ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-6874-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ICASSP49660.2025.10888359

SiQA: A Large Multi-Modal Question Answering Model for Structured Images Based on RAG

Jiawang Liu*, Ye Tao*, Fei Wang*, Hui Li*, and Xiugong Qin‡
*School of Information Science and Technology, Qingdao University of Science and Technology, Qingdao, China
Email: liujiawang@mails.qust.edu.cn, ye.tao@qust.edu.cn, wangfei@mails.qust.edu.cn, lipeilin1984xyz@163.com
‡Beijing Research Institute of Automation for Machinery Industry Co., Ltd., Beijing, China
Email: 13121990213@163.com

***Abstract*—Existing Large Multimodal Models (LMMs) demonstrate excellent performance in handling visual tasks in everyday scenarios. However, they still face challenges in understanding structured images, such as flowcharts and organizational charts, which are characterized by text-rich and complex hierarchical components. In this paper, we propose SiQA, a knowledge construction and Retrieval-Augmented Generation(RAG)-based multimodal Question-Answering model designed for Structured Images. SiQA operates in three stages: Knowledge Graph (KG) generation, retrieval-augmented, and answer generation. First, a KG representing the semantics of the structured images is generated through component analysis. We then performed similarity retrieval between the KG and queries, using a node-first algorithm to construct the most relevant subgraph. Finally, after performing an encoding alignment on the multimodal information, it is fed into the LLM to generate the answer. Additionally, we introduce a new dataset, OCQA[1], which includes 5,112 questions derived from 1,000 Organizational Charts. We evaluated SiQA's structured image detection and question-answering capabilities on the FD-DETR (a flowchart dataset) and SCQA, and verified its effectiveness and strong generalization ability through comparisons with existing state-of-the-art (SOTA) methods.**

***Index Terms*—multimodal, RAG, structured images, QA**

I. Introduction

The performance of Large Language Models(LLMs) across various Natural Language Processing(NLP) tasks has been well demonstrated [1]–[3]. Recently, attention has shifted towards LMMs, which integrate visual information into LLMs through visual encoders, enabling these models to understand both text and images simultaneously. Models such as BLIP2 [4] and GPT-4V [5] have demonstrated exceptional performance in visual tasks within everyday scenarios. However, the internet is filled with text-rich, structured images which differ from typical everyday scenes [6], [7].

To process image data, some existing LMMs directly use pre-trained visual encoders (such as VIT [8]) to encode images and then align these encodings with LLMs through a projection layer [9]. While this approach is convenient, it has limited ability to capture the global information required by LLMs. Some studies [10], [11] have focused on improving how LLMs receive visual information. For instance, BLIP2 developed a Q-Former that is fine-tuned alongside the LLM to better align visual features. However, traditional LMMs still struggle to understand text-rich structured images, and they may even produce hallucinations due to the inherent commonsense knowledge within the LLM. This is primarily because the answers to questions are often located in specific regions of the image and require reasoning about the image's structural relationships.

For question-answering on structured images, a direct approach [12]–[14] is to use a similar concept of Chain-of-Thought (COT) [15], [16] from LLM research, applying image COT engineering on top of a visual encoder to fine-grain locate the image regions that approximately contain the answer based on the input question, but this requires additional models to be trained separately on a large dataset of high-resolution image-text pairs [17]. Another approach is RAG [18]–[21]. In this method, a document is first generated for the target image using existing tools (such as Image Captioning or OCR). The most relevant passages are then retrieved based on the question, and the retrieved text is concatenated with the initial question as input to the LLM. However, traditional RAG methods fall short of enabling LLMs to comprehend the structural relationships within images fully.

To address these challenges, we propose a multimodal framework for question answering on structured images(SiQA), as illustrated in Fig. 1. First, in the SI2KG step, we designed a Separated Feature Cascaded Attention Block (SFCA-Block) tailored to the characteristics of various structured images, enhancing the efficiency of detecting structural components, and we developed a KG generator that creates a KG for each image, enabling the LLM to understand the internal structural relationships within the image. Subsequently, we designed a KG-RAG framework that retrieves the most relevant nodes and edges based on the question and constructs relevant subgraphs from the complete KG. Finally, the data from the three modalities—image, KG, and text—are encoded and aligned before being fed into the LLM to provide comprehensive information for answering the questions. Furthermore, we introduce a dataset called OCQA, which is designed to evaluate the model's multimodal question-answering capabilities. To the best of our knowledge, this is the first Chinese structured images question-answering

This work was supported in part by the National Key R&D Program of China under Grant 2023YFF0612102, and in part by the Key technology research and industrialization demonstration projects in Qingdao 24-1-2-qljh-19-gx.

