

# ViTextVQA: A Large-Scale Visual Question Answering Dataset and an Novel Multimodal Feature Fusion Method for Vietnamese Text Comprehension in Images

Quan Van Nguyen<sup>1,2</sup>, Dan Quang Tran<sup>1,2</sup>, Huy Quang Pham<sup>1,2</sup>,  
Thang Kien-Bao Nguyen<sup>1,2</sup>, Nghia Hieu Nguyen<sup>1,2</sup>,  
Kiet Van Nguyen<sup>1,2\*</sup>, Ngan Luu-Thuy Nguyen<sup>1,2</sup>

<sup>1</sup>Faculty of Information Science and Engineering, University of  
Information Technology, Ho Chi Minh City, Vietnam.

<sup>2</sup>Vietnam National University, Ho Chi Minh City, Vietnam.

\*Corresponding author(s). E-mail(s): [kietnv@uit.edu.vn](mailto:kietnv@uit.edu.vn);

Contributing authors: [21521333@gm.uit.edu.vn](mailto:21521333@gm.uit.edu.vn);

[21521917@gm.uit.edu.vn](mailto:21521917@gm.uit.edu.vn); [21522163@gm.uit.edu.vn](mailto:21522163@gm.uit.edu.vn);

[21521432@gm.uit.edu.vn](mailto:21521432@gm.uit.edu.vn); [nghiangh@uit.edu.vn](mailto:nghiangh@uit.edu.vn); [ngannlt@uit.edu.vn](mailto:ngannlt@uit.edu.vn);

## Abstract

Visual Question Answering (VQA) is a complicated task that requires the capability of simultaneously processing natural language and images. This task was initially researched with a focus on developing methods to help machines understand objects and scene contexts in images. However, some scene text that carries explicit information about the full content of the image is not mentioned. Along with the continuous development of the AI era, there have been many studies on the reading comprehension ability of VQA models in the world. Therefore, we introduce the first large-scale dataset in Vietnamese specializing in the ability to understand scene text, we call it ViTextVQA (**Vietnamese Text-based Visual Question Answering dataset**) which contains **over 16,000** images and **over 50,000** questions with answers. To tackle this task efficiently, we propose ViTextBLIP-2, an novel multimodal feature fusion Method, which optimizes Vietnamese OCR-based VQA by integrating a frozen Vision Transformer, Swin-TextSpotter OCR, and ViT5 LLM with a trainable Q-Former for multimodal feature fusion. Through experiments with various state-of-the-art models, we uncover the significance of the order in which tokens in OCR text are processed

and selected to formulate answers. This finding helped us significantly improve the performance of the baseline models on the ViTextVQA dataset. Our dataset is available \*for research purposes.

**Keywords:** Text VQA, Scene Text VQA, OCR Arrangement, Transformer, Vietnamese

## 1 Introduction

In recent years, Visual Question Answering (VQA) has garnered significant attention from researchers in Computer Vision (CV) and Natural Language Processing (NLP). The appeal of VQA has surged with the advent of powerful models like Flamingo [1], GPT-4 [2], and Gemini [3], which integrate advanced image-based question-answering capabilities. This has propelled VQA into a phase of robust global development, particularly in resource-rich languages such as English and Chinese. Numerous VQA datasets, including [4–7], have been released, accompanied by innovative methods showcasing remarkable performance, driven largely by neural and transformer architectures [8–10], which extend beyond traditional approaches.

In the realm of Vietnamese VQA research, Tran et al. [11] introduced ViVQA, the first dataset tailored for this task, comprising 15,000 samples derived semi-automatically from VQA v2. Despite its pioneering role, ViVQA faced scrutiny for its limitations, as later detailed by Nguyen et al. [12]. Addressing these issues, the OpenViVQA dataset emerged in 2023, designed for open-ended Vietnamese VQA with 11,199 images and 37,914 manually crafted question-answer (QA) pairs. This dataset marked a shift toward open-ended questions and answers, opening new research avenues. Subsequently, the ViOCRVQA dataset [13] scaled up efforts with nearly 30,000 images and over 120,000 QA pairs, emphasizing text-based question answering in images.

While ViVQA largely overlooks scene text, OpenViVQA incorporates it in about half its samples, and ViOCRVQA fully focuses on text-based VQA. However, these datasets still exhibit limitations. First, their sample sizes may not suffice for a comprehensive evaluation of scene text understanding. Second, the diversity of text in images might not fully capture real-world complexity. Lastly, OpenViVQA’s free-form answers complicate objective performance assessment, while ViOCRVQA, despite its scale, reveals gaps in addressing specialized needs.

To overcome these shortcomings, we introduce the **Vietnamese Text-based Visual Question Answering (ViTextVQA)** dataset, featuring 16,762 images and 50,342 QA pairs. ViTextVQA prioritizes extracting information from scene text to tackle challenges traditional VQA models face (see Figure 1). Beyond offering diversity, it lays a foundation for evaluating and enhancing VQA models adept at handling text, particularly in Vietnamese. To further advance this task, we propose **ViTextBLIP-2**, an extension of BLIP-2, optimized for Vietnamese text-based VQA by integrating pretrained vision and language models with a trainable Q-Former, as detailed in Section 4.

---

\*<https://github.com/minhquan6203/ViTetVQA-Dataset>



**Figure 1:** Samples extracted from the ViTextVQA dataset.

Our main contributions are described as follows:

- First, we create the first high-quality, large-scale dataset for text-based VQA in Vietnamese, emphasizing scene text in images.
- Then, we conduct analysis of ViTextVQA challenges through OCR system performance evaluation, underscoring its pivotal role in VQA models. Our findings reveal that sorting text from top-left to bottom-right markedly boosts performance.
- Finally, we propose ViTextBLIP-2, a novel method enhancing Vietnamese VQA efficiency, leveraging frozen pretrained components and a trainable Q-Former for multimodal integration.

This article is structured as follows: Section 2 reviews related studies. Section 3 details the ViTextVQA creation process and analysis. Section 4 our proposed ViTextBLIP-2. Section 5 presents baseline models and outlines experimental design and results. Section 6 delves into result analysis. Section 7 explores Vietnamese-specific traits. Finally, Section 8 concludes with future directions.

## 2 Related work

### 2.1 Visual Question Answering datasets

#### 2.1.1 Former Visual Question Answering Datasets

**Table 1:** Overview of the former VQA datasets in English.

Dataset	Type	Image Source	Annotation Method	Language
DAQUAR [14]	Non text	NYU-DepthV2	Human annotation	English
VQA v1 [15]	Non text	MS COCO	Human annotation	English
VQA v2 [16]	Non text	MS COCO	Human annotation	English
TextVQA [4]	Text	MS COCO	Human annotation	English
OCR-VQA [17]	Text	Book covers, movie covers	Semi-Automation	English
ST-VQA [5]	Text	Multiple sources	Human annotation	English
DocVQA [18]	Text	Scanned documents, receipts etc.	Human annotation	English
VisualMRC [19]	Text	Internet sources	Human annotation	English

The field of Visual Question Answering (VQA) has been significantly advanced by several large-scale, high-quality datasets in English, which serve as valuable resources and inspiration for our ViTextVQA dataset development. The DAQUAR dataset, introduced by Malinowski and Fritz [14], was an early VQA benchmark that initiated the Visual Turing Challenge, assessing models' abilities to process both image data and natural language questions. Antol et al. [15] revolutionized the field by proposing free-form questions in the VQA v1 dataset, attracting numerous researchers and methodologies such as MCB, MUTAN, and NMN. To address shortcomings in VQA v1, Goyal et al. [16] published VQA v2, enhancing the dataset with paired similar images leading to different answers for the same question. Recognizing the importance of scene text, Singh et al. [4] developed the TextVQA dataset with 45,336 questions requiring scene text as answers, while also introducing the LoRRA model incorporating OCR capabilities. Similarly, Mishra et al. [17] proposed the OCR-VQA task and created the OCR-VQA-200K dataset with 207,572 book cover images and over 1 million question-answer pairs. Biten et al. [5] addressed limitations in scene text understanding with the ST-VQA dataset, specifically designed for questions requiring knowledge of text present in images. Moving beyond natural scenes, Mathew et al. [18] introduced DocVQA, focusing on document images like invoices and contracts, with 50,000 question-answer pairs based on 12,767 document images. Finally, Tanaka et al. [19] advanced document-based VQA with VisualMRC, an abstractive task featuring over 30,000 question-answer pairs and 10,000 diverse document images, requiring models to generate natural language responses based on comprehensive document understanding.

### 2.1.2 Visual Question Answering Datasets in Vietnamese

**Table 2:** Overview of the former VQA datasets in Vietnamese.

Dataset	Type	Image Source	Annotation Method	Language
ViVQA [11]	Non text	MS COCO	Semi-Automation	Vietnamese
EJVQQA [20]	Non text	Internet sources	Human annotation	English, Japanese, Vietnamese
OpenViVQA [12]	Text & Non text	Internet sources	Human annotation	Vietnamese
ViOCRVQA [13]	Text	Internet sources - bookcover	Semi-Automation	Vietnamese
ViTextVQA(ours)	Text	Internet sources, hand-craft	Human annotation	Vietnamese

In developing ViTextVQA, we draw upon both English resources and existing Vietnamese VQA datasets to better understand Vietnamese language structure and cultural context. The ViVQA dataset by Tran et al. [11] represents the first Vietnamese VQA dataset, containing 10,320 images and 15,000 question-answer pairs, though its semi-automatic creation method results in questions that don't fully capture Vietnamese linguistic nuances and omits text elements in images. The EVJVQQA dataset introduced by Nguyen et al. [20] offers a unique multilingual collection of human-annotated visual question-answer pairs using images from Vietnam, comprising approximately 33,790 question-answer pairs and over 5,000 images, with 11,689 pairs in Vietnamese and the remainder translated into English and Japanese. This dataset served as the primary resource for the Multilingual Visual Question Answering task at VLSP 2022. Addressing the limitations of ViVQA, Nguyen et al. [12] created OpenViVQA, the first large-scale Vietnamese VQA dataset with open-ended answers, featuring 11,199 images with 37,914 question-answer pairs in natural language forms rather than as a classification task. Finally, the ViOCRVQA dataset by Pham et al. [13] focuses specifically on OCR-VQA, providing Vietnam's largest VQA dataset with nearly 30,000 images and over 120,000 question-answer pairs centered on scene text from Vietnamese book covers, created through semi-automatic labeling. Referencing these Vietnamese datasets helps us create questions and answers that accurately reflect Vietnam's cultural context and characteristics while overcoming existing limitations.

The existing Vietnamese VQA datasets have several common limitations. ViVQA lacks natural language fluency in its questions and does not include text in images, while EVJVQQA misses out on scene text, overlooking valuable OCR-related information. OpenViVQA, though designed for open-ended answers, faces challenges in classification tasks, and ViOCRVQA, despite being large and specialized, focuses primarily on text from book covers, lacking diversity in image types. These limitations highlight the need for more comprehensive, diverse, and culturally relevant datasets to improve the performance of VQA models in the Vietnamese context.

## 2.2 Visual Question Answering Methods

In the field of deep learning and the scope of this research, based on VQA models from previous research works, we classify them into three main groups: CNN-RNN, CNN-LM, and ViT-LM. In these groups, CNN-RNN is one of the early and important

developments of VQA. CNN-RNN models combine two types of networks: Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). CNN models are often used to extract features from images, while RNN models are used for natural language processing. When combined, the model is capable of understanding both forms of data and generating or classifying answers to image-based questions. However, using models based on CNN-RNN methods has some limitations, including the inability to effectively handle long questions or questions that are too complex. Therefore, to achieve higher performance and solve increasingly complex challenges, later research has expanded and developed ideas from CNN-RNN. Other branches, such as CNN-LM (combination of CNN and Transformer Language Model) and ViT-LM (combination of Vision Transformer Model and Transformer Language Model), have been proposed to improve the ability to understand and process diverse information from images and questions. However, both ViT-LM-based method and CNN-LM-based method require significant computational resources. For detailed computational resources and training times for models based on these two methods on our ViTextVQA dataset, refer to Section 5. The following Section presents related studies to these types:

### **2.2.1 Convolutional Neural Network - Recurrent Neural Network Based Method**

The deep learning method based on CNN-RNN gained prominence in the years before the advent of transformer architecture. To tackle VQA task, leveraging the strengths of Convolutional Neural Network (CNN) for image understanding and Recurrent Neural Network (RNN) to process sequential information in natural language. In the context of VQA, models that combine CNN with LSTM are employed to extract relevant visual features from the input image and textual features from the input question, respectively. Notable related works include VGGnet [21] and LSTM [22] in VIS+LSTM [23], or another powerful CNN architecture is GoogLeNet [24] and BoW [25] in SMem-VQA [26], or another very deep CNN architecture is ResNet [27] and GRU [28] in Up-Down [29], which have been proven effective in integration facilitates a holistic understanding of both visual and textual modalities, enabling the model to generate accurate answer based on the given image and question.

