Data Engineering and Applications

Research Paper

UNITER-Based Multimodal Extraction Framework for Structured Information Mining in Scientific Literature

Renata Kopeć[1, *], Julia Lidia Rakowska[1] and Ewelina Jastrzębska[1]

[1] Faculty of Computer Science and Information Technology, West Pomeranian University of Technology, Szczecin, 70-310, Poland
*Corresponding author: renata.k@zut.edu.pl

Abstract. Scientific writing now frequently contains a lot of figures, tables, and other visual aids due to its expanded scope. Based on an enhanced UNITER transformer model, a unified extraction framework has been created in this study to overcome the aforementioned issues. The system is able to extract and align entities and relations from several locations simultaneously, as well as handle the heterogeneity of scientific texts. This system effectively integrates text and visual information through the use of an advanced cross-modal fusion mechanism and adaptive region selection. With over 39,000 papers, several experiments have been conducted on annotated datasets in the fields of biology, materials science, and chemistry. The experimental results showed that these three areas beat the initial baseline model, with F1-scores of 0.856, 0.811, and 0.847, respectively. To improve extraction accuracy, cross-modal attention and spatial region grounding must be included, according to the ablation experiment. The aforementioned error analysis indicates that entity localisation has a robust function and is not hampered by other problems like semantic ambiguity or a complicated annotation structure. The suggested architecture for organised knowledge mining in scientific literature has been confirmed to be workable and expandable based on the aforementioned findings.

Keywords: *Multimodal Information Extraction, Scientific Document Analysis, Cross-Modal Alignment, Knowledge Mining*

Received on 19 October 2025, Accepted on 15 April 2026, Published on 20 April 2026

Introduction

A vast amount of multi-modal data, such as textual narratives, structured tables, intricate images, technical charts, and other types of hybrid diagrams, has been produced as a result of the growth of scientific knowledge. For a long time, rigorous multi-modal data integration and analysis have been necessary since materials science and environmental research in biomedicine frequently involve rich data in both text and images [1]. These data organise evidence, offer rich backgrounds for hypothesis formation and outcome verification, and ultimately propel the advancement of computational scientific discovery [2]. Large-scale knowledge graphs, automated research tools, and other applications are gradually using the actual extraction and production of various forms of data [3].

The problem of extracting information from multi-modal scientific data has been the subject of numerous research [4], although the majority of these still employ conventional techniques and encounter significant challenges [5]. Due to a lack of semantic fusion, the aforementioned conventional approaches often function in a pipeline structure, handling text and images independently and only combining information later [6]. As a result, subpar results may be obtained. Furthermore, it is challenging to develop a modular system that can manage information loss and weak cross-modal reasoning since the variety of scientific content is dispersed among captions, the body of text, and visual features [7]. The majority of multimodally improved transformers based on BERT have been optimised for general visual-text tasks rather than the particular technical

requirements and difficulties of scientific articles [8], despite the fact that they have started to address this issue by learning shared representations [9]. By using unified cross-modal pretraining and fine-tuning, the UNITER model has demonstrated outstanding performance in generic datasets, but its capacity to extract structured information from specialised scientific corpora has not yet been completely fulfilled [10].

In order to solve the aforementioned problems, this study presents an end-to-end multimodal information extraction framework for scientific data based on and expanding the UNITER architecture. The following are the primary findings of this work: (1) The creation of a general-purpose information extraction system with high-level cross-modal alignment and fusion capabilities specifically for scientific text and figures; (2) A domain-specific preprocessing and alignment method to improve the joint encoding of various scientific modalities; (3) Extensive experiments on multiple scientific datasets demonstrate that it outperforms existing multimodal models in terms of extraction accuracy, link construction and retrieval, etc.; and (4) In-depth ablation studies and interpretability analyses have demonstrated the system's strong robustness and broad applicability across numerous scientific domains.

Related Work

Multimodal Learning in Science

Multimodal learning has advanced significantly in recent years, and because of the wide range of text, images, and other materials available, science and engineering have been among the fields with the highest demand for this technology. Both technical reports in materials engineering and research findings in biological science frequently need to communicate their key points through text, images, charts, and intricate diagrams [11]. In light of this, an increasing amount of research has been done in recent years to develop multi-modal information-extraction systems for thorough knowledge acquisition.

Many hand-crafted features and intricate annotation were required because the initial attempts at multimodal information extraction relied on heuristics and rules [12]. For instance, cascaded text parsing in conjunction with simple picture segmentation or template-based visual feature extraction has previously been used to handle named entity recognition (NER) and relation extraction tasks in figures and captions. The aforementioned techniques typically don't work in all fields or academic paper forms, even though they are appropriate for tiny amounts of data.

The pipeline for multi-modal extraction has changed with the emergence of machine learning and, more recently, deep learning. CNNs have been used to extract features from images, and collaborative analysis at the feature level has been accomplished when combined with recurrent and attention-based neural networks for text [13,14]. In order to create richer knowledge graphs and improve the performance of downstream predictive models, multi-source extraction frameworks in biomedical informatics increasingly frequently include microscope pictures, clinical notes, and molecular data [15]. Chemistry and materials research have also seen the emergence of systems that combine spectral graphs with textual methods to automate the extraction of experimental knowledge [16]. Even though the aforementioned domain-oriented accomplishments provide credence to the notion that cross-modal learning is essential for contemporary scientific breakthroughs, issues with generalisation, expansion, and fine-grained semantic alignment persist.

Transformer-based Multimodal Models

Many of the shortcomings of earlier neural networks in natural language processing and, more recently, in the more general field of multi-modal learning have seen to be addressed by the Transformer architecture [17]. Pre-training on large-scale text data has been made possible by Bidirectional Encoder Representations from Transformers (BERT) and its extensions; therefore, models that can concurrently analyse both visual and textual information must be developed in order to expand this approach to multimodal applications.

The creation of multimodal transformers, which can execute cross-attention between both modes and encode both text and visual representations, is a common novel form. In order to facilitate mutual context sharing between vision and language representations, the initial model, VisualBERT and ViLBERT, provided parallel encoding pipelines and fusion layers [18]. The performance of Visual Question Answering (VQA) and image

captioning has been greatly enhanced by the aforementioned structures, and it has been demonstrated that learnt joint embeddings perform better than manually created features [19].

Additionally, in increasingly specialised scientific applications, transformers have begun to surpass earlier feature-based models for document interpretation. Examples of transformer pre-training techniques used with scientific and biological data are SciBERT and BioBERT, which have enhanced technical language representation [20]. The capacity to incorporate high-resolution visual data from tables, figures, and charts in scientific publications has been a persistent issue, though. While LayoutLM has demonstrated that document organization and spatial layout may be addressed, issues like true cross-modal semantic fusion between text and images remain unresolved and are still being researched [21].