[1]Dataset available at:https://github.com/a824705518/OCQA

dataset. Experimental results demonstrate that SiQA exhibits excellent performance in question answering on structured images.

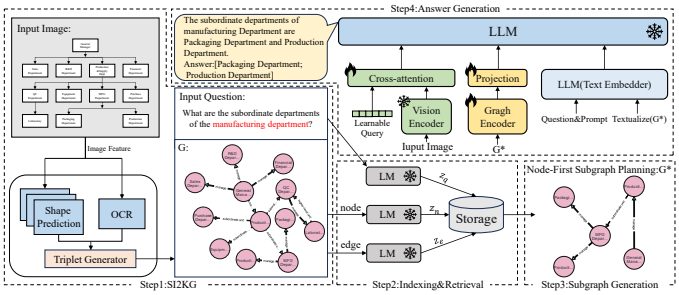

Fig. 1. Overview of the proposed SiQA model framework.

II. Method

A. SI2KG

There are various ways to represent structured images, such as flowcharts, organizational charts, and framework diagrams. To construct knowledge graphs that accurately capture the meaning of each type of structural diagram, we begin by predicting the components. This task is defined as an object detection problem, where we aim to predict the set of all shapes in the image, denoted as S. Each shape $s \in S$ includes a bounding box $s_b \in \mathbb{R}^4$ and a class label s_c. We base our component prediction task on RT-DETR [22], a model that has demonstrated its effectiveness across various detection tasks. However, considering the need to detect small and elongated elements in the input images, such as lines and arrows, as well as the overall execution efficiency of the SiQA model, we have made modifications to the RT-DETR's backbone network and its Attention-based Intra-scale Feature Interaction (AIFI) module to meet these requirements better. Specifically, we propose a Separated Feature Cascaded Attention Block (SFCA-Block) to replace the Bottleneck structure in ResNet [23], which is traditionally composed of stacked Convolutional Neural Networks (CNNs). The structure of the SFCA-Block is illustrated in Fig. 2. It is composed of a combination of CNN, Depth-Wise Convolutional Neural Network (DW-CNN), and Separated Feature Cascaded Attention(SFCA). Experimental results show that this structure retains the high-efficiency local feature learning capability of CNNs (targeting each component within the structural diagram) while also possessing the dynamic global feature learning ability of Transformers (targeting the global connectivity of components within the structural diagram). In the SFCA, each attention head processes different segments of the complete feature set, explicitly decoupling the computations between traditional attention heads. This optimization addresses the issue of redundant feature learning caused by head redundancy in the Multi-Head Self-Attention (MHSA) mechanism. Formally, SFCA can be expressed as:

$$\begin{aligned} X_{ij}^{'} &= X_{ij} + \widetilde{X}_{i(j-1)}, \\ \widetilde{X}_{ij} &= Att(X_{ij}^{'}Q_{ij}, X_{ij}^{'}K_{ij}, X_{ij}^{'}V_{ij}), \\ G_i &= Linear(Concat[\widetilde{X}_{i1}, \widetilde{X}_{i2}, \cdots, \widetilde{X}_{in}]). \end{aligned} \tag{1}$$

where, SFCA has n attention heads, where X_{ij} represents the feature obtained by the $j-th$ attention head after dividing the input feature $X_i \in \mathbb{R}^{C\times H\times W}$ into n parts. $\widetilde{X}_{ij}$ denotes the output of the $j-th$ attention head after applying the self-attention mechanism. G_i is the final output of the SFCA.

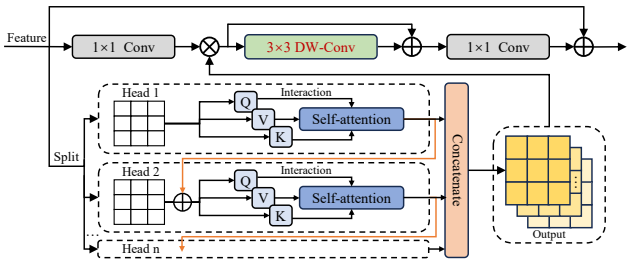

Fig. 2. Illustration of the SFCA-Block module.

RT-DETR selects the last three stages $\{s_3, s_4, s_5\}$ of the backbone network as inputs to the encoder, with the AIFI module in the encoder processing only the features from s_5. Similar to the SFCA-Block, we replace the MHSA in the AIFI with the SFCA, which is more suitable for this task. Additionally, we incorporate a Learned Positional Encoding, which is trained alongside the model, making the model more robust when detecting small and elongated components in structural diagrams.

