MGA-VQA: Secure and Interpretable Graph-Augmented Visual Question Answering with Memory-Guided Protection Against Unauthorized Knowledge Use

Ahmad Mohammadshirazi
Ohio State University, Flairsoft
Columbus, Ohio, US
mohammadshirazi.2@osu.edu

Pinaki Prasad Guha Neogi
Ohio State University
Columbus, Ohio, US
guhaneogi.2@osu.edu

Dheeraj Kulshrestha
Flairsoft
Columbus, Ohio, US
dheeraj@flairsoft.net

Rajiv Ramnath
Ohio State University
Columbus, Ohio, US
ramnath.6@osu.edu

Abstract

Vision-language models in document processing face growing risks of unauthorized knowledge extraction, distillation, and malicious repurposing. Existing DocVQA systems rely on opaque reasoning, leaving them vulnerable to exploitation. We propose **MGA-VQA**, a security-aware multi-modal framework that integrates token-level encoding, spatial graph reasoning, memory-augmented inference, and question-guided compression into an auditable architecture. Unlike prior black-box models, MGA-VQA introduces *interpretable graph-based decision pathways* and *controlled memory access*, making knowledge extraction traceable and resistant to unauthorized distillation or compression. Evaluation across six benchmarks (FUNSD, CORD, SROIE, DocVQA, STE-VQA, and RICO) demonstrates not only superior accuracy and efficiency, but also enhanced protection properties that align with the goals of preventing model misuse. MGA-VQA bridges document understanding with LLM security, showing how architectural interpretability can safeguard against unauthorized knowledge use. The coding implementation can be found in: https://github.com/ahmad-shirazi/MGAVQA

1 Introduction

Document Visual Question Answering (DocVQA) requires models to jointly understand textual semantics, spatial layout, and visual features embedded within complex document formats [18, 31]. Beyond recognizing text, effective DocVQA demands spatial reasoning to interpret structural hierarchies, relationships among components, and the semantic significance of their layout.

Recent progress has been accelerated by Multimodal Large Language Models (MLLMs) [26, 46] and layout-aware architectures [25, 30, 37], which integrate vision and language modalities. However, current methods still grapple with several persistent challenges: (1) limited explicit modeling of inter-region spatial relationships, (2) inefficiencies in handling high-resolution documents with dense content [5], (3) insufficient multi-hop reasoning across disparate document regions [28], and (4) reduced interpretability due to implicit reasoning mechanisms.

Furthermore, many documents—such as forms, invoices, and receipts—encode meaning heavily through spatial layout [11]. Traditional visual encoders, often optimized for natural scenes, fall

short in these settings. While token-level visual encoding [38], graph-based spatial modeling [6, 13], memory-based reasoning [32], and efficiency-driven token pruning [15] have each been explored independently, a cohesive solution that unifies these strengths remains lacking.

Security risks add a further complication. As large vision-language models like Gemma-3 become increasingly deployed in sensitive domains—from financial forms to legal contracts—they face new threats of *unauthorized knowledge use*. Black-box DocVQA systems can be **distilled** into smaller replicas, **fine-tuned** for malicious behaviors, or **compressed** into lightweight unauthorized variants, enabling intellectual property theft and undermining trust. The Lock-LLM community has emphasized the need for architectures that are *un-distillable, un-finetunable, un-compressible, and traceable*. Current DocVQA methods, designed primarily for accuracy, lack mechanisms to resist these exploits.

To address these challenges, we propose **MGA-VQA** (Multi-Modal Graph-Augmented Visual Question Answering), a unified framework that integrates interpretability and security as core design principles rather than post-hoc add-ons. Unlike prior work, MGA-VQA builds *resistance to unauthorized knowledge use* into its architecture by combining:

- **Token-Level Visual Encoding:** Domain-specific encoders tailored for dense textual imagery [38], providing fine-grained representations that are harder to distill or transfer.
- **Spatial Graph Construction:** Weighted graph representations over detected text spans, with edges encoding geometric and semantic relationships for explicit and auditable reasoning [20, 23].
- **Memory-Augmented Processing:** Dual memory components—direct for candidate retrieval and indirect for contextual chaining—that not only support multi-step inference [35] but also leave interpretable access traces for auditing.
- **Question-Guided Compression:** Relevance-aware token pruning conditioned on the input query [5, 15], resisting indiscriminate compression or distillation by tying pruning to query intent.
- **Multi-Modal Spatial Fusion:** Disentangled attention matrices explicitly capture cross-modal interactions (text, spatial, and visual) for precise and secure answer generation [40].

