Improving Multimodal Fusion Learning via Visual Knowledge Enrichment: A Study on Vietnamese Scene Text Visual Question Answering

Triet Minh Thai[*,1,3], **Nhan Duc Nguyen**[*,2,3], **Huy M. Le**[*,2,3], **Khang Gia Le**[2,3], **Ngan Tran-Thu Vo**[2,3] **and Son T. Luu** [1,3]

[1] Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Vietnam
[2] Faculty of Computer Science, University of Information Technology, Ho Chi Minh City, Vietnam
[3] Vietnam National University, Ho Chi Minh City, Vietnam
{19522397,21520373,20521394,21522189,21520069}@gm.uit.edu.vn
sonlt@.uit.edu.vn

Abstract

The VLSP - Visual Reading Comprehension for Vietnamese (VLSP-ViRC 2023) introduced a challenging dataset for the VQA task that includes the information pertaining to the scene text in the question. The dataset contains 9,129 images with more than 30K question-answer pairs. To solve the VQA challenge in this task, we enhance the multimodal fusion learning by enriching the visual knowledge including objects counting, image captioning, scene text recognition, and hints extractions, and integrating them into robust image and text encoders. Our proposed method achieved the 3^{rd} place in the VLSP-ViRC 2023 challenge. The results consolidate the advantages of visual knowledge enrichment in improving the multimodal fusion learning toward Vietnamese scene text VQA task.

1 Introduction

Visual Question Answering (VQA) is a multi-discipline task that combines Computer Vision (CV) and Natural Language Processing (NLP) [2] with the aim of acquiring the computer's ability to understand and extract valuable information from complex data such as text and pictures. The VQA task can be defined as giving an image as context, and the question in textual form. The purpose of the computer is to return the appropriate answer to the question.

However, when dealing with complex data structures like text and picture combination (we called this multimodal data), the unimodal is stuck in the representation of the semantic information because each individual unimodal representation is significantly different and it is hard to align these representations into a consistent form [39]. Multimodal Learning can solve the challenge of learning with multimodal data by utilizing the unimodal representation for each type of data first. Then, the model uses the fusion method to combine the representation of each unimodal with a suitable strategy before fitting to the learning algorithm for the downstream task. Multimodal fusion is a potential approach for the VQA task because it can leverage the knowledge from one modality, e.g., the text or image, to support the model trained on another modality. This can enhance the ability of the model in both visual and textual understanding for the VQA task.

In the upcoming VLSP 2023 challenge on Visual Reading Comprehension for Vietnamese (VLSP-ViRC 2023), a notable challenge surfaces-images may contain scene text, and questions might pertain to this textual information within the scenes. The organizer provides the OpenViVQA dataset [30], in which its questions and answers are open-ended. Moreover, the distribution of answer length shows the diversity and complication of answers of the OpenViVQA dataset, most of the answers have lengths that fall between 2 and 10. Fig 1 illustrates several examples from the dataset provided by the organizer.

In response to these challenges, our study unfolds with a focused approach, addressing the intricacies of Scene Text VQA and the complexities introduced by open-ended questions. The paper navigates through three primary contributions, each tailored to tackle these challenges:

- **Specialized Multimodal Fusion:** Fusion strategies are tailored to effectively handle the nuances of scene text integration within the open-ended VQA framework.

- **Visual Knowledge Enrichment:** Leveraging Visual Knowledge Enrichment (VKE), which encompasses Scene Text Recognition (STR), Visual Recognition (VR), and VQA hints, to

[*] These authors contributed equally to this work

enhance the performance of the multimodal fusion model.

- **Ablation Study on Dataset-specific Feature Enrichment:** Conducting a meticulous ablation study to dissect the impact of individual feature enrichment components, specifically tailored to the demands of OpenViVQA. This study aims to provide nuanced insights into the dataset-specific contributions of each component. Particularly, the STR component significantly enhances results, and its combination with other components further increases overall performance."

The paper is structured as follows. Section 1 introduces the task and summarizes our contributions. Section 2 takes a brief survey about previous works for the VQA task, especially general VQA and Vietnamese VQA. Section 3 describes our proposed solution. Section 4 overviews the OpenViVQA dataset and devotes to experiements and ablation studies. Finally, section 5 concludes our works and presents feature studies.

2 Related Work

Visual question answering (VQA) is a challenging task that requires a model to understand both the visual content of an image and the semantics of a natural language question. VQA aims to predict the correct answer to the question, given the image and the question.

2.1 VQA Datasets

In computer vision, the research purpose for VQA is to make computers understand the semantic context of images. The Microsoft COCO dataset [22] is one of the large-scale datasets that impact many studies in computer vision tasks, including object detection, image classification, image captioning, and visual question answering. Several VQA datasets are built on the MS-COCO in different languages, such as the VQA [2] in English, the Japanese VQA [37] for Japanese, and the ViVQA [43] for Vietnamese. There are also other two benchmark datasets for training and fine-tuning VQA methods, including Visual Genome (VG-QA) [20] and GQA [15]. VG-QA is a VQA dataset that contains real-world photographs. It is designed and constructed to emphasize the interactions and relationships between natural questions and particular regions on the images. The creation of VG-QA lays the groundwork for building GQA, another large VQA collection that make use of Visual Genome scene graph structures to feature compositional question answering and real-world reasoning. Besides, in the natural language processing field, the SQuAD dataset [34] has boosted many studies in question-answering and natural language understanding. Based on SQuAD, many corpora are created in different languages like DuReader [12] for Chinese and ViQuAD [18, 27] for Vietnamese.