For detecting and recognizing text blocks in structural images, we used the open-source tool PP-OCRv4 [24] and fine-tuned it on the OCQA dataset. Finally, we developed a graph generator designed to identify relationships between the various shapes based on component and OCR predictions and convert these shapes and relationships into triplets in the form (sub, pre, ob). The process begins by iterating through all entity boxes based on their positional information. A heuristic algorithm is then used to find all arrows connected to the current entity box and another entity. At this point, triples serve as a source of knowledge for the LLM to answer questions.

B. Indexing & Retrieval

Directly matching triples with a query based on similarity may lead to issues such as redundant knowledge retrieval and disconnected subgraphs, which can result in incorrect reasoning. To address this, when designing the KG-RAG framework, we used a Pre-trained Language Model (PLM) to separately encode the nodes and edges of the KG into z_n and z_e, storing them in adjacent memory spaces. We also used the same PLM to encode the question, creating a query index z_q. The retrieval task is modelled as a search problem. After generating the query vector, we use inner product similarity in Faiss [25] to search the node vectors z_n and edge vectors z_e stored in the adjacent spaces, ultimately returning the top-k nodes and edges by similarity ranking.

$$N = \underset{n\in G_n}{\arg\max}\, F(z_n, z_q) \in \mathbb{R}^k. \tag{2}$$

where, F represents the cosine similarity function. The $\arg\max$ function maps a set of corresponding entity nodes

from the $top{-}k$ similar vectors returned by the F. The formula for the edges E is similar.

C. Subgraph Generation

After retrieval, we obtain a set of entity nodes and a set of edges. Our goal is to generate a subgraph $G*$ from the KG constructed in sectionII-B, which is most relevant to the query. This step helps pre-filter out triples unrelated to the natural language query, as excessive irrelevant information can divert the focus of the LLM, potentially leading to incorrect answers. Additionally, reducing the number of input tokens significantly enhances the efficiency of both the graph encoder and the LLM. Therefore, we propose a Node-First Subgraph Planning algorithm to construct the subgraph most relevant to the question, as shown in Algorithm 1.

Algorithm 1 Node-First Subgraph Planning

Input: KG G, Node group N, Edge group E, Threshold η
Output: Subgraph $G*$

for $n_i \in N(n_i$ of confidence $c = \langle f(q) \cdot f(n_i) \rangle > \eta)$ **do**
  $\exists t \in T$ Such that $t = (n_a, e, n_b) \in G$ and $t \notin G*$ and ($n_i = n_a$ or $n_i = n_b$)
  for $t_i \in T$ **do**
    if $(n \neq n_i) \in N$ and $t_i \notin G*$ **then** ▷ n is the node of the triple t_i
      add t_i to $G*$
    else if $e \in E$ and $t_i \notin G*$ **then** ▷ e is the node of the triple t_i
      add t_i to $G*$
    end if
  end for
end for
for $e_i \in E(e_i$ of confidence $c = \langle f(q) \cdot f(e_i) \rangle > \eta)$ **do**
  if $t_e = (n_a, e_i, n_b) \notin G*$ and ($n_a \in G*$ or $n_b \in G*$) **then**
    add t_e to $G*$
  end if
end for

D. Answer Generation

To leverage the powerful reasoning capabilities of the LLM, we align the encoded information from three different modalities before feeding it into the LLM to infer the answer to the question. The encoding for each modality is as follows:

(1)Textual Embedding: We combined the prompt, query, and Textualize($G*$) according to a template, then encoded them using the LLM's pre-trained embedder to serve as the input text vectors.

(2)KG Embedding: To provide the LLM with complete subgraph structure information, we not only textualize $G*$ and include it as part of the text data embedding but also independently model $G*$ using a graph encoder, and finally aligned the graph embeddings with the LLM's vector space through a projection layer to obtain the graph embedding vector E_g. The formula is as follows:

$$\begin{aligned} \widetilde{E_{\mathrm{g}}} &= \mathrm{GCN}_{\partial_1}(\mathrm{G}*) \in \mathbb{R}^{d_g}, \\ E_g &= MLP_{\partial_2}(\widetilde{E_{\mathrm{g}}}) \in \mathbb{R}^{d}. \end{aligned} \tag{3}$$

where, we use a Graph Convolutional Network (GCN) [26] as the graph encoder, $\partial 1$ and $\partial 2$ are the learnable parameters of the graph encoder and the projection layer, respectively.