### **2.2.2 Convolutional Neural Network - Transformer Language Model Based Method**

With the emergence of transformer architecture, especially BERT [30], CNN-RNN based method no longer seems to be the optimal choice. Instead of using RNN, which is limited in capturing the context of the question, Transformer Language Model is becoming a new alternative. CNN-LM based method became attractive for VQA task, combining Convolutional Neural Network (CNN) for image processing and Transformer Language Model to better understand the meaning of the question. Models like ViLBERT [31], VisualBERT [32], Unicoder-VL [33], LXMERT [34], VL-BERT [35], UNITER [36], OSCAR [37] are good examples, they use Faster R-CNN [38] and BERT to embed images and questions respectively, but apply different fusion techniques and attention mechanism to improve model performance.

### 2.2.3 Vision Transformer Model - Transformer Language Model Based Method

Advances in transformers are not limited to the field of NLP but also open new doors with promising potential in the field of CV. Research by Dosovitskiy et al. [39] demonstrated that the Vision Transformer (ViT) architecture provides impressive results in feature extraction from images, opening up new and promising perspectives for the development of transformers pressure in this field. Models based on ViT and combined with the Transformer Language Model (ViT-LM) emerged as a cutting-edge approach to improve the performance of VQA models. Recent studies have applied this approach and achieved notable achievements, such as BLIP [40], BLIP-2 [41], GIT [42], mPlug [43], Flamingo [1], BeiT-3 [44], PaLI [45] deliver not only outstanding results but also open up new possibilities in practical applications. However, it is essential to note that this advancement often comes at a significant training cost. The learning process of these models requires extensive computational resources and long training time.

## 3 Task and Dataset

First, we define text-based visual question answering (text-based VQA) task (see Section 3.1). Then, we introduce a rigorous process of data annotation, involving the participation of sixteen native Vietnamese speakers. The process of building the dataset is presented in Section 3.2. The prominent and intriguing features of the dataset, which we have identified through the labeling process, are detailed in Section 3.3. This Section includes an in-depth analysis of elements such as Question, Answer, and Object.

### 3.1 Task Definition

The Text-based Visual Question Answering (Text-based VQA) task is a subdomain of Visual Question Answering (VQA), which lies at the intersection of Computer Vision (CV) and Natural Language Processing (NLP). The primary goal is to develop a model capable of answering questions based on information from images. Specifically, given an input consisting of an image and a related question, the model must generate an accurate answer by leveraging the visual data present in the image.

In the case of Text-based VQA, the requirement is more specific: the image must contain text, and the question-answering process must rely heavily on the textual content embedded within the image. The Text-based VQA task is defined as following.

- **Input of this task includes an image and a question.**

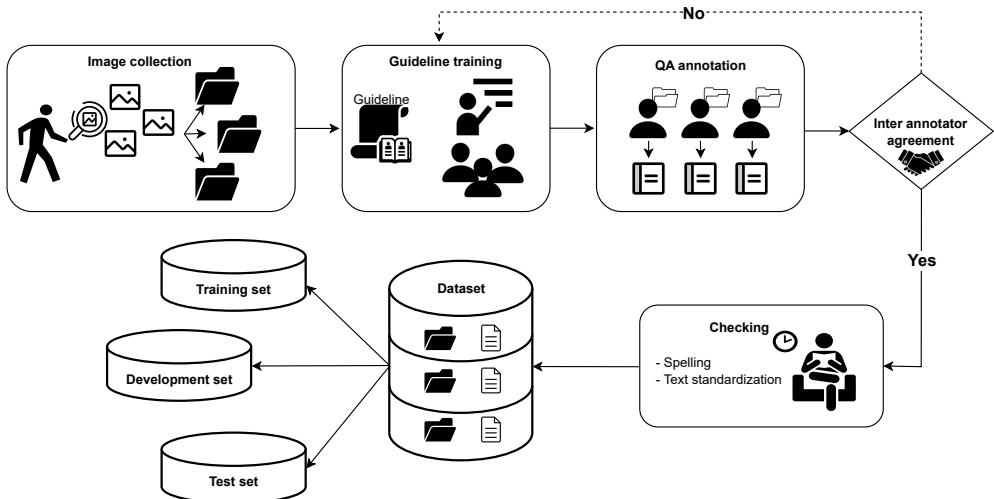
- **Image:** An image that contains the necessary information for the model to analyze in order to answer the question. The image can belong to various categories, such as landscapes, people, or street scenes, but the crucial condition is the presence of text within the image.

- **Question:** A question related to the content of the image. The question should be designed to align with the task’s purpose, which is to derive an answer grounded in the textual content present in the image.

- **Output of this task is a suitable answer to the question.**

- **Answer:** Given the input image and question, the model generates a corresponding answer by utilizing the textual information embedded in the image.

### 3.2 Dataset Creation



**Figure 2:** The overview process of creating our dataset ViTextVQA.

The process of creating the ViTextVQA dataset can be outlined as depicted in the diagram presented in Figure 2. After coming up with the idea, we started by collecting images on open platforms such as Google<sup>1</sup>, Facebook<sup>2</sup>, Pinterest<sup>3</sup>, etc. and taking photos manually somewhere near the area where we live. Then, pre-processing of the images is applied to remove unwanted images such as duplicates, images without text, etc., and select the images that have the most Vietnamese elements that we can perceive. After that, we divided these images into subsets to facilitate annotation. Once the image preparation process is finished, we then design a special guideline and tool and base it on the annotation process. We also use this guideline for training annotators to start the questions and answers (QAs) creation. The main idea of the guideline is to require the annotators to ask any question related to the image, and the answer must be the text that they see in the image. In the whole process, we hired sixteen ungraduated students for this work, and the Inter-Annotator Agreement was

---

<sup>1</sup><https://www.google.com/>

<sup>2</sup><https://www.facebook.com/>

<sup>3</sup><https://www.pinterest.com/>

measured twice. This is for evaluating whether the guideline is good or not as well as closely monitoring the annotators to ensure the dataset has the best quality. The next step involves checking and adjusting the dataset, which includes a thorough review of spelling errors and ensuring alignment with the desired quality standards. The final step of this process is to divide all samples into training, development, and test sets.

### 3.2.1 Image Collection

First of all, we created a list containing many keywords related to images of big cities in Vietnam such as Hanoi, Ho Chi Minh City, Da Nang, Hai Phong, Hoi An, Can Tho, etc. For example, “bảng hiệu ở Thành phố Hồ Chí Minh” (“signs in Ho Chi Minh City”), “khu du lịch ở Hội An” (“tourist areas in Hoi An”), “bảng chỉ dẫn đường ở Hà Nội” (“road signs in Hanoi”), “banner quảng cáo” (“advertising banners”), “bảng tên đường ở Cần Thơ” (“street signs in Can Tho”), and many more another keywords.

To collect image data, we employed Selenium<sup>4</sup>, a widely-used tool for web data crawling, to automatically search and retrieve images from Google Image and Pinterest through the keyword lists we built. We also conducted a manual collection from Facebook and Instagram. However, the performance was not as expected.

To bring more diversity and richness to our image resources, the decision was made to take photos in the neighborhoods where we live. This not only helps us control the quality of our data but also brings great diversity to our datasets.

By the end, we collected more than 40,000 bearing beauty and characteristic images in Vietnam. However, crawling this data cannot avoid noise and unwanted images. For duplicate images, we use the hash algorithm to remove them, but we still cannot completely resolve them because some images have different resolutions but are same image. Therefore, we encourage annotators to make a strong decision to remove images that they feel they have previously annotated. To filter images containing scene text, we used SwinTextSpotter [46] - one of the powerful OCR tools that supports Vietnamese.

Finally, we obtained more than 22,000 high-quality images. To facilitate the annotation process, we organized these images into many different subset folders, with each folder containing about 300 images.

### 3.2.2 Annotation Tool

To increase the convenience and efficiency of data annotation, we have developed a labeling support tool specifically for the ViTextVQA dataset named “VQALabeling”<sup>5</sup>. Utilizing the Qt framework and C++ programming language, we crafted a desktop application that operates seamlessly across multiple platforms including Windows, Ubuntu, and CentOS. This tool provides a user-friendly interface (see Figure 3), which is designed to facilitate efficient data management at the image folder level, ensuring organized and effective annotation for the ViTextVQA dataset. Upon completion of annotating a folder of images, the associated annotations information is stored in a \*.json file (see Figure 4) structured as follows:

---

<sup>4</sup><https://www.selenium.dev/>

<sup>5</sup><https://github.com/hieunghia-pat/VQALabeling.git>

- **annotations:** List of question-answer pairs of images. Each question-answer pair includes the corresponding question and answer.
- **delete:** During the image collection process, some images, although meeting size requirements, are not suitable due to being broken, blurred, or cannot be used to ask questions, etc. We use this field to mark whether an image is used in the official dataset or not.
- **filename:** Name of the image.
- **filepath:** The folder path containing the images.



**Figure 3:** User interface of VQALabeling.

```
{
  "annotations": [
    {
      "answer": "bà chiểu",
      "question": "khu chợ này có tên là gì?"
    },
    {
      "answer": "cán bộ - công nhân viên - tiểu thương chợ bà chiểu",
      "question": "ai xây dựng chợ vân minh thương nghiệp?"
    }
  ],
  "delete": false,
  "filename": "10019.jpg",
  "filepath": "st_images"
}
```

**Figure 4:** \*.json file storage structure.

### 3.2.3 Question-Answer Creation

After completing the image collection process, in the early stages of the annotation process, we developed a guideline for annotating. The goal of the guideline is to ensure the validity of question-answer pairs annotated with images. This is a difficult stage, requiring focus and effort because the guideline must be continuously updated throughout the annotation process. We pay special attention to handling unexpected situations and perform Inter-Annotator Agreement measurements periodically to ensure dataset quality. Furthermore, we utilize the “VQALabeling” outlined in Section 3.2.2 for creating question-answer pairs on different images.

#### General Rules

General Rules are established to ensure that the process of creating questions and answers for images is done carefully and systematically. This regulation helps ensure consistency and quality of data, from ensuring the number of questions and answers per image.

- **Rule 1:** For each subset, the annotator created at least 600 pairs of questions and answers.
- **Rule 2:** The annotator is encouraged to use their personal vocabulary to ask questions.

- **Rule 3:** Each image contains at most 10 pairs of questions and answers.
- **Rule 4:** All letters are in lowercase and in the scope of the standard keyboard.

## Rules for creating questions

The rules for creating questions provide specific instructions on how to ask questions so that they are specific, not vague, and ensure that the question leads to an answer that is already in the image. This way, not only does it create valuable questions, but it also creates convenience for participants to think and reason creatively and logically.

- **Rule 1:** Question must be answered by text that appears only in the image.
- **Rule 2:** Do not ask questions with mathematical implications.
- **Rule 3:** Do not ask where the text is located.
- **Rule 4:** Do not ask options (yes/no).
- **Rule 5:** Do not ask the color of the text.
- **Rule 6:** Do not ask the number of text that the annotator has to count.
- **Rule 7:** Avoid asking questions that are too short.
- **Rule 8:** Avoid asking multiple questions that have the same answer in the same image.
- **Rule 9:** Avoid asking general questions that lead to many different answers.
- **Rule 10:** Take the direction of the annotator seeing the image as the reference.

## Rules for creating answers

The rule for creating answers guides annotators to make sure that annotated answers are correct and do not contain extraneous information. This ensures the consistency of the dataset and makes the models easier to process.

- **Rule 1:** The answer is only the words that appear in the image.
- **Rule 2:** Keep the content intact, convert everything to lowercase.
- **Rule 3:** If the text has a line, replace it with a space.
- **Rule 4:** If the answer has some words that are obscured, only annotate what the annotator sees.