2.2 General VQA approach

VQA methods include three main components: the external information embedding module, the multi-source information fusion module, and the answer classifier or answer generator module. The development of VQA methods concentrates on improving the external information embedding and multi-source information fusion modules.

Initial approaches used pre-trained image models such as ResNet [10] to extract features from images [17]. Anderson et al. [1] proposed the Bottom-up Top-down mechanism for the VQA task by using FasterRCNN [36] to extract region features from images. This way of feature extraction is effective as it reduces noises introduced by the regions of images that are not relevant to given questions. Jiang et al. [16] conducted experiments to prove that when using grid features as the input features of images for the attention-based deep learning method they can perform approximately the same performance as the Bottom-up Top-down attention mechanism. Recently there are pre-trained models that leverage the information extraction for VQA tasks such as OSCAR [21] or VinVL [52], such models enhance significant results of VQA methods on various datasets.

Former methods used pre-trained word embedding such as fastText [4] or GloVe [31] together with LSTM [13] network to extract linguistic features. Recent studies [50, 14] used large language models (LLM) like BERT to leverage the linguistic feature extraction and achieved positive results.

Besides the way of extracting information from multiple sources, fusing information is also a crucial part of output features that select or construct appropriate answers. Most VQA methods concentrated on the attention mechanism [25, 3] to perform information fusion, or in other words, they perform the attention mechanism to determine the correlation among multiple sources of information.

Question: cậu bé mặc áo cam đang làm gì? (what is the boy in orange doing?)
Answer: đang dắt cho đi dạo (walking a dog)

(a) Non-text QA

Question: có những ai đang đứng ở trong sạp hàng số 371? (who are standing at the kiosk number 371?)
Answer: có một người đàn ông và một người phụ nữ (a man and a woman)

Question: 20 000 đồng mua được bao nhiêu viên há cảo? (how many balls of har gow can be bought with 20 000 dong?)
Answer: năm viên há cảo (five balls of har gow)

(b) Text QAs

Figure 1: Typical examples of Non-text QA and Text QA.

According to the survey study [51], the attention strategy for information fusion can be divided into two categories based on the number of attention layers: single-hop attention and multi-hop attention. Previous works implemented single-hop attention such as [17, 24]. However, these methods did not obtain promising results while multi-hop attention methods such as [50] achieved better ones. These results showed that the attention module with a single layer can not model properly the reasoning ability required in VQA tasks. Recent studies enhance this viewpoint by performing co-attention mechanisms with multiple transformer layers [45] such as ViLBERT [23] or LXMERT [40].

2.3 State-of-the-art VQA approach

BEIT-3 [49]: BEIT-3 is known for its exceptional performance in cross-modal tasks. This model excels at understanding the relationship between images and text, making it a powerful choice for VQA tasks. Its bidirectional encoders and transformers allow it to capture rich contextual information from both text and images, resulting in accurate and context-aware answers to visual questions. BEIT's multimodal capabilities, large-scale pretraining, and adaptability through fine-tuning have solidified its position as a top choice in VQA research

PALI [7]: PALI is a multilingual language-image model designed for a range of languages, making it highly adaptable for global applications. Its joint language-image architecture enables it to process textual and visual information effectively, making it a versatile choice for VQA tasks across various languages. PALI's scalability, combined with fine-tuning options, allows it to excel in understanding and generating textual answers to visual questions. This model's cross-lingual capabilities and adaptability make it a top contender in the field of multilingual VQA.

2.4 VQA on Vietnamese Language

2.4.1 The Datasets

In 2021, Tran et al [43] launched the first Visual Question Answering (VQA) dataset in Vietnamese, known as ViVQA. This dataset, which drew from established English VQA benchmarks and COCO-QA [35], has since been widely used as a benchmark in related research. To populate the dataset, the team used a semi-automatic annotation system that translated English question-answer pairs into Vietnamese. The ViVQA dataset includes 10,328 images and 15,000 corresponding question-answer pairs, based on the images visual content. The data was randomly split into training and test sets at an 8:2 ratio. The dataset categorizes questions into four types: Object, Number, Color, and Location, which account for 41.55%, 14.81%, 20.82%, and 22.82% of total questions, respectively.

Given the shortcomings of machine translation and the lack of validation, the semi-automatic method proposed in ViVQA may not match human performance in translating an English VQA dataset to Vietnamese. As such, the ViVQA dataset cannot

Figure 2: Three largest Vietnamese ViVQA dataset. From top to bottom, left to right: VQA, EVJVQA and OpenViVQA.

be considered a reliable benchmark for constructing experiments, exploring, or evaluating VQA systems [30].

However, Nguyen et al. [28] have noted that the context of images taken in Vietnam presents more complexities than those from English VQA benchmarks. This is due to the crowded scenes and the presence of ”out-of-common” objects, specifically objects not typically encountered outside Vietnam. To address these issues and further the development of a VQA in Vietnamese, they created a new dataset comprising images manually collected from everyday life in Vietnam.

Furthermore, they have incorporated multilingual question-answer pairs into a dataset to stimulate and challenge the research community. This addition is designed to prompt the research community to develop an effective system capable of responding to questions penned in a variety of languages. The result is UIT-EVJVQA [28], the first trilingual Visual Question Answering dataset encompassing English, Vietnamese, and Japanese. This dataset features over 33,000 manually annotated question-answer pairs across approximately 5,000 images captured in Vietnam. The answers are derived from the corresponding image and input question. In addition to diverse question types, the answers are constructed in a free-form structure, posing a considerable challenge for VQA systems. To perform effectively on UIT-EVJVQA and yield satisfactory results, typical VQA systems must be capable of identifying and predicting correct free-form answers to multilingual questions, reflecting the unique characteristics of this dataset [41].