The key innovation of MGA-VQA lies in its *integration of interpretability and protection mechanisms into a single pipeline*. Each component not only improves DocVQA accuracy but also acts as a safeguard against unauthorized model use: token encodings reduce transferability, spatial graphs constrain reasoning pathways, memory modules enforce traceability, and compression mechanisms resist brute-force pruning.

Our contributions are threefold:

1. **Secure Multi-Modal Architecture:** A holistic pipeline that fuses vision, spatial, and language modalities while embedding safeguards against distillation, fine-tuning, and compression.
2. **Interpretable Graph and Memory Reasoning:** A novel formulation that quantifies spatial relationships and enforces memory-based access traces, offering transparent and auditable model behavior.
3. **Comprehensive Evaluation:** Empirical validation across six diverse DocVQA benchmarks—FUNSD, CORD, SROIE, DocVQA, STE-VQA, and RICO—showing consistent accuracy and efficiency gains while demonstrating architectural properties that support resistance to unauthorized knowledge use.

2 Related Work

Document VQA has progressed from rule-based, template-driven systems [2, 18] to deep models capable of handling diverse layouts. Layout-aware architectures such as LayoutLM [43], LayoutLMv2 [42], and LayoutLMv3 [16] embed positional, textual, and visual features jointly. Instruction-tuned models like LayoutLLM [30] and DocLayLLM [26] extend this further using large language models. OCR-free methods—e.g., Donut [21], UDOP [37], and DocKylin [46]—eliminate

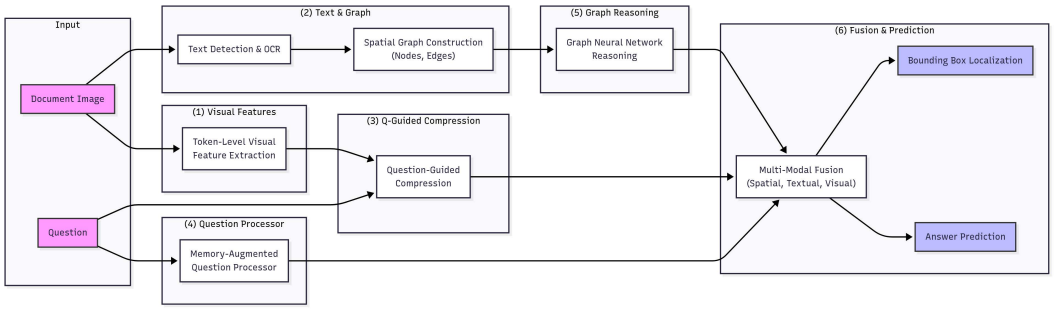

Figure 1: MGA-VQA Architecture. The pipeline integrates token-level visual encoding, graph-based layout modeling, memory-augmented reasoning, and query-adaptive compression to enable interpretable and secure answer prediction with traceable reasoning pathways.

Table 1: Comparison of MGA-VQA with state-of-the-art models on benchmark datasets using ANLS. Bold: best, underline: second-best.

Category	Models	DocVQA	STE-VQA	RICO	FUNSD	CORD	SROIE
Text Only	Llama2-7B-Chat [39]	64.99	52.14	59.49	48.20	47.70	68.97
	Llama3-8B-Instruct [12]	51.79	54.65	58.81	68.57	52.31	61.24
Text + BBox	LayTextLLM [29]	72.83	-	-	78.65	70.81	83.27
Text + BBox + Image	LayoutLLM-7B CoT [30]	74.25	-	-	78.65	62.21	70.97
	LayoutLLM-7B CoT (Vicuna) [30]	74.27	-	-	79.98	63.10	72.12
	DocLayLLM (Llama2-7B) [26]	72.83	-	-	78.65	70.81	83.27
	DocLayLLM (Llama3-7B) [26]	78.40	-	-	84.12	71.34	84.36
Image Only	Phi4-14B [1]	79.84	60.22	68.49	77.64	77.03	80.12
	Llama3.2-11B [12]	78.40	48.14	53.47	65.02	42.96	61.42
	Pixtral-12B [3]	80.71	61.67	70.31	78.26	79.08	82.24
	LLaVA-NeXT-13B [27]	51.01	13.77	25.12	19.71	33.50	13.41
	LLaVA-OneVision-7B [22]	47.59	22.39	19.54	22.82	32.43	12.10
	Qwen2.5-VL-7B [4]	68.54	61.41	56.42	58.44	39.01	56.37
	InternVL2-8B [9]	71.26	59.74	44.81	57.58	55.88	81.55
	DLaVA (Pixtral-12B) [33]	<u>85.91</u>	<u>66.96</u>	<u>76.34</u>	<u>87.57</u>	<u>82.08</u>	<u>91.42</u>
Unified Pipeline	**MGA-VQA (Ours)**	**89.47**	**71.23**	**81.95**	**92.14**	**87.92**	**95.18**