Next, we recruited 16 undergraduate students. Then, we trained them to understand the annotation guideline and asked them to annotate on about 100 test images. After ensuring that all students understood and applied the annotation guideline correctly, we appointed one person from our team for every four students to act as a supervisor. Annotation supervisors are responsible for monitoring and supporting the annotators, ensuring that everyone is performing according to the required process and quality.

During the annotation process, we also measured the Inter-Annotator Agreement which is mentioned in Section 3.2.4. If they meet the requirements, the data they annotated will be checked by annotation supervisors for spelling errors and standardized with the same font, punctuation, etc. If not, we will retrain them to follow the guideline to avoid unwanted errors and show them samples that need fixing.

Finally, after six weeks of non-stop work, we and our passionate annotators completed more than 50,000 samples. Then, all annotated data were randomly split into

training, development, and test sets corresponding to the ratio of 7:1:2 for future evaluation and analysis.

### 3.2.4 Inter-Annotator Agreement

When measuring the Inter-Annotator Agreement, our team, those who understand the guideline best, carefully annotate 100 question-answer pairs on about 30 images. Next, we hide the answers in these samples and ask annotators to re-annotate these hidden answers. This was done regularly every two weeks during the annotation process. Due to the nature of this VQA task, we are not aiming for a cheat answer like a multi-label classification task, so the dataset does not have a specific label space. That makes it impossible to measure the Inter-Annotator Agreement using traditional methods such as Cohen’s kappa [47], Fleiss’s kappa [48], and F1-score [49]. Therefore, to address this problem, we choose the F1-score commonly used in the Question Answering task as [50].

In this case, F1-score is computed over the individual words in the prediction(answer annotated by annotators) against those in the true answer (answer annotated by annotation supervisors). The number of shared words between the prediction and the truth is the basis of this metric. F1-score includes precision is the ratio of the number of shared words to the total number of words in the prediction, and recall is the ratio of the number of shared words to the total number of words in the ground truth. The F1-score formula is defined in Section 5.4.2. The requirement we set for annotators was an F1 score greater than 70%.

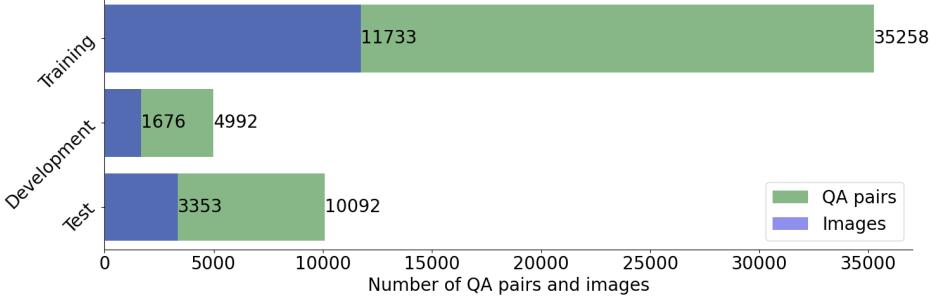
**Table 3:** Results of two rounds of Inter-Annotator Agreement measurements.

	First Time	Second Time
min F1-score (%)	73.91	83.36
max F1-score (%)	87.73	87.05
avg F1-score (%)	81.83	85.34

After measuring the Inter-Annotator Agreement through the F1-score twice, we simplify the results into Table 3. Based on the table, it can be seen that in both evaluations, the lowest F1-score value is relatively high (73.91 and 83.36), this shows that the level of uniformity between annotators is not low. In addition, the average F1-score also increased from the first to the second time, from 81.83 to 85.34, showing an improvement in consistency between annotators. This suggests that the annotated dataset has improved in quality over sessions, and this measurement can be considered reliable to confirm the quality of the dataset.

## 3.3 Dataset Analysis

The ViTextVQA dataset includes 16,762 images accompanied by 50,342 QA pairs. Figure 5 depicts the size of training, development and test sets in our dataset. This dataset offers a diverse range of visual scenes as well as scene texts and corresponding questions and answers. Each image is annotated with some question-answer pairs,



**Figure 5:** Statistic of images and question-answer pairs.

providing valuable insights into scene text and object required for accurate model predictions where the answer cannot be outside OCR text in the image and this makes the dataset more special in the VQA task. Moreover, we conducted statistics to compare our dataset with other common datasets in this field, and more details can be seen in Table 4.

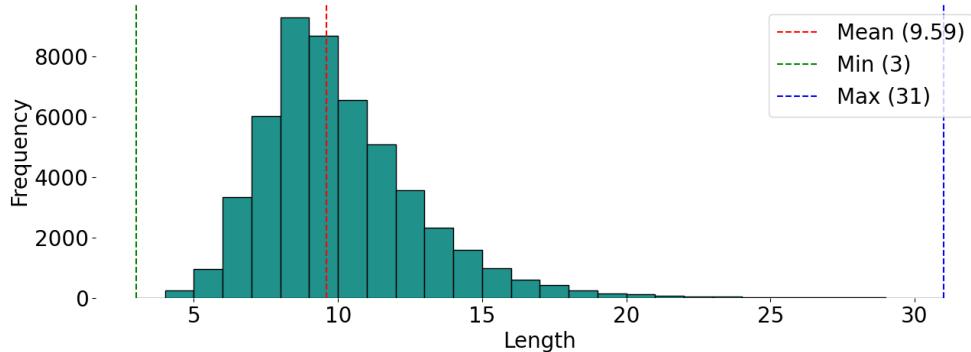
**Table 4:** Comparison between well-known VQA datasets.

Dataset	Language	Images	Questions	Answers
TextVQA [4]	English	28,408	45,336	453,360
ST-VQA [5]		23,038	31,791	31,791
DocVQA [18]		12,767	50, 000	50, 000
OCR-VQA-200k [17]		207,572	1,002,146	1,002,146
InfographicVQA [51]		5, 485	30,035	30,035
VisualMRC [19]		10,197	30,562	30,562
VizWiz [52]		32,842	265,420	265,420
OK-VQA [53]		14,031	14,055	14,055
UIT-EVJVQA [20]	Multilingual	4,879	33,790	33,790
MCVQA [54]		-	369,861	369,861
ViVQA [11]	Vietnamese	10,328	15,000	15,000
OpenViVQA [12]		11,199	21,271	21,271
ViCLEVR [55]		26,000	30,000	30,000
ViOCRVQA [13]		28,282	123,781	123,781
<b>ViTextVQA (Ours)</b>		<b>16,762</b>	<b>50,342</b>	<b>50,342</b>

### 3.3.1 Question Length

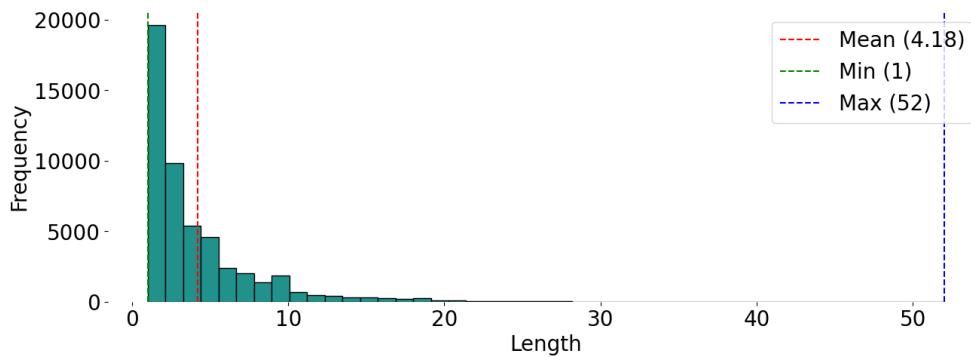
We performed a statistical examination of the question lengths in Figure 6. In the dataset, question lengths are not only concentrated in a fixed range but are distributed fairly wide range, from very short of only 3 tokens to very long of up to 31 tokens. This creates a considerable variety in terms of question length. The short questions are used often which aim to ask about basic information. On the contrary, there are also questions with considerable length, up to 31 tokens. These questions often involve complex issues, requiring a level of detailed information or multiple interrelated issues. On average, the question length was 9.59 tokens indicating that most questions were

not too short but not too long either. This may reflect the attempt of annotators to provide enough information without making the question too cumbersome.



**Figure 6:** Distribution of question length.

### 3.3.2 Answer Length

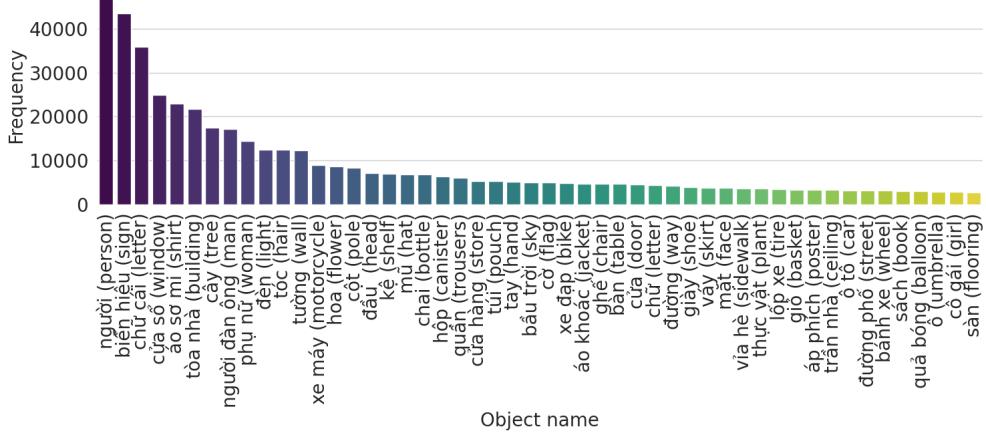


**Figure 7:** Distribution of answer length.

Figure 7 shows a large variation in the length of the answers in the dataset, from very short of only 1 token to very long of up to 52 tokens. This creates a significant difference in length between the answers. The average length of each answer is approximately 4.18 tokens, showing that the majority of answers are short. The reason for this phenomenon can be explained by the fact that annotators are constrained not to use tokens that do not appear in the image, leading to answers focusing mainly on from 1 to 3 tokens.

### 3.3.3 Object in Image

By Utilizing VinVL [56], we extracted various details about objects within the image, including location coordinates, object names, attributes, and more. This object information enhances our understanding of the visual content depicted in the image.



**Figure 8:** Distribution of the 50 most frequent objects in images.

Figure 8 shows that “person” ranks first with an occurrence count of over 47,000, showing a trend that most images in our dataset contain human presence. This deeply reflects the importance of humans in the context of the dataset. It can be said that human play a central and important role in the images we collect.

Additionally, “sign” and “letter” ranked second and third with about 43,000 and 36,000 appearances, respectively. This highlights the importance of signs and letters in the dataset, emphasizing our focus on scene text.

Our dataset not only reflects the diversity of people but also the diversity of landscapes, with objects such as “building”, “tree”, “sky”, “street”, and “sidewalk” appeared popular. Each of these objects has thousands of occurrences, creating a rich and diverse scene and highlighting deep and multidimensional images in the dataset.

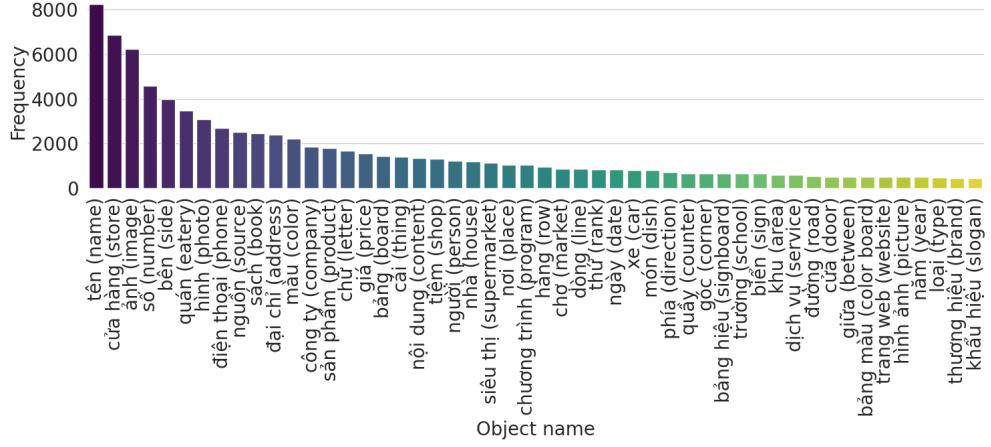
Besides, “shirt”, “trousers”, “hat”, and “jacket” are objects representing clothing and accessories that appear quite frequently, demonstrating the importance of this element in the dataset. This could be related to fashion, personal style, or the specific setting in the images.

