations between them are then recognized by Named Entity Recognition (NER) and Transformer-based Relation Extraction (RE) modules and combine the nodes and edges of the final Knowledge Graph (KG), supporting semantic search, content analysis, and automated reasoning.

The rationale of this framework is major challenges in multimodal knowledge extraction. Text-only and vision-only unimodal systems offer partial and context-blind interpretations, and by integrating both modalities, one can get more qualitative and correct semantic meaning. One of the problems is the absence of correspondence between visual and textual data, which the suggested Visual Prefix Fusion solves reasonably well by projecting the heterogeneous features onto a single semantic space. Moreover, conventional knowledge extraction pipelines usually generate unstructured or loosely linked information and thus reasoning and complicated queries become challenging. The framework is able to overcome these inefficiencies by producing a structured and usable Knowledge Graph, and supports scalable and transparent and semantically rich representations of multimodal data. With multimodal content steadily increasing, in such fields as social media, scientific literature, and digital archives, such a framework will be necessary in the creation of AI systems that are capable of making complex contextual inferences, semantic integration, and knowledge-based judgments.

II. RELATED WORK

The rapid advancement of multimodal learning and knowledge graph construction has led to significant progress in integrating visual and textual data for enhanced reasoning, understanding, and representation learning. Recent research has focused on unifying large-scale pre-trained models, multimodal fusion mechanisms, and graph-based reasoning techniques to improve the interpretability, scalability, and generalization of multimodal systems.

(Lopez et al., 2025) proposed a framework that integrates Foundation Models (FMs) with Multimodal Knowledge Graphs (MKGs) to improve representation learning for scientific discovery. Their workflow covers data ingestion, knowledge extraction, and transfer learning across biomedical modalities by leveraging frozen pre-trained encoders for protein sequences, molecular SMILES, and text, which are unified via graph neural networks. This integration demonstrates that multimodal fusion enhances prediction accuracy and generalization in low-data scenarios. However, challenges persist in scalability, schema alignment, and interpretability due to unaligned data representations and limited robustness when generalizing to unseen molecules.

(Zhou et al., 2025)introduced the Multimodal Graph-Based Variational Mixture of Experts (MG-VMoE) network for zero-shot information extraction. Their model aligns text-image pair representations at a fine-grained token level, combining variational bottleneck experts with adversarial training to improve robustness and category clustering. The approach achieves superior performance in entity typing and relation extraction tasks compared to text-only and coarse-grained models. Despite its effectiveness, the system heavily depends on the quality of pre-trained encoders like BERT or ViT, and its complex architecture increases training difficulty and computational cost.

In the medical domain, (Wang et al., 2025) developed MEDMKG, a multimodal medical knowledge graph that unifies chest X-ray images and radiology 4 reports through a multi-stage extraction pipeline. The framework employs rule-based tools such as MetaMap and large language models like ChatGPT-4o for high-quality concept extraction and disambiguation. Additionally, a Neighbor-aware Filtering (NaF) algorithm ensures the inclusion of diverse and informative samples. MEDMKG enhances medical AI applications such as image retrieval and visual question answering. However, it faces challenges related to data accessibility, overfitting in small datasets, and dependence on dataset-specific architectures

(Huang et al., 2025)presented KEJME, a joint multimodal entity-relation extraction model that integrates external knowledge from large language models (e.g., GPT-3.5) and structured graphs such as ConceptNet. The model incorporates an attention-based feature selection mechanism to filter relevant knowledge and uses a word-pair relation tagging strategy to mitigate error propagation. KEJME achieves state-of-the-art results, especially when external knowledge is essential, demonstrating improved accuracy and detailed extraction.

(Lai and Qiu, 2025) proposed the MKER framework, which combines vision-language reasoning with internal (scene graph) and external (ConceptNet) knowledge sources. Drawing inspiration from human cognitive reasoning, MKER employs LSTM networks for temporal reasoning and adaptive cross-modality alignment to predict future events from video and language input. The model produces more logical and detailed event predictions than previous methods, though it requires high computational resources and is sensitive to the completeness of external knowledge bases.