Besides, Nghia et al. [30] claims that most VQA datasets treat the VQA task as a classification task is not a reasonable or accurate reflection of human capabilities. Humans typically answer questions using various forms of natural language, including single words, phrases, or sentences, which a classification approach does not capture well. They redefined the VQA task as open-ended VQA, characterized by open-ended questions and answers to address these constraints. Based on this innovative definition, they introduced a Vietnamese open-ended VQA dataset, OpenViVQA. This dataset comprises 11,199 images and 37,914 manually annotated question-answer pairs. Unique to OpenViVQA, the included images were taken in Vietnam, providing a valuable resource for researchers to investigate and understand the differences between images captured in Vietnam and those from other regions outside Vietnam. This region-specific image set can stimulate the development of appropriately pre-trained models tailored to this dataset. Example of all three datasets in Figure 2

2.4.2 The Approaches

In 2022, Thai et al. [41] transformed the VQA problem into a sequence-to-sequence task. They utilize SOTA (state-of-the-art) vision-language models to provide a richer understanding of the dependencies between the question and image in the input sequence. The method is executed in two primary steps. During the initial phase, they extract miltitude hints from the question-image pairs

using pre-trained vision-language models. Subsequently, these extracted hints are combined with the question and visual features to create a sequence representation. This combined sequence representation is used as input for their proposed Seq2Seq model (ConvS2S), which generates corresponding answers in free-form natural language. The ConvS2S model utilizes 512 hidden units for both encoders and decoders. All embeddings, inclusive of the output generated by the decoder before the final linear layer, maintain a dimensionality of 768. This configuration enables the encoders to concatenate with patch embeddings derived from the ViT model.

In 2023, Nguyen et al. [29] put forth a novel attention scheme known as Parallel Attention. This scheme is a form of multi-hop attention and sets itself apart from recent methods. Further enhancing this, in order to take advantage of the linguistic features of Vietnamese, they equip Parallel Attention with a hierarchical feature extractor for questions. This gives rise to a novel method we term the Parallel Attention Transformer (PAT). Their experiments validate that this hierarchical extractor is indeed indispensable. This method comprises four primary components: the Hierarchical Linguistic Feature Extractor, the Image Embedding module, the Parallel Attention module, and the Answer Selector, details are described in [29].

Besides, Tran et al. [44] introduced BARTphoBEIT, which takes inspiration from two pre-trained models BARTpho and BEIT-2 [11]. In this paper, authors used BARTpho for tokenizing text data and used BEITv2 for image data. After that, the tokenized data will be passed through a Multi-Head Self Attention layer, then through the Language/Vision Forward Neural Network model, and then through another Multi Head Self Attention layer. Finally, through a Vision - Language Forward Neural Network, the model gets the desired output.

2.5 Modality Fusion

The VQA algorithms can be divided into three main phases: extracting image features, extracting question features, and integrating image and text features to produce answers. The third of these phases which is how to fuse image features with text features better has been a focal point in enhancing the model's ability to comprehend and respond to complex queries. Different methods have been proposed to perform modality fusion, such as concatenation, addition, multiplication, attention, or transformer-based models. However, most of these methods use modality-specific representations in their fusion modules instead of joint representation learning. This may limit the ability of the model to discover the underlying relation between both the image and question modality.

To address this issue, some recent works have proposed to use joint representation learning for modality fusion, such as multi-modal fusion transformer (MFT) [38], dynamic intra-modality attention flow (DyIntraMAF) [38], or visual BERT (VB) [8]. These methods leverage the power of transformer networks to learn cross-modal interactions and alignments between the image and question features. They also use self-attention mechanisms to capture the intra-modality dependencies within each modality.

2.5.1 Feature Enrichments

Another aspect that can affect the performance of modality fusion is the quality and diversity of the input features. Most of the existing VQA systems use pre-trained convolutional neural networks (CNNs) to extract visual features from the input image, such as ResNet [10] or Faster R-CNN [36]. However, these features may not be optimal or sufficient for answering complex questions that require high-level semantic understanding or reasoning. Therefore, some works have proposed to enrich the visual features with additional information, such as location context [46], scene text [53] [47], or discriminative correlation analysis (DCA) [26]. These methods aim to enhance the visual representation with more relevant and discriminative information that can help the model to answer the questions more accurately.

3 Proposed Method

Inspired by the approach of Thai et al. [41] on the VLSP-EVJVQA dataset, to produce a competitive result for the VLSP-ViVRC 2023 challenge, we proposed an efficient multimodal fusion method that leverages a Visual Knowledge Enrichment (VKE) module that captures the visual content and integrates it with the ViT5 [32] - a version of T5 model pre-trained on Vietnamese, to improve cross-modality learning. The overall multimodal fusion architecture for tackling the challenge is described in Figure 3.