In our dataset, a particularly striking and interesting point about the culture and characteristics of Vietnamese people is the popularity of motorbikes compared to cars. Specific Figure 8 shows that the presence of “motorbike” is significantly more than “car”. With nearly 9,000 occurrences, motorbikes far outnumber more than 3,000 occurrences of cars. This highlights a notable ratio, with motorbikes constituting approximately three times the frequency of cars in the dataset and it suggests that the priority and preference for using motorbikes in daily travel and in the transportation

system as well as partly reflecting the development of current Vietnamese culture and lifestyle.

### 3.3.4 Object in Question

We employed VNCoreNLP [57] to extract POS tags from questions, enabling us to compile statistics on the frequency of mentioned objects. This facilitates a deeper analysis of the questions, shedding light on the prevalence and significance of specific objects in questions of the dataset.



**Figure 9:** Distribution of the 50 most frequent objects in questions.

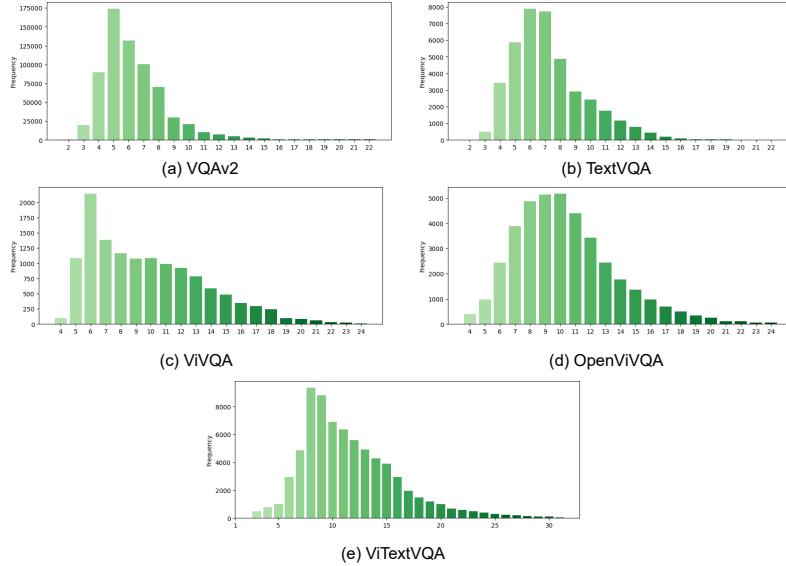
The detailed data in Figure 9 clearly illustrates the variety and richness of information that the annotator is interested in “store” and “name” appear especially frequently, more than 8,000 and nearly 7,000 times, respectively, highlighting their interest in specifically identifying the store and the names of the objects in the image.

Objects related to visual content such as “photo”, and “image” appear quite a lot, with a total of about 10,000 occurrences for both, showing special interest in with details of the image. At the same time, attention to information about “number” and “price” is also clearly shown through more than 4,500 and 1,500 appearances, respectively.

Culinary object group such as “eatery” and “dish” attract interest, especially with the desire to learn about the cuisine in the photo. At the same time, group of objects about people such as “person”, “man”, and “woman” also play an important role with a relatively high number of appearances.

Meanwhile, in addition to “store”, information about shopping and trading locations is shown through the significant appearance of “shop”, “supermarket” and “market”, highlighting the interest of the annotators with places in the image.

### 3.4 Comparison with Other Visual Question Answering Datasets

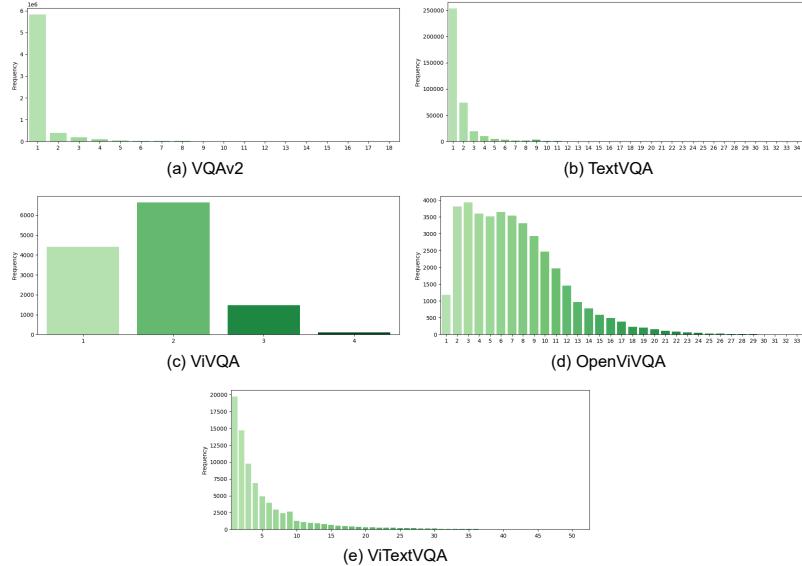


**Figure 10:** Comparison of question length among VQA datasets.

In order to contextualize our dataset within the broader landscape of Visual Question Answering (VQA), we compare it with several widely used VQA datasets based on key textual characteristics: question length and answer length. These attributes influence model performance, as shorter questions may be simpler to parse, while longer questions can provide more context or require complex reasoning. Similarly, answer length can affect model expressiveness and comprehension.

ViTextVQA is distinguished by its ability to accommodate substantially longer questions compared to existing datasets (see Figure 10), making it particularly suitable for tasks that require deeper reasoning and detailed query formulation. In contrast to VQAv2, TextVQA, and ViVQA, which predominantly feature shorter questions, ViTextVQA maintains a consistent distribution of questions exceeding 15 words. This characteristic enhances its applicability to complex visual question-answering (VQA) tasks that demand more comprehensive language understanding. While OpenViVQA provides a broader range of question lengths, ViTextVQA further extends this diversity by sustaining a high frequency of longer queries, thereby facilitating more intricate reasoning processes.

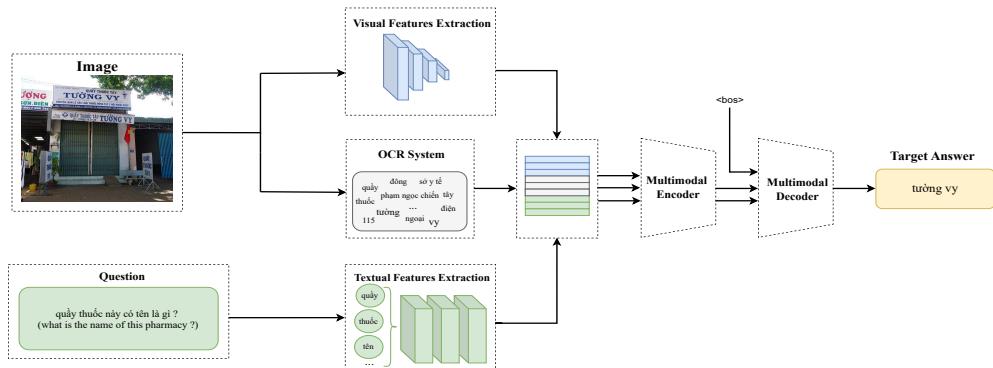
In terms of answer length, ViTextVQA allows for substantially longer responses compared to VQAv2, ViVQA, and TextVQA, providing greater flexibility for VQA tasks (see Figure 11). While OpenViVQA maintains a more balanced answer length distribution, ViTextVQA surpasses it by supporting answers of up to 50 words, enabling more detailed and comprehensive textual responses. This capability reinforces



**Figure 11:** Comparison of answer length among VQA datasets.

ViTextVQA as a valuable resource for advancing research in VQA, particularly for tasks that require complex reasoning based on textual information extracted from images.

## 4 Methodology

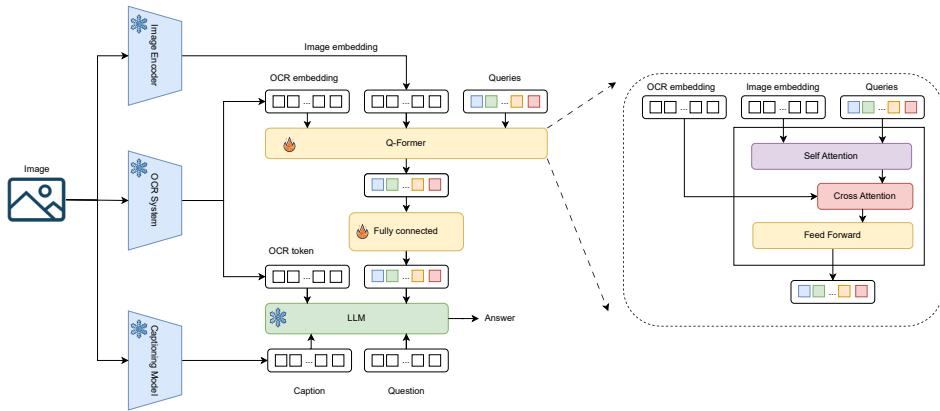


**Figure 12:** The overview of text-based VQA models structure.

Our text-based VQA method employs three key components to tackle image-based questions: Visual Features Extraction, Textual Features Extraction, and OCR System

(see Figure 12). In this method, the Visual Features Extraction extracts vital information from images, while the Textual Features Extraction encapsulates questions into features vectors. Meanwhile, OCR System undertakes the crucial task of recognizing and extracting text portrayed in images. Through the seamless integration of these three components, the text-based VQA method can comprehend and answer questions from images with enhanced accuracy and efficiency.

We propose ViTextBLIP-2 is an extension of BLIP-2 (see Figure 13), designed to overcome computational and data limitations in Vietnamese OCR-based Visual Question Answering (OCR-VQA). Instead of training from scratch, the model focuses on the second stage of BLIP-2, optimizing language generation by leveraging a frozen Large Language Model (LLM). To enhance Vietnamese text comprehension, ViTextBLIP-2 integrates additional modules, including an OCR system and an image captioning model. During training, only the Q-Former is optimized, while other components remain frozen to reduce computational costs.



**Figure 13:** The structure of ViTextBLIP-2.

**Image Encoder:** A pretrained Vision Transformer (ViT) is employed and kept frozen to retain its learned visual representations.

$$E_{\text{img}} = \text{ViT}(I) \quad (1)$$

where  $E_{\text{img}}$  is the image embedding, and  $I$  is the input image.

**OCR System:** The model incorporates SwinTextSpotter, maintaining its original architecture and parameters. The extracted text tokens and embeddings are directly passed to Q-Former.

$$\{T_{\text{OCR}}, E_{\text{OCR}}\} = \text{SwinTextSpotter}(I) \quad (2)$$

where  $T_{\text{OCR}}$  represents OCR tokens and  $E_{\text{OCR}}$  denotes OCR embeddings.

**Captioning Model:** The Qwen2-VL model is utilized to generate automatic image captions, providing additional contextual information. This module remains frozen during training.

$$C = \text{Qwen2VL}(I) \quad (3)$$

where  $C$  represents the generated caption.

**Q-Former:** This component is trained to integrate multimodal features from the Image Encoder, OCR System, and Captioning Model using Self-Attention and Cross-Attention mechanisms.

$$Q = \text{Q-Former}(E_{\text{img}}, E_{\text{OCR}}, Q_{\text{init}}) \quad (4)$$

**Large Language Model (LLM):** The ViT5 Phan et al. [58] model, a pretrained Vietnamese LLM, is incorporated as the language generation component. Its parameters are kept frozen, while the system optimizes its capability to generate responses based on the processed multimodal embeddings.

$$A = \text{ViT5}(Q, Q_{\text{ques}}, T_{\text{OCR}}, C) \quad (5)$$

where  $A$  is the generated answer, and  $Q_{\text{ques}}$  is the input question.