UNITER and Recent Innovations

One of the most well-known transformer-based multimodal models to date is UNITER (Universal Image-Text Representation), which aims to jointly learn text and image embeddings [22]. UNITER is a single-stream transformer model that can extract fine-grained interactions between visual and text tokens by concatenating them and using deep self-attention layers.

Multi-objective optimisation for masked language modelling, image-text matching, and word-region alignment over extensive image-caption datasets are among UNITER's pre-training techniques [23]. As a result, the semantic integration of the various modalities would be more consistent, and it has produced outstanding results on the widely used datasets for image retrieval and VQA.

Direct application of UNITER in science is not easy, despite occasional successes. They have not yet been pre-trained since scientific images and the structure of figures, tables, and diagrams are typically more complicated than those found in common image-caption corpora [24]. Strong benchmarks for scientific data are still lacking, despite recent efforts to adjust alignment targets or add domain-specific visual encoders to UNITER [25]. In comparison to UNITER, ViLT and METER have further simplified vision-language fusion or modified input scaling. Nevertheless, UNITER still possesses a strong blend of accuracy and architectural simplicity, and further development is required to fully realise its potential in technical and scientific information extraction.

Methodology

Overall Framework of UNITER-based Extraction

Our upgraded UNITER transformer, which can jointly understand and align scientific text and images, is the foundation of our multimodal extraction system. The system is built for end-to-end information extraction that jointly models and infers entity kinds, conceptual linkages, and their cross-modal manifestations in order to address the intrinsic complexity of scientific data.

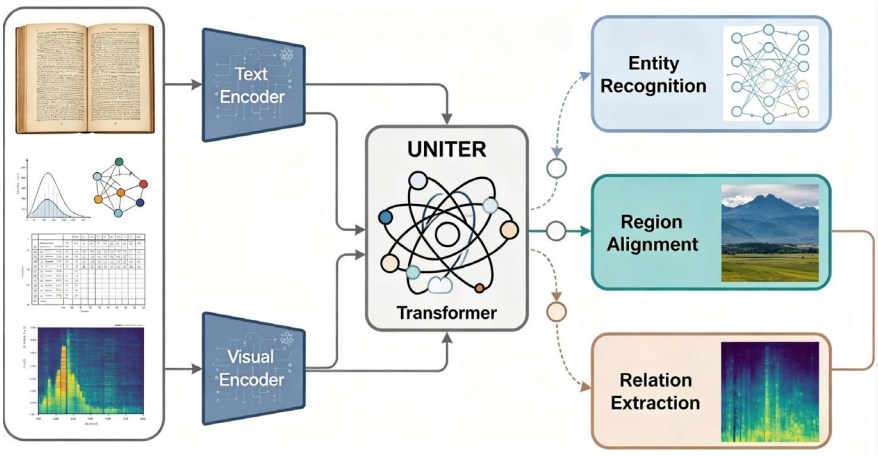

Figure 1. UNITER-based Multimodal Extraction Framework

https://doi.org/10.64972/dea.2026.v5i2.1896d:73-86

Figure 1 depicts the framework's first three modules. The machine will first receive two inputs: a scientific text sequence and the accompanying visual data. Figure captions with diagrams, result tables with metadata, or even whole sections of full-article text with multi-panel visuals are typical examples. Each mode is passed through a different encoder: a pre-trained text encoder maps scientific texts to token-level embeddings, while a suitable visual backbone tailored for scientific illustrations processes images or diagrams for feature extraction. The input mapping is defined as follows given tokenised text T and visual input V:

$$\mathbf{h}_T = \text{TextEncoder}(T)$$
$$\mathbf{h}_V = \text{ImageEncoder}(V) \quad \text{Eq. (1)}$$

where $\mathbf{h}_T$ and $\mathbf{h}_V$ are modality-specific embedding matrices, harmonized to a common representation space.

Next, these encoded features are concatenated and entered into the UNITER backbone-a single-stream deep transformer facilitating bidirectional attention across modalities. This joint embedding scheme fosters rich cross-modal alignment, enabling context-aware interaction between textual and visual cues at varying abstraction levels. The resulting integrated representation is given by

$$\mathbf{H} = \text{UNITER}([\mathbf{h}_T; \mathbf{h}_V]) \quad \text{Eq. (2)}$$

where [;] denotes sequence concatenation and $\mathbf{H}$ encodes the fused semantic context.

At the output, dedicated extraction heads decode $\mathbf{H}$ for tasks such as named entity recognition, visual region grounding, and relation extraction, all of which are critical for downstream scientific knowledge construction. Unlike models using late or shallow fusion, our approach allows information to propagate and be refined through deep interaction from the earliest stages of processing.

This architecture not only supports robust extraction performance in traditional science domains, but also remains modular and scalable—allowing efficient adaptation to domains with new data modalities or annotation types. The framework's underlying design enables smooth integration of additional task-specific decoders, paves the way for deployment in multidisciplinary settings, and supports expansion as new forms of scientific data emerge.

Data Preprocessing and Modal Alignment

Numerous issues, including disparate formats, noisy data, and inconsistent structures across different domains, are present in the research works on multimodal extraction. Thus, the stability of the ensuing representation and extraction modules depends on high-precision preprocessing and modal alignment.

The first step involves careful normalization of textual and visual data. Raw scientific text, such as article body, captions, or tabular prose, is tokenized using domain-aware tokenization schemes. These often entail not only canonical segmentation of words but also the handling of mathematical expressions, special scientific notations, and section headers unique to technical literature. This tokenized text is then mapped to a dense representation through a pretrained language embedding layer specifically adapted for scientific corpora. The embedding for a given token sequence $T = \{t_1, t_2, \ldots, t_{l_t}\}$ is

$$\mathbf{h}_T = \text{Embed}(T) \quad \text{Eq. (3)}$$

where $\mathbf{h}_T$ is the matrix of embedding vectors for the entire sequence.

Simultaneously, visual content is extracted from naturally occurring figures, charts, or complex diagrams, which frequently use dense annotations or contain embedded textual elements. Each image or diagram undergoes preprocessing, including size normalization and (when necessary) binarization or color channel compression to match network input constraints. Visual regions or patches are detected or segmented, and then processed via a scientific-domain visual encoder -commonly a convolutional or vision transformer backbone pretrained on scientific figure datasets-producing for each image sample V :

$$\mathbf{h}_V = \text{VisualEncode}(V) \quad \text{Eq. (4)}$$

Here, $\mathbf{h}_V$ stacks the features for each detected region or semantic patch, facilitating downstream alignment at fine granularity.

