A Graph Attention Network Combining Multifaceted Element Relationships for Full Document-Level Understanding

computers

MDPI

Article

A Graph Attention Network Combining Multifaceted Element Relationships for Full Document-Level Understanding

Lorenzo Vaiani, Davide Napolitano and Luca Cagliero *

Dipartimento di Automatica e Informatica, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Torino, Italy; lorenzo.vaiani@polito.it (L.V.); davide.napolitano@polito.it (D.N.)
* Correspondence: luca.cagliero@polito.it; Tel.: +39-011-090-7179

Abstract

Question answering from visually rich documents (VRDs) is the task of retrieving the correct answer to a natural language question by considering the content of textual and visual elements in the document, as well as the pages' layout. To answer closed-ended questions that require a deep understanding of the hierarchical relationships between the elements, i.e., the full document-level understanding (FDU) task, state-of-the-art graph-based approaches to FDU model the pairwise element relationships in a graph model. Although they incorporate logical links (e.g., a caption refers to a figure) and spatial ones (e.g., a caption is placed below the figure), they currently disregard the semantic similarity among multimodal document elements, thus potentially yielding suboptimal scoring of the elements' relevance to the input question. In this paper, we propose GRAS-FDU, a new graph attention network tailored to FDU. GATS-FDU is trained to jointly consider multiple document facets, i.e., the local, spatial, and semantic elements' relationships. The results show that our approach achieves superior performance compared to several baseline methods.

Keywords: full document-level understanding; document element recognition; visually rich documents; multimodal learning

Academic Editor: Paolo Bellavista

Received: 22 July 2025
Revised: 13 August 2025
Accepted: 18 August 2025
Published: 1 September 2025

Citation: Vaiani, L.; Napolitano, D.; Cagliero, L. A Graph Attention Network Combining Multifaceted Element Relationships for Full Document-Level Understanding. *Computers* **2025**, *14*, 362. https://doi.org/10.3390/computers14090362

1. Introduction

Querying visually rich documents (VRDs), such as PDF files or scanned document images, is particularly challenging because they include a mix of textual and visual elements (e.g., textual paragraphs, headers, captions, figures, diagrams) arranged according to arbitrarily complex layouts. Given a question posed in natural language, question answering (QA) from VRDs entails retrieving the elements that are most relevant to the input question and then generating the corresponding answer [1].

When dealing with closed-ended, multiple-choice questions, QA from VRDs can be reformulated as a multilabel classification task [2]. Specifically, when the answer consists of one or more elements appearing on the document page (e.g., Which figures contain a boxplot?), the QA problem consists of document element recognition (DER) and can be addressed as a content retrieval task. The relevance of every document element is estimated first, and then the elements are ranked by decreasing relevance; finally, the top-ranked ones are returned [3].

In this work, we are interested in questions that have a sequence of answers that occur on multiple pages of the full document. Answering these questions also involves identifying elements that are hierarchically related to those mentioned in the question

(e.g., Which section describes a given figure?). The latter problem, commonly represented by full document-level understanding (FDU), is a DER specialization that requires the accurate detection of logical element relationships (e.g., A caption describes a given figure) and spatial ones (e.g., A textual paragraph appears next to a given figure) [4].

State-of-the-art graph-based approaches to FDU model VRD elements and element relationships as graph nodes and edges, respectively [3]. A graph neural network (GNN) is trained to encode key information about document elements. Given a question, the GNN returns the probability of every candidate element for prediction.

Alternative FDU approaches [5–9] rely on multimodal learning frameworks to estimate the similarities between textual questions and multimodal document elements. Although multimodal learning solutions are established for image–text retrieval tasks, they neglect the logical and spatial relationships modeled by graph-based approaches, thus struggling with documents characterized by complex layout structures.

In this work, we bridge the gap between graph-based and multimodal learning approaches to FDU. We present Graph Attention Network-Based Semantics-Aware FDU (GATS-FDU), a new graph attention network that integrates logical, spatial, and semantic element relationships. First, we represent multipage documents as composite graphs that encapsulate both content and layout information. Then, we train a GAT to predict, for every candidate element within a target pair (question and VRD), the multifaceted relevance score. Experiments conducted on a benchmark VRD collection demonstrate that our method significantly outperforms five baseline methods in terms of exact matching accuracy.

The rest of this paper is organized as follows. Section 2 gives an overview of the related work. Section 3 formalizes the FDU task addressed in the present work. Section 4 thoroughly describes GATS-FDU. Section 5 summarizes the main experimental results. Finally, Sections 6 and 7 summarize the main findings, draw conclusions, highlight the main limitations of the proposed method, and discuss the future research extensions of the present work.

