Journal of Computer Science and Cybernetics, V.39, N.3 (2023), 237–258
DOI no. 10.15625/1813-9663/39/3/18157

EVJVQA CHALLENGE: MULTILINGUAL VISUAL QUESTION ANSWERING

NGAN LUU-THUY NGUYEN[1,2,*], NGHIA HIEU NGUYEN[1,2], DUONG T. D. VO[1,2], KHANH QUOC TRAN[1,2], KIET VAN NGUYEN[1,2] *

[1] *Faculty of Information Science and Engineering, University of Information Technology, Ho Chi Minh City, Viet Nam*

[2] *Vietnam National University, Ho Chi Minh City, Vietnam*

Abstract. Visual question answering (VQA) is a challenging task of natural language processing (NLP) and computer vision (CV), attracting significant attention from researchers. English is a resource-rich language that has witnessed various developments in datasets and models for visual question answering. Visual question answering in other languages also would be developed for resources and models. In addition, there is no multilingual dataset targeting the visual content of a particular country with its own objects and cultural characteristics. To address the weakness, we provide the research community with a benchmark dataset named EVJVQA, including 33,000+ pairs of question-answer over three languages: Vietnamese, English, and Japanese, on approximately 5,000 images taken from Vietnam for evaluating multilingual VQA systems. EVJVQA is used as a benchmark dataset for the challenge of multilingual visual question answering at the 9th Workshop on Vietnamese Language and Speech Processing (VLSP 2022). This task attracted 62 participant teams from various universities and organizations. In this article, we present details of the organization of the challenge, an overview of the methods employed by shared-task participants, and the results. The highest performances were 0.4392 in F1-score and 0.4009 in BLUE on the private test set. The multilingual QA systems proposed by the top 2 teams used ViT for the pre-trained vision model and mT5 for the pre-trained language model, a powerful pre-trained language model based on the transformer architecture. EVJVQA is a challenging dataset that motivates NLP and CV researchers to further explore the multilingual models or systems for visual question-answering systems.

Keywords. Visual question answering; Vision-language understanding; MultiModal learning; Information fusion; Transformer.

1. INTRODUCTION

Visual or Image-Based Question Answering is a challenging task that requires knowledge of two most attractive AI fields: natural language processing and computer vision. Specifically, querying the information of images through human-language questions is a friendly and natural approach to searching for information, meeting the needs of people extracting information in many domains such as life, education, work, etc. However, the existing studies have mainly focused on resource-rich

*Corresponding author.
E-mail addresses: ngannlt@uit.edu.vn (N.L.T.Nguyen); 19520178@gm.uit.edu.vn (N.H. Nguyen); 19520483@gm.uit.edu.vn (D.T.D. Vo); khanhtq@uit.edu.vn (K.Q.Tran); kietnv@uit.edu.vn (K.V.Nguyen).

languages such as English. In this challenge, we aim to extend visual question answering to more languages, including rich and low-resource languages.

English has witnessed numerous benchmarks for evaluating and developing visual question answering models or systems. Recently, researchers designed datasets with goal-oriented evaluations. Firstly, VQA datasets [1–4] were created on general images. Soon after, a more complex dataset based on reasoning was discovered by [1]. In addition, VQA [2] is also geared towards support applications for seemingly-impaired and blind people. [3] showed that VQA has the ability to read text on photos. More challenging, VQA requires external, common sense, or world knowledge to predict more correct answers [4, 5]. Besides, Changpinyo et al. [4] proposed a multilingual dataset for visual question answering in 13 languages. However, this dataset is done automatically based on the auto-translation and verification method. A human-annotated multilingual dataset including three languages: English (resource-rich language), Japanese, and Vietnamese (resource-poor language) is presented in this paper. More specifically, three main contributions of this paper are described as follows.

1. Firstly, we constructed UIT-EVJVQA, a multilingual dataset for evaluating the visual question-answering systems or models, which comprises 33.790 question-answer pairs in three languages: English, Vietnamese, and Japanese.

2. Secondly, we organized the VLSP2022-EVJVQA Challenge for evaluating multilingual VQA models (Vietnamese, English, and Japanese) at the VLSP 2022. Our baseline system obtains 0.3346 in F1-score and 0.2275 in BLEU on the public and private test sets, respectively, and there are no models of participating teams that pass 0.44 (in F1-score) on the private test set, which indicates that our dataset is challenging and requires the development of multilingual VQA models.