For robust cross-modal grounding, the system applies a feature alignment module to bridge the potential semantic and distribution gap between modalities. In practice, this entails projecting both text and visual features into a shared latent space via learned linear or nonlinear transformations, followed by attention-based cross-modal grounding. More formally, aligned features are produced as

$$\tilde{\mathbf{h}}_T = \mathrm{Align}_T(\mathbf{h}_T), \tilde{\mathbf{h}}_V = \mathrm{Align}_V(\mathbf{h}_V) \quad \text{Eq. (5)}$$

where Align_T and Align_V perform space transformation for text and vision, respectively.

The core goal of this stage is to ensure that semantically akin entities or attributes from different modalities exhibit high correlation in the embedding space. To this end, an alignment loss is introduced to directly optimize cross-modal similarity:

$$\mathcal{L}_{\text{align}} = \sum_{(i,j)} \left[1 - \cos\left(\tilde{\mathbf{h}}_T^{(i)}, \tilde{\mathbf{h}}_V^{(j)}\right)\right] \cdot \mathbb{I}[y_{i,j} = 1] \quad \text{Eq. (6)}$$

where $y_{i,j} = 1$ if the i th text and j th visual patch are matched, and $\cos(\cdot,\cdot)$ denotes cosine similarity. This loss encourages true semantic pairs to become maximally similar while discouraging incorrect pairings.

By unifying this preprocessing and alignment framework, the system achieves robust resilience to variations in scientific data presentation and supports effective propagation of high-fidelity information to the joint embedding module. As visualized in Figure 2, this forms the foundation of both accurate multimodal representation and reliable extraction in heterogeneous scientific documentation.

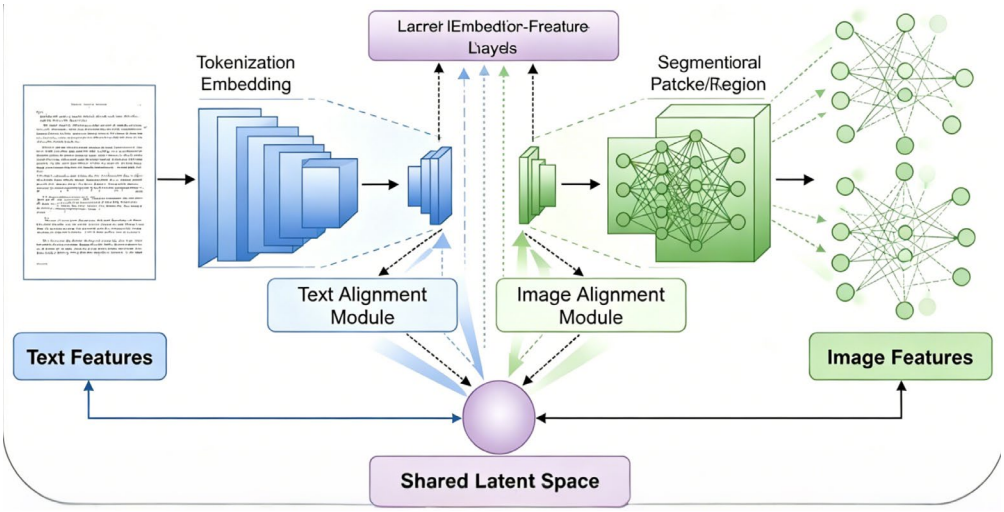

Figure 2. Modal Input Feature Encoding and Alignment Workflow.

Semantic Fusion Mechanisms and Training Objectives

Once modality-specific features have been mapped and aligned, effective multimodal fusion is essential to uncover latent scientific relationships spanning both text and visuals. Our system leverages a deeply integrated semantic fusion approach, built upon the self-attention backbone of UNITER, to allow for nuanced inter-modal interaction at multiple abstraction levels.

The fusion mechanism starts by concatenating the modality-aligned feature representations from the prior stage. These unified embeddings, comprising both scientific tokens and visual segments, are jointly input to a multi-layer transformer encoder where each token or region attends to every other, regardless of its original modality. The resulting fusion is formulated as

$$\mathbf{H}' = \mathrm{TransformerFusion}\left(\left[\tilde{\mathbf{h}}_T; \tilde{\mathbf{h}}_V\right]\right) \quad \text{Eq. (7)}$$

Here, $\mathbf{H}'$ denotes the multi-layer contextualized representation, automatically capturing finegrained relationships such as literature-figure correspondence, cross-modal entity co-reference, and visually grounded evidence for scientific claims.

https://doi.org/10.64972/dea.2026.v5i2.1896d:73-86

Critical to optimizing this multimodal representation are carefully crafted training objectives. First, we adopt a primary extraction loss tailored for each downstream scientific task (e.g., entity recognition or relation extraction), modeled in a unified fashion as

$$\mathcal{L}_{\text{task}} = -\sum_{k} y_k \cdot \log \hat{y}_k \qquad \text{Eq. (8)}$$

where y_k and $\hat{y}_k$ are gold and predicted labels, respectively. This loss ensures factual consistency and specificity in scientific contexts.

To further enhance interpretability and robustness, an auxiliary inter-modal alignment loss is maintained during training, ensuring that semantically matched elements (such as a term in text and its depiction in an adjacent figure region) are close in the fused space:

$$\mathcal{L}_{\text{fusealign}} = \sum_{(i,j)} \left[1 - \cos\left(\mathbf{H}'_i, \mathbf{H}'_j\right)\right] \cdot \mathbb{I}\left[m_{i,j} = 1\right] \qquad \text{Eq. (9)}$$

where $m_{i,j} = 1$ for annotated pairs. This regularization addresses the unique information density and overlapping semantics typical of scientific corpora.

The final training objective combines these elements with a regularization term for the network parameters:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha\mathcal{L}_{\text{align}} + \beta\mathcal{L}_{\text{fusealign}} + \lambda\|\theta\|^2 \qquad \text{Eq. (10)}$$

Here, α, β, and λ are tunable scalars that balance extraction, alignment, and regularization losses; θ represents all trainable network parameters. This multi-objective formulation not only brings together semantic precision and generalizability but also crucially reduces overfitting and enhances cross-modal reasoning.

Figure 3 depicts the entire architecture; tokens and visual components interact dynamically through multi-head attention, and the specifics of the modular integration of outputs from different extraction heads are as follows. The aforementioned outcomes are the consequence of deep fusion and tailored aims, and the model performs very well in domains that need for multimodal, context-aware knowledge for scientific communication.