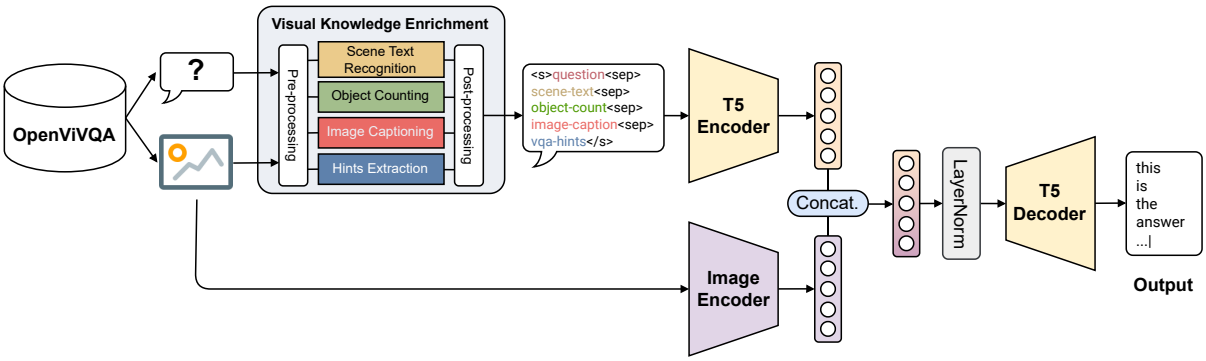

Figure 3: The overall architecture of the proposed method for VLSP2023-ViVRC challenge

3.1 Visual Knowledge Enrichment

The Visual Knowledge Enrichment (VKE) module proposed in this study takes advantage of various pre-trained vision-language models that serve three main functions: Scene-text Recognition, Visual Recognition, and VQA Hints Extraction to enrich the internal interaction between visual and textual features. As described in Figure 3, the question-image pair is first pass through an initial preprocessing block, in which proper transformation steps are performed separately for each modality. This is an important process, especially for Vietnamese data in order to adapt to the framework constraints that lack pre-trained availability for the language. For questions, the steps include noise removal, proofreading, and English translation. While the image also needs preprocessing steps such as resizing and transform based on each model requirements to produce an acceptable results. After this procedure, the pre-processed input is passed into each VKE components, which specifically described below, to extract the features of interest, namely the scene text, the counting of objects, image caption and VQA hints. Figure 4 presents two examples of VKE output for question-image pairs in each Non-text and Text QA scenarios after obtained and translated back to Vietnamese. Though there are number of mistakes occur in the output, such as the word "chung tay" in Text QA example, those newly obtained features together form the meaningful visual knowledge that is essential for our main model in the later stage.

3.1.1 Scene Text Recognition

- **EasyOCR:** EasyOCR is an open-source Python library that simplifies the process of extracting text from images. It boasts the ability to recognize text in over 80 languages, including English, Chinese, and Vietnamese. Built on the PaddlePaddle deep learning framework[1], EasyOCR[2] delivers accurate text recognition with easy implementation.

- **VietOCR:** The fusion of PaddleOCR[3] and VietOCR [4] technologies presents an innovative approach for extracting text from outdoor images. PaddleOCR utilizes a DL framework and CNNs, enabling proficient detection of diverse fonts and orientations in varied lighting conditions. Meanwhile, VietOCR, through its Transformer-based architecture, specializes in the recognition of Vietnamese script with high accuracy. This combination results in a dynamic method where PaddleOCR's robust detection capabilities are complemented by VietOCR's precise script recognition. The synergy of these models is particularly effective in complex outdoor environments, ensuring efficient image processing. They work together to detect text regions and reliably extract textual content accurately. This solution offers versatility and effectiveness in handling different types of texts and backgrounds. It demonstrates its applicability across a wide range of outdoor text extraction tasks. The approach underscores the potential of combining specialized OCR technologies for improved text extraction. This methodology is significant for its adaptability and relevance in the field of photo question-and-answer research.

[1] https://pypi.org/project/paddlepaddle/
[2] https://github.com/JaidedAI/EasyOCR
[3] https://github.com/PaddlePaddle/PaddleOCR
[4] https://github.com/pbcquoc/vietocr

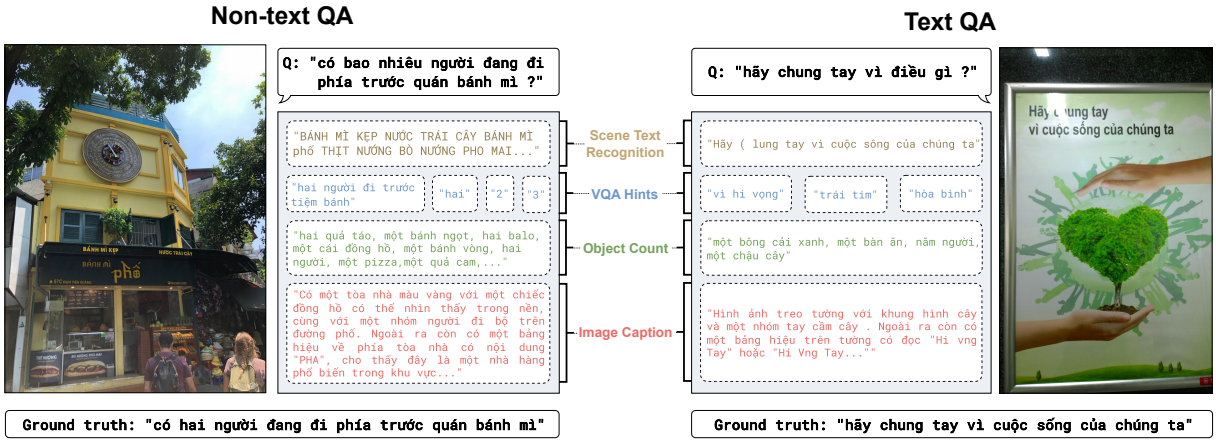

Figure 4: An illustration of extracted features from the VKE module including scene text, hints, object count, and image caption in Non-text QA and Text QA scenarios.