#### 4.1 Textual Features Extraction

To achieve a better match with the Vietnamese language and simultaneously enhance performance on the dataset, we opted for pre-trained language models to extract textual features on questions and OCR text for baseline models. These pre-trained models not only capture the nuances of the language but also enhance the capacity to process and comprehend text deeply within the Vietnamese context. Refer the Table 5 mentioned in Section 5.1 to see specific pre-trained language models utilized by baseline models. Pre-trained language models that we chose as follows:

**ViT5:** The Vietnamese Text-to-Text Transformer named ViT5 in the paper [58], stands as a cutting-edge, preeminent model in the realm of Transformer-based encoder-decoder architectures tailored specifically for the Vietnamese language. Conceived and crafted by VietAI, this state-of-the-art model undergoes extensive training on an extensive and varied corpus of high-quality Vietnamese texts, employing a T5-style self-supervised pretraining methodology. This allows ViT5 to generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.

**mBERT-cased:** BERT multilingual base cased [59] is a variant of BERT [30], this is a pre-trained multilingual Transformer-based masked language modeling (MLM) model developed by Google. It is a cased model, which means that it distinguishes between different cases of letters, such as uppercase and lowercase. The model is pre-trained on a massive corpus of text data, including Wikipedia articles, books, and news articles in 104 different languages, including Vietnamese. This training allows the model to capture the nuances of multiple languages and perform well on a variety of downstream tasks in multiple languages.

## 4.2 Visual Features Extraction

ViT introduced by Dosovitskiy et al. [39], a pre-trained image model based on transformer architecture, has been trained on a large amount of image data, playing an important role in the English baseline models mentioned in Section 5.1. In the baseline adaptation to Vietnamese, we kept the same visual feature extraction using ViT. Meanwhile, When adapting Vietnamese to conduct experiments on the ViTextVQA dataset with baseline models based on CNN-LM based method, we chose VinVL [56], a model based on Faster R-CNN [38] to extract object features. This choice also brings convenience when we use VinVL to analyze object in Section ??.

**ViT:** Short for Vision Transformer [39] is a new approach to image recognition that utilizes the Transformer architecture. The first step involves dividing the image into smaller squares called patches. These patches are fed into a Transformer encoder, similar to those used in translating languages. This excels at representing relationships between different parts of the input image and helps the model understand the spatial context between various objects. Finally, the rich features are used to classify the image, determining what the image contains. ViT has shown promising results in image processing tasks, achieving superior performance to traditional CNN while potentially requiring less training data.

**VinVL:** Prior research mainly concentrated on perfecting the way these models fuse visual and language information, neglecting advancements in object detection itself. The authors proposed VinVL [56], a system that emphasizes better visual representations by enhancing the object detection model. They develop a new object detection model with a broader range of object and attribute categories and training on a larger dataset. This improved object detection model is trained on a significantly larger dataset, allowing it to capture more nuanced visual details. VinVL achieves state-of-the-art performance on various VL tasks, including Visual Question Answering (VQA), Image Captioning, Visual Commonsense Reasoning (VCR), etc. VinVL highlights the importance of strong visual representations in Vision Language models. By prioritizing advancements in object detection, the model achieved superior performance across different vision language task.

## 4.3 OCR system

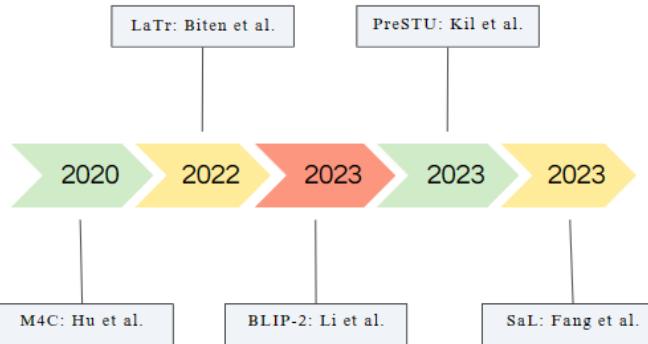
In contrast to commercial closed source English OCR systems for VQA baseline models which are introduced in Section 5.1 such as Rosetta-en, Amazon OCR, etc., SwinTextSpotter [46] stands out for its open source Vietnamese language support. This capability makes it an ideal tool for extracting vital OCR features tailored specifically to the Vietnamese language. Consequently, we opted for SwinTextSpotter to extract essential OCR features for our baseline models adapted for Vietnamese. This choice ensures that our baseline models are well-equipped for better performance in Vietnamese text.

**SwinTextSpotter:** Current state-of-the-art OCR approaches typically share a common backbone for both text detection and recognition tasks. However, this does not fully exploit the potential interaction between the two. Therefore, the authors introduced a new framework named SwinTextSpotter [46]. It is an end-to-end system

that aims to achieve better synergy between detection and recognition. SwinTextSpotter achieves better performance in scene text spotting tasks compared to previous methods in many different languages, including Vietnamese.

## 5 Experiment and Results

### 5.1 Baseline Models



**Figure 14:** Timeline of Visual Question Answering method.

**Table 5:** Features extraction and OCR system of baseline models.

Model	Visual Features	Textual Features	OCR System	Method Type
M4C	VinVL	mBERT-cased	SwinTextSpotter	CNN-LM
BLIP-2	ViT	ViT5	SwinTextSpotter	ViT-LM
LaTr	ViT	ViT5	SwinTextSpotter	ViT-LM
SaL	VinVL	ViT5	SwinTextSpotter	CNN-LM
PreSTU	ViT	ViT5	SwinTextSpotter	ViT-LM

To evaluate the difficulty of the ViTextVQA dataset, we selected several state-of-the-art VQA models as baselines according to their historical development 14. These include M4C [60], which pioneered treating VQA as a generative task using BERT for question embedding and introducing the Dynamic Pointer Network for integrating vocabulary and scene text features; LaTr [8], which adopts an encoder-decoder transformer architecture based on T5 with three primary modules focusing on document layouts, spatial embedding, and visual features extraction; PreSTU [10], which uniquely arranges OCR text by position and uses T5 architecture for pre-training on scene text image data; BLIP-2 [41], which combines frozen image encoders and large language models with a query transformer bridge; and SaL [61], which addresses the neglect of semantic connections and spatial relationships between words through its Text Semantic Separation and Spatial Circle Position modules. While these baselines

were originally designed for English, we adapted them to Vietnamese while preserving their core functionalities. Table 5 provides a comprehensive overview of these models' feature extraction methods and OCR systems, with additional details on their pre-trained models and methodologies available in the referenced sections.

## 5.2 Large Language Models (LLMs)

Currently, large language models (LLMs) are entering a race to develop advanced techniques to solve the Visual Question Answering (VQA) problem. Leading models such as GPT-4o [2], Gemini-1.5-Flash [3], and Qwen2-VL [62] are continuously being improved to enhance performance in analyzing, understanding, and reasoning from image data, helping to improve user experience. These advancements not only help improve visual information processing capabilities but also create many potential applications in academia, industry, and daily life.

GPT-4o enhances VQA capabilities through image analysis, excelling in object recognition, OCR, and chart interpretation. Gemini-1.5 Flash, Google's multimodal chatbot, processes images and text simultaneously, delivering real-time answers to visual queries with optimized speed. Qwen2-VL, an open-source multimodal model from Alibaba Research, handles complex tasks like chart analysis and multilingual text recognition, utilizing Naive Dynamic Resolution and Multimodal Rotary Position Embedding (M-ROPE) for improved visual-textual integration.

## 5.3 Experimental Configuration

The baseline models were trained using the Adam optimization algorithm [63] on an RTX 4090 GPU with 24GB of memory. Each model underwent 10 epochs, with approximately 7 hours required for completion using ViT-LM based method and 6 hours using CNN-LM based method. Hyperparameters for the baseline model were set as follows: a learning rate of  $3.0e^{-5}$ , dropout rate of 0.2, batch size of 16, and weight decay of  $1.0e^{-4}$ .

For baseline models with large version, training was extended to 10 hours over 10 epochs using ViT-LM based method and 9 hours with CNN-LM based method. All other hyperparameters remained consistent with the configuration in base version.

## 5.4 Evaluation Metrics

### 5.4.1 Exact Match

Exact Match (EM) score is a metric commonly used in NLP tasks, particularly in Question Answering systems. It measures the percentage of predictions made by a model that exactly match the target or ground truth answers.

In the context of question answering, an answer is considered to be correct only if it matches the reference answer exactly. The EM score is a straightforward evaluation metric that provides a clear indication of how often a model produces answers that are exactly right.

The formula for EM score is:

$$\text{EM} = \frac{\text{Count of answers predicted exactly}}{\text{Count of answers}} \quad (6)$$

#### 5.4.2 F1-score

F1-score is a common way to measure how well a system performs in question-answering task. It looks at how much the answers of system match the correct ones. Precision focuses on the accuracy of the answers of system, Recall looks at how many correct answers the system can find, and F1 combines both precision and recall for a more comprehensive evaluation. The purpose of this metric is to see how good a system is at providing accurate and complete answers.

The formula for Precision, Recall, and F1 Score are determined as follows:

$$\text{Precision} = \frac{\text{Count of tokens predicted exactly}}{\text{Count of tokens predicted}} \quad (7)$$

$$\text{Recall} = \frac{\text{Count of tokens predicted exactly}}{\text{Count of gold standard tokens}} \quad (8)$$

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

## 5.5 Main Results

**Table 6:** Results of models on ViTextVQA test set.

Model	Version	Method Type	F1-score (%)	EM (%)
M4C	base	CNN-LM	25.46	10.26
BLIP-2	base	ViT-LM	36.87	13.59
BLIP-2	large	ViT-LM	38.47	15.58
LaTr	base	ViT-LM	40.41	17.85
LaTr	large	ViT-LM	42.34	20.94
SaL	base	CNN-LM	43.49	20.11
SaL	large	CNN-LM	43.39	21.48
PreSTU	base	ViT-LM	43.81	20.85
PreSTU	large	ViT-LM	44.93	22.64
<b>ViTextBLIP-2 (Ours)</b>	<b>base</b>	<b>ViT-LM</b>	<b>53.95</b>	<b>25.48</b>

Table 6 presents a comprehensive overview of the results achieved by individual models, showcasing their F1-score and EM metrics. The findings underscore significant disparities in performance among the various models, providing valuable insights into their respective strengths and weaknesses. Notably, within the baseline models, the large version of PreSTU, based on the ViT-LM method, stands out as the top performer with an impressive F1-score of 44.93 and an EM of 22.64. More notably, the proposed ViTextBLIP-2 model achieves an outstanding F1-score of 53.95 and

an EM of 25.48, demonstrating a remarkable improvement over the baseline models. These results highlight the superior potential of the proposed model in enhancing the performance of text-based VQA tasks in Vietnamese.

## 5.6 Performance of LMMs Compared to Human

**Table 7:** LLM results.

Model	F1-score (%)				EM (%)			
	0-shot	1-shot	3-shot	5-shot	0-shot	1-shot	3-shot	5-shot
GPT-4o	55.51	<b>69.55</b>	<b>76.67</b>	<b>75.62</b>	25.00	<b>45.00</b>	<b>56.00</b>	<b>55.00</b>
Gemini-1.5-flash	<b>57.75</b>	59.09	62.37	64.40	26.00	29.00	35.00	39.00
QwenVL-7b	53.80	54.10	54.85	54.74	<b>28.00</b>	36.00	36.00	35.00
ViTextBLIP-2	56.43				26.00			
Human	<u>96.42</u>				<u>88.00</u>			

We randomly selected 100 samples from the test set for evaluation. Specifically, the group used large language models (LLMs) to perform zero-shot and few-shot (including 1, 3, and 5 shots) measurements and compared the results with human-generated answers for the same 100 samples.

From Table 7, it can be seen that the most advanced large language models (LLMs) today, such as GPT-4o, Gemini-1.5-flash, and QwenVL-7b, still lag behind humans in terms of providing answers in visual question answering (VQA) tasks based on images containing text, especially in Vietnamese. Although there is an improvement from 0-shot to few-shot, the F1-score and Exact Match (EM) scores of these models are still much lower than humans. Humans achieved scores of 96.42 (F1) and 88.00 (EM), while the models only reached scores from 53.80 to 75.62 (F1) and 25.00 to 55.00 (EM). This shows that LLMs are still not capable of generating answers as accurate and nuanced as humans in VQA tasks involving Vietnamese text.