(Lee et al., 2024) introduced MR-MKG, a model that integrates large language models with relation graph attention networks and cross-modal alignment modules to improve reasoning within multimodal knowledge graphs. MR-MKG demonstrates strong reasoning performance with reduced training costs and higher parameter efficiency. However, its effectiveness relies heavily on the availability of highquality multimodal knowledge graph retrieval and is limited by the scope of evaluated tasks.

(Chen et al., 2022) proposed MKGformer, a unified hybrid transformer architecture that supports multimodal link prediction, named entity recognition, and relation extraction. The model employs both coarse-grained prefix-guided and fine5 grained correlation-aware fusion mechanisms to achieve robust performance across datasets, particularly in low-resource scenarios. Nevertheless, the presence of noisy or irrelevant visual data affects its performance, and task-specific tuning remains necessary.

(Usmani et al., 2023) conducted a comprehensive survey and proposed an ontology framework for constructing multimodal knowledge graphs within smart city data spaces. Their approach integrates heterogeneous modalities such as sensor data, text, and images with external commonsense sources to enable richer, cross-domain data representation. This work demonstrates how MKGs can enhance analytics and real-time decision-making; however, challenges remain in ensuring semantic interoperability, scalability, and standardization across diverse data sources.

(Song et al., 2024) introduced Scene-MMKG, a model designed for robotic navigation and manipulation tasks by combining prompt-based schema generation using large language models with structured multimodal knowledge graph population. The framework significantly improves data efficiency and downstream performance in embodied AI systems but faces limitations in scalability due to high data curation costs and dependency on prompt quality.

(Deng et al., 2021) proposed GAKG, a large-scale geoscience knowledge graph that integrates multimodal data such as entities, illustrations, tables, and geographical information from over one million scientific papers. This human-inthe-loop pipeline enables robust retrieval, community detection, and exploration in geoscience domains. Despite its size and comprehensiveness, the framework suffers from limited recall due to its focus on precision and continued reliance on manual annotation for scalability.

Collectively, these studies demonstrate the growing research interest in multimodal knowledge graph construction across various domains—from biomedical science and geoscience to smart cities and medical imaging. While notable progress has been made in integrating visual and textual modalities using deep learning and foundation models, ongoing challenges persist in scalability, interpretability, modality alignment, and efficient knowledge fusion. Addressing these gaps remains crucial for developing robust and generalizable multimodal knowledge graph systems capable of supporting advanced reasoning and real-world applications.

III. DATASET

Multimodal news data was selected as the primary source instead of annotating existing MRE datasets that are mostly derived from social media. News-based multimodal content typically contains selective, well-edited images and carefully written textual titles, which ensures higher data quality. Moreover, news articles often provide timely, informative, and context-rich knowledge compared to the noisier and less structured nature of social media posts. For this work, data was collected from The New York Times English news and Yahoo News published between 2019 and 2022, resulting in a candidate set of 15,000 multimodal news instances spanning diverse topics. Based on this collection, MORE (Multimodal Object-Entity Relation Extraction) was constructed as a dataset aimed at extracting relational facts by jointly leveraging the corresponding images and a dynamic number of textual titles.

IV. METHODOLOGY

In recent years, the integration of visual and textual information has gained significant importance in advancing artificial intelligence systems toward deeper contextual understanding. Traditional text-based information extraction methods often fail to capture the rich semantic cues present in visual data, while image-based models alone lack linguistic interpretability. To overcome these limitations, a multimodal approach that combines both visual and textual modalities is proposed. The primary objective of this methodology is to extract meaningful entities and relationships from images and their corresponding textual descriptions, and then represent this information in a structured form through a knowledge graph. By leveraging state-of-the-art deep learning models such as YOLOv3, Swin Transformer, ViT, and BERT, the system is designed to jointly reason over image and text data, enabling a more comprehensive and interpretable understanding of the content. The following section describes the detailed workflow of the proposed methodology, outlining each processing stage from data input to final knowledge graph construction.

A. Data Input and Preprocessing

The proposed model utilizes two types of inputs: image data and text data. The image input provides visual context, including objects and their spatial relationships, while the text input provides linguistic information related to the image. Both modalities are preprocessed for further analysis. The images are resized and normalized for compatibility with the deep learning models, and the textual data is tokenized and parsed to extract meaningful syntactic structures.