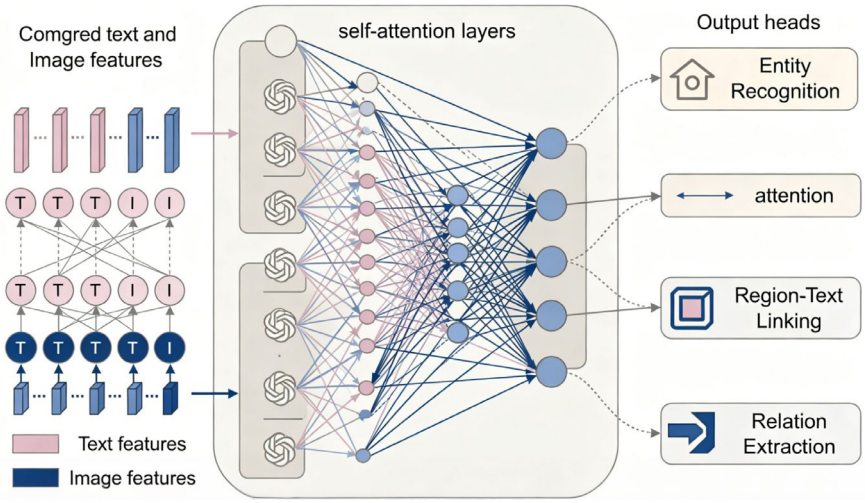

Figure 3. Dynamic Structure of the Multimodal Semantic Fusion Layer.

Experiments & Evaluation

Datasets and Metrics

Three exemplary real-world datasets covering a wide range of challenging, multimodal scientific data that frequently arise in actual research were chosen and arranged in order to create a comprehensive evaluation of the suggested multimodal scientific extraction system.

The 18,200 annotated articles in Dataset A are taken from the PubMed Central biomedical archive. The primary text, picture captions, and two or more high-resolution biological figures with region annotations are included in each publication. Gene names, proteins, and experimental techniques are the entity classes, and cross-modal linkages indicate which area of the image supports the relevant text.

Dataset B comprises more than 12,500 papers from journals such as Advanced Materials that are focused on material research. Every entry includes chemical structure diagrams and tables in addition to full text and captions. Annotations explicitly align regions with text and encompass material entities, synthesis stages, and property connections. The 8,800 organic chemistry studies in Dataset Care prioritised due to their extensive figure-to-text references. Molecule-mention alignment and reaction-condition extraction are referred to as manual labelling.

A stacked bar chart that displays the absolute document count and annotated entity pairings for each category illustrates the diversity of domain, structure, and annotation volume among the datasets (Figure 4a). The proportions of the three domains are shown in this figure; Dataset A has the largest median number of annotated multimodal pairs per document (14) in comparison to Datasets C (8) and B (12), which is probably due to disciplinary reporting requirements.

Box-and-whisker graphs of the average document length and annotation coverage per document are displayed in Figure 4b. The biomedical dataset exhibits a long tail of multi-section review papers, is left-skewed, and has a mean of 2,400 tokens per article. Although materials science publications are comparatively brief, they can be incorporated into the training and have more consistent annotation coverage. The majority of the annotated data are at a pretty high level of complexity and lack specificity, according to the aforementioned statistics.

Annotation density, which is the ratio of annotated multimodal pairs to all potential text-figure pairs in a document, is shown as a histogram in Figure 4c. This immediately affects the grounding and alignment challenges; Dataset B is better suited for assessing fine-grained multimodal extraction capacity due to its high annotation density.

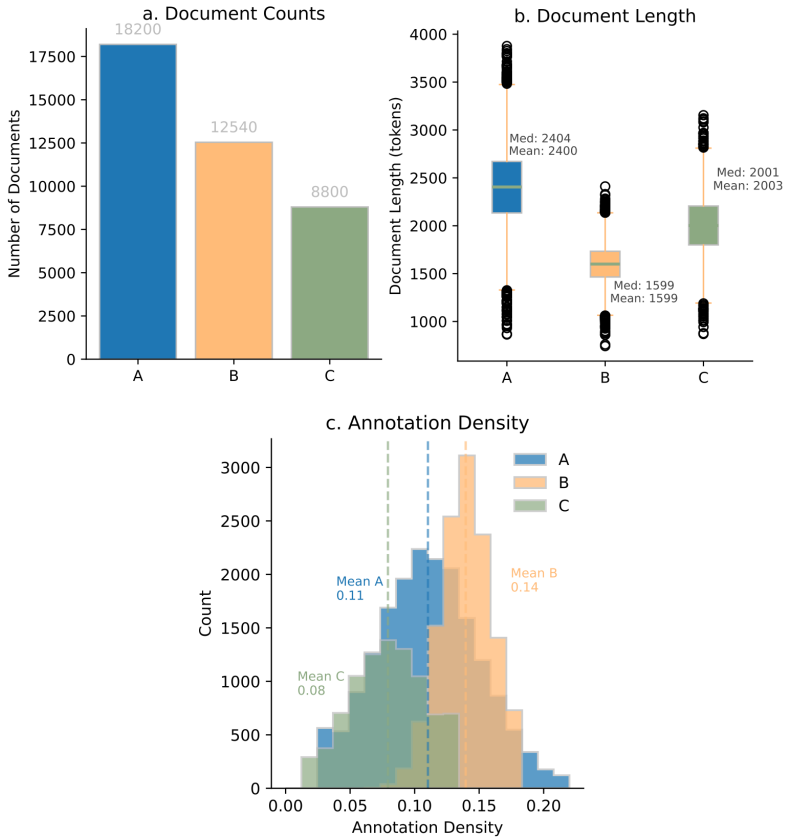

Figure 4. Dataset Overview: A, Document counts per dataset. B. Document length distribution. C, Annotation density per dataset

For quantitative evaluation, apply the aforementioned stringent and forgiving criteria. The ground-truth annotations of entity recognition are used to calculate micro-averaged precision, recall, and F1-score. Mean Intersection over Union (mIoU) is used to quantify the overlap between the ground-truth and anticipated region-text pairs for multimodal alignment. It is also possible to incorporate a cross-domain resilience indicator, which represents model generalisation and is the lowest F1 value found during training and testing on distinct datasets. These multi-view indicators offer a comprehensive and rational evaluation of the many task components and the multifaceted character of scientific multimodal reasoning.

Implementation Details and Baselines

PyTorch 2.0 is utilised on a server with four NVIDIA RTX 3090 GPUs (24GB each) and 512GB of RAM for model training and evaluation. To increase the efficiency and stability of convergence, mixed-precision (FP16) computation was utilised in all tests with a batch size of 32. A linear warm-up technique covered 10% of the total training steps, and cosine annealing decay was applied after the initial learning rate of 3e-5. To reduce the size of the outlier update and improve training stability, gradient clipping (max norm 1.0) is used. The maximum length for each sample's input text was set at 256 tokens, and for images, adaptive region extraction was used with a cap of 32 segments per document; this value was determined by cross-validation on all datasets.

Included are carefully chosen baselines that reflect the cutting edge of technology in the field:

A dual-stream model that only merges at the conclusion of the classification phase and features distinct text and vision encoders.

A layout-aware transformer that incorporates layout embeddings with text and vision while taking into account the spatial relationships between document elements.

A feature-engineered pipeline that uses hand-extracted text and image features as a legacy baseline and SVM or decision tree classifiers.