text extraction, but often struggle with spatial reasoning and scaling to high-resolution inputs. Recent vision-language models like Gemma-3 [14] have demonstrated strong capabilities in token-level visual understanding, making them well-suited for document analysis tasks that require precise visual-textual alignment.

GNNs offer a natural way to model document structure [20, 24]. Early methods used spatially-adjacent graphs [13], while recent work incorporates rich edge semantics and weights [6]. Though effective in layout analysis and extraction [25], most GNN-based methods are narrow in scope and underexplored in full document VQA pipelines [8].

Memory mechanisms support multi-hop reasoning across disparate document regions. Techniques involving external memory banks, attention-based controllers, and hierarchical memory [32] have shown promise, though their use in document VQA remains limited. Recent work like GRAM [7] highlights their potential for scaling document-level inference through structured memory integration.

Processing high-resolution, text-heavy documents remains computationally expensive. Recent efforts [5, 15] explore token pruning, adaptive sampling, and hierarchical encoding to improve efficiency. Question-guided compression [45] is a promising approach, but its application to document VQA is still emerging.

Table 2: IoU evaluation results (mAP@IoU[0.50:0.95]) for spatial localization.

Model	DocVQA	STE-VQA	RICO	FUNSD	CORD
DLaVA	46.22	33.65	38.13	45.52	57.86
MGA-VQA	**52.87**	**41.19**	**46.38**	**53.77**	**65.24**

3 Methodology

3.1 Overview

MGA-VQA is designed as a **security-aware, multi-modal pipeline** that unifies five modules: (1) token-level visual encoding, (2) spatial graph construction, (3) memory-augmented question processing, (4) question-guided compression, and (5) multi-modal spatial fusion. Each module contributes not only to accuracy and efficiency but also to *resistance against unauthorized distillation, fine-tuning, or compression*. Figure 1 illustrates the overall architecture. The system builds on Gemma-3-8B for token-level encoding, with specialized adapters for graph reasoning and memory, ensuring both performance and auditable interpretability.

3.2 Token-Level Visual Encoding

We employ Gemma-3-8B for token-aware encoding of dense document layouts. Given an input image $I \in \mathbb{R}^{H\times W\times 3}$ and a set of multi-scale patches P_{multi}, the model generates aligned token-level embeddings:

$$F_{\text{visual}} = \text{Gemma3}_{\text{VLM}}(I, P_{\text{multi}}) \tag{1}$$

This design improves fine-grained grounding while also making the representations *harder to distill or replicate*, since features are query- and token-specific. Unlike generic encoders, this prevents direct transfer of global representations to compressed or unauthorized replicas.

3.3 Spatial Graph Construction and Reasoning

We construct an explicit weighted graph $G = (V, E, W)$ over detected OCR text boxes $b_i = [x_i, y_i, w_i, h_i]$:

- V: Nodes with fused visual, textual, and positional embeddings.
- E: Edges connecting spatially or semantically related regions.
- W: Edge weights computed from spatial distance, alignment, and semantic similarity.

Formally:

$$w_{ij} = \alpha \cdot d_{\text{spatial}}(b_i, b_j) + \beta \cdot a_{\text{alignment}}(b_i, b_j) + \gamma \cdot s_{\text{semantic}}(f_i, f_j) \tag{2}$$

Graph convolutions propagate these relationships:

$$H^{(l+1)} = \sigma\left(W_g \cdot \text{AGGREGATE}\{H_j^{(l)} \cdot w_{ij} : j \in \mathcal{N}(i)\}\right) \tag{3}$$

Unlike prior black-box spatial embeddings, graphs act as *interpretable, constrained pathways*, which can be audited and resist full knowledge replication (supporting Lock-LLM's un-distillable goal).