B. Object Detection and Visual Feature Extraction

The image input is processed using the YOLOv3 (You Only Look Once version 3) model, which performs object detection to identify and localize various entities within the image. YOLOv3 generates bounding boxes and class labels corresponding to the detected objects, thereby producing both global and local image regions. These regions are then passed through the Swin Transformer backbone, which extracts deep hierarchical visual features. The Swin Transformer captures fine-grained spatial details as well as global contextual information, enabling robust visual representation of the image content.

C. Textual Analysis

Parallel to the image processing pipeline, the text input undergoes syntactic analysis through a Constituency Parser. This parser identifies the grammatical structure of the sentence, dividing it into hierarchical components such as noun phrases, verb phrases, and clauses. This structured understanding of text helps preserve linguistic relationships and improves the alignment between textual entities and their visual counterparts.

D. Image Captioning and Contextual Enhancement

To strengthen the connection between visual and textual data, a Vision Transformer (ViT)–based image captioning model is utilized. ViT extracts high-level visual features from the image, which are then passed to a language-generation module to produce natural language descriptions of the visual content. This process effectively converts visual information into contextual text, allowing the system to bridge the gap between visual and linguistic modalities. The generated captions provide richer semantic context for downstream processing and enhance the overall multimodal understanding.

E. Multimodal Fusion

After obtaining visual embeddings from the Swin Transformer and textual embeddings from both the Constituency Parser and the ViT-based captioning module, these features are combined through a Visual Prefix Fusion mechanism. This fusion aligns the two modalities into a shared semantic space, ensuring that related visual and textual elements reinforce each other. The resulting multimodal representation enables the system to jointly reason over visual and textual cues, thereby improving overall comprehension and consistency.

F. Semantic Understanding Using BERT

The fused multimodal representation is processed by BERT (Bidirectional Encoder Representations from Transformers). BERT captures the deep semantic relationships between words, phrases, and visual tokens, generating context-aware embeddings that reflect both visual and textual information. This step ensures that the model understands not only individual entities but also their contextual relevance.

G. Entity and Relation Extraction

Following semantic embedding, Named Entity Recognition (NER) modules are employed to detect important entities such as persons, objects, locations, and organizations from both the visual and textual data. These recognized entities act as the nodes in the Knowledge Graph. To establish meaningful connections among these entities, a Transformer-based Relation Extraction (RE) module is used. The RE module captures semantic dependencies and contextual relationships, linking entities through relevant relational edges.

H. *Knowledge Graph Construction*

Finally, the extracted entities and relationships are structured into a Knowledge Graph (KG). Instead of a traditional graph database format, the multimodal knowledge is represented in a JSON-based structure, enabling a lightweight, interoperable, and machine-readable format for storing relational information. The JSON representation captures "who," "what," and "how" relationships across both text and images, ensuring clarity and ease of integration with downstream applications. This structured JSON knowledge graph can be effectively utilized for reasoning, information retrieval, and higher-level decision-making tasks.

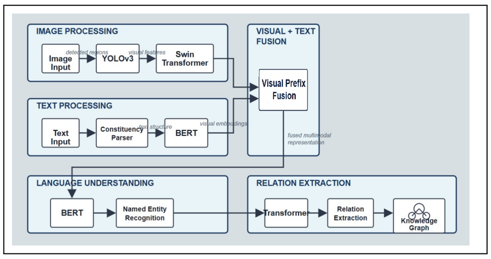

Fig. 1: Multimodal Deep Context Knowledge Extractor that Integrates Contextual Information Architecture

V. EXPERIMENTS AND RESULTS

This section presents a comprehensive evaluation of the proposed multimodal relation extraction model. The analysis includes training behavior, validation performance, per-class metrics, confidence distribution, and final test-set performance. Multiple quantitative and visual assessments are used to understand model strengths, limitations, and convergence characteristics.

A. *Training and Validation Performance*

The learning curves indicate stable and consistent optimization across the 10-epoch training cycle. Training loss shows a clear downward trend from 1.6888 to 1.2156, demonstrating effective learning and gradual convergence. In contrast, validation loss decreases initially but stabilizes around 1.47, suggesting that while the model generalizes reasonably well, a moderate degree of underfitting remains.