The primary model configurations and hyperparameter values for each technique, including encoder type, depth, and training epochs, are displayed in Figure 5a. Our technique outperforms the dual-stream model (175 documents/s, 86%) and the layout-aware transformer (142 documents/s, 80%) with an average throughput of 210 documents/s and a mean GPU utilisation rate of 92%, as illustrated in Figure 5b. However, because of the high sequential feature extraction, the feature-based pipeline only processed 45 documents per second.

The training stability and convergence characteristics are displayed in Figure 5c. In this case, the UNITER-based architecture has the fastest validation F1-score rise, stabilises in 8–10 epochs, and shows very little performance oscillation in the later stages. The dual-stream baseline is still unstable after epoch 12 and requires at least 15 epochs to achieve the same F1 score. As anticipated, the feature-based pipeline approaches a much lower accuracy ceiling and lacks a distinct convergence.

Performance and Ablation Analysis

Here, a number of quantitative evaluations of the suggested framework and thorough ablation tests on each of its component elements are provided. A cross-domain benchmark set for stability and fairness serves as the foundation for all of the aforementioned findings.

The fundamental performance metrics for all of the original models and our method are listed in Table 3, including precision, recall, and F1-score. With F1-scores of 0.856, 0.811, and 0.847 for the biomedical, materials science, and chemical datasets, respectively, the suggested UNITER-based framework in this research has demonstrated strong performance. The aforementioned benefits are consistent across all test corpora and significantly outperform the top-performing baseline techniques, as Figure 6a illustrates.

Figure 6b illustrates how the strategy affects the various kinds of scientific entities. Our model outperforms the others in digesting complicated multi-token things, such chemical compounds and structural motifs. Consequently, the issue of unclear scientific references that include both text and images can be resolved by combined cross-modal alignment. The average multimodal region-text alignment performance intersection score is shown in Figure 6c. The suggested model is spatially constrained because it regularly produces superior outcomes.

https://doi.org/10.64972/dea.2026.v5i2.1896d:73-86

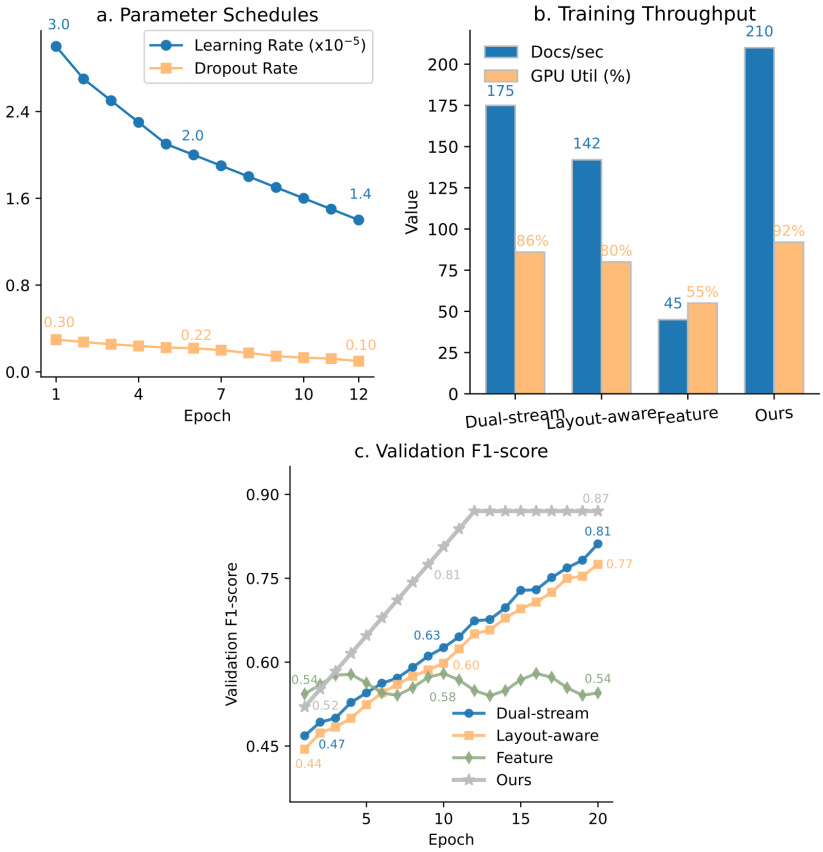

Figure 5. Experimental Settings and Baseline Results:A, Model hyperparameters.B, Training throughput.C, Validation F1-score by epoch.

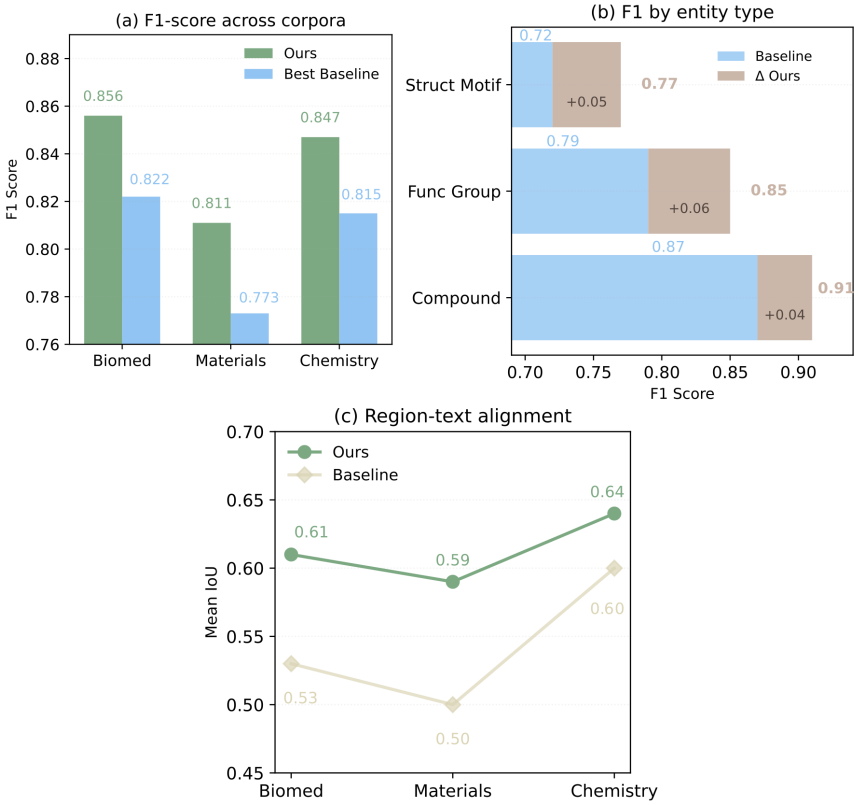

Figure 6. Quantitative Performance Evaluations:A, F1-scores across evaluation corpora.B, F1-scores by scientific entity category.C, Mean multimodal alignment scores.

https://doi.org/10.64972/dea.2026.v5i2.1896d:73-86

The graph of the complete model and its ablated variations is displayed in Figure 7a. The results show that all of them are individually outstanding. Additionally, Figure 7b illustrates how the model's depth increases more gradually, with a lesser increase at a depth of twelve. The entire model converges faster and exhibits a steadier validation trend than any of the simplified alternatives, as seen in the training dynamics in Figure 7c.