3.1.2 Visual Recognition

- **Image Captioning:** introduced at NeurIPS in 2023, InstructBLIP[5] marks a notable advancement in AI-driven image captioning. Its primary strength is in discerning and articulating complex inter-object relationships within images. Utilizing a Transformer-based architecture, InstructBLIP skillfully merges visual and textual data for a nuanced, context-aware captioning approach. This method is efficient, eschewing region-specific features and convolutional networks for a direct interpretation of visual elements. Consequently, InstructBLIP not only identifies objects but also crafts captions reflecting a deeper understanding of the scene. This innovative strategy yields captions that are accurate and contextually rich. InstructBLIP thus represents a significant step in narrowing the gap between AI capabilities and human-like perception in image understanding and description. This model's ability to generate insightful captions underscores its potential as a transformative tool in the field of vision-language models.

- **Object Recognition and Counting**: Detectron2[5] developed by Facebook AI Research, represents a leap forward in object recognition technology. Its flexible modular architecture utilizes Convolutional Neural Networks (CNNs) for proficiency in various object recognition tasks. The framework includes algorithms like Faster R-CNN and Mask R-CNN, adept at instance segmentation and object detection with high accuracy and efficiency. Notable for its rapid and precise image processing, Detectron2 is well-suited for real-world applications in autonomous driving, surveillance, and augmented reality. Being open-source and supported by an active community, it remains accessible and continually evolving. This positions Detectron2 as a leading tool in AI-driven object recognition.

3.1.3 VQA Hints Extraction

- **OFA:** OFA [48] is a unified multimodal framework that supports a comprehensive set of vision-and-language, vision-only, and language-only tasks. The architecture uses a Transformer-based with the Encoder-Decoder framework. Pre-trained on 20 million image-text pairs, OFA achieves state-of-the-art performance on various vision-and-language downstream tasks and demonstrates comparable performance to specialized pre-trained models on unimodal tasks.

- **ViLT:** introduced at ICML in 2021, ViLT [19] is designed to be particularly good at understanding the relationships between objects in the image. ViLT ingeniously leverages Transformer to extract and process visual features, this was achieved without the need for region features or convolutional visual embeddings, thereby ensuring inherent efficiency in terms of runtime and parameters. The design of the ViLT model treats the task as a classification task. Consequently, the model's output is a set of potential keyword answers, each associ-

[5] https://github.com/facebookresearch/detectron2

ated with a probability representing the confidence of the model in that particular answer. This unique approach of processing both visual and textual data together, combined with the model's ability to generate probabilistic outputs.

3.2 Multimodal Fusion Architecture

The nature of the VLSP2023-ViVRC dataset consists of long answers, in which scenario classification methods are not feasible. Due to this, we approach the challenge as a generative VQA task. The proposed multimodal fusion architecture comprises two main components, an image encoder to capture visual features from input image, which are then combined with textual data encoded in a language model to output the answer. This approach potentially leverages the power of language models for cross-modality learning and allows us to generate correct and precise answer of different length given textual input and image.

3.2.1 Image Encoder

Some of recent SOTA vision models are applied in this study including Vision Transformer (ViT), BEiTv2 and DeiT, which all have been pre-trained and fine-tuned on large image datasets, to extract visual features from the input image.

- **Vision Transformer (ViT):** Vision Transformer (ViT) [6] revolutionizes image processing by applying transformer architectures originally designed for natural language understanding to visual data. Instead of relying on conventional convolutional layers, ViT breaks images into fixed-size patches and linearly embeds them. This allows for more effective long-range interactions among pixels, enhancing the model's ability to capture global context. ViT's self-attention mechanism enables it to attend to all patches simultaneously, making it versatile for various image-related tasks. Its success underscores the adaptability of transformer models beyond text, shaping the landscape of modern computer vision.

- **BEiT-v2:** BEiT-v2 [11] is an enhanced version of the BEiT (Vision Transformer with Token Expansion) [9] architecture that further improves the concept of token expansion to achieve new heights in visual understanding. BEiT-v2's novel techniques enable a more sophisticated and context-aware representation of image content. At the heart of BEiT-v2 lies an enhanced token expansion mechanism that allows the model to capture intricate visual patterns with unprecedented precision. By adapting token sizes to local image complexity, BEiT-v2 allocates more computational resources to areas rich in detail. This dynamic approach to tokenization not only enhances the model's ability to discern fine-grained details but also fosters a deeper understanding of contextual relationships within the image. BEiT-v2's ability to capture the interplay between different image elements leads to a more comprehensive representation of the overall scene.

- **DeiT (Data-efficient Vision Transformer):** DeiT [42] addresses the data efficiency challenge in training large-scale vision transformers. Leveraging a novel distillation technique, DeiT achieves remarkable performance even with limited labeled data. It employs a teacher-student framework where the teacher, a larger transformer, imparts its knowledge to the student, a smaller model. This process allows DeiT to distill rich information from a pre-trained teacher, enabling effective transfer learning. DeiT's approach marks a significant stride in democratizing the use of powerful vision transformers, making them accessible for applications with resource constraints.

3.2.2 T5 Self-supervised learning and Multimodal Fusion

With a unified text-to-text Transformer-based framework, T5 [33] is one of the most prominent architecture in the field of Natural Language Processing (NLP) for language understanding tasks. It has achieved SOTA results on many benchmarks covering text summarization, question answering or text classification, and has undeniable potentials in the VQA task.

Our approach apply the pretrained version of this model on Vietnamese language - the ViT5 [32] . With Transformer-based encoder-decoder as backbone and T5-style self-supervised learning, ViT5 has been pre-trained on a large dataset that contains high quality Vietnamese texts, and achieves competitive results on two downstream text generation tasks namely Abstractive Text Summarization and Named Entity Recognition. In order for ViT5 to

capture the interaction between modalities during training, we perform a simple fusion method that concatenate the obtained features from image encoder with the last hidden states of ViT5 encoder along the sequence dimension. The combined features is then pass into a layer normalization module before feeding to ViT5 decoder for calculating loss.