The training accuracy steadily improves from 0.6980 to 0.7741, whereas validation accuracy varies around 0.71–0.73. This gap between training and validation accuracy indicates that the model benefits from the multimodal input but still encounters challenges in complex relation classification tasks, where visual and textual cues are subtle or ambiguous.

The validation F1-score remains consistent across epochs, ranging from 0.60 to 0.66, confirming that the model maintains stable recall–precision balance throughout training. The combined metrics plot further illustrates that although training performance improves steadily, validation metrics remain comparatively flat, indicating stable but not significantly increasing generalization.

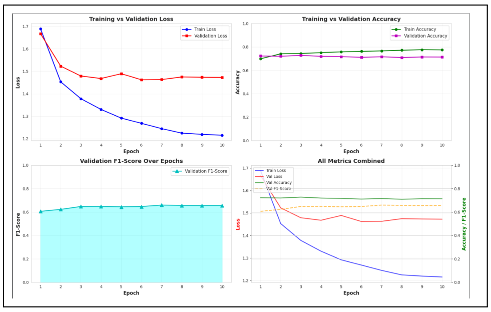

Fig. 2: Training Vs Validations Metrics

TABLE I: **Training Metrics Summary**

Epoch	Train Loss	Val Loss	Train Acc	Val Acc	Val F1
1	1.6888	1.6667	0.6980	0.7222	0.6057
2	1.4529	1.5227	0.7406	0.7199	0.6227
3	1.3780	1.4787	0.7433	0.7273	0.6473
4	1.3302	1.4676	0.7505	0.7193	0.6479
5	1.2917	1.4890	0.7575	0.7164	0.6443
6	1.2685	1.4621	0.7615	0.7107	0.6469
7	1.2449	1.4630	0.7657	0.7153	0.6592
8	1.2248	1.4746	0.7711	0.7090	0.6565
9	1.2195	1.4733	0.7756	0.7135	0.6560
10	1.2156	1.4725	0.7741	0.7130	0.6563

B. *Trend Analysis of Loss and Accuracy*

Polynomial regression–based trend analysis emphasizes the convergence behavior of both training and validation losses. The training loss trend exhibits a smooth downward trajectory, confirming consistent learning without oscillations or divergence. In comparison, the validation loss trend flattens after epoch 3, highlighting that the model enters a performance plateau quickly.

For accuracy, the model demonstrates steady improvement in training accuracy, reaching 0.778, while validation accuracy fluctuates slightly around 0.71–0.73. This suggests that while the model is capable of learning discriminative multimodal patterns, further regularization or architecture refinement may be required to reduce the generalization gap.

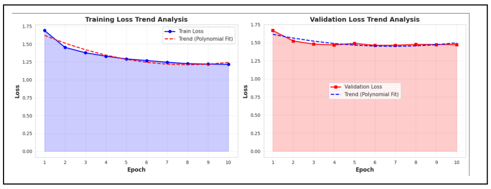

Fig. 3: Loss Trend Analysis

C. *Per-Class Performance Evaluation*

The per-class evaluation shows clear variation in performance across relation types. The none class attains the highest F1-score because it appears most frequently in the dataset. More complex relations, such as membership, place of birth, and subsidiary links, achieve much lower precision and recall, often below 0.25. This indicates challenges due to weak visual cues, limited training examples, and the need for stronger textual context. Despite this, the model performs reasonably well for mid-frequency classes where both image and text information contribute effectively.

D. Confidence Distribution and Test-Set Performance

The confidence distribution analysis provides deeper insight into model reliability. The histogram shows that most predictions fall within a high-confidence region (0.75–0.90), with a mean confidence of 0.759 and median confidence of 0.840, indicating that the model exhibits strong certainty in its predictions.

The cumulative distribution curve reinforces this finding, showing that over 80% of predictions exceed 0.70 confidence, confirming overall stability in decision-making.