## 5.7 Compare with Other Dataset

Table 8 presents the results of traditional methods such as MCAN, LoRRA, and M4C based on standard evaluation metrics like CIDEr and Avg BLEU on the OpenViVQA dataset. M4C, with a CIDEr score of 1.5073 and Avg BLEU of 0.2941, pioneered integrating reading comprehension with visual question answering (VQA). The proposed models (FST, QuMLPAG, and MLPAG) were also evaluated, where QuMLPAG achieved a strong performance with a CIDEr score of 1.7082 and average BLEU of 0.2651.

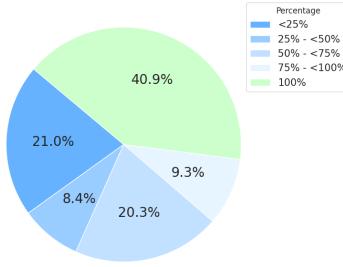
Notably, ViTextBLIP-2 outperformed all models with a CIDEr score of 3.2129 and Avg BLEU of 0.4717, surpassing both traditional VQA models and the proposed approaches. This result highlights its superiority not only on the authors's dataset but also on other benchmarks.

**Table 8:** Performance comparison of different models.

Model	Version	Method Type	CIDEr	Avg BLEU
MCAN	base	ViT-LM	1.0613	0.1699
LoRRA	base	CNN-RNN	0.8005	0.1349
M4C	base	CNN-LM	1.5073	0.2941
FST	base	CNN-LM	0.6141	0.1050
QuMLPAG	base	CNN-LM	1.7082	0.2651
MLPAG	base	CNN-LM	1.6104	0.2739
<b>ViTextBLIP-2 (our)</b>	<b>base</b>	<b>ViT-LM</b>	<b>3.2129</b>	<b>0.4717</b>

## 6 Results Analysis

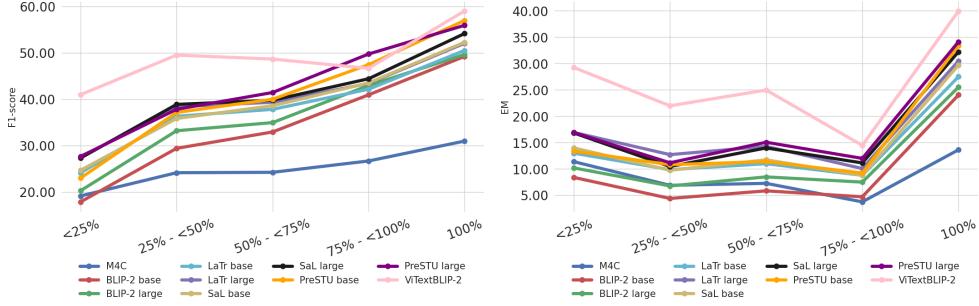
### 6.1 Effect of OCR System Performance and Challenge

**Figure 15:** The ratio of percentage answer token appears in the OCR text in test set.

Based on experiments, we observed a correlation between the performance of the OCR system SwinTextSpotter and the efficacy of the models. To analyze this relationship, we partitioned the test set data based on the percentage of answer tokens present in the OCR text extracted by SwinTextSpotter. The evaluation was carried out on segments with varying ratios: less than 25%, 25% to less than 50%, 50% to less than 75%, 75% to less than 100%, and 100%. Refer to Figure 15 for a detailed breakdown of these ratios.

To evaluate the impact of the performance of the OCR system on the overall performance of the VQA models, we conduct a detailed analysis in the context of both the VQA model and the ratio appearance of answer tokens in OCR text extracted by SwinTextSpotter.

Figure 16 clearly depicts that despite the diversity in the models, the general trend in F1-score is similar. Performance increases as the number of answer tokens appearing in the OCR text increases. This clearly demonstrates the importance of information in OCR text, providing important context and supporting data for the VQA model to predict accurately.



**Figure 16:** Results of models based on OCR system performance.

Figure 16 also illustrates in detail the EM results of the basic models on the occurrence rates of answer tokens in OCR text. It shows a general zig-zag trend, where performance drops when the proportion of answer tokens appearing in OCR text is between 25% and below 50%. Performance increases as the token ratio increases from 50% to less than 75%, but decreases when the token ratio is 75% to less than 100%. In particular, performance increases sharply when the ratio reaches 100%.

It should be noted that, even with a 100% rate of answer tokens appearing in OCR text, the EM and F1-score performances are still only low, lower 35.00 and lower 60.00 respectively. This highlights the major challenge that VQA models face with the ViTextVQA dataset, which requires the ability to synthesize information from multiple sources and the ability to remove noise from OCR text to be able to predict answer correctly.

## 6.2 Effect of Answer and Question Length

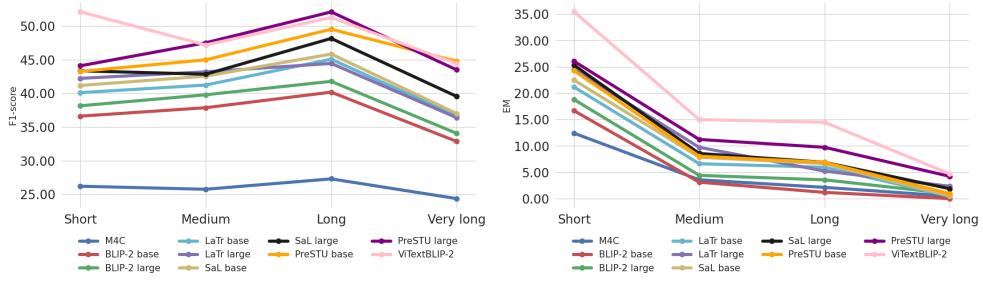
**Table 9:** Group of answer and question length in test set.

Group	Length (n)	Answer Samples	Question Samples
Short	$n \leq 5$	7853	225
Medium	$5 < n \leq 10$	1541	6834
Long	$10 < n \leq 15$	421	2617
Very long	$n > 15$	213	352

We divided questions and answers based on their length in the number of tokens. This classification helps us evaluate the performance of the underlying models more flexibly and accurately. The number of samples in the test set based on different lengths is detailed in Table 9. Classification is done as follows:

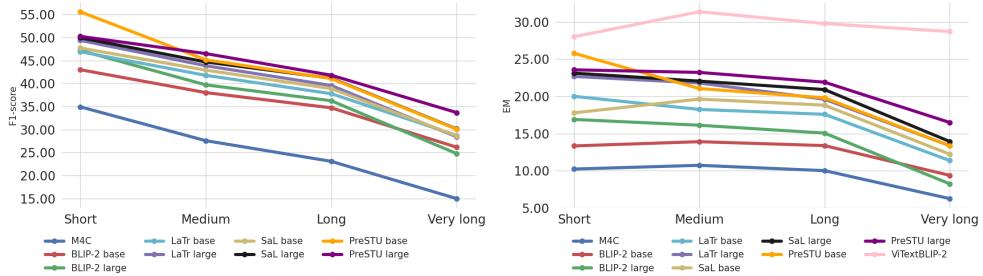
- Short question (and short answer): These are questions and answers that do not exceed 5 tokens in length. These are usually simple questions with concise answers.
- Medium question (and medium answer): This group includes questions and answers ranging from 6 to 10 tokens in length. Questions can be more complex, requiring more detailed information, while answers provide complex answers concisely.

- Long question (and long answer): Questions and answers in this group range from 11 to 15 tokens in length. These questions require detailed information or are more complex. Answers in this group also contain rich and detailed information.
- Very long question (and very long answer): The last group includes questions and answers that are more than 15 tokens. These are often very complex questions, requiring models with strong generality and performance to handle them. Answers in this group also contain a lot of detailed information about the context and details in the image.



**Figure 17:** The results of models based on answer length.

Figure 17 illustrates that models have the same trend in performance with different answer lengths, from short to very long. The results from models clearly demonstrate that as the length of the answer increases, there is a corresponding decrease in performance, as evidenced by a gradual reduction in EM scores. Conversely, with the F1 metric, short answers do not achieve the highest scores like EM. For medium and long answers, the F1 scores tend to be higher. However, this does not mean that longer answers always result in higher F1-score. It clearly shows that the longer the answer, the more F1-score decreases.



**Figure 18:** The results of models based on question length.

As can be seen in Figure 18, in terms of the F1-score metric, the highest scores are observed for short questions, gradually decreasing as the question length increases

and reaching the lowest scores for very long questions across all models. However, for the EM metric, the performance is relatively stable across different question lengths, from short to long. Notably, the SaL base and PreSTU base exhibit different behaviors. While the SaL base appears to increase in performance as the question length increases from short to long, the PreSTU base shows a significant decrease. Particularly, the performance of models becomes increasingly challenging as the question length becomes very long, resulting in a notable decrease in scores.

### 6.3 Effect of OCR text size

We evaluated the performance of the models on different OCR text sizes to better understand this aspect. To ensure the accuracy of the evaluation, we need to select samples where all tokens in the answer appear in the OCR text. This is important because if any token falls outside the range of recognizable OCR text, we cannot know what size the token is. Therefore, we selected samples where 100% tokens in the answer appeared in the OCR text as in Section 6.1. First, we determine the total area of the

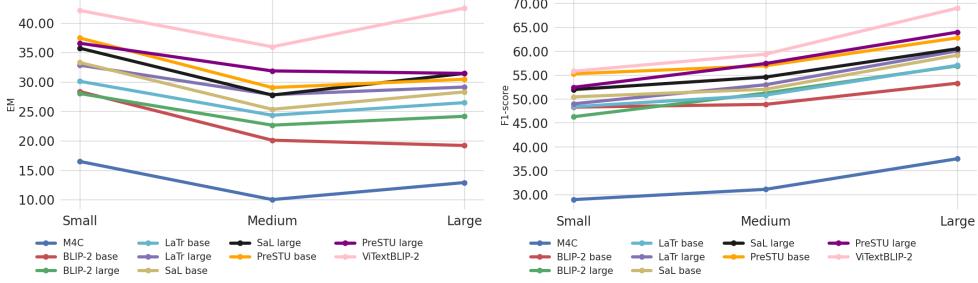


**Figure 19:** Samples of OCR text size group.

tokens in the answer that also appear in the OCR text through the coordinates of the bounding boxes that we used VinVL [56] to extract. We then calculate the percentage of this total area covered in the image. Finally, we determine the group to which these samples belong. With a ratio less than 1%, we classify it into the small group, from 1% to less than 5%, we classify it into the medium group, and the remaining, greater than 5%, we classify it into the large group (see Figure 19).

As depicted in Figure 20, there is a direct correlation between the increase in OCR text size and the performance of models. As results of the F1-score metric, all models share a common trait that when the text size increases, so does the performance of these models.

Interestingly, although this performance increases as the text is enlarged in the F1-score metric, there is a difference when compared with the EM metric. Specifically, the performance of models, measured by the EM metric, decreases when transitioning from small to medium text sizes but tends to rise slightly again when moving to the large size group.



**Figure 20:** Results of models based on OCR text size.

#### 6.4 Effect of OCR Arrangement

In other words, since we often observe that scene text tends to follow a Top-Left To Bottom-Right pattern, we decided to explore whether sorting the OCR outputs in different ways might improve the model’s performance. By experimenting with various sorting strategies, we wanted to see if organizing the text in a specific order could help the VQA model process the information more effectively.

In this research, we conducted an intricate technical analysis to examine the impact of OCR arrangement using three distinct methods on the performance of VQA baseline models applied to the ViTextVQA dataset. Our experimentation involved three sorting approaches: no sorting (maintaining the original order of OCR), sorting based on confidence scores (decrease) extracted from SwintTextSpotter based on the study of Biten et al. [5], and sorting inspired by the top-left to bottom-right strategy as utilized in the PreSTU model by Kil et al. [10].

**Table 10:** Results of models with different OCR arrangements.  $\Delta$  indicates the increase ( $\uparrow$ ) or decrease ( $\downarrow$ ) compared with the original (\*).