2. Related Works

Existing approaches to FDU can be classified as (1) Transformer-based architectures and vision LLMs [8,10–14]; (2) text–image retrieval architectures [5–7,15,16]; and (3) graph neural networks [3].

2.1. Transformer-Based Architectures and Vision LLMs

Layout-aware language models, such as LayoutLM [8] and LXMERT [10], are capable of understanding both semantic similarity and layout information but require a computationally intensive fine-tuning stage. For example, Hi-VT5 [12] relies on a hierarchical encoder–decoder framework based on T5 [17], where each page is encoded independently. GRAM [18], instead, extends an existing single-page model [13] to enable interpage interactions. Pix2Struct [14] fine-tunes the single-page Doc2VQA model to select the most relevant page including the answer based on question–page matching. More recently, LLM-based models, such as LayoutLLM [19] and HRVDA [20], have been pretrained for complex semantic understanding from VRDs and then specialized via instruction tuning. A more extensive review of Transformer-based and vision LLM-based approaches can be found in [2].

2.2. Text-Image Retrieval

Image–text retrieval models leverage multimodal representations of images and text to retrieve elements in the document that are semantically similar to those in the question. Contrastive-based models, such as CLIP [7], are, in general, less computationally expensive to fine-tune on domain-specific VRDs than Transformers.

To evaluate the similarity between natural language questions and VRD content, previous works have adopted (1) unimodal encoders, which separately embed text (e.g., BERT [21], RoBERTa [22]) and visual content (e.g., VIT [23]), or (2) multimodal encoders, which jointly process visual and textual elements (e.g., CLIP [7], EVA-CLIP [24], BLIP [25], E5-V [26], VisualBERT [15], ViLT [27], LXMERT [10]). MemSum-DQA [6] is, instead, an extractive summarization model adapted to solve the FDU task. It first adds prefixes to each text element in the VRD with the provided question and question type and then selectively extracts text blocks as answers from documents. To perform the extractive step, in [6], the authors explore the use of several cross-modal encoders.

2.3. Graph Neural Networks

LoSpa [3] is, to the best of our knowledge, the only graph-based model that incorporates both logical and spatial information. Although it is competitive against both image–text retrievers and Transformer-based models, it ignores the semantic relationships in the graph. Our solution, instead, opportunistically combines the semantic information conveyed by multimodal learning models with hierarchical and layout-related document features.

2.4. Position of Our Work

To the best of our knowledge, the integration of multimodal encoders into graph-based networks tailored to FDU has never been investigated so far.

3. Full Document-Level Understanding

Let D be a VRD with a fixed layout structure. D's elements $E = \{e_1, e_2, \ldots, e_n\}$ consist of sections, subsections, and other elements, such as tables, figures, table captions, and figure captions. Pairs of elements e_x and e_y in D may have hierarchical logical relationships, such as a subsection is part of a section or a caption refers to a figure, or spatial relationships, depending on the layout of the document's pages. Examples of spatial relationships are top, bottom, left, right, top-left, top-right, bottom-left, and bottom-right [3] (e.g., a figure is on the left-hand side of a section).

Given a closed-ended question Q on D (e.g., Which figures contain a boxplot?), the document element recognition task aims to identify the correct elements in E to return as the final answer. VRD processing encapsulates both spatial, logical, and semantic information.

DER in multipage documents can involve reviewing the full document's contents to identify the elements that are hierarchically related to the queried elements in the question [3]. For example, the question Which section describes a given figure? requires the identification of the Section element whose content explicitly refers to the Figure element mentioned in the question. This subtask, namely full document-level understanding (FDU), is the main problem addressed in this work.

4. Methodology

We present GAT-Based Semantics-Aware Full Document-Level Understanding (GATS-FDU), a new graph-based approach to tackle FDU in multipage VRDs. A sketch of the architecture is depicted in Figure 1.

GATS-FDU consists of four main steps:

1. *Multimodal element encoding*, whose goal is to generate vector representations of the VRD's elements in a shared latent space;
2. *Graph modeling*, which aims to represent logical, spatial, and semantic entity relationships in a composite graph;

3. *Multifaceted graph attention network training*, which learns from training triples of questions, VRDs, and candidate elements and determines the relevance scores to be assigned according to logical, spatial, and semantic facets;
4. *Elements' retrieval*, whose aim is to apply the GAT trained for FDU and return a shortlist of the most likely elements.