(3)Image Embedding: To enable the SiQA model to have rich visual soft prompts, we constructed a cross-attention module that includes a set of learnable vectors and a single-layer cross-attention mechanism. The cross-attention is randomly initialized and operates on the fixed-dimensional output features from the visual encoder, with the learnable vectors serving as the query vectors for the cross-attention. We kept the visual encoder frozen throughout the process and only trained the cross-attention module. The formula is as follows:

$$\begin{aligned} \widetilde{E_i} &= VIT(I) \in \mathbb{R}^{d_i}, \\ E_i &= CrossAtt_{\partial_q}(\widetilde{E_i}) \in \mathbb{R}^{256}. \end{aligned} \tag{4}$$

where d_i represents the output dimension of the VIT, and ∂_q represents the learnable vectors in the cross-attention module. We fixed the image feature dimension to 256.

Finally, we concatenate E_t, E_g and E_i and input them into the LLM's self-attention layer to generate the answer. When the LLM is frozen, the ∂_1 and ∂_2 parameters in the graph encoder and the ∂_q parameters in the cross-attention module are trained using standard backpropagation.

III. Experiment

Datasets. SCQA is a proposed dataset comprising 1,000 organizational charts from Chinese-listed companies and 5,112 questions requiring image comprehension for accurate answers. We manually annotated the Si2KG portion, while for the multimodal question-answering part, we adopted a semi-automated two-stage approach to generate useful questions and provide corresponding answers for each image, with a minimum of five questions per image. In the first stage, we used existing image understanding tools [5], [33] to generate initial questions and answers for each image. These were then reviewed and corrected by experts to address errors and inconsistencies. The dataset was split into training, validation, and test sets in a 6:2:2 ratio. FR-DETR [27] is a detection dataset comprising over 1,000 flowcharts, containing 20k line segment instances and 25k box-shaped symbols. We use this dataset to validate detection models' effectiveness and generalization capabilities.

Evaluation Metrics. In the Si2KG step, we use mAP@IoU (50 and 50-95) and F_1 score as the evaluation metrics for the detection model. For the natural language question answers generated by the multimodal LLM, we use Exact Match (EM) and F_1 score as the evaluation criteria.

Implementation Details. In the detection model, we use ResNet-50 and ResNet-101 [23] as backbone networks. Each dataset was trained for a total of 200 epochs, with the learning

rate initialized at 1e-4 and multiplied by 0.1 every 20 epochs. We used AdamW as the optimizer with a weight decay of 5e-4. In the indexing and retrieval steps, we use SentenceBERT [28] as the language model to encode both the natural language questions and the KG. During the answer generation step, we employ a GCN [26] to encode the KG. The visual encoder is a ViT [8], initialized with the pre-trained weights from VIT-bigG [29]. We use QWEN [30] as the LLM, and all experiments were conducted on 1 NVIDIA A100-PCIE-40GB GPU.

TABLE I
COMPARISON OF THE SIQA MODEL WITH SOTA METHODS FOR DETECTING STRUCTURED IMAGES. **Bold** INDICATES THE BEST PERFORMANCE, WHILE THE SECOND-BEST IS UNDERLINED.

Method	FR-DETR			OCQA		
	50	50-95	F_1	50	50-95	F_1
Arrow R-CNN [31]	84.9	64.8	87.8	86.1	66.7	90.2
FR-DETR$_{R101}$ [27]	89.3	69.2	93.5	91.2	70.3	92.7
RT-DETR$_{R101}$ [22]	**93.7**	75.2	95.1	94.1	74.8	95.9
DiagramNet [32]	90.6	72.3	92.8	92.7	72.0	93.4
SiQA$_{R50}$	92.8	74.5	94.7	93.8	74.3	95.7
SiQA$_{R101}$	93.2	**75.6**	**95.2**	**94.4**	**76.3**	**96.8**

TABLE II
COMPARISON OF THE PERFORMANCE OF SIQA AND OTHER MULTIMODAL MODELS ON THE OCQA TEST SET. **Bold** INDICATES THE BEST PERFORMANCE, WHILE THE SECOND-BEST IS UNDERLINED.