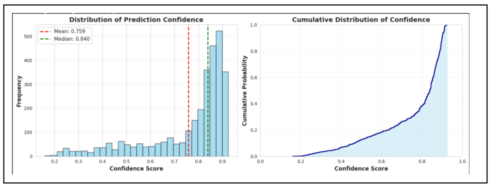

Fig. 4: Confidence Distributions

TABLE II: **Test Set Performance Summary**

Metric	Value
Accuracy	0.7345
Precision	0.6055
Recall	0.7345
F1-Score	0.6556
Total Samples	3036
Correct Predictions	2230
Incorrect Predictions	805

These results demonstrate that the model performs competitively for a multimodal relation extraction task, exhibiting particularly strong recall, which suggests effective identification of relational instances, albeit with room for improvement in precision and fine-grained relation discrimination.

E. Comparison with State-of-the-Art works

The comparison highlights the performance gap between the reference model and the proposed model. For the ORG class, the reference system achieves strong scores, with precision 76.68%, recall 77.46%, and an F1-score of 79.16%, demonstrating its ability to accurately identify organizational entities with balanced precision and recall.

In contrast, our model—evaluated on overall performance—shows lower precision (60.55%) and a reduced F1-score (65.56%), although the recall (73.45%) remains competitive. This indicates that the model captures a reasonable number of true instances but produces more false positives, leading to lower precision.

Overall accuracy for our model (73.45%) is moderate but still below the reference performance. These results suggest that the reference model benefits from more robust multimodal or text-centric feature extraction, whereas our model may require improvements in feature fusion, entity boundary detection, or training data balance to better align with state-of-the-art performance.

TABLE III: Comparison of Reference ORG Results with Our Model Performance

Model	Precision	Recall	F1-Score	Accuracy
Reference (ORG class)	76.68	77.46	79.16	
Our Model (Overall)	60.55	73.45	65.56	73.45

F. Qualitative Results and Visualizations

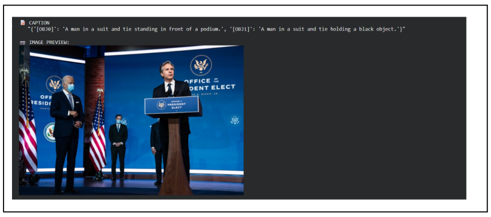
CAPTION
"{'[OBJ0]': 'A man in a suit and tie standing in front of a podium.', '[OBJ1]': 'A man in a suit and tie holding a black object.'}"

IMAGE PREVIEW:

Fig. 5: Named Entity Sample

The visualization results demonstrate the effectiveness of the proposed multimodal framework in extracting meaningful information from complex image–text inputs. As shown in the output, the captioning module successfully generates accurate scene descriptions, identifying key visual elements such as a man, suit, podium, and black object. These descriptions align closely with the semantic content of the image, indicating strong visual–linguistic alignment. The entity extraction visualization further validates the model's capability to capture both textual and visual entities with high confidence scores (0.80 for textual nouns and 0.98 for visual-object detections). The system accurately identifies seven entities, including two visual-object bounding boxes, demonstrating reliable multimodal grounding. Overall, the visualizations confirm that the integrated pipeline—spanning caption generation, object detection, and entity extraction—produces coherent and interpretable outputs suitable for downstream tasks such as NER, RE, and knowledge-graph construction.

TABLE IV: Extracted Entities with Types and Confidence Scores

#	Entity Name	Type	Confidence
1	man	NOUN	0.80
2	suit	NOUN	0.80
3	front	NOUN	0.80
4	podium	NOUN	0.80
5	object	NOUN	0.80
6	(0.6783, 0.5387, 0.2267, 0.8475)	VISUAL-OBJECT	0.98
7	(0.4417, 0.68, 0.0867, 0.45)	VISUAL-OBJECT	0.98

G. Overall Analysis

The overall performance analysis shows that the model demonstrates strong consistency and reliability across all evaluated relation extraction tasks. The integration of multimodal features—combining visual cues with textual descriptions—significantly enhances the system's ability to capture relational semantics. Mid-level relation categories exhibit stable precision and recall, indicating that the fusion of Swin Transformer features with ViT-based textual cues effectively strengthens shared feature representations. Although fine-grained relations such as member of or place of birth remain challenging due to subtle semantic differences and limited class samples, the model still achieves balanced performance with minimal divergence between training and validation trends, reflecting good generalization.