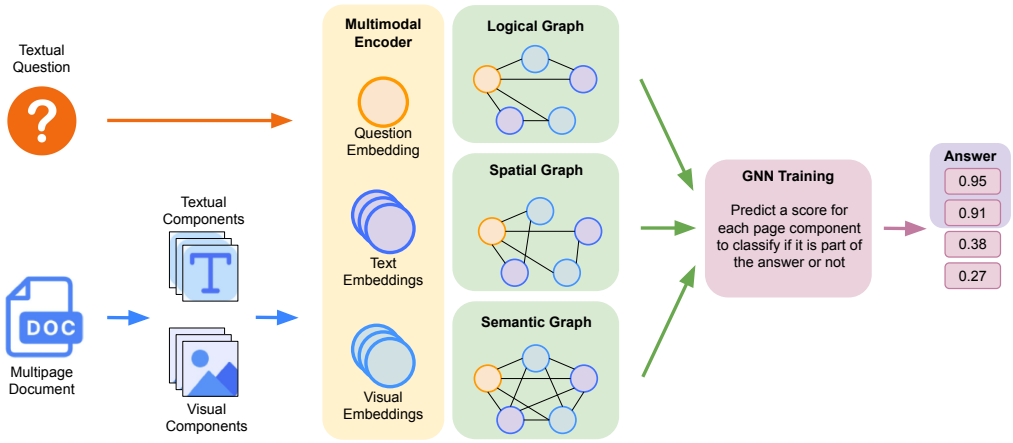

Figure 1. Sketch of the proposed GATS-FDU approach to full document-level understanding.

A more detailed description of each step follows.

4.1. Multimodal Element Encoding

Textual and visual elements are encoded into a shared latent space to allow the GAT to capture semantic similarities between a multimodal VRD's elements. We segment the elements' text to conform to the maximum token length. For scanned/PDF versions of the documents, we extract tables and figures by detecting and leveraging their bounding boxes. When an image is not associated with any descriptive text, we synthetically generate the image caption using Qwen2.5VL [28].

GATS-FDU supports a number of different multimodal encoders, such as CLIP [7], JINA-CLIP [29], EVA-CLIP [30], BLIP [25], and E5-V [26]. Based on our preliminary experimental comparisons, we will consider E5-V as the default encoder unless otherwise specified. The text of each question is encoded using the same encoder adopted for document elements.

4.2. Multifaceted Graph Modeling

We build a composite graph $\mathcal{G}$ consisting of the three subgraphs G_{lo}, G_{sp}, G_{se}, which, respectively, represent the logical, spatial, and semantic relationships among the elements in E. Notice that G_{lo}, G_{se} incorporate both intra- and interpage elements' relationships and also account for relationships between questions and VRD elements.

As logical and spatial relationships, hereafter, we will consider those defined in the PDFVQA benchmark [3]:

- *Logical relationship types*: parent and child relationships (e.g., title of section, caption of figure, caption of table, list in text) and the opposite child relationships.
- *Spatial relationship types*: top, bottom, left, right, top-left, top-right, bottom-left, and bottom-right.

Since the GATS-FDU framework is general, the above-mentioned list of relationship types can be straightforwardly extended.

Oriented edges in G_{lo} and G_{sp}, respectively, indicate the presence of a parent/child or spatial relationship.

G_{se}'s edges indicate the pairwise similarity between pairs of multimodal elements in the embedding space. Given G_{se}'s nodes e_x and e_y, the bidirectional edge connecting them is weighted by sim(e_x,e_y), where sim($\cdot$) is the cosine similarity between the elements' embeddings.

To avoid generating fully connected similarity graphs, which would likely include noisy information and be costly in training GNNs, we prune the initial graph by maintaining, for each node G_{se}, only the top-K outgoing edges in order of decreasing similarity, where K is a user-specified parameter.

The question embedding is included as an additional node of the graph. Regardless of the type of subgraph, the question encoding is always linked to all other nodes to ensure that the document element embeddings are aware of the question's semantic content.

4.3. Multifaceted Graph Attention Network

GATS-FDU employs a graph attention network (GAT) to process the composite graph, including complementary, heterogeneous facets (i.e., logical, spatial, and semantic relationships). GAT training leverages the attention mechanism to process all subgraphs as a shared representation space, where the contents of both VRDs and questions are embedded [31]. The tight integration between multifaceted subgraphs, modeling logical, spatial, and semantic graphs, ensures effective information fusion and alignment.

Given a VRD D, a question Q, and a candidate target element $e \in E$, the GAT is trained to jointly return three separate probability scores corresponding to the following relevance estimates:

- score$_{lo}$(D,Q,e): the relevance score of element e to Q on D according to the logical relationships facet;
- score$_{sp}$(D,Q,e): the relevance score of element e to Q on D according to the spatial relationships facet;
- score$_{se}$(D,Q,e): the relevance score of element e to Q on D according to the semantic relationships facet.