4 Experiments

4.1 Exploratory Data Analysis

The VLSP-ViRC 2023 offers a comprehensive mix of images and question-answer (QA) pairs in Table 1. It comprises 9,129 images and 30,833 QA pairs in the training set, and 1,070 images and 3,545 QA pairs in the test set. Notably, the ratio of QA pairs to images is approximately 3.37 in the training set and 3.31 in the test set, suggesting that each image is associated with multiple QAs.

	Train set	**Dev set**
Overall		
Number of images	9,129	1,070
Number of QA pairs	30,833	3,545
No. QA/No. images	3.377	3.313
Question		
Question vocab size	6,133	2,073
Avg. question length	8.642	8.527
Std. question length	2.586	2,622
Min question length	3	3
Med. question length	8.0	8.0
Max question length	27	21
Answer		
Answer vocab size	9,317	2,963
Avg. answer length	5.768	6.041
Std. answer length	3.406	3.564
Min answer length	1	1
Med. answer length	5.0	6.0
Max answer length	48	30

Table 1: Statistical information of VLSP-ViRC 2023 Visual Reading Comprehension Dataset

In terms of language complexity, the dataset showcases a diverse vocabulary of over 6,000 unique words in the training set and more than 2,000 in the test set. The average question length stands at 8.64 words, while the average answer length is around 5.85 words. These statistics indicate the dataset's potential for training robust models capable of handling varied linguistic constructs.

Furthermore, the dataset presents an intriguing blend of QA types: those associated with images containing text and those without. Data have images with text like in Figure 5 promote the development of visual literacy skills, enabling model focus on various aspects of visual understanding, such as spatial reasoning, and object identification to interpret and extract meaning from visual content effectively. However, this type of data requires more effort to capture the key concept of questions because the model must rely solely on visual cues to extract meaning. On the other hand, in Figure 6, data with image with text will provide a richer and more comprehensive context because of the combination of text and image. With it, the necessity to incorporate a scene text task requires the model to interpret the visual context and comprehend and extract information from text embedded within images. Thus, the dataset offers a multifaceted challenge, driving the development of models capable of intricate visual and textual understanding.

4.2 Experiment Settings

To evaluate the performance of the proposed method toward VLSP2023-ViVRC challenge, we carry out different experiments of multimodal fusion models based on the presence VKE module. The final results of each experiment are then calculated using BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEU average (BLEU) and CIDer metrics.

We set up the same hyperparameters for all experiments to have the comparative result. The fusion model is trained in 3 epochs with batch size of 16 using Adam optimizer with initial learning rate of 2.5e-5, warm up ratio of 0.05 and weight decay 0.01. After each epoch, the performance loss on the train and development sets is calculated using the Cross-Entropy Loss function.

The VKE components and proposed architecture are implemented in Python 3.10.12, PyTorch framework and trained with hardware specifications: Intel(R) Xeon(R) CPU @ 2.00GHz; GPU Tesla P100 16 GB with CUDA 11.4.

4.3 Experimental Results

Table 2 presents the experimental results of our proposed approach on the development set (public test set) of VLSP2023-ViVRC challenge.

Without VKE, ViT5 achieves the initial CIDER value 2.7829 using only question. When incorporating the image information from pre-trained

Sample id: 14035
Image id: 3872
Question: người phụ nữ chèo ghe đang đội vật gì trên đầu ?
Groundtruth: người phụ nữ chèo ghe đang đội một chiếc nón lá trên đầu

Sample id: 12596
Image id: 3650
Question: người phụ nữ đội nón lá đang đứng trên thuyền để làm gì ?
Groundtruth: để chèo đò

Figure 5: Example of non-text QA case

Sample id: 16717
Image id: 504
Question: đoạn đường hay xảy ra gì ?
Groundtruth: xảy ra tai nạn

Sample id: 23621
Image id: 326
Question: trường nào có dạy mầm non ?
Groundtruth: trường pascal

Figure 6: Example of text QA case

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	BLEU	CIDEr
w/o VKE						
ViT5	0.4293	0.3761	0.3351	0.3023	0.3607	2.7829
ViT-ViT5	0.4399	0.3858	0.3438	0.3100	0.3699	2.8567
BEiTv2-ViT5	0.4399	0.3856	0.3437	0.3102	0.3699	2.8351
DeiT-ViT5	0.4377	0.3835	0.3413	0.3077	0.3675	2.8421
with VKE						
ViT5	0.5385	0.4817	0.4347	0.3964	0.4628	3.6492
ViT-ViT5	0.5424	**0.4854**	0.4381	0.3995	0.4664	3.7007
BEiTv2-ViT5	0.5407	0.4835	0.4363	0.3977	0.4646	3.6930
DeiT-ViT5	**0.5425**	**0.4854**	**0.4384**	**0.3997**	**0.4665**	**3.7037**

Table 2: Empirical Results of VQA models on the VLSP ViRC 2023 dataset

vision models, the VQA performance tend to improve. Based on CIDER, ViT-ViT5 fusion achieve the better performance with score of 2.8567. When adding VKE features to the input, the performance continues to improve. The DeiT-ViT5 outperforms others with BLEU score of 0.4665 and CIDEr score of 3.7037.

Based on the experimental results, DeiT-ViT5 fusion with VKE module has show a competitive performance and is chosen to give the final prediction in private test phase of the challenge. Officially, our method achieve the 3rd best performance compare to other teams's solution with CIDEr of 3.4121 and BLEU of 0.4457, as shown in Table 3. Our approach is able to achieve up to 3.5205 CIDEr and 0.4576 BLEU score on the private test set with some minor data correction. Though, this result is not yet submitted during the private test phase. The results consolidate the advantages of VKE module in improving the multimodal fusion learning toward Vietnamese scene text VQA task.