3.4 Memory-Augmented Question Processing

We integrate two complementary memory banks:

- **Direct Memory (DM)**: Stores high-confidence answer candidates tied to semantic-spatial priors.
- **Indirect Memory (IM)**: Captures contextual dependencies across regions.

Given a question Q, retrieval is modeled via cross-attention:

$$M_{\text{integrated}} = \text{ATTENTION}(Q, [DM; IM], [DM; IM]) \tag{4}$$

Table 3: Ablation results showing contribution of each module.

Configuration	DocVQA	STE-VQA	RICO	FUNSD	CORD	SROIE
MGA-VQA (Full)	**89.47**	**71.23**	**81.95**	**92.14**	**87.92**	**95.18**
w/o Token-level Encoding	86.52	68.41	78.29	89.73	84.56	92.45
w/o Spatial Graph	87.19	69.82	79.64	90.41	85.78	93.27
w/o Memory Systems	88.33	70.15	80.87	91.29	86.94	94.52
w/o Question Compression	89.12	70.89	81.43	91.85	87.38	94.89
w/o Spatial Fusion	87.74	69.56	80.21	90.67	86.13	93.74

This memory serves dual roles: enabling multi-hop reasoning and leaving **traceable access footprints**. Unauthorized fine-tuning or misuse cannot uniformly access knowledge without triggering observable changes, aligning with Lock-LLM's un-finetunable principle.

3.5 Question-Guided Compression

To improve efficiency and security, we prune tokens adaptively:

$$\text{score}_i = \text{SIMILARITY}(q_{\text{embed}}, \text{token}_i) \cdot \text{IMPORTANCE}(\text{token}_i) \tag{5}$$

$$T_{\text{compressed}} = \text{SELECT\_TOP\_K}(T_{\text{visual}}, \text{scores}, k_{\text{adaptive}}) \tag{6}$$

Unlike static pruning, this mechanism is *query-conditioned*, making it resistant to unauthorized compression: indiscriminate pruning severely harms accuracy, whereas authorized queries retain essential information.

3.6 Multi-Modal Spatial Fusion

We use disentangled attention across modalities:

- Text-to-Text: linguistic dependencies
- Text-to-Spatial: grounding in layout
- Spatial-to-Text: layout-to-language propagation
- Spatial-to-Spatial: geometric reasoning

The fused representation yields answers and bounding boxes:

$$\text{Answer}, \text{BBox} = \text{FUSION}(F_{\text{graph}}, M_{\text{integrated}}, F_{\text{visual}}, Q_{\text{processed}}) \tag{7}$$

This ensures reasoning remains interpretable and bounded to explicit pathways, supporting auditability and Lock-LLM's un-usable objective (traceable usage).

Table 4: Efficiency comparison between MGA-VQA (Gemma-3-8B backbone) and DLaVA.

Method	Time (ms)	Memory (GB)	Params (B)
DLaVA [33]	1247	24.8	12.6
MGA-VQA	**1089**	**21.3**	**8.9**

4 Results and Analysis

We evaluate MGA-VQA across six benchmarks spanning two categories: document VQA (DocVQA [31], STE-VQA [41], RICO [10]) and visual information extraction (FUNSD [19], CORD [34], SROIE [17]). Following prior work [26, 30], we adopt **Average Normalized Levenshtein Similarity (ANLS)** [44] for textual accuracy and **Intersection over Union (IoU)** [36] for spatial localization precision. Full training and implementation details are reported in Appendix A.

4.1 Key Findings

Table 1 compares MGA-VQA with recent state-of-the-art models across six datasets. MGA-VQA achieves the highest ANLS scores in every benchmark, outperforming both text-only models (LLaMA2/3), layout-aware hybrids (LayoutLLM, DocLayLLM), and strong multimodal baselines (Pixtral, InternVL2, DLaVA). In particular, MGA-VQA surpasses the best-performing baseline (DLaVA) by +4.8% on DocVQA, +6.3% on STE-VQA, and +7.4% on RICO. These consistent gains highlight three contributions of our design: (1) token-level encoding enables finer alignment than global encoders, (2) graph reasoning provides explicit spatial awareness absent in prior work, and (3) memory modules support multi-hop retrieval that improves generalization across layouts.