3. Lastly, combined with other VQA datasets for analysis, UIT-EVJVQA could potentially be a useful resource for multilingual research.

The rest of this paper is organized as follows. In Section 2, a brief overview of the background and relevant studies is presented. The VLSP 2022-EVJVQA Challenge is introduced in Section 3. Our new dataset (UIT-EVJVQA) is presented in detail in Section 4. Section 5 presents the systems and results proposed by participating teams. In Section 6, we provide further analysis of the challenge results. Finally, Section 7 summarizes the findings of the VLSP 2022-EVJVQA Challenge and suggests future research directions.

2. BACKGROUND AND RELATED WORKS

Visual question-answering (VQA) is a challenging task that has significant value not only in the research community but also in daily life. VQA task was first introduced by [5]. The authors were successful in creating a novel dataset and fundamental English methodologies. Inspired by that success, various further studies have been created and implemented in a variety of languages [6] including Chinese [7], Japanese [8], and Vietnamese [9].

VQA has gained more attention from researchers recent years and has shown significant growth. The studies were introduced not only in monolingual but also in multilingual applications [10–13]. This stage contributes significantly to the creation of multilingual VQA (mVQA) systems. Some typical research works in this approach can be mentioned such as [10] with the study that the proposed

Figure 1: Overview of the multilingual visual question answering task.

model is capable of predicting responses from questions in Hindi, English, or Code-mixed (Hinglish: Hindi-English) languages; Changpinyo et al. [4] with a translation-based mVQA dataset in 7 distinct languages; Gao et al. [11] constructed a Freestyle Multilingual Image Question Answering (FM-IQA) dataset containing over 150,000 images and 310,000 freestyle Chinese question-answer pairs and their English translations.

In this study, the first dataset for the task of mVQA on English-Vietnamese-Japanese (EVJVQA) is created and it is expected to open up new research areas and aid in evaluating multilingual VQA models.

3. EVJVQA CHALLENGE

3.1. Task definition

The main difference between the former VQA task [5], or monolingual VQA in the context of this paper, and multilingual VQA [4] is that monolingual VQA requires the VQA methods to find the answer from a given image for a question, while in multilingual VQA, VQA methods are required to first determine the language of the given question, then give the answers in the language of given question based on the information in the image.

The multilingual VQA task in this shared-task challenge is defined as follows (Figure 1):

- **Input**: Given an image and a question that can be answerable.
- **Output**: An answer where there can be a span related to the image's content.

Language selection. Three languages are selected in which the main language is Vietnamese, and the other two popular other languages (English and Japanese) in the pictures taken from Vietnam.

3.2. Evaluation metrics

In this challenge, we use two evaluation metrics: $F1$ and BLUE. Based on [12], the $F1$ score of each answer is calculated based on tokens of the gold answer (GA) and tokens of the predicted answer (PA). The overall $F1$ is averaged across all questions of each set. For Vietnamese and

English languages, $F1$ is calculated based on tokens, whereas $F1$ is calculated based on characters for Japanese.

$$Precision(P) = \frac{GA \cap PA}{PA},$$

$$Recall(R) = \frac{GA \cap PA}{GA},$$

$$F1 = \frac{2PR}{P+R}.$$

Inspired by [13], the Bilingual Evaluation Understudy (BLEU), a popular evaluation metric in machine translation, computes the n-gram co-occurrence between human-generation answers and system-generation answers. The best performances were estimated by averaged BLEU-based performances (BLEU-1, BLEU-2, BLEU-3, and BLEU-4) on the public test and private test sets. Both evaluation metrics ignore punctuations.

The difference between $F1$ and BLEU is that the $F1$ metric does not take into account the position of tokens in the sentence as they treat the sentence as a set of tokens. On the other hand, BLEU takes into account the constraint of position between tokens in the sentence because of calculation using n-gram tokens. Accordingly, BLEU is stricter than $F1$, hence we can see in Subsection 5.3 that the scores in BLEU are not as significant as those in $F1$.

3.3. Overview summary

Table 1 presents an overview of the participating teams who joined the VLSP2022-EVJVQA.