The separate scores are optimized during training using the binary cross-entropy with logits loss function and averaged to estimate the overall element relevance.

4.4. Element Retrieval

For each question Q and VRD D, the retrieval stage leverages the inference capabilities of the trained GAT to estimate the relevance of every candidate element in E. All candidates whose estimated probability of being part of the correct answer is greater than 50% are returned.

5. Experimental Results

This section summarizes the main results achieved in the empirical evaluation of the GATS-FDU model.

5.1. Hardware Settings

To run the experiments, we exploited a machine equipped with an 18-core Intel Core i9-10980XE processor, an Nvidia A6000 GPU, and 128 GB of RAM.

5.2. VRD Contents

We analyze the collection of VRDs and questions released in the PDF-VQA challenge [4]. The benchmark consists of the PDF versions of VRDs retrieved from the PubMed Central (PMC) Open Access Subset. The documents are annotated according to the in-

dications reported in [3]. The main purpose is to accurately predict the index(es) of the elements that can provide correct answers to the given questions.

The dataset consists of three main splits: training, validation, and testing. For each split, it contains the questions and their corresponding ground-truth answers. Furthermore, for each VRD, it includes the bounding box coordinates of each document layout component, the textual contents inside each bounding box, and parent–child relationships.

A list of the PDFVQA dataset's features (https://www.kaggle.com/competitions/pdfvqa/data (latest access: 20 July 2025)) is given as follows:

- *pmcid*: identifier of the question–answer pair;
- *question*: text of natural language question;
- *question type*: logical relationship (parent or child relationship);
- *answer*: list of answer text contents;
- *global id*: global identifier of corresponding answers—if it is set to -1, then it means that *No subsection* or *No section* can be found to answer this question.

5.3. Baseline Methods

We compare our approach against the following state-of-the-art methods supporting multimodal element retrieval:

- *Multimodal learning architectures*: ViLT [27], VisualBERT [15], CLIP [5,7];
- *Graph-based approaches*: LoSpa [3];
- *Layout-aware vision language models*: LXMERT [10], M4C [11].

To ensure a fair comparison with our approach, for each baseline method, we perform parameter tuning and consider the result of the best-performing settings. For all models, we run the experiments with the corresponding open-source model version. Specifically, for ViLT, VisualBERT, CLIP, and LXMERT, we rely on the latest HuggingFace model versions, whereas, for LoSpa and M4C, we follow the indications provided by the respective authors.

5.4. Performance Metrics

Similarly to [3], we adopt the exact matching accuracy (EMA). It counts the number of questions for which the FDU approach correctly predicts *all* elements associated with the corresponding questions in the ground truth. Note that, for every question, the expected elements may not be unique and could even be zero. Note also that, since each VRD potentially contains hundreds of elements, in the classical FDU setting (which is an extreme classification task), the EMA is, in general, preferable to per-class metrics, such as precision and recall [1].

5.5. Reproducibility

The official repository of the research project is available online at https://github.com/VaianiLorenzo/gat-doc (accessed on 20 July 2025). The repository includes all code sources, the references to open data and tools, the settings used in the experiments, and the main empirical results.

5.6. Configuration Settings

For GATS-FDU, we exploit AdamW [32] and tune the network hyperparameters to define the best configuration settings: learning rate 5×10^{-5}, decay 1×10^{-5}, loss function BCEWithLogitsLoss, learning rate $\gamma = 0.5$ every 10 epochs, learning rate scheduler with step decay 10, combined weight decay $\lambda = 10^{-5}$, and batch size 1.

We also test different encoding strategies and empirically verify whether the contents of the tables and figures correctly match the corresponding captions. E5-V [26] is the most effective encoder as it accurately matches all image captions and most table captions.

The experiments have been executed across five different runs to compute statically relevant metrics. We report the average results (with the corresponding confidence intervals). More details are given in the public repository.

5.7. Results

Table 1 compares the EMA results for the best configurations of both GATS-FDU and the baseline methods. It is worth noting that CLIP achieves competitive results despite not being specifically designed for the FDU task. This performance can likely be attributed to the strength of its joint visual–textual embeddings, which outperform the BERT- and ResNet-based representations employed by LoSpa. As a result, CLIP is able to retrieve relevant document elements effectively even without modeling logical or spatial relationships explicitly. Compared to LoSpa, our proposed GATS-FDU further improves the performance by integrating semantic relationships into the graph representation, in addition to the logical and spatial ones. While LoSpa can effectively capture hierarchical and layout-related dependencies, it cannot exploit the semantic similarity between multimodal elements, which is crucial when the connection between a question and its answer extends beyond explicit structural links.