#	Teams Name	CIDEr	avg BLEU
1	ICNLP	3.6384	0.4663
2	linh	3.4293	0.4609
3	**DS@ViVRC (our)**	**3.4121**	**0.4457**
4	DS@UIT - Multimodal Team	3.3172	0.4742
5	NTQ Solution	3.2926	0.4876

Table 3: The final leaderboard results on the private test from the VLSP-ViRC 2023 shared task of the Top-5 participated teams

4.4 Ablation Study

In order to comprehensively assess the impact of the proposed Visual Knowledge Enrichment (VKE) module on the model's performance, we conducted an ablation study. This involved systematically removing different components of the VKE module and evaluating the model's performance on the OpenViVQA dataset [30]. The results of this ablation study are detailed in Table 4.

Scene Text Recognition		Visual Recognition		VQA Hints					BLEU	CIDEr	ΔBLEU	ΔCIDEr
EasyOCR	VietOCR	Object Count	Image Caption	OFA	ViLT@1	ViLT@3	ViLT@5	ViLT@10				
✓	✗	✗	✗	✗	✗	✗	✗	✗	0.4111	3.1792	0.0436	0.3371
✗	✓	✗	✗	✗	✗	✗	✗	✗	0.4182	3.2390	0.0507	0.3969
✓	✓	✗	✗	✗	✗	✗	✗	✗	0.4326	3.3766	0.0651	0.5345
✗	✗	✓	✗	✗	✗	✗	✗	✗	0.3739	2.8573	0.0064	0.0152
✗	✗	✗	✓	✗	✗	✗	✗	✗	0.3866	2.9857	0.0191	0.1436
✗	✗	✓	✓	✗	✗	✗	✗	✗	0.3876	3.0273	0.0201	0.1852
✗	✗	✗	✗	✓	✗	✗	✗	✗	0.3772	2.9045	0.0097	0.0624
✗	✗	✗	✗	✗	✓	✗	✗	✗	0.3760	2.9124	0.0085	0.0703
✗	✗	✗	✗	✗	✗	✓	✗	✗	0.3788	2.9476	0.0113	0.1055
✗	✗	✗	✗	✗	✗	✗	✓	✗	0.3799	2.9378	0.0124	0.0957
✗	✗	✗	✗	✗	✗	✗	✗	✓	0.3832	2.9406	0.0157	0.0985
✗	✗	✗	✗	✓	✗	✗	✗	✓	0.3881	3.0349	0.0206	0.1928
✓	✓	✓	✓	✓	✓	✓	✓	✓	**0.4665**	**3.7037**	0.0990	0.8616

Table 4: Experimental results of DeiT-ViT5 fusion model using different combination of VKE features

As observed in Table 4, the VKE module consistently enhances the model's performance across a spectrum of evaluation metrics. To delve into the nuanced impact of individual components, we systematically enabled each submodule to gauge their specific contributions.

- **Visual Recognition Component:** Enabling only the image description component of the VKE module results in a noteworthy improvement. There's a 0.1852 enhancement in CiDEr and a 0.0201 increase in the BLEU average, translating to values of 0.3876 for the average BLEU and 3.0273 for CiDEr. This highlights the pivotal role of image descriptions in refining the model's understanding and interpretative capabilities.
- **OFA and ViLT@10 Components:** When the OFA and ViLT@10 components are activated, we observe a modest yet discernible improvement. A 0.0206 increase in CiDEr and a 0.1928 boost in the average BLEU signify the added value these components bring to the model's holistic comprehension.
- **Scene Text Recognition Component:** The incorporation of the scene text recognition component marks a substantial leap in performance. Achieving 3.3766 in CiDEr and 0.4326 in the average BLEU, this component significantly enhances the model's ability to extract meaningful information from scene text, a crucial aspect in the context of Vietnamese VQA.
- **All Feature Enrichment Components:** Activating all feature enrichment components culminates in the model's peak performance. Advancements of 3.7037 in CiDEr and 0.4665 in the average BLEU underscore the synergistic effect of combining various visual knowledge sources, demonstrating the comprehensive capabilities of the complete VKE module.

5 Conclusion and future work

We employed a sophisticated multimodal fusion architecture for our solution in the VLSP-ViVRC task. This architecture incorporates an Image Encoder and a ViT5 Encoder-Decoder, seamlessly integrated with the Visual Knowledge Enrichment module. This module plays a pivotal role, serving three key functions: Scene-text Recognition, Visual Recognition, and VQA Hints Extraction. Together, these components synergize to enhance the overall performance of our proposed system. The final results reveal an CIDEr score of 3.5015 on the public test set and 3.4121 on the private test set. Additionally, our model achieved an average BLEU score of 0.4556 on the public test and 0.4457 on the private test. From the result, we placed the 3rd rank in the competition. In summary, several factors may exert significant impacts on our solution for the ViVRC task, including data pre-processing, the efficacy of Visual Knowledge Enrichment modules, the performance of image models like DeiT, and the generation capability of the core ViT5 model.

Our future research for this task is to explore advanced techniques such as constrastive learning between image and text, Optimal Transportation and the application of diffusion models. These avenues could potentially contribute to further en-

hancing the capabilities of our system in tackling the complexities of the ViVRC task. Investigating these methods holds the promise of advancing the state-of-the-art in VQA and may unveil some insights for future research endeavors.

Acknowledgments

We would like to thank and give special respect to VLSP organizers for providing the valuable dataset for this challenge.

References

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, , and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv*, 2014.