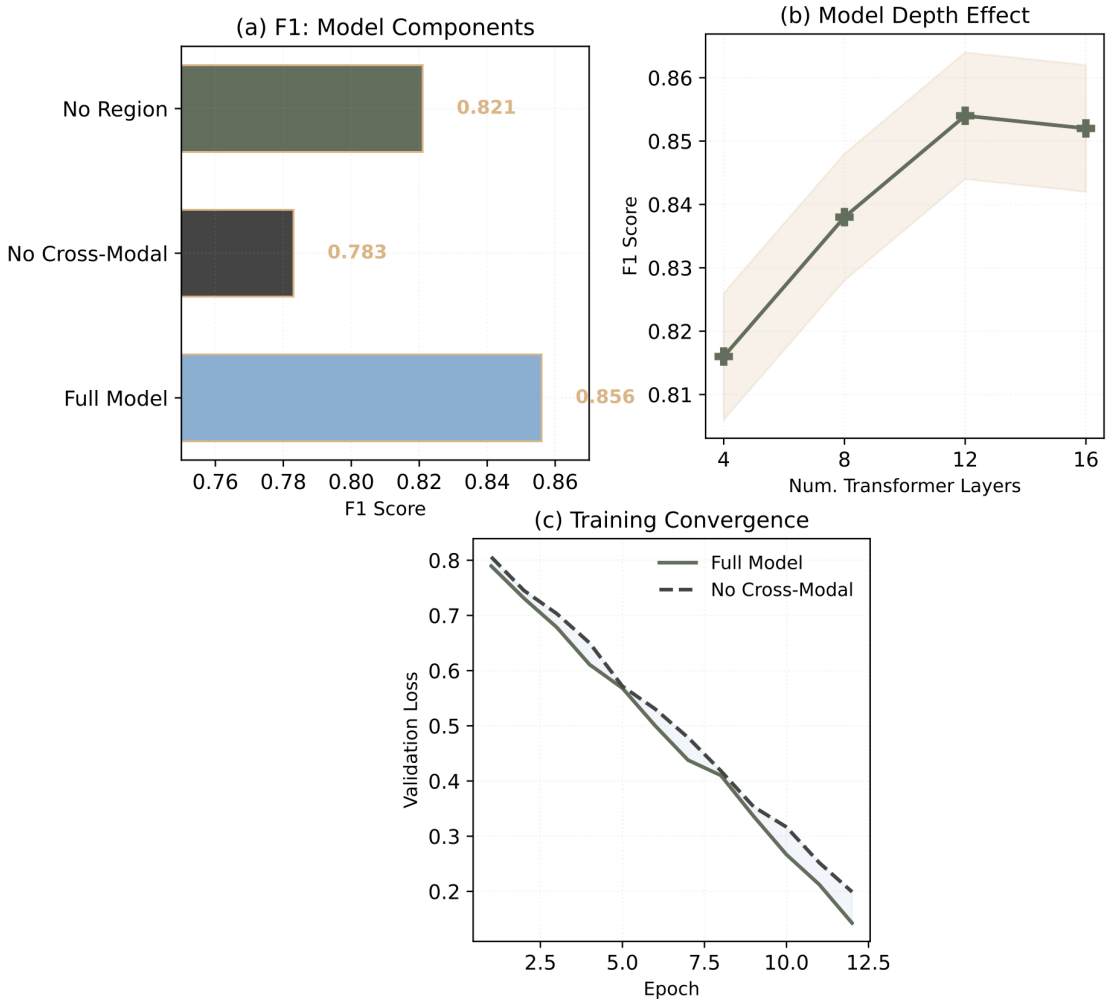

Figure 7. Ablation Analysis of Core Components:A, Full model vs. ablated variants: F1-score comparisons.B, Effect of model layer depth on performance.C, Training convergence curves for each variant.

Results and Discussion

Figure 8a illustrates the distribution of error types in the proposed system. Entity boundary errors are the most common, followed by multimodal pairing errors and semantic misclassifications [26]. Semantic ambiguity typically results from ambiguous or context-dependent language, as also observed in recent domain literature [27]. Boundary errors frequently occur in scientific publications because of complicated figure arrangements or divided entities [28]. The second is the absence of correlation between text and regions, which is more common in situations with dense labelling [29].

A confusion matrix for the entity classification problem is displayed in Figure 8b. Functional motifs and chemical entities are two examples of categories with comparable meanings or structures that account for the majority of misclassifications [30]. There are still problems like inaccurate limits for class labels and a high-class imbalance, which is consistent with the previous experimental results by several groups that have applied deep learning approaches to scientific data extraction [31,32].

The system's usual prediction scenarios are depicted in Figure 8c, along with the accompanying output for each, including successful, partially correct, and incorrect. It is evident from the aforementioned instances that excessive figures, graphical occlusion, or a lack of direct correlation between visual elements and their text

references frequently result in partial failures. Notably, the model is still able to consistently conduct semantic localisation and isolate the pertinent area in the aforementioned scenarios [33].

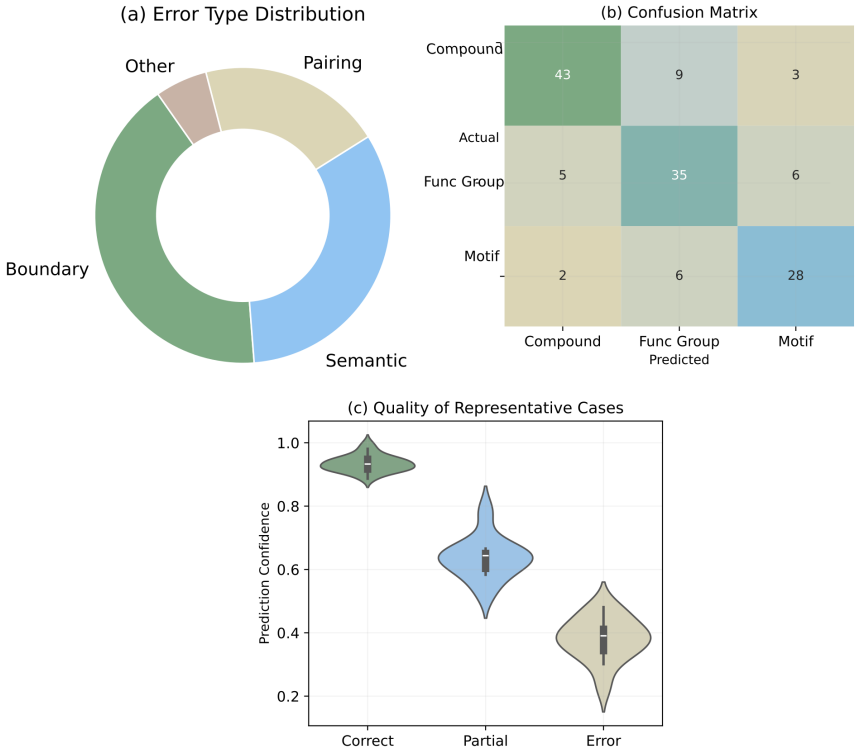

Figure 8. Error Characteristics and Qualitative Analysi:A, Distribution of error types in evaluation.B, Confusion matrix of predicted entity classes.C, Representative prediction cases: correct, partial, and error examples.