Method	Type	EM	F_1
BLIP2-FLanT5$_{XXL}$ [4]	MLLM	44.5	49.7
BLIVA-Vicuna$_{XXL}$ [10]	MLLM	50.2	55.6
GPT-4V [5]	MLLM	52.4	56.9
Qwen-VL$_P$ [33]+TextCoT [12]	MLLM+COT	51.9	55.4
Cantor$_{Gemini}$ [34]	MLLM+COT	53.7	58.2
VTQA [18]	MLLM+RAG	48.3	52.7
SiQA$_{7B}$	MLLM+KG-RAG	52.3	56.5
SiQA$_{14B}$	MLLM+KG-RAG	**55.9**	**59.2**

Results. Our method demonstrates significant improvements over existing SOTA approaches in detecting structural components in images and answering questions. Detecting structural components is fundamental to the SI2KG approach, as shown in TABLE I. Given our SFCA-Block, which considers both local details and global structure, SiQA$_{R101}$ outperforms RT-DETR [22] in tasks involving structured images, showing substantial improvements across both detection datasets. Additionally, we utilized prompt engineering to conduct question-answering experiments on OCQA with various existing LMMs and categorized their types, as shown in TABLE II. Given that we generated a KG representing the image's meaning and aligned information from three modalities as input to the LLM, SiQA's question-answering performance surpassed that of mainstream LMMs or LMM-enhanced methods. Compared to GPT-4V [5], SiQA$_{14B}$'s EM and F_1 scores improved by 3.5% and 2.3%, respectively.

Ablation study. To validate the effectiveness of RAG and multimodal information encoding in SiQA, we conducted ablation experiments on 7B, as shown in TABLE III. SuG represents the subgraph construction module, GE is the graph encoder module that transfers KG information to the LLM, and CrA refers to the cross-attention mechanism used to compress visual information and transfer it to the LLM. Experiments 1-3 demonstrate that constructing subgraphs and aligning graph encodings before inputting them into the LLM significantly aids the model in finding the basis for answering questions and dramatically reduces the occurrence of knowledge hallucinations. In Experiment 4, we replaced the cross-attention module with an MLP [35] to align the visual encoder's output with the large model's encoding. Based on the results of Experiment 5, relying solely on a pre-trained visual encoder and prompt engineering for question-answering on structured images does not yield satisfactory results.

TABLE III
ABLATION STUDY OF DIFFERENT MODULES IN THE SIQA MODEL ON THE OCQA DATASET.

Experiment	SuG	GE	GrA	EM	F_1
1	✓	×	✓	50.2	54.9
2	×	✓	✓	51.4	55.7
3	×	×	✓	49.5	54.0
4	✓	✓	×	50.8	55.2
5	✓	×	×	48.6	53.4

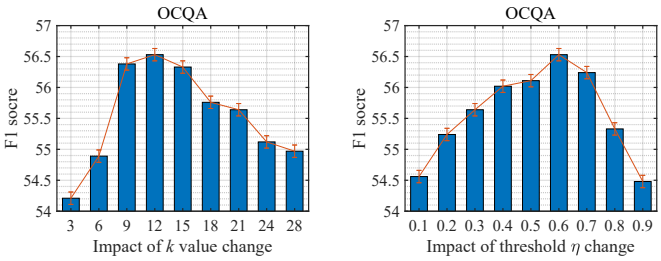

Fig. 3. Performance variations of SiQA on the OCQA test set with changes in the k and threshold η.

Analysis of top-k and threshold. We explored the critical parameter k during retrieval and the threshold η in Algorithm1 on 7B. In the SCQA dataset, each image generates an average of 28 triples using the Si2KG method, so the k value is set to be less than 28. We use a $one-fixed, one-variable$ approach to enhance experimental efficiency, with results shown in Fig. 3. Setting the k value excessively high can result in an overload of irrelevant information, potentially diverting the focus of the LLM. Conversely, setting the k value meager can lead to an incomplete KG, causing the LLM to fill in gaps on its own, which may result in hallucinations. Similar observations were made in experiments with the threshold η. Ultimately, we fixed k and η at 12 and 0.6, respectively.

IV. CONCLUSION

This paper proposes a novel large multimodal model, SiQA, designed for question-answering on structured images, and introduces a Chinese structured images question-answering dataset, SCQA. Experimental results demonstrate that each component of SiQA is both practical and efficient, outperforming current state-of-the-art methods on the FR-DETR and OCQA datasets. In the future, we plan to expand the dataset to include more diverse types of structured images.

REFERENCES

[1] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[2] Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang, "GLM: General language model pretraining with autoregressive blank infilling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland, May 2022, pp. 320–335, Association for Computational Linguistics.

[3] OpenAI, "Gpt-4 technical report," 2023.