Table 1. Performance comparison among different FDU approaches, with 95% confidence intervals. The results of the best-performing methods are written in boldface.

Model	Visual	Text	Logical	Spatial	Semantic	Val	Test
VisualBERT	✓	×	×	×	×	21.55% (±1.20)	18.52% (±1.15)
ViLT	✓	×	×	×	×	10.21% (±0.85)	9.87% (±0.80)
LxMERT	✓	×	×	×	×	16.37% (±1.05)	14.41% (±0.95)
M4C	✓	✓	×	✓	×	12.14% (±0.90)	13.77% (±1.00)
CLIP	✓	✓	×	×	×	29.95% (±1.25)	34.52% (±1.35)
LoSpa	✓	✓	✓	✓	×	30.21% (±1.10)	28.99% (±1.20)
GATS-FDU (ours)	✓	✓	✓	✓	✓	**31.50%** (±1.05)	**38.72%** (±1.15)

5.8. Graph Ablation Study

We carry out an ablation study to analyze the effects of including different combinations of relationship facets in the graph (logical, spatial, semantic). Figure 2 shows the average results achieved for different combinations, together with the corresponding confidence levels. The results obtained on the PDFVQA benchmark show that the combination of logical and semantic information is the most beneficial, although the performance gap is not statistically significant. Spatial relationships appear to be less discriminating in predicting the correct answer, likely due to the high variability in the layout structure of the documents analyzed.

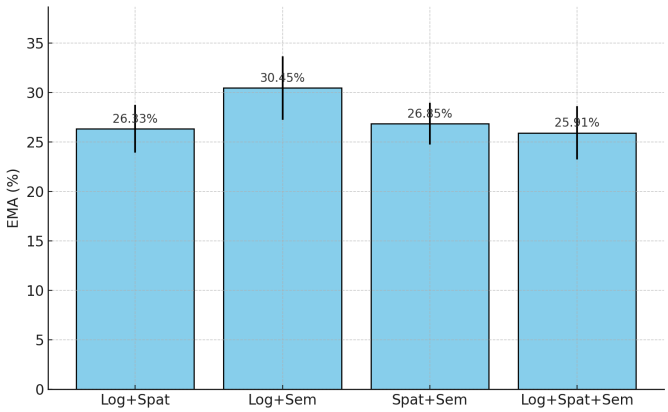

Figure 2. Ablation study of the facet combinations (logical+spatial, logical+semantic, spatial+semantic, logical+spatial+semantic) represented in the composite graph.

5.9. Impact of Logical Graph Scope

To further investigate the impact of the logical graph's scope, in our proposed solution, we extend the original LoSpa configuration, which models logical connections only at the page level, by including document-level connections between nodes. This enhancement allows the model to capture hierarchical relationships spanning across different pages, thus better reflecting the nature of multipage VRDs. As shown in Table 2, expanding the scope from the page level to the document level yields a modest yet consistent improvement in EMA (+0.86%), suggesting that cross-page logical dependencies contribute to a more complete representation of the document structure and can benefit FDU performance.

Table 2. Comparison between page-level and document-level scope for the G_{lo} graph.

Scope	Mean Acc
Page level	26.33%
Document level	27.19%

5.10. Impact of Number of Attention Heads

We also analyze the effect of varying the number of attention heads in the GAT architecture, as reported in Table 3. Increasing the number of heads from one to four yields the highest exact matching accuracy, suggesting that a moderate degree of multihead attention is beneficial in integrating heterogeneous graph facets. Conversely, using a single head limits the model's ability to capture diverse relational patterns, while excessively increasing the number of heads (8 or 16) slightly degrades the performance, likely due to the overfragmentation of the attention space and the introduction of redundant computations. These results indicate that balancing the diversity of attention mechanisms with computational efficiency is crucial for optimal FDU performance.

Table 3. Validation performance of global GNN with G_{lo} and G_{se} graphs for different numbers of attention heads, including 95% confidence intervals. The result of the best setting is written in boldface.

Number of Heads	Mean Acc
1	30.45% ($\pm$1.10)
4	**31.50%** ($\pm$1.05)
8	30.46% ($\pm$1.12)
16	30.29% ($\pm$1.15)

5.11. Parameter Analysis

We analyze the effects of the selection of the top-K most relevant elements for each input question. The results are reported in Figure 3. We vary the value of K between 1 and 10, achieving the best performance when K ranges between 3 and 6. When setting smaller values, the system tends to discard candidate elements that are highly similar to the input question (i.e., the average cosine similarity with the returned elements is around 70% when k = 3). Conversely, when increasing the value of K, the selected candidates also include outliers (i.e., the average cosine similarity decreases to 60% when K = 10).