Model	Not Sorted		Confidence Score Sorted		Top-Left To Bottom-Right Sorted	
	F1-score (%)	EM (%)	F1-score (%)	EM (%)	F1-score (%)	EM (%)
M4C	25.46*	10.26*	18.41	8.29	30.04	11.60
BLIP-2 base	36.87*	13.59*	35.21	12.97	37.78	15.01
BLIP-2 large	38.47*	15.58*	36.74	14.15	42.16	18.56
LaTr base	40.41*	17.85*	38.76	16.32	43.13	20.42
LaTr large	42.34*	20.94*	39.72	18.88	44.07	22.27
SaL base	43.49*	20.11*	43.66	20.28	44.89	20.97
SaL large	43.39*	21.48*	43.71	20.47	44.74	21.18
PreSTU base	41.74	19.15	40.09	17.74	43.81*	20.85*
PreSTU large	42.85	21.40	40.13	18.43	44.93*	22.64*
ViTextBLIP-2 (Ours)	51.16	24.10	53.95	25.48	<b>52.51*</b>	<b>24.83*</b>
Avg	40.62	18.45	39.04	17.30	42.81	19.83

Experimental results have shown different performance measures between VQA models when applying different alignment methods for OCR. For details, see Table 10, numbers marked with an asterisk (\*) are understood to be the results of the original model.

**Top-left to bottom-right sort:** All models exhibit an upward performance trend when employing the top-left to bottom-right sorting method. This phenomenon can be attributed to the linguistic characteristics of the Vietnamese language, where characters are typically organized from left to right and top to bottom and accurately reflect the way Vietnamese people read and understand text. This sequential arrangement seems to enhance the semantic context of OCR text, helping the models can easily understand and process information from text in the most natural way.

**Confidence score sort:** Results found that sorting OCR according to confidence score reduces performance compared to not sort and top-left to bottom-right sort. It indicates that using this strategy may disturb the position of tokens in OCR text, cause the semantic context of OCR text to be lost, and create difficulties for the VQA model in reading and understanding the content.

**SaL and ViTextBLIP-2:** These two models exhibit stable performance regardless of the OCR token sorting method, whether arranged top-left to bottom-right, by confidence score, or unsorted. For the SaL model, this stability is due to its TSS (Token Selection Strategy) and SCP (Semantic Context Processing) modules, which help it focus on meaningful tokens rather than their spatial arrangement. Similarly, the ViTextBLIP-2 model achieves comparable robustness through its Q-Former module, which collaborates with the OCR system to filter out noise and extract essential features. By leveraging these mechanisms, both models minimize dependence on OCR token order, ensuring consistent performance across different sorting strategies.

In general, arranging OCR text from top-left to bottom-right appears to be a beneficial approach for VQA models when dealing with Vietnamese text. On the contrary, sorting based on confidence scores can introduce ambiguity in placement and result in decreased performance. However, the SaL model stands out as an exception due to the unique capabilities of its two special modules, allowing it to focus on crucial text irrespective of its specific location.

## 6.5 Ablation Study

**Table 11:** Ablation study on different model components.

Image Encoder	OCR System	Captioning Model	F1-score (%)	EM (%)
✓	✓	✓	0.5251*	0.2483*
✓	✓	✗	0.3914 ( $\downarrow 0.1337$ )	0.1694 ( $\downarrow 0.0789$ )
✓	✗	✓	0.3953 ( $\downarrow 0.1298$ )	0.1794 ( $\downarrow 0.0689$ )
✓	✗	✗	0.1417 ( $\downarrow 0.3834$ )	0.0469 ( $\downarrow 0.2014$ )
✗	✓	✓	0.4167 ( $\downarrow 0.1084$ )	0.1882 ( $\downarrow 0.0601$ )
✗	✓	✗	0.3668 ( $\downarrow 0.1583$ )	0.1332 ( $\downarrow 0.1511$ )
✗	✗	✓	0.3512 ( $\downarrow 0.1739$ )	0.1267 ( $\downarrow 0.1216$ )
✗	✗	✗	0.0016 ( $\downarrow 0.5235$ )	0.0000 ( $\downarrow 0.2483$ )

The goal of this experiment is to explore the performance of the ViTextBLIP-2 model when combining different components, including the Image Encoder, OCR

System, and Image Captioning Model. Understanding the impact of each component on model performance helps optimize image and text processing systems to improve recognition and caption generation accuracy. Details of this ablation analysis are presented in Table 11.

**Full Component Integration (✓ ✓ ✓):** When all three components are included in the model, the highest performance is achieved, with an F1-score of 0.5251 and an EM score of 0.2483. This indicates that the full integration of all systems enables the model to maximize information utilization, leading to the best performance. Combining all three components allows the model to not only understand the image but also generate more accurate answers.

**Image Encoder and OCR System Only (✓ ✓ ✗):** When the Image Encoder is combined with the OCR System but without the Image Captioning Model, the F1-score drops to 0.3914 and the EM score decreases to 0.1694. Although OCR assists in extracting textual information, the absence of an image captioning model limits the model’s ability to generate semantically rich descriptions, resulting in lower performance.

**Image Encoder and Image Captioning Model Only (✓ ✗ ✓):** When the Image Encoder is combined with the Image Captioning Model but without OCR, the performance further declines, with an F1-score of 0.3953 and an EM score of 0.1794. While the model can generate descriptive captions for images, the lack of OCR support prevents it from recognizing text within the image, leading to missing textual information that could enhance the response. The absence of OCR reduces the model’s ability to handle text-heavy images.

**Image Encoder Only (✓ ✗ ✗):** When only the Image Encoder is used without other components, the performance drops significantly, with an F1-score of just 0.1417 and an EM score of 0.0469. This suggests that the Image Encoder alone is insufficient for generating accurate answers, as it lacks both OCR-derived text information and image-captioning semantic understanding.

**No Components Used (✗ ✗ ✗):** When none of the modules are utilized, the F1-score and EM score drop to their lowest values, at just 0.0016 and 0.0000, respectively. This demonstrates that the model is completely unable to produce meaningful results. The reason is that without any image-related information, the model cannot process the context needed to understand the query or generate accurate predictions. This clearly highlights the importance of image-based information for the ViTextBLIP-2 model to function effectively.

## 7 Discussion on Characteristics Language for Vietnamese Text-based Visual Question Answering

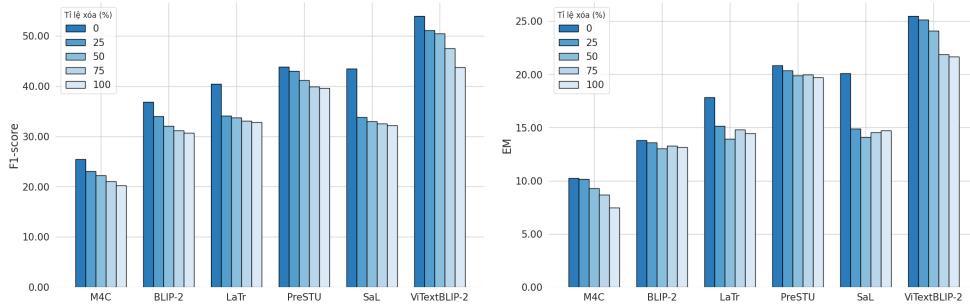
For a language like Vietnamese, there are unique characteristics that need to be considered to achieve good performance in the VQA task. In this Section, we analyze and discuss two main aspects of the Vietnamese language: Vietnamese diacritics, Vietnamese word segmentation and their relationship to this task.

## 7.1 Vietnamese Diacritics

The modern Vietnamese language, with its diacritic system, has undergone a complex development process. Initially, ancient Vietnamese did not have diacritics to distinguish phonemes. The influence of Nom and Chinese characters in the history of Vietnamese has contributed to promoting the proposal and development of diacritics. The colonial period marked a great step forward when French intellectuals contributed to the systematization of diacritics. Although there were debates about the necessity of diacritics, they eventually became an indispensable part of Vietnamese, helping to clearly distinguish the sounds and meanings of words.

Today, diacritics are an indispensable part of the Vietnamese national language, including seven letters (ă, â, ê, ô, ù, ð) and five marks to designate tone(as in à, á, â, ã, á) [64]. They not only distinguish phonemes but also help readers understand and pronounce words correctly. This has made learning and using Vietnamese easier for everyone. An interesting thing is that there are some words in Vietnamese that only differ in tone marks, but have completely different meanings. For example, “má” means “mother”, while “mà” means “but”. This demonstrates the importance of using diacritics in distinguishing words and their meanings.

To underscore the importance of diacritics in Vietnamese, especially in the VQA task, we conducted a series of experiments on our ViTextVQA dataset. In these experiments, we examined diacritic removal levels of 0%, 25%, 50%, 75%, and 100%, respectively, and analyzed the impact of these effects on the performance of the model.



**Figure 21:** The results of models (base version) based on diacritics remove rate.

The results of the experiments shown in Figure 21 show that removing diacritics significantly affects the ability of models to understand and process language. The performance gradually decreases as the diacritic removal rate increases, as measured by the F1-score and EM. This highlights the role of diacritics in providing important information that helps the models understand questions and predict answers based on the image.

There is a notable difference in the F1-score and EM. In the F1-score, there is observation of a clear and consistent decrease in performance as the diacritic removal rate increases. In contrast, the EM score fluctuates unevenly. Except for the M4C model, the base models experience the steepest decrease at the diacritic removal ratio

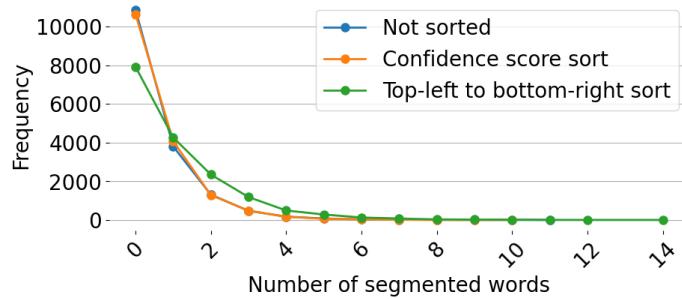
of 50%, and then have a slight increase or no discernible change as the ratio increases to 75% and 100%.

It is worth noting that the LaTr and SaL models experience a significant drop when the diacritic removal rate increases from 0% to 25%. This suggests that these two models rely heavily on the presence of diacritics.

In general, diacritics not only provide information about context and intonation but also help VQA models understand and interact with the language accurately and effectively.

## 7.2 Vietnamese Word Segmentation

One exciting aspect of the Vietnamese language is word segmentation, which is complex due to the absence of explicit word boundaries in the written language. Unlike English, white space is a weak indicator of word boundaries in Vietnamese because when written, it is also used to separate syllables that constitute words [65]. The Vietnamese language often forms compound words by combining multiple individual syllables. For example, “thời gian” (“time”) consists of two syllables in a word “thời\_gian”. This shows that if the words are segmented exactly, many syllables put together will form a meaningful word.



**Figure 22:** Distribution of the number of words segmented in OCR text.

We assume that the different arrangement of text from OCR when applying word segmentation will affect the results of the models. That means which arrangement makes more words segmented in OCR text, the more semantic enhancement in context, thereby improving the performance of the models.

To clarify this assumption, we performed word segmentation analysis for different OCR text arrangements. We used the VnCoreNLP tool [57] to perform word segmentation. Refer to Table 12 and Figure 22 for detailed results.

Table 12 clearly shows that top-left to bottom-right sort gives the best results. This is illustrated by the fact that this arrangement achieves the highest average number of segmented words with 1.06 words, the highest number of segmented words in an OCR text at 14 words, and has the best average performance in F1-score and EM at 41.73 and 19.28 respectively. Arranging OCR text from top-left to bottom-right creates a more systematic context. When words are arranged in order from left to right and top

**Table 12:** Statistics of word segmentation in OCR text with different arrangement types.

Arrangement Method	Min	Max	Avg	Avg EM (%)	Avg F1-score (%)
Not sorted	0	11	0.56	18.45	40.62
Confidence score sorted	0	10	0.56	17.30	39.04
Top-left to bottom-right sorted	0	14	1.06	19.83	42.81

to bottom, the model can easily capture the structure of the text and the relationship between words. This makes word segmentation more accurate, thereby improving the understanding and prediction capabilities of the underlying models.

Focusing on increasing the number of segmented words through OCR text arrangements can be considered a good strategy to improve model performance in many natural language processing applications, especially VQA task, where word-by-word precision plays an important role.

## 8 Conclusion and Future Work

In this article, we presented a new dataset for the VQA task in Vietnamese to the global research community, especially Vietnam. Specifically, our ViTextVQA dataset includes more than 16,000 images and more than 50,000 question-answer pairs focusing on exploiting . We believe that the effort to create our ViTextVQA dataset will contribute to improving the diversity and enriching resources for the Vietnamese language.