Table 1: Participation summary of the VLSP 2022 - EVJVQA challenge

Metric	**Value**
#Registration Teams	62
#Joined Teams	57
#Signed Data Agreements	36
#Submitted Teams	8
#Paper Submissions	5

4. CORPUS CREATION

A previous work [14] inherited assets from the well-known VQA benchmark in English and the COCO-QA [15], then they proposed a semi-automatic annotating system by using machine translation to translate question-answer pairs from English to Vietnamese. On the other hand, we argue that the context in images captured in Vietnam is more complicated than in images coming from VQA benchmarks in English because of its crowded scene and "out-of-common" objects, or in particular, objects that are not commonly used outside of Vietnam. Moreover, using such a machine translation system as [14] is hard to ensure the natural aspect of using language, which can cause lots of confusion while evaluating VQA methods in Vietnamese. To overcome the above flaws and develop a VQA system, particularly for the Vietnamese, we constructed the novel dataset with images collected

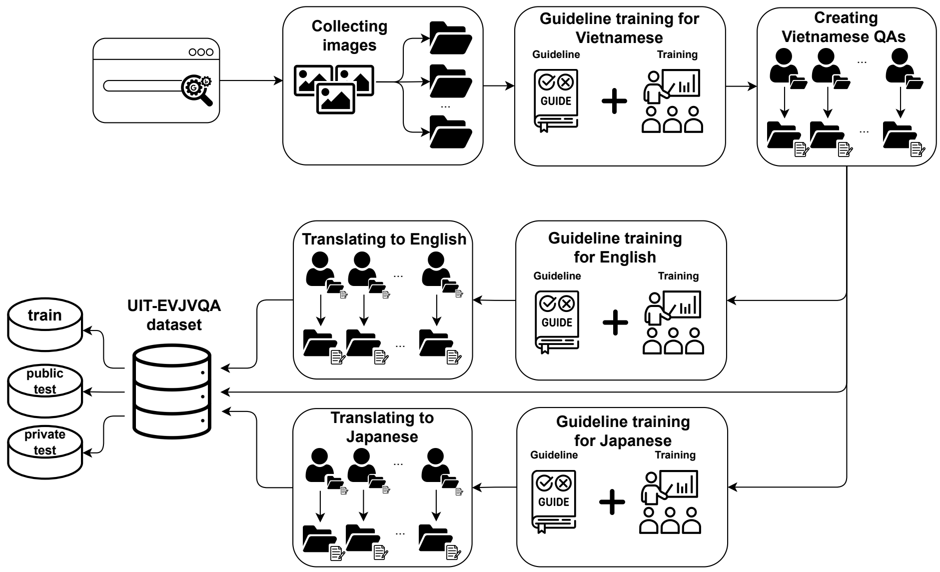


Figure 2: Overall pipeline process for creating the UIT-EVJVQA dataset

manually and relevant to daily life in Vietnam. In addition, to address and challenge the research community, we provided our dataset of ***multilingual*** question-answer pairs to encourage the research community to explore and propose an effective system that can answer questions written in various languages.

4.1. Image collection

To build a VQA dataset in the Vietnamese context, we searched for images with a diverse and open-domain set of keywords. We first selectively prepared various keywords which were relevant to Vietnamese locations and daily life activities or result images that specifically contained targeted objects in Vietnamese scenes. For instance, the keywords can be đường phố ở Việt Nam (Vietnamese streets), chợ (markets), ăn vặt đường phố (sidewalk eateries), du lịch (travel), phố cổ (ancient corner street), phương tiện giao thông (means of transport), lễ hội dân gian (traditional festivals), etc. Some keywords were appended with Vietnamese location names like Hanoi or Saigon for more variation in geological and cultural context. We then used these keywords to scrap images from Google Images. The image-scraping process was facilitated by the Octoparse tool.

After collecting the images, we proceeded with the filtering stage. The images originally came in various sizes. However, we must ensure the details in them are visible. Therefore, we only kept images with widths and heights greater than 500 and 400, respectively. We also filtered out GIF files and other file formats apart from JPEG and PNG. As a result, we obtained 4,909 images with their size varying in the range of 500 - 6000 pixels in terms of width and the range of 400 - 4032 pixels in terms of height.