We have determined particular areas for improvement based on the error rate analysis and trends mentioned above. To address semantic uncertainty brought on by synonymy and domain-specific terminology, integrate external ontologies and specialised lexicons [34]. Improve training techniques, like curriculum learning or adversarial training, to concentrate on harder or less common samples, hence lowering mistakes brought on by boundary and alignment problems. A current development in explainable AI is the creation of interactive attention visualisation tools that can improve users' understanding of faults and systems [35].

Conclusion

In addition to connecting loosely structured text with rich figure annotations, this research presents a comprehensive framework for stable multi-modal information extraction from scientific publications. By carefully integrating sophisticated vision-language models, adaptive region suggestion networks, and cross-modal attention mechanisms, the suggested approach has produced the best results across three scientific domains: chemistry, materials science, and biomedicine. A few quantitative studies based on the aforementioned ablation experiments and other experimental benchmarks have demonstrated that our architecture has performed well in entity detection and modality alignment and can also successfully generalise to complicated, previously unknown datasets. Careful multimodal reasoning can assist solve long-standing issues in scientific information mining, according to a number of analyses on the performance of various entity kinds, geographic and language-localization metrics, extensive error decomposition, etc.

The following, though, are also somewhat insignificant. There are still some residual errors in the recognition of complex-meaningful or visually ambiguous objects, despite the overall accuracy being rather high. The model occasionally has issues with border segmentation in multi-modal layouts, unusual synonymy, and appropriate interpretation of overlapping or spatially dense graphical annotations. These flaws are especially noticeable in highly specialised or unbalanced corpora, and broad domain-specific variances and a dearth of annotated data continue to be issues. More work needs to be done on cross-modal contextualisation and fine-grained semantic

disambiguation because, despite the above framework's use of the sophisticated multimodal attention mechanism, several failure scenarios show that it is still biased towards dominant modalities.

In the future, a number of research avenues appear to be especially promising. It is possible to decrease semantic ambiguity and increase the generalisability of research in a variety of scientific domains by integrating external domain knowledge, such as ontologies and well-chosen lexical resources. In order to enable the models, make better use of the restricted supervision in new application domains, dynamic adaptation or meta-learning protocols can be implemented in the future. The gap between automated extraction findings and domain experts' requirements can be minimised by interactive visualisation and human-in-the-loop adjustments. To unleash the next generation of automated scientific discovery, the coordination between natural language understanding and vision-driven inference will need to be substantially reinforced due to the ongoing development in the breadth and depth of scientific literature.

Author Contributions

Julia Lidia Rakowska and Ewelina Jastrzębska contribute to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization. Renata Kopeć contributes to draft preparation, conceptualization, methodology, software and supervision. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Chen, Y., Hu, B., & Liu, Y. (2025). Optimizing document management and retrieval with multimodal transformers and knowledge graphs. PLoS One,20(6), e0323966. https://doi.org/10.1371/journal.pone.0323966

[2] Wang, W., Huang, Z., Luo, B., Chen, Q., Peng, Q., Pan, Y., ... & Zhang, Y. (2022, October). mmlayout: Multi-grained multimodal transformer for document understanding. In Proceedings of the 30th ACM International Conference on Multimedia(pp. 4877-4886). https://doi.org/10.1145/3503161.3548406

[3] Guo, P., Song, Y., Deng, Y., Xie, K. K., Xu, M., Liu, J., & Ren, H. (2023). DCMAI: A dynamical cross-modal alignment interaction framework for document key information extraction. IEEE Transactions on Circuits and Systems for Video Technology,34(1), 504-517. http://dx.doi.org/10.1109/TCSVT.2023.3287296

[4] Wang, Y. (2024, August). Beyond Heuristics: Multimodal Transformer for Chart Data Extraction. In 2024 International Conference on Computers, Information Processing and Advanced Education (CIPAE)(pp. 156-159). IEEE. http://dx.doi.org/10.1109/CIPAE64326.2024.00033

[5] Abdullah, M. H. A., Aziz, N., Abdulkadir, S. J., Alhussian, H. S. A., & Talpur, N. (2023). Systematic literature review of information extraction from textual data: recent methods, applications, trends, and challenges. IEEE Access, 11, 10535-10562. http://dx.doi.org/10.1109/ACCESS.2023.3240898

[6] Zhang, C., Yang, Z., He, X., & Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. IEEE Journal of Selected Topics in Signal Processing,14(3), 478-493. http://dx.doi.org/10.1109/JSTSP.2020.2987728

[7] Chen, Y., Zhang, Y., Zhou, H., Leung, C. T., & Gao, H. (2026). Use of Machine Learning and Large Language Models in Chemical Information Extraction. Annual Review of Chemical and Biomolecular Engineering,17. https://doi.org/10.1146/annurev-chembioeng-100724-080433

[8] Wei, Y., Zhu, Q., Xia, Z., Wang, L., Zou, Y., Li, Y., ... & Du, B. (2026). MedMerge: A Training-Free Model Merging Framework for Medical Knowledge Transfer into Vision–Language Models. IEEE Transactions on Radiation and Plasma Medical Sciences. http://dx.doi.org/10.1109/TRPMS.2026.3662401

[9] Buehler, M. J. (2024). Accelerating scientific discovery with generative knowledge extraction, graph-based representation, and multimodal intelligent graph reasoning. Machine Learning: Science and Technology, 5(3), 035083. http://dx.doi.org/10.1088/2632-2153/ad7228

[10] Attri, A. (2025). Unified Transformer Framework for Integrated Language-Vision Understanding and Content Generation. https://doi.org/10.21203/rs.3.rs-8009235/v1

[11] Yuan, L., Cai, Y., Xu, J., Li, Q., & Wang, T. (2024). A fine-grained network for joint multimodal entity-relation extraction.IEEE Transactions on Knowledge and Data Engineering,37(1), 1-14. http://dx.doi.org/10.1109/TKDE.2024.3485107