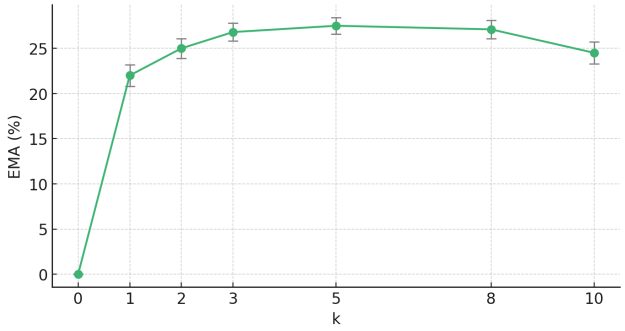

Figure 3. Effects of top-*K* selection.

6. Discussion

State-of-the-art FDU approaches currently face challenges in dealing with multipage VRDs [3] for the following reasons:

- The performance is influenced by the document layout structure, which could vary from one page to another and from one document to another. Visual language models that are fine-tuned on specific layout structures are not easily portable to different structures.
- The logical hierarchical relationships among VRD elements can occur both within a single page and across different pages. While existing approaches primarily focus on intrapage relationships, they often overlook those that span multiple pages.
- The similarity among VRD elements' content is not embedded in the graph-based VRD representation [3].
- Text-only FDU approaches are not capable of retrieving answers that include visual content only.
- Image–text retrieval modules ignore the impacts of logical and spatial relationships.

In this work, we pave the way for new graph-based approaches combining logical, spatial, and semantic element relationships. The key advantages are summarized below:

- Graph models are particularly effective in representing the inherent complexity of multipage VRDs.
- Graph attention networks are suited to heterogeneous graph structures, modeling multiple document facets.
- GNN training requires a more limited number of examples than visual LLM fine-tuning.

The main limitations of the current work are enumerated below:

- GATS-FDU has been tested on an FDU benchmark consisting of scientific articles. Its portability to other domains has not yet been investigated and will be addressed in future work.
- There is room for the extension of the proposal to the more complex DER task, i.e., when the questions do not refer to hierarchical element relationships, but this was beyond the scope of the present work.
- The documents were all written in English and contained a limited number of visual items. The applicability of the method to multilingual documents and to more complex multimodal scenarios (e.g., presentation slides) is still under investigation.

7. Conclusions and Future Work

This paper presented a graph-based approach to full document-level understanding in visually rich documents. The proposed solution outperforms existing multimodal strategies, including former graph-based methods, thanks to the integration of multifaceted

document relationship information, such as logical relationships, spatial relationships, and semantic relationships. The integration of semantic similarity scores derived by multimodal learning architectures boosts the model's performance, particularly when the connection between the posed questions and the candidate elements is ambiguous.

The empirical results demonstrate significant advantages in integrating different document facets. First, the use of multimodal encoders addresses the limitations of text-only methods. Secondly, combining logical and semantic information helps to prevent the retrieval of logically relevant but irrelevant answers, thereby improving the model's precision. Finally, spatial relationships can sometimes be misleading, especially in the case of multipage VRDs. In these complex scenarios, the fusion of spatial and semantic relationships helps to maintain the quality of the extracted information.

In future work, we plan to pursue the following research lines:

- Explore new application scenarios (e.g., financial documents [33], news documents [34], legal documents [35]) and document types (e.g., charts [36,37], receipts [9,38], reports [39]);
- Extend the scope of the research to the more general document element recognition task [9];
- Address challenging aspects related to FDU, such as scalability, real deployment, and memory footprint;
- Design new graph-based architectures leveraging deep learning and reinforcement learning strategies for content retrieval and refinement;
- Evaluate the portability of GATS-FDU to diverse languages and linguistic styles.

Author Contributions: Conceptualization, L.V., D.N., L.C.; methodology, L.V., D.N., L.C.; software, L.V., D.N.; validation, L.V., D.N.; formal analysis, L.V., D.N., L.C.; investigation, L.V., D.N.; resources, L.C.; data curation, L.V., D.N.; writing—original draft preparation, L.C.; writing—review and editing, L.V., D.N., L.C.; visualization, L.V., D.N.; supervision, L.C.; project administration, L.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Restrictions apply to the availability of these data. This research used data released by third parties. Data were obtained from the Workshop on Document Intelligence Understanding Challenge (https://kaggle.com/competitions/pdfvqa (accessed on 20 July 2025)) [4].