4.2. Question and answer creation process

The questions and answers (QAs) of the UIT-EVJVQA dataset were first created in Vietnamese throughout the set of images. Afterward, these QAs were translated into English and Japanese. Both stages were conducted by the source of crowd workers. QAs in all three languages were then merged according to the images and eventually constituted the final corpus. The overall pipeline of the aforementioned process is visualized in Figure 2.

4.2.1. Base Vietnamese QAs creation

We first employed five crowd workers for the Vietnamese question and answer creation stage. For each image, the workers were asked to formulate 3-5 question-answer pairs based on the details and objects that appeared in the visible scene. The workers are required to conform to the following guidelines:

- Encourage using phrases or full sentences to give the answers.
- Restrict the use of single words as answers.
- No selective question or yes/no question allowed.
- Numbers must be typed in alphabet characters rather than numeric characters, and they must not be greater than 10.
- For colors, only use provided colors such as black, white, red, orange, yellow, green, blue, pink, purple, brown, and grey. If these colors cannot exactly describe the true color of the object, ignore such color property in the sentence.
- In the case of mentioning the side of an object, if following that direction words is an object, then such direction is defined based on that object, otherwise using the perspective of the annotator to define the direction.

Eventually, this stage yields 11,689 pairs of question and answer in Vietnamese for our dataset.

4.2.2. English and Japanese QAs human translation

All Vietnamese QAs are henceforth passed through the human translation stages. These stages demand the employment of qualified crowd workers for the translation of questions and answers. For English translation, the workers must have at least IELTS certification with an overall band score of 6.5. Meanwhile, translators for the Japanese translation task must achieve an N3 proficiency level or above. Overall, there were seven and nine translators working simultaneously for the English and Japanese translation process, respectively. The English and Japanese translation stages also have their corresponding guidelines.

English translation guideline: The English QAs are translated from the Vietnamese ones, with as many entities and attributes in the sentences retained during the translation as possible. The translators are encouraged to use phrases or full sentences as the translated answers. Apostrophes in sentences are restricted. Thus, translators should use the uncontracted form or other valid grammar formulations.

Japanese translation guideline: The Japanese QAs are translated from the Vietnamese ones, with as many entities and attributes in the sentences retained during the translation as possible. For transcribing Vietnamese proper nouns or other foreign words, the katakana syllabary is adopted. The polite form is utilized for writing translated questions and answers that contain verbs. In the case of complex Vietnamese questions and answers with multiple relevant information that may not be easily translated into continuous Japanese sentences, translators can use commas to split the sentences into smaller parts and then translate them subsequently. For example, the question "What products does the woman wearing a helmet go to the store to buy?" can be translated to Japanese as "ヘルメットをかぶった女性が店に買いに行って、どうの商品を買いますか?", which literally means "A woman wearing a helmet goes shopping at a store, what products does she buy?".

After the human-translation process, we obtained 10,539 question-answer pairs in English and 11,562 question-answer pairs in Japanese.

4.3. Pre-processing and splitting

We normalized the Vietnamese and English QAs into lowercase. Latin characters in Japanese words, "Tシャツ" (T-shirt) for instance, were also normalized similarly. After that slight pre-processing step, we merged all the QAs in all three languages into one dataset.

To prepare for the VLSP 2022 - EVJVQA Challenge, the dataset was split into the training set, public test set, and private test set. 3,763 images were used for the training set, while each test set comprised 558 images. Since no image is placed in multiple sets at once, we ensure that all QAs for a particular image only appear in the set to which that image belongs. The corresponding number of QAs in each language for each of the sets is shown in Table 2.

Table 2: Number of QAs in each language in our UIT-EVJVQA dataset

	Training	**Public test**	**Private test**	**Total**
Vietnamese	8,334	1,685	1,670	11,689
English	7,189	1,679	1,671	10,539
Japanese	8,262	1,651	1,649	11,562
Total	23,785	5,015	4,990	33,790

4.4. Statistics

As our dataset was constructed from three languages: Vietnamese, English, and Japanese, we conducted statistics to deeply observe the characteristic of each language as well as gain clear insights into the three languages.