[4] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 2017.

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, 2023.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[7] Xi Chen et al. Pali: A jointly-scaled multilingual language-image model. In *ICLR*, 2023.

[8] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra- and inter-modality attention flow for visual question answering. *CVPR*, 2019.

[9] Jiaqing He, Zhe Hao, Heng Zhang, and Shaoqing Ren. Beit: Bert pre-training of image transformers. In *CVPR*, pages 2053–2062, 2021.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[11] Tongxin He, Hang Mao, Shiming Gan, Junwei Dai, Yiming Sun, and Yujie Wen. Beit-v2: Advancing visual transformer architectures with enhanced token expansion. *arXiv*, 2023.

[12] W. He, K. Liu, J. Liu, Y. Lyu, S. Zhao, X. Xiao, Y. Wang Y. Liu, H. Wu, Q. She, X. Liu, T. Wu, and H. Wang. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *ACL*, 2018.

[13] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, pages 1735–1780, 1997.

[14] R. Hu, A. Singh, T. Darrell, and M. Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, 2020.

[15] D. A. Hudson and C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019.

[16] H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen. In defense of grid features for visual question answering. In *CVPR*, 2020.

[17] V. Kazemi and A. Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv*, 2017.

[18] N. V. Kiet, T. Q. Son, N. T. Luan, H. V. Tin, L. T. Son, and N. L.-T. Ngan. Vlsp 2021 - vimrc challenge: Vietnamese machine reading comprehension. *VNU Journal of Science: Computer Science and Communication Engineering*, 2022.

[19] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

[20] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vision*, 2017.

[21] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[22] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, D. Ramanan P. Perona, P. Doll´ar, , and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014.

[23] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 2019.

[24] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image coattention for visual question answering. In *Advances in Neural Information Processing Systems*, 2016.

[25] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv*, 2015.

[26] Sruthy Manmadhan and Binsu Kovoor. *Optimal Image Feature Ranking and Fusion for Visual Question Answering*, pages 103–113. 2020.

[27] K. Nguyen, V. Nguyen, A. Nguyen, and N. Nguyen. A vietnamese dataset for evaluating machine reading comprehension. In *COLING*, 2020.

[28] Ngan Luu-Thuy Nguyen, Nghia Hieu Nguyen, Duong T. D. Vo, Khanh Quoc Tran, and Kiet Van Nguyen. VLSP2022-EVJVQA challenge: Multilingual visual question answering. *CoRR*, 2023.

[29] Nghia Hieu Nguyen and Kiet Van Nguyen. Pat: Parallel attention transformer for visual question answering in vietnamese. In *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6, 2023.

[30] Nghia Hieu Nguyen, Duong T. D. Vo, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. Openvivqa: Task, dataset, and multimodal fusion models for visual question answering in vietnamese. *Inf. Fusion*, 2023.

[31] J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

[32] Long Phan, Hieu Tran, Hieu Nguyen, and Trieu H. Trinh. ViT5: Pretrained text-to-text transformer for Vietnamese language generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 136–142, Hybrid: Seattle, Washington + Online, July 2022. Association for Computational Linguistics.

[33] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

[34] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP*, 2016.

[35] Mengye Ren, Ryan Kiros, and Richard S. Zemel. Exploring models and data for image question answering. In *NIPS*, 2015.

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks, 2015.

[37] N. Shimizu, N. Rong, and T. Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In *COLING*, 2018.

[38] Tim Siebert, Kai Norman Clasen, Mahdyar Ravanbakhsh, and Begüm Demir. Multi-modal fusion transformer for visual question answering in remote sensing. *Image and Signal Processing for Remote Sensing XXVIII*, 2022.

[39] William C. Sleeman, Rishabh Kapoor, and Preetam Ghosh. Multimodal classification: Current landscape, taxonomy and future directions. *ACM Comput. Surv.*, 2022.

[40] H. Tan and M. Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv*, 2019.

[41] Triet Minh Thai and Son T. Luu. Integrating image features with convolutional sequence-to-sequence network for multilingual visual question answering. *CoRR*, 2023.

[42] Hugo Touvron, Eric Cosatto, Matthieu Cord, and Jakob Verbeek. Data-efficient image transformer. In *CVPR*, pages 13628–13637. IEEE, 2021.

[43] K. Q. Tran, A. T. Nguyen, A. T.-H. Le, and K. V. Nguyen. Vivqa: Vietnamese visual question answering. In *PACLIC*, 2021.

[44] Khiem Vinh Tran, Kiet Van Nguyen, and Ngan Luu Thuy Nguyen. Bartphobeit: Pre-trained sequence-to-sequence and image transformers models for vietnamese visual question answering. In *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pages 1–6, 2023.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[46] Jiří Vyskočil and Lukáš Picek. Vinvl+l: Enriching visual representation with location context in vqa. In *Computer Vision Winter Workshop*, CEUR Workshop Proceedings, 2023.

[47] Jun Wang, Mingfei Gao, Yuqian Hu, Ramprasaath R. Selvaraju, Chetan Ramaiah, Ran Xu, Joseph F. JaJa, and Larry S. Davis. Tag: Boosting text-vqa via text-aware visual question-answer generation, 2022.

[48] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.

[49] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv*, 2022.

[50] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, 2019.

[51] D. Zhang, R. Cao, and S. Wu. Information fusion in visual question answering: A survey. *Information Fusion*, 2019.

[52] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021.

[53] Yongxin Zhu, Zhen Liu, Yukang Liang, Xin Li, Hao Liu, Changcun Bao, and Linli Xu. Locate then generate: Bridging vision and language with bounding box for scene-text vqa, 2023.