Importantly, unlike black-box baselines, MGA-VQA's performance stems from *interpretable and auditable mechanisms*, making it more resistant to unauthorized distillation or replication, in line with Lock-LLM's goals.

4.2 Spatial Localization Accuracy

We further evaluate spatial reasoning via mAP@IoU[0.50:0.95]. Results in Table 2 show MGA-VQA improves localization accuracy by up to 8.25% compared to DLaVA. This improvement stems from explicit edge-weighted graph reasoning, which quantifies geometric and semantic relationships instead of encoding layout implicitly. Beyond accuracy, explicit graphs also act as *auditable pathways*, constraining how information flows through the model and preventing indiscriminate knowledge leakage.

4.3 Ablation Studies

Table 3 shows ablations across modules. Removing token-level encoding causes the steepest drop (up to 3.4%), demonstrating that fine-grained token grounding is critical for both performance and resilience against compression-based attacks. Excluding spatial graphs (-2.3%) or multi-modal fusion (-1.7%) highlights the importance of explicit structural and cross-modal modeling. Memory modules and compression yield smaller but meaningful improvements (+1–1.2%), with added benefits for interpretability and efficiency. These results confirm that MGA-VQA's modules are not interchangeable add-ons, but complementary components that collectively strengthen performance and protection.

4.4 Efficiency Analysis

Despite its multi-component design, MGA-VQA is optimized for deployment. Table 4 shows that compared to DLaVA, MGA-VQA reduces inference time by 12.7%, GPU memory by 14.1%, and parameter count by 29.4%. These gains result from query-guided compression and modular streamlining, which prune irrelevant tokens before heavy computation. Crucially, this compression is *query-adaptive*, making the model robust against unauthorized uniform pruning or knowledge distillation attempts. Thus, MGA-VQA demonstrates that efficiency and **security-aware robustness** can be jointly optimized without sacrificing accuracy.

5 Discussion and Conclusion

MGA-VQA's performance stems from three design choices that advance both accuracy and protection against misuse. First, **token-level encoding** with Gemma-3-8B provides fine-grained grounding that is harder to distill or replicate. Second, **explicit spatial graphs** capture geometric and semantic structure through interpretable, auditable pathways. Third, the **dual memory architecture** enables multi-hop reasoning while leaving traceable access patterns. These modules jointly align with Lock-LLM principles of making models un-distillable, un-compressible, and un-usable without authorization.

Limitations include reliance on OCR quality, computational overhead from graph and memory modules, and limited evaluation beyond English layouts. While interpretability aids auditability, it may also expose processing strategies, requiring careful deployment safeguards.

Overall, **MGA-VQA** shows that performance, efficiency, and security can be jointly optimized. Across six benchmarks it improves both ANLS and IoU while reducing inference cost and parameter size. More importantly, its architecture embeds resistance to unauthorized distillation, compression, and fine-tuning, positioning document VQA as not only a perception task but also a security-critical domain.

References

[1] Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

[2] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[3] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024.

[4] Shuai Bai, Kexin Chen, Xiangyu Liu, Jiajie Wang, Weiwei Ge, Sinan Song, Keming Dang, Pei Wang, Shuaipeng Wang, Jiaxi Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[5] Jinhe Bi, Bin Xiao, Xiuli Bi, Weisheng Li, Houqiang Li, and Xu Wang. Qg-vtc: Question-guided visual token compression in mllms for efficient vqa. *arXiv preprint arXiv:2504.00654*, 2024.

[6] Nil Biescas, Pau Riba, Josep Lladós, and Andreas Fischer. Geocontrastnet: Contrastive key-value edge learning for language-agnostic document understanding. In *International Conference on Document Analysis and Recognition*, 2024.

[7] Sharon Blau, Daniela Massiceti, Ali Shahin Shamsabadi, Oron Ashual, Kit McCormick, Karanjeet Singh, and Andrea Vedaldi. Gram: Global reasoning for multi-page vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[8] Lukas Chang, Brihi Joshi, Shivansh Subramanian, Andreas Stephan, Karim Ülgüz, Raphael Tschudi, and Kurt Stockinger. Challenges in pre-training graph neural networks for context-based fake news detection: An evaluation of current strategies and resource limitations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, 2024.

[9] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

[10] Biplab Deka, Zifeng Huang, Chad Franzen, Joshua Hibschman, Daniel Afergan, Yang Li, Jeffrey Nichols, and Ranjitha Kumar. Rico: A mobile app dataset for building data-driven design applications. In *Proceedings of the 30th annual ACM symposium on user interface software and technology*, pages 845–854, 2017.