Table 3: Statistic of question and answer in the UIT-EVJVQA dataset

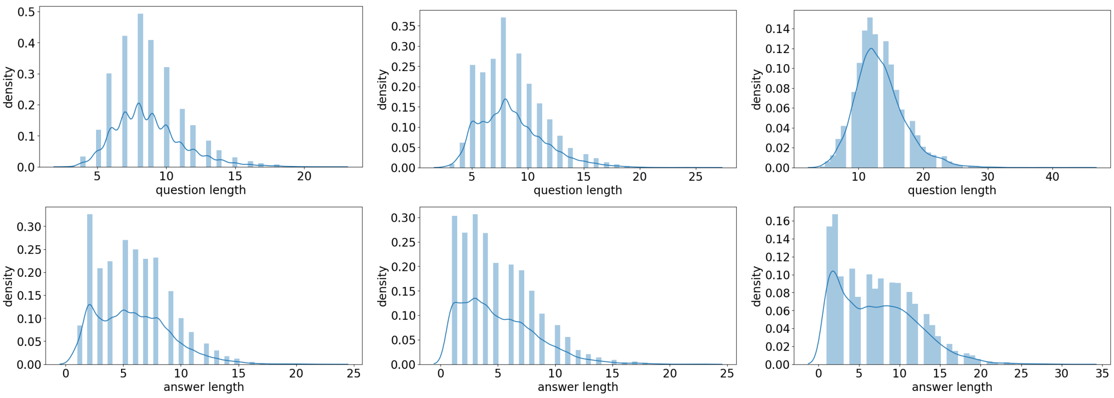
	Question			**Answer**		
	Max.	**Min.**	**Avg.**	**Max.**	**Min.**	**Avg.**
Vietnamese	22	3	8.7	32	1	7.2
English	26	3	8.6	23	1	5.0
Japanese	45	4	13.3	23	1	5.9

To conduct statistics on Vietnamese, we use the word segmentation method from the VnCoreNLP [16] as in Vietnamese, a word may have more than one token (for instance, "cửa hàng tạp hóa" is formed from two Vietnamese words "cửa hàng" and "tạp hóa", which is in turn formed from more than one token). For English QAs, we achieved tokens by splitting sentences using space. For Japanese QAs, like Vietnamese, Japanese uses hieroglyphs to form their word, hence each Japanese word may have more than one hieroglyph. We use the janome* library to perform word segmentation on Japanese QAs.

(a) Vietnamese (b) English (c) Japanese

Figure 3: Word cloud of tokens in three language partitions of the UIT-EVJVQA dataset

Japanese in general use more words to describe than Vietnamese and English. This implies another challenge for the VQA method when tackling Japanese text beside the complexity of multilingualism in our dataset. Moreover, Vietnamese and English have the same distribution of length in terms of questions (according to Table 3), while English has shorter answers compared with those in Vietnamese.



(a) Vietnamese (b) English (c) Japanese

Figure 4: Statistics of the length of QAs in each language partition of the UIT-EVJVQA dataset

Answers of the three languages share the same characteristic where the most appearance of length is two. This indicates humans while giving answers, prefer saying in short statements, and this behavior leads to the classification approach on the VQA dataset [17]. However, such short answers are not always preferable in our dataset as the context of images taken in Vietnam is complicated because of the crowded scenes with multiple entities and various activities happening (imagine describing the location of a person in a red shirt on a populous street). Hence, short answers cannot bear enough

https://github.com/mocobeta/janome

information to answer questions that enquire about a specific detail or attribute in complex scenes. Moreover, we aim to emphasize the language aspect of the VQA task, which means we want to guide the community to research and propose a system that can give answers flexibly and naturally as a human does, not the way of "selecting" answers from a defined set as most approaches in the VQA dataset [5, 17]. To this end, the answers given in our dataset are diverse in length and complicated in terms of level (word, phrase, or sentence level). Interestingly, while annotating answers, we found that giving a phrase or sentence as an answer is more fluent and human-like than giving only words or phrases as in the VQA dataset of [5].

Another factor to consider while annotating the UIT-EVJVQA dataset is the language prior phenomenon [18]. This is the phenomenon where the VQA methods try to learn patterns between questions and answers, such as questions starting with "how many" usually go with the answer "two". As analyzed in [18], the language priors in the VQA dataset are the result of the classification approach proposed for the VQA task [17] and cause the model to learn the way of recognizing answers based on the question rather than the way to make use of the image to answer the given question. Therefore while constructing the guideline to annotate the UIT-EVJVQA dataset, we propose to give answers using words, phrases, or sentences. In this way, we can first eliminate the traditional classification approach proposed for the VQA task in English as well as avoid the language priors in our dataset.