Additionally, we evaluated the performance of state-of-the-art VQA models on our dataset. We found that while these models have proven effective on popular English VQA datasets, similar performance has not been achieved on the ViTextVQA dataset. This prompted us to conduct extensive analyses of the performance aspects of our models, identifying limitations that appear during the experiment, thereby finding valuable insights that lay the groundwork for future research. The results analysis also revealed that reordering OCR text from top-left to bottom-right enhances performance and boosts the accuracy of baseline models in Vietnamese.

In our endeavors moving forward to the future, we recognize the multifaceted utility of our ViTextVQA dataset beyond its application solely in VQA task. One promising avenue lies in leveraging this dataset for the task of Visual Question Generation (VQG). We will aim to use this dataset in VQG task, which the goal is to automatically generate questions about what is happening in the image when given an answer [66–68].

Additionally, we will continue to explore prompting-based methods with large language models. By combining image features with large language models, prompting-based methods such as [69, 70] have demonstrated a significantly effective, even when trained with fewer parameters than fully fintuning.

Furthermore, our plan also includes developing a chatbot capable of answering image-based questions like Flamingo [1], GPT4 [2], Gemini [3]. This chatbot will be trained on our ViTextVQA dataset and combined with current available VQA datasets

in Vietnamese. This will provide an interesting and useful application, helping users interact with the chatbot more naturally and flexibly in Vietnamese.

## Acknowledgements

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under the grant number DS2024-26-01.

## Author Contributions Statement

**Quan Van Nguyen:** Conceptualization; Formal analysis; Investigation; Methodology; Validation; Visualization; Writing - original draft.

**Dan Quang Tran:** Conceptualization; Data curation; Formal analysis; Investigation; Validation; Visualization; Writing - original draft.

**Huy Quang Pham:** Conceptualization; Data curation; Investigation; Methodology; Writing - original draft.

**Thang Kien-Bao Nguyen:** Conceptualization; Data curation; Investigation; Methodology; Writing - original draft.

**Nghia Hieu Nguyen:** Conceptualization; Data curation; Investigation; Methodology; Writing - review & editing.

**Kiet Van Nguyen:** Conceptualization; Formal analysis; Investigation; Methodology; Validation; Supervision; Writing - review & editing.

**Ngan Luu-Thuy Nguyen:** Conceptualization; Formal analysis; Investigation; Methodology; Validation; Supervision; Writing - review & editing.

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

## Data Availability

Data will be made available on reasonable request.

## References

- [1] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., *et al.*: Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* **35**, 23716–23736 (2022)
- [2] OpenAI, :, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, H., Kiros, J., Knight, M., Kokotajlo, D., Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mely, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O’Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., Avila Belbute Peres, F., Petrov, M., Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotstед, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng,

- J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: GPT-4 Technical Report (2023)
- [3] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
- [4] Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8317–8326 (2019)
- [5] Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4291–4301 (2019)
- [6] Wang, B., Lv, F., Yao, T., Ma, J., Luo, Y., Liang, H.: Chiqa: A large scale image-based real-world question answering dataset for multi-modal understanding. In: Proceedings of the 31st ACM International Conference on Information & Knowledge Management, pp. 1996–2006 (2022)
- [7] Qi, L., Lv, S., Li, H., Liu, J., Zhang, Y., She, Q., Wu, H., Wang, H., Liu, T.: Dureadervis: A: A chinese dataset for open-domain document visual question answering. In: Findings of the Association for Computational Linguistics: ACL 2022, pp. 1338–1351 (2022)
- [8] Biten, A.F., Litman, R., Xie, Y., Appalaraju, S., Manmatha, R.: Latr: Layout-aware transformer for scene-text vqa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16548–16558 (2022)
- [9] Geigle, G., Jain, A., Timofte, R., Glavaš, G.: mblip: Efficient bootstrapping of multilingual vision-lmms. arXiv preprint arXiv:2307.06930 (2023)
- [10] Kil, J., Changpinyo, S., Chen, X., Hu, H., Goodman, S., Chao, W.-L., Soricut, R.: Prestu: Pre-training for scene-text understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 15270–15280 (2023)
- [11] Tran, K.Q., Nguyen, A.T., Le, A.T.-H., Van Nguyen, K.: Vivqa: Vietnamese visual question answering. In: Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation, pp. 683–691 (2021)
- [12] Nguyen, N.H., Vo, D.T., Van Nguyen, K., Nguyen, N.L.-T.: Openvivqa: Task,

dataset, and multimodal fusion models for visual question answering in vietnamese. *Information Fusion* **100**, 101868 (2023)

- [13] Pham, H.Q., Nguyen, T.K.-B., Van Nguyen, Q., Tran, D.Q., Nguyen, N.H., Van Nguyen, K., Nguyen, N.L.-T.: Viocrvqa: novel benchmark dataset and visionreader for visual question answering by understanding vietnamese text in images. *Multimedia Systems* **31**(2), 106 (2025)
- [14] Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems* **27** (2014)
- [15] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2425–2433 (2015)
- [16] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6904–6913 (2017)
- [17] Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952 (2019). IEEE
- [18] Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209 (2021)
- [19] Tanaka, R., Nishida, K., Yoshida, S.: Visualmrc: Machine reading comprehension on document images. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 13878–13888 (2021)
- [20] Nguyen, N.L.-T., Nguyen, N.H., Vo, D.T., Tran, K.Q., Van Nguyen, K.: Vlsp 2022–evjvqa challenge: Multilingual visual question answering. *arXiv preprint arXiv:2302.11752* (2023)
- [21] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations (ICLR 2015)*, 1–14 (2015)
- [22] Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural computation* **9**(8), 1735–1780 (1997)
- [23] Strobelt, H., Gehrmann, S., Pfister, H., Rush, A.M.: Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics* **24**(1), 667–676 (2017)

- [24] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
- [25] Zhang, Y., Jin, R., Zhou, Z.-H.: Understanding bag-of-words model: a statistical framework. *International journal of machine learning and cybernetics* **1**, 43–52 (2010)
- [26] Xu, H., Saenko, K.: Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In: Computer Vision - 14th European Conference, ECCV 2016, Proceedings, pp. 451–466 (2016). Springer
- [27] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
- [28] Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. In: NIPS 2014 Workshop on Deep Learning, December 2014 (2014)
- [29] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)
- [30] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)
- [31] Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* **32** (2019)
- [32] Li, L.H., Yatskar, M., Yin, D., Hsieh, C.-J., Chang, K.-W.: What does bert with vision look at? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5265–5275 (2020)
- [33] Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11336–11344 (2020)
- [34] Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. 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), pp. 5100–5111 (2019)

- [35] Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: ViL-bert: Pre-training of generic visual-linguistic representations. In: International Conference on Learning Representations (2019)
- [36] Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European Conference on Computer Vision, pp. 104–120 (2020). Springer
- [37] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., *et al.*: Oscar: Object-semantics aligned pre-training for vision-language tasks. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16, pp. 121–137 (2020). Springer
- [38] Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580–587 (2014)
- [39] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., *et al.*: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)
- [40] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning, pp. 12888–12900 (2022). PMLR
- [41] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: International Conference on Machine Learning, pp. 19730–19742 (2023). PMLR
- [42] Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., Wang, L.: GIT: A generative image-to-text transformer for vision and language. Transactions on Machine Learning Research (2022)
- [43] Li, C., Xu, H., Tian, J., Wang, W., Yan, M., Bi, B., Ye, J., Chen, H., Xu, G., Cao, Z., Zhang, J., Huang, S., Huang, F., Zhou, J., Si, L.: mPLUG: Effective and efficient vision-language learning by cross-modal skip-connections. In: Goldberg, Y., Kozareva, Z., Zhang, Y. (eds.) Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 7241–7259 (2022)
- [44] Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O.K., Singhal, S., Som, S., *et al.*: Image as a foreign language: Beit pretraining for vision and vision-language tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19175–19186 (2023)

- [45] Chen, X., Wang, X.: Pali: Scaling language-image learning in 100+ languages. In: Conference on Neural Information Processing Systems (NeurIPS) (2022)
- [46] Huang, M., Liu, Y., Peng, Z., Liu, C., Lin, D., Zhu, S., Yuan, N., Ding, K., Jin, L.: Swintextspotter: Scene text spotting via better synergy between text detection and text recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4593–4603 (2022)
- [47] Cohen, J.: A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**(1), 37–46 (1960)
- [48] Fleiss, J.L.: Measuring nominal scale agreement among many raters. *Psychological bulletin* **76**(5), 378 (1971)
- [49] Hoang, P.G., Luu, C.D., Tran, K.Q., Nguyen, K.V., Nguyen, N.L.-T.: ViHOS: Hate speech spans detection for Vietnamese. In: Vlachos, A., Augenstein, I. (eds.) *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 652–669. Association for Computational Linguistics, Dubrovnik, Croatia (2023)
- [50] Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: SQuAD: 100,000+ questions for machine comprehension of text. In: Su, J., Duh, K., Carreras, X. (eds.) *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392. Association for Computational Linguistics, Austin, Texas (2016)
- [51] Mathew, M., Bagal, V., Tito, R., Karatzas, D., Valveny, E., Jawahar, C.: Infographicvqa. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1697–1706 (2022)
- [52] Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617 (2018)
- [53] Marino, K., Rastegari, M., Farhadi, A., Mottaghi, R.: Ok-vqa: A visual question answering benchmark requiring external knowledge. In: *Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition*, pp. 3195–3204 (2019)
- [54] Gupta, D., Lenka, P., Ekbal, A., Bhattacharyya, P.: A unified framework for multilingual and code-mixed visual question answering. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pp. 900–913 (2020)

- [55] Tran, K.V., Phan, H.P., Van Nguyen, K., Nguyen, N.L.T.: Viclevr: A visual reasoning dataset and hybrid multimodal fusion model for visual question answering in vietnamese. arXiv preprint arXiv:2310.18046 (2023)
- [56] Zhang, P., Li, X., Hu, X., Yang, J., Zhang, L., Wang, L., Choi, Y., Gao, J.: Vinvl: Revisiting visual representations in vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5579–5588 (2021)
- [57] Vu, T., Nguyen, D.Q., Dras, M., Johnson, M., *et al.*: Vncorenlp: A vietnamese natural language processing toolkit. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 56–60 (2018)
- [58] Phan, L., Tran, H., Nguyen, H., Trinh, T.H.: 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, pp. 136–142 (2022)
- [59] Pires, T., Schlinger, E., Garrette, D.: How multilingual is multilingual bert? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4996–5001 (2019)
- [60] Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9992–10002 (2020)
- [61] Fang, C., Li, J., Li, L., Ma, C., Hu, D.: Separate and locate: Rethink the text in text-based visual question answering. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 4378–4388 (2023)
- [62] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
- [63] Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR) (2015)
- [64] Nguyen, H.T.M., Nguyen, H.V., Ngo, Q.T., Vu, L.X., Tran, V.M., Ngo, B.X., Le, C.A.: Vlsp shared task: Sentiment analysis. Journal of Computer Science and Cybernetics **34**(4), 295–310 (2019)
- [65] Nguyen, D.Q., Nguyen, D.Q., Vu, T., Dras, M., Johnson, M.: A fast and accurate Vietnamese word segmenter. In: Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno,

- A., Odijk, J., Piperidis, S., Tokunaga, T. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Miyazaki, Japan (2018)
- [66] Zhang, S., Qu, L., You, S., Yang, Z., Zhang, J.: Automatic generation of grounded visual questions. In: International Joint Conference on Artificial Intelligence, pp. 4235–4243 (2017)
  - [67] Fan, Z., Wei, Z., Wang, S., Liu, Y., Huang, X.-J.: A reinforcement learning framework for natural question generation using bi-discriminators. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 1763–1774 (2018)
  - [68] Xu, X., Wang, T., Yang, Y., Hanjalic, A., Shen, H.T.: Radial graph convolutional network for visual question generation. IEEE transactions on neural networks and learning systems **32**(4), 1654–1667 (2020)
  - [69] Jia, M., Tang, L., Chen, B.-C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.-N.: Visual prompt tuning. In: European Conference on Computer Vision, pp. 709–727 (2022). Springer
  - [70] Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)