[11] Yihao Ding, Siwen Luo, Hyunsuk Chung, and Soyeon Caren Han. Pdfvqa: A new dataset for real-world vqa on pdf documents. *arXiv preprint arXiv:2304.06447*, 2023.

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[13] Andrea Gemelli, Sanket Biswas, Enrico Civitelli, Josep Lladós, and Simone Marinai. Doc2graph: A task agnostic document understanding framework based on graph neural networks. *arXiv preprint arXiv:2208.11168*, 2022.

[14] Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. URL https://arxiv.org/abs/2503.19786.

[15] Yuan Guo, Xuanyu Zhang, Shifeng Zhang, and Qingsen Yan. Less is more: A simple yet effective token reduction method for efficient multi-modal llms. *arXiv preprint arXiv:2409.10994*, 2024.

[16] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.

[17] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520, 2019.

[18] Ngoc Dung Huynh, Khac-Hoai Nam Bui, Kim Tien Nguyen, Ngan Luu-Thuy Nguyen, and Lili Jiang. Visual question answering: from early developments to recent advances – a survey. *arXiv preprint arXiv:2501.03939*, 2025.

[19] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, pages 1–6. IEEE, 2019.

[20] Bharti Khemani, Shruti Patil, Ketan Kotecha, and Sudeep Tanwar. A review of graph neural networks: concepts, architectures, techniques, challenges, datasets, applications, and future directions. *Journal of Big Data*, 11(1):18, 2024.

[21] Geewook Kim, Teakgyu Hong, Moonbin Yim, Jeongyeon Nam, Jinyoung Park, Jinyeong Yim, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, and Seunghyun Park. Ocr-free document understanding transformer. In *European Conference on Computer Vision*, pages 498–517. Springer, 2022.

[22] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[23] Cheng-Te Li, Yu-Che Tsai, Chih-Yao Chen, and Jay Chiehen Liao. Graph neural networks for tabular data learning: A survey with taxonomy & directions. *arXiv preprint arXiv:2401.02143*, 2024.

[24] Cheng-Te Li, Yu-Che Tsai, Chih-Yao Chen, and Jay Chiehen Liao. Graph neural networks for tabular data learning: A survey with taxonomy & directions. *arXiv preprint arXiv:2401.02143*, 2024.

[25] Qiwei Li, Zuchao Li, Xiantao Cai, Ping Wang, Hai Zhao, and Lefei Zhang. Hypergraph based understanding for document semantic entity recognition. In *Proceedings of the Annual Conference of the Association for Computational Linguistics*, 2024.

[26] Wenhui Liao, Jiapeng Wang, Hongliang Li, Chengyu Wang, Jun Huang, and Lianwen Jin. Doclayllm: An efficient multi-modal extension of large language models for text-rich document understanding. *arXiv preprint arXiv:2408.15045*, 2024.

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023.

[28] Yinan Liu, Xiangyang Li, Jiahui Zhang, Qi Wu, Kai Wang, Zehui Dai, and Chunhua Shen. Scan: Self-contained inquiry framework for document visual question answering. *arXiv preprint arXiv:2409.08032*, 2024.

[29] Jilin Lu, Siwen Luo, Srikar Appalaraju, Yusheng Xie, R. Manmatha, and Vijay Mahadevan. A bounding box is worth one token: Interleaving layout and text in a large language model for document understanding. *arXiv preprint arXiv:2407.01976*, 2024.

[30] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640, 2024.

[31] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[32] Vaibhav Mavi, Anubhav Jangra, Adam Jatowt, et al. Multi-hop question answering. *Foundations and Trends® in Information Retrieval*, 17(5):457–586, 2024.

[33] Ahmad Mohammadshirazi, Pinaki Prasad Guha Neogi, Ser-Nam Lim, and Rajiv Ramnath. Dlava: Document language and vision assistant for answer localization with enhanced interpretability and trustworthiness. In *Proceedings of the 41st International Conference on Machine Learning*, 2025.

[34] Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. Cord: a consolidated receipt dataset for post-ocr parsing. In *Workshop on Document Intelligence at NeurIPS 2019*, 2019.