5. METHODS AND RESULTS

The challenge aims to evaluate the quality of the teams' approaches to multilingual visual question-answering systems.

5.1. Baseline method

Our baseline method includes two components: the feature extraction component and the answer generator component. Following Changpinyo et al. [19], we construct the feature extraction using Vision Transformer (ViT) [20] to obtain visual features from images and mBERT [21] to retrieve textual features from questions for our baseline method. Both pre-trained ViT and mBERT models are initialized from HuggingFace[†] [‡] checkpoints.

Suppose that f_v and f_q are features extracted from images and questions, respectively. To model the ability to generate answers based on information in images for the given questions, we follow the architecture of the decoder of the transformer [22] to perform an attention mechanism for the answer generator component. In particular, the answer generator consists of N layers of the decoder module. Each decoder module contains three components: a multi-head attention to perform self-attention, a multi-head attention to perform cross-attention, and a feed-forward component [22]. Moreover, the target answers are shifted right in order that we want to train the baseline to have the ability to predict the next token based on previous tokens. In the inference step, the model begins to predict the first token when it receives the [GO] token as the starting signal and then sequentially generates the completed answers.

We trained the baseline with a batch size of 64 and adapted the learning rate scheduler from Vaswani et al. [22] to reduce the learning rate gradually after several iterations. The training process was interrupted automatically if the evaluation scores did not increase after 5 epochs.

---

https://huggingface.co/docs/transformers/model_doc/vit
https://huggingface.co/bert-base-multilingual-uncased

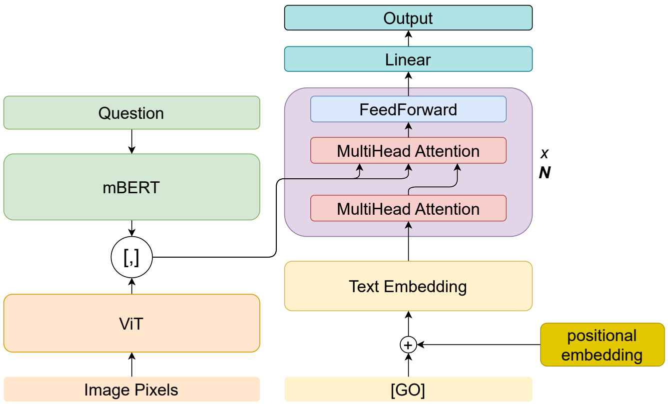


Figure 5: The baseline model architecture at the VLSP2022-EVJVQA Challenge

5.2. Challenge submission

The competition was hosted offline where the organizers provided the training set and public test set to the participant teams to evaluate and fine-tune their methods. When the competition came into the private test phase, the submission policy allowed each participant team to submit up to 3 *different* methods during a submission time lasting 3 days. After the three-day private test phase, we evaluated their submitted results and obtained their $F1$ score as well as average. BLEU score on the private test set. The final score of each participant team was the highest score among their submitted models on the private test set and their rank was indicated based on the $F1$ score (BLUE as a secondary score when there is a tie).

5.3. Winning results

Table 4: Final results of submitted methods

No.	Team name	Model type	Models	Public Test		Private Test	
				$F1$	BLEU	$F1$	BLEU
1	**CIST AI**	Single	ViT + mT5	0.3491	0.2508	**0.4392**	**0.4009**
2	**OhYeah**	Single	ViT + mT5	**0.5755**	**0.4866**	0.4349	0.3868
3	**DS-STBFL**	Ensemble	CNN-Seq2Seq + ViT + OFA	0.3390	0.2156	0.4210	0.3482
4	FCoin	Single	ViT + mBERT	0.3355	0.2437	0.4103	0.3549
5	VL-UIT	Single	BEiT + CLIP + Detectron-2 + mBERT + BM25 + FastText	0.3053	0.1878	0.3663	0.2743
6	BDboi	Ensemble	ViT + BEiT + SwinTransformer + CLIP + OFA + BLIP	0.3023	0.2183	0.3164	0.2649
7	UIT-squad	Ensemble	VinVL+mBERT	0.3224	0.2238	0.3024	0.1667
8	VC-Internship	Single	ResNet-152 + OFA	0.3017	0.1639	0.3007	0.1337
9	*Baseline*	Single	ViT + mBERT	*0.2924*	*0.2183*	*0.3346*	*0.2275*

There are 8 participant teams submitted in total. Only 5 of them have $F1$ scores higher than the baseline system. The results of the top-5 highest solutions are indicated in Table 4.