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

VRD	Visually Rich Document
QA	Question Answering
VQA	Visual Question Answering
DER	Document Entity Recognition
FDU	Full Document-Level Understanding
LLM	Large Language Model
VLLM	Vision Large Language Model
EMA	Exact Matching Accuracy
MML	Multimodal Learning

References

1. Ding, Y.; Han, S.C.; Lee, J.; Hovy, E. Deep Learning based Visually Rich Document Content Understanding: A Survey. *arXiv* **2025**, arXiv:2408.01287.

2. Barboule, C.; Piwowarski, B.; Chabot, Y. Survey on Question Answering over Visually Rich Documents: Methods, Challenges, and Trends. *arXiv* **2025**, arXiv:cs.CL/2501.02235. [CrossRef]
3. Ding, Y.; Luo, S.; Chung, H.; Han, S.C. PDF-VQA: A New Dataset for Real-World VQA on PDF Documents. In Proceedings of the Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track, Turin, Italy, 18–22 September 2023; De Francisci Morales, G., Perlich, C., Ruchansky, N., Kourtellis, N., Baralis, E., Bonchi, F., Eds.; Springer: Cham, Switzerland, 2023; pp. 585–601.
4. Han, S.C.; Ding, Y.; Luo, S.; Poon, J.; Yoon, H.; Huang, Z.; Duuring, P.; Holden, E.J. Workshop on Document Intelligence Understanding. *arXiv* **2023**, arXiv:cs.IR/2307.16369. [CrossRef]
5. Napolitano, D.; Vaiani, L.; Cagliero, L. Enhancing BERT-Based Visual Question Answering through Keyword-Driven Sentence Selection. *arXiv* **2023**, arXiv:cs.CL/2310.09432.
6. Gu, N.; Gao, Y.; Hahnloser, R.H.R. MemSum-DQA: Adapting An Efficient Long Document Extractive Summarizer for Document Question Answering. *arXiv* **2023**, arXiv:cs.CL/2310.06436.
7. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. *arXiv* **2021**, arXiv:abs2103.00020. [CrossRef]
8. Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; Wei, F. LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking. *arXiv* **2022**, arXiv:cs.CL/2204.08387.
9. Ding, Y.; Vaiani, L.; Han, S.C.; Lee, J.; Garza, P.; Poon, J.; Cagliero, L. 3MVRD: Multimodal Multi-task Multi-teacher Visually-Rich Form Document Understanding. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand, 11–16 August 2024; Ku, L., Martins, A., Srikumar, V., Eds.; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 15233–15244. [CrossRef]
10. Tan, H.; Bansal, M. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. *arXiv* **2019**, arXiv:cs.CL/1908.07490. [CrossRef]
11. Hu, R.; Singh, A.; Darrell, T.; Rohrbach, M. Iterative Answer Prediction With Pointer-Augmented Multimodal Transformers for TextVQA. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Los Alamitos, CA, USA, 13–19 June 2020; pp. 9989–9999. [CrossRef]
12. Tito, R.; Karatzas, D.; Valveny, E. Hierarchical multimodal transformers for Multipage DocVQA. *Pattern Recogn.* **2023**, *144*, 109834. [CrossRef]
13. Chen, J.; Lv, T.; Cui, L.; Zhang, C.; Wei, F. XDoc: Unified Pre-training for Cross-Format Document Understanding. *arXiv* **2022**, arXiv:cs.CL/2210.02849.
14. Kang, L.; Tito, R.; Valveny, E.; Karatzas, D. Multi-Page Document Visual Question Answering using Self-Attention Scoring Mechanism. *arXiv* **2024**, arXiv:cs.CV/2404.19024.
15. Li, L.H.; Yatskar, M.; Yin, D.; Hsieh, C.J.; Chang, K.W. VisualBERT: A Simple and Performant Baseline for Vision and Language. *arXiv* **2019**, arXiv:cs.CV/1908.03557. [CrossRef]
16. Peng, Q.; Pan, Y.; Wang, W.; Luo, B.; Zhang, Z.; Huang, Z.; Hu, T.; Yin, W.; Chen, Y.; Zhang, Y.; et al. ERNIE-Layout: Layout Knowledge Enhanced Pre-training for Visually-rich Document Understanding. *arXiv* **2022**, arXiv:cs.CL/2210.06155.
17. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *arXiv* **2023**, arXiv:cs.LG/1910.10683.
18. Blau, T.; Fogel, S.; Ronen, R.; Golts, A.; Ganz, R.; Avraham, E.B.; Aberdam, A.; Tsiper, S.; Litman, R. GRAM: Global Reasoning for Multi-Page VQA. *arXiv* **2024**, arXiv:cs.CL/2401.03411. [CrossRef]
19. Luo, C.; Shen, Y.; Zhu, Z.; Zheng, Q.; Yu, Z.; Yao, C. LayoutLLM: Layout Instruction Tuning with Large Language Models for Document Understanding. *arXiv* **2024**, arXiv:cs.CV/2404.05225. [CrossRef]
20. Liu, C.; Yin, K.; Cao, H.; Jiang, X.; Li, X.; Liu, Y.; Jiang, D.; Sun, X.; Xu, L. HRVDA: High-Resolution Visual Document Assistant. *arXiv* **2024**, arXiv:cs.CV/2404.06918. [CrossRef]
21. Devlin, J.; Chang, M.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:abs/1810.04805.
22. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:cs.CL/1907.11692.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:abs/2010.11929.
24. Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; Cao, Y. EVA: Exploring the Limits of Masked Visual Representation Learning at Scale. *arXiv* **2022**, arXiv:cs.CV/2211.07636. [CrossRef]
25. Li, J.; Li, D.; Xiong, C.; Hoi, S.C.H. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv* **2022**, arXiv:abs/2201.12086.
26. Jiang, T.; Song, M.; Zhang, Z.; Huang, H.; Deng, W.; Sun, F.; Zhang, Q.; Wang, D.; Zhuang, F. E5-V: Universal Embeddings with Multimodal Large Language Models. *arXiv* **2024**, arXiv:cs.CL/2407.12580. [CrossRef]