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning*. 23–29 Jul 2023, vol. 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742, PMLR.

[5] OpenAI, "Gpt-4v(ision) system card.," 2023.

[6] Meixuan Qiao, Jun Wang, Junfu Xiang, Qiyu Hou, and Ruixuan Li, "Structure diagram recognition in financial announcements," in *Document Analysis and Recognition - ICDAR 2023*, Cham, 2023, pp. 20–44, Springer Nature Switzerland.

[7] Shreyanshu Bhushan and Minho Lee, "Block diagram-to-text: Understanding block diagram images by generating natural language descriptors," in *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*. Nov. 2022, pp. 153–168, Association for Computational Linguistics.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*. 2023, vol. 36, pp. 34892–34916, Curran Associates, Inc.

[10] Wenbo Hu, Yifan Xu, Yi Li, Weiyue Li, Zeyuan Chen, and Zhuowen Tu, "Bliva: A simple multimodal llm for better handling of text-rich visual questions," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, pp. 2256–2264, Mar. 2024.

[11] Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie, "Lion: Empowering multimodal large language model with dual-level visual knowledge," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 26540–26550.

[12] Bozhi Luan, Hao Feng, Hong Chen, Yonghui Wang, Wengang Zhou, and Houqiang Li, "Textcot: Zoom in for enhanced multimodal text-rich image understanding," *arXiv preprint arXiv:2404.09797*, 2024.

[13] Jiajin Tang, Ge Zheng, Jingyi Yu, and Sibei Yang, "Cotdet: Affordance knowledge prompting for task driven object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 3068–3078.

[14] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[15] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al., "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[16] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michał Podstawski, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler, "Graph of Thoughts: Solving Elaborate Problems with Large Language Models," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 16, pp. 17682–17690, Mar 2024.

[17] Zijun Long, George Killick, Richard McCreadie, and Gerardo Aragon Camarasa, "Multiway-adapter: Adapting multimodal large language models for scalable image-text retrieval," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6580–6584.

[18] Kang Chen and Xiangqian Wu, "Vtqa: Visual text question answering via entity alignment and cross-media reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 27218–27227.

[19] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim, "Ma-lmm: Memory-augmented large multimodal model for long-term video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 13504–13514.

[20] Pankaj Joshi, Aditya Gupta, Pankaj Kumar, and Manas Sisodia, "Robust multi model rag pipeline for documents containing text, table and images," in *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, 2024, pp. 993–999.

[21] He Zhu, Ren Togo, Takahiro Ogawa, and Miki Haseyama, "Interpretable visual question answering referring to outside knowledge," in *2023 IEEE International Conference on Image Processing (ICIP)*, 2023, pp. 2140–2144.

[22] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen, "Detrs beat yolos on real-time object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024, pp. 16965–16974.

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[24] Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al., "Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system," *arXiv preprint arXiv:2206.03001*, 2022.

[25] Jeff Johnson, Matthijs Douze, and Hervé Jégou, "Billion-scale similarity search with gpus," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[26] Thomas N Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[27] Lianshan Sun, Hanchao Du, and Tao Hou, "Fr-detr: End-to-end flowchart recognition with precision and robustness," *IEEE Access*, vol. 10, pp. 64292–64301, 2022.

[28] Nils Reimers and Iryna Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, Eds., Hong Kong, China, Nov. 2019, pp. 3982–3992, Association for Computational Linguistics.

[29] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev, "Reproducible scaling laws for contrastive language-image learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 2818–2829.

[30] Jinze Bai, Shuai Bai, and Yunfei Chu et al., "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[31] Bernhard Schäfer, Margret Keuper, and Heiner Stuckenschmidt, "Arrow r-cnn for handwritten diagram recognition," *International Journal on Document Analysis and Recognition (IJDAR)*, vol. 24, no. 1, pp. 3–17, 2021.

[32] Bernhard Schäfer and Heiner Stuckenschmidt, "Diagramnet: Hand-drawn diagram recognition using visual arrow-relation detection," in *Document Analysis and Recognition – ICDAR 2021*, Josep Lladós, Daniel Lopresti, and Seiichi Uchida, Eds., Cham, 2021, pp. 614–630, Springer International Publishing.

[33] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023.

[34] Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al., "Cantor: Inspiring multimodal chain-of-thought of mllm," *arXiv preprint arXiv:2404.16033*, 2024.

[35] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy, "Mlp-mixer: An all-mlp architecture for vision," in *Advances in Neural Information Processing Systems*, 2021, vol. 34, pp. 24261–24272.