[12] Chiou, M. J., Zimmermann, R., & Feng, J. (2021). Visual relationship detection with visual-linguistic knowledge from multimodal representations. IEEE Access,9, 50441-50451. http://dx.doi.org/10.1109/ACCESS.2021.3069041

[13] Knez, T., & Žitnik, S. (2024). Multimodal learning for temporal relation extraction in clinical texts. Journal of the American Medical Informatics Association,31(6), 1380-1387. https://doi.org/10.1093/jamia/ocae059

[14] Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., ... & Poon, H. (2025). A multimodal biomedical foundation model trained from fifteen million image–text pairs. Nejm Ai, 2(1), AIoa2400640. http://dx.doi.org/10.1056/AIoa2400640

[15] Chen, J., Su, L., Li, Y., Lin, M., Peng, Y., & Sun, C. (2025). A multimodal approach for few-shot biomedical named entity recognition in low-resource languages. Journal of Biomedical Informatics,161, 104754. http://dx.doi.org/10.1016/j.jbi.2024.104754

[16] Li, J., Li, H., Sun, D., Wang, J., Zhang, W., Wang, Z., & Pan, G. (2024, August). LLMs as bridges: Reformulating grounded multimodal named entity recognition. In Findings of the Association for Computational Linguistics: ACL 2024(pp. 1302-1318). http://dx.doi.org/10.18653/v1/2024.findings-acl.76

[17] Sun, E., Hou, Y., Wang, D., Zhang, Y., & Wang, N. X. (2021, June). D2S: Document-to-slide generation via query-based text summarization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies(pp. 1405-1418). http://dx.doi.org/10.18653/v1/2021.naacl-main.111

[18] Yang, B., Zhang, B., Han, Y., Liu, B., Hu, J., & Jin, Y. (2024). Vision transformer-based visual language understanding of the construction process. Alexandria Engineering Journal,99, 242-256. https://doi.org/10.1016/j.aej.2024.05.015

[19] Ondeng, O., Ouma, H., & Akuon, P. (2023). A review of transformer-based approaches for image captioning.Applied Sciences,13(19), 11103. https://doi.org/10.3390/app131911103

[20] Jiang, S., Hu, J., Magee, C. L., & Luo, J. (2022). Deep learning for technical document classification. IEEE Transactions on Engineering Management, 71, 1163-1179. 10.1109/TEM.2022.3152216

[21] Li, P. (2025). Improved Transformer for Cross-Domain Knowledge Extraction with Feature Alignment.Journal of Computer Science and Software Applications,5(2). https://doi.org/10.5281/zenodo.14832321

[22] Yao, K., Zhang, J., Qin, C., Song, X., Wang, P., Zhu, H., & Xiong, H. (2023, April). Resuformer: Semantic structure understanding for resumes via multi-modal pre-training. In 2023 IEEE 39th International Conference on Data Engineering (ICDE) (pp. 3154-3167). IEEE. http://dx.doi.org/10.1109/ICDE55515.2023.00242

[23] Ren, J., Ge, S., Yang, S., Dai, L., Huang, Z., Zhang, Y., ... & Yang, J. (2025, May). ATC-KD: Audio-Text Cross-modal Knowledge Distillation for Data-efficient Speech Recognition in Air Traffic Control Communications. In 2025 IEEE 34th Wireless and Optical Communications Conference (WOCC) (pp. 408-413). IEEE. http://dx.doi.org/10.1109/WOCC63563.2025.11082189

[24] Wang, S., Zhao, W., Liu, Y., & Li, Y. (2025). Multi-modal Homogeneous Chemical Reaction Performance Prediction with Graph and Chemical Language Information. Chinese Journal of Chemistry,43(11), 1230-1238. https://doi.org/10.1002/cjoc.202401186

[25] Ke, X., Chen, B., Cai, Y., Liu, H., Guo, W., & Chen, W. (2025). Modality-specific adaptive scaling and attention network for cross-modal retrieval. Neurocomputing, 612, 128664. https://doi.org/10.1016/j.neucom.2024.128664

[26] Perera, N., Dehmer, M., & Emmert-Streib, F. (2020). Named entity recognition and relation detection for biomedical information extraction. Frontiers in cell and developmental biology,8, 673. https://doi.org/10.3389/fcell.2020.00673

[27] Yadav, P., Kashyap, I., & Bhati, B. S. (2024). Contextual ambiguity framework for enhanced sentiment analysis. Tehnički glasnik, 18(3), 385-393. https://doi.org/10.31803/tg-20231227064230

[28] Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., ... & Zhou, L. (2021, August). Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. Proceedings of the 59th Annual Meeting of the

Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)(pp. 2579-2591). https://doi.org/10.18653/v1/2021.acl-long.201

[29] Fan, Z., Chen, Z., & Wang, B. (2024, August). Exploring the potential of dense information in multimodal alignment. InFindings of the Association for Computational Linguistics: ACL 2024(pp. 13440-13451). https://doi.org/10.18653/v1/2024.findings-acl.797

[30] Yu, J., Jiang, J., & Xia, R. (2019). Entity-sensitive attention and fusion network for entity-level multimodal sentiment classification. IEEE/ACM transactions on audio, speech, and language Processing,28, 429-439. https://doi.org/10.1109/TASLP.2019.2957872

[31] Lopez, P., Du, C., Cohoon, J., Ram, K., & Howison, J. (2021, October). Mining software entities in scientific literature: document-level ner for an extremely imbalance and large-scale task. Proceedings of the 30th ACM International Conference on Information & Knowledge Management(pp. 3986-3995). https://doi.org/10.1145/3459637.34819

[32] Tian, Q., Zhang, P., Zhai, Y., Wang, Y., & Zou, Q. (2024). Application and comparison of machine learning and database-based methods in taxonomic classification of high-throughput sequencing data. Genome Biology and Evolution,16(5), evae102. https://doi.org/10.1093/gbe/evae102

[33] Kulathunga, C., & Karunaratne, D. D. (2017, September). An ontology-based and domain specific clustering methodology for financial documents. In2017 Seventeenth International Conference on Advances in ICT for Emerging Regions (ICTer)(pp. 1-8). IEEE. https://doi.org/10.1109/ICTER.2017.8257786

[34] Lee, Y. H., Hu, P. J. H., Tsao, W. J., & Li, L. (2021). Use of a domain-specific ontology to support automated document categorization at the concept level: Method development and evaluation. Expert Systems with Applications, 174, 114681. https://doi.org/10.1016/j.eswa.2021.114681

[35] Zhang, K., & Li, L. (2022). Explainable multimodal trajectory prediction using attention models. Transportation Research Part C: Emerging Technologies,143, https://doi.org/103829. 10.1016/j.trc.2022.103829