27. Kim, W.; Son, B.; Kim, I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision. *arXiv* **2021**, arXiv:stat.ML/2102.03334.
28. Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. Qwen2.5-VL Technical Report. *arXiv* **2025**, arXiv:2502.13923. [CrossRef]
29. Koukounas, A.; Mastrapas, G.; Günther, M.; Wang, B.; Martens, S.; Mohr, I.; Sturua, S.; Akram, M.K.; Martínez, J.F.; Ognawala, S.; et al. Jina CLIP: Your CLIP Model Is Also Your Text Retriever. *arXiv* **2024**, arXiv:cs.CL/2405.20204. [CrossRef]
30. Sun, Q.; Fang, Y.; Wu, L.; Wang, X.; Cao, Y. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv* **2023**, arXiv:cs.CV/2303.15389. [CrossRef]
31. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. *arXiv* **2018**, arXiv:stat.ML/1710.10903.
32. Loshchilov, I.; Hutter, F. Fixing Weight Decay Regularization in Adam. *arXiv* **2017**, arXiv:abs/1711.05101.
33. Islam, P.; Kannappan, A.; Kiela, D.; Qian, R.; Scherrer, N.; Vidgen, B. FinanceBench: A New Benchmark for Financial Question Answering. *arXiv* **2023**, arXiv:cs.CL/2311.11944. [CrossRef]
34. D'Amico, L.; Napolitano, D.; Vaiani, L.; Cagliero, L. PoliTo at MULTI-Fake-DetectiVE: Improving FND-CLIP for Multimodal Italian Fake News Detection. In Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), Parma, Italy, 7–8 September 2023; Lai, M., Menini, S., Polignano, M., Russo, V., Sprugnoli, R., Venturi, G., Eds.; CEUR-WS: Aachen, Germany, 2023; Volume 3473.
35. Sharma, E.; Li, C.; Wang, L. BIGPATENT: A Large-Scale Dataset for Abstractive and Coherent Summarization. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; Korhonen, A., Traum, D., Màrquez, L., Eds.; ACL: New York, NY, USA, 2019; pp. 2204–2213. [CrossRef]
36. Masry, A.; Long, D.; Tan, J.Q.; Joty, S.; Hoque, E. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, 22–27 May 2022; ACL: New York, NY, USA, 2022; pp. 2263–2279. [CrossRef]
37. Dai, Y.; Han, S.C.; Liu, W. Graph-Based Multimodal Contrastive Learning for Chart Question Answering. In Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, 13–18 July 2025; Ferro, N., Maistro, M., Pasi, G., Alonso, O., Trotman, A., Verberne, S., Eds.; ACM: New York, NY, USA, 2025; pp. 2658–2663. [CrossRef]
38. Huang, Z.; Chen, K.; He, J.; Bai, X.; Karatzas, D.; Lu, S.; Jawahar, C.V. ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019; Springer Nature: Gewerbestrasse, Switzerland, 2019; pp. 1516–1520. [CrossRef]
39. Van Landeghem, J.; Borchmann, L.; Tito, R.; Pietruszka, M.; Jurkiewicz, D.; Powalski, R.; Joziak, P.; Biswas, S.; Coustaty, M.; Stanisawek, T. ICDAR 2023 Competition on Document UnderstanDing of Everything (DUDE). In Proceedings of the ICDAR 2023, San José, CA, USA, 21–26 August 2023; Springer Nature: Gewerbestrasse, Switzerland, 2023.