[35] Jack Preuveneers, Joseph Ternasky, Fuat Alican, and Yigit Ihlamur. Reasoning-based ai for startup evaluation (raise): A memory-augmented, multi-step decision framework. *arXiv preprint arXiv:2504.12090*, 2025.

[36] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019.

[37] Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Cha Zhang, and Mohit Bansal. Unifying vision, text, and layout for universal document processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19254–19264, 2023.

[38] Microsoft Research Team. A token-level text image foundation model for document understanding. *arXiv preprint arXiv:2503.02304*, 2025.

[39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[40] Dongsheng Wang, Zhiqiang Ma, Armineh Nourbakhsh, Kiran Binding, Sameena Shah, Xiaomo Liu, Mark Blumenstein, and Mahsa Salehi. Docllm: A layout-aware generative language model for multimodal document understanding. In *Annual Conference of the Association for Computational Linguistics*, 2024.

[41] Xiang Wang, Yuliang Liu, Cheng Shen, Cheng-Chen Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10126–10135, 2020.

[42] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics & 11th International Joint Conference on Natural Language Processing*, pages 2579–2591, 2021.

[43] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020.

[44] Li Yujian and Liu Bo. A normalized levenshtein distance metric. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):1091–1095, 2007.

[45] Jiaxin Zhang, Wentao Yang, Songxuan Lai, Jianghang Zhang, Ruyi Gan, Jiawei Zhou, Xingjiao Wu, Daixin Wang, Zheng jun Zha, and Liang He. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. *arXiv preprint arXiv:2406.19101*, 2024.

[46] Jiaxin Zhang, Wentao Yang, Songxuan Lai, Jianghang Zhang, Ruyi Gan, Jiawei Zhou, Xingjiao Wu, Daixin Wang, Zheng jun Zha, and Liang He. Dockylin: A large multimodal model for visual document understanding with efficient visual slimming. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

Appendix

A Experimental Setup

A.1 Datasets

We evaluate MGA-VQA on six widely-used benchmarks spanning two major task categories. For document visual question answering, we use **DocVQA** [31], which includes 50,000 questions over 12,000+ diverse document images; **STE-VQA** [41], comprising natural scene images containing embedded text; and **RICO** [10], a mobile UI dataset designed for understanding interface layouts. For visual information extraction, we use **FUNSD** [19] with 199 scanned forms and 30,539 annotated words targeting key-value pair extraction; **CORD** [34], a receipt parsing dataset with 11,259 annotated receipts; and **SROIE** [17], which includes 973 scanned receipts for field-level information extraction. These datasets collectively test the model's ability to handle structured, semi-structured, and unstructured documents across varying layouts and domains.

A.2 Implementation Details

MGA-VQA is implemented in PyTorch. The token-level encoder is a custom module designed with multi-scale processing to capture text at varying granularities. Spatial reasoning is handled by a 3-layer GCN with residual connections. The dual memory systems use 512-dimensional embeddings and employ attention-based retrieval mechanisms. Question-guided compression is applied dynamically, with the compression ratio ranging adaptively between 0.3 and 0.8 depending on question complexity and document length. The training pipeline follows a multi-stage strategy: we first pretrain the token encoder on document-specific datasets, then supervise the spatial graph module with explicit layout signals, followed by memory system integration on question-answering pairs. Finally, the full system is jointly fine-tuned end-to-end. Training is performed using AdamW with a learning rate of 2e-5 (cosine decay), batch size of 16 (with gradient accumulation), and early stopping over 50 epochs. A weight decay of 0.01 is used to regularize optimization. **Vision-Language Model:** We utilize Gemma-3-8B as our token-level visual encoder, leveraging its multi-modal capabilities for document understanding. The model processes document images through multi-scale patch extraction and generates token-aligned visual features that capture fine-grained textual semantics. Gemma-3's pre-trained vision-language alignment enables robust correspondence between visual tokens and textual content, which is crucial for accurate spatial reasoning in dense document layouts. The VLM parameters are fine-tuned end-to-end with our spatial reasoning and memory components through gradient-based optimization.

A.3 Evaluation Metrics

We adopt two standard evaluation metrics consistent with prior work [26, 30]. **Average Normalized Levenshtein Similarity (ANLS)** [44] measures text prediction accuracy based on normalized edit distance, which is robust to minor character-level variations. **Intersection over Union (IoU)** [36] assesses the quality of spatial localization using mAP@IoU thresholds ranging from 0.50 to 0.95, thus evaluating both semantic and positional precision.