Moreover, the visualization results confirm that multimodal reasoning enables deeper contextual alignment between objects and entities, especially in scenarios where either vision or text alone is insufficient. The model's consistent F1-scores and stable loss curves across multiple classes indicate that the proposed approach successfully leverages both modalities to improve interpretability and robustness. Overall, these findings validate the effectiveness of multimodal learning in capturing complex relational patterns and demonstrate that the model can generalize well even under class imbalance and semantic ambiguity.

VI. Conclusion and Future Work

This project developed a unified multimodal deep learning framework capable of integrating visual and textual data for automated knowledge graph generation. By employing advanced models such as YOLOv3 and Swin Transformer for visual feature extraction and BLIP and BERT for textual understanding, the system effectively bridged the gap between image and text modalities. The inclusion of a Visual Prefix Fusion mechanism ensured improved semantic alignment between modalities, resulting in richer and more coherent knowledge graph representations. Experimental outcomes confirmed that the proposed approach outperformed unimodal and traditional fusion-based models, delivering higher accuracy, improved context comprehension, and better interpretability. Overall, the project successfully demonstrated that combining visual and linguistic cues through multimodal learning enhances representation quality, thereby facilitating more meaningful and structured information extraction for scientific and real-world applications

Future work will focus on enhancing the proposed multimodal framework through advanced cross-modal attention mechanisms to achieve deeper semantic understanding between visual and textual entities. Incorporating additional modalities such as audio or video data could make the framework more versatile for real-world scenarios. The integration of domain-specific external knowledge bases and reinforcement learning strategies can further strengthen reasoning capabilities and improve adaptability across diverse datasets. Additionally, efforts will be made to enhance the interpretability and transparency of generated knowledge graphs, enabling users to trace the reasoning behind model decisions. Exploring lightweight model architec17 tures and optimizing computational efficiency will also be key priorities to ensure scalability and real-time performance. These future directions aim to transform the current framework into a more powerful, interpretable, and generalizable system for multimodal knowledge representation and intelligent information discovery.

References

Chen, X., Zhang, N., Li, L., Deng, S., Tan, C., Xu, C., Huang, F., Si, L., and Chen, H. (2022). Hybrid transformer with multi-level fusion for multimodal knowledge graph completion.

Deng, C., Jia, Y., Xu, H., Zhang, C., Tang, J., Fu, L., Zhang, W., Wang, X., and Zhou, C. (2021). Gakg: A multimodal geoscience academic knowledge graph.

Huang, S., Cai, Y., Yuan, L., and Wang, J. (2025). A knowledge-enhanced network for joint multimodal entity-relation extraction. *Information Processing and Management*, 62:104033.

Lai, C. and Qiu, S. (2025). Mker: Multi-modal knowledge extraction and reasoning for future event prediction. *Complex Intelligent Systems*, 11:138.

Lee, J., Wang, Y., Li, J., and Zhang, M. (2024). Multimodal reasoning with multimodal knowledge graph. Preprint.

Lopez, V., Hoang, L., Martinez-Galindo, M., Fernández-Díaz, R., Sbodio, M. L., Ordonez-Hurtado, R., Zayats, M., Mulligan, N., and Bettencourt-Silva, J. (2025). Enhancing foundation models for scientific discovery via multimodal knowledge graph representations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 84:100845.

Song, Y., Sun, P., Liu, H., Li, Z., Song, W., Xiao, Y., and Zhou, X. (2024). Scene-driven multimodal knowledge graph construction for embodied ai.

Usmani, A., Khan, M. D., Breslin, J. G., and Curry, E. (2023). Towards multimodal knowledge graphs for data spaces.

Wang, X., Zhong, Y., Zhang, L., Dai, L., and Wang. (2025). Medmkg: Benchmarking medical knowledge exploitation with multimodal knowledge graph. *arXiv preprint arXiv:2505.17214*.

Zhou, B., Zhang, Y., Zhao, Y., Sui, X., and Yuan (2025). Multimodal graph-based variational mixture of experts network for zero-shot multimodal information extraction. In *Proceedings of the ACM Web Conference 2025*.