5.4. Winning solutions

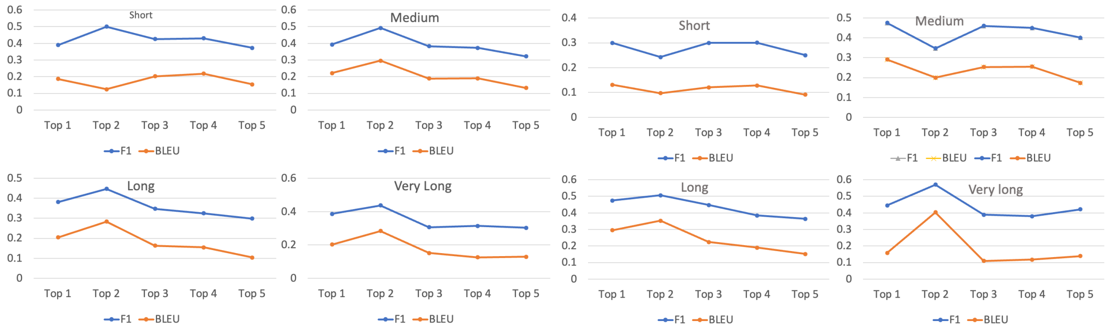
According to Table 4, 8 participant teams submitted their results to the private test phase.

As the requirement of the shared task as well as for academic purposes, all top-5 participants were asked to open their source code for result verification and write an academic paper. The ideas of two approaches among them are summarized as follows:

- **CIST AI:** CIST AI team proposed to use ViT [20] to extract visual features from the input images and fined-tuning mT5 [23] as the underlying model for multimodal learning. That is, they used ViT to obtain the visual features and then combine them with the textual features of questions (extracted through an embedding layer) to form the fused features. These features were then forwarded to the mT5 model to construct the answers as any typical text-to-text tasks. Specially, to enhance the ability to distinguish different languages of given questions to give appropriate answers, the CIST AI team proposed to prefix the given questions with the prompt "Answer in ⟨lang⟩ : ⟨question⟩" where "⟨lang⟩" is one of {"Vietnamese", "English", "Japanese"}, and "⟨question⟩" is the given question.
- **OhYeah**: OhYeah proposed a method based on PaLI [24]. At first they used polyglot to detect the language of questions. The detected languages were then encoded as language token, such as ⟨eng⟩ for English, ⟨vi⟩ for Vietnamese and ⟨jp⟩ for Japanese, and prefixed to the given questions as the signal which language the model should use to give answers. Same as CIST AI, ViT [20] was used to extracted visual features. These features were concatenated with prefixed questions to yield the input.
- **The DS-STBFL:** DS-STBFL team obtained the 3rd rank in the shared task by proposing an ensemble approach. In particular, the DS-STBFL team designed the CNN-Seq2Seq module to extract the linguistic features from questions. The visual features of images were achieved using ViT [20]. All extracted features from questions and images were then passed to OFA [25] to obtain the fused features for the answer generator. Moreover, the DS-STBFL team shared the same color as the CIST AI team where they prefixed the question prompts before forwarding through the CNN-Seq2Seq module. The difference from the CIST AI is that the prompt used by DS-STBFL is a list of detected objects in images. Particularly, tags of detected objects from ViT were listed and concatenated by colons to form a list and then prefixed to the questions.
- **VL-UIT**: VL-UIT team proposed to explore the advantage of pretrained models. In particular, they used BEiT [26] and CLIP [27] to extract features at patch level of images, and Detectron2 to extract features at pixel level. These visual features were concatenated to form the ultimate visual features for given images. To obtained the linguistic features of given questions, the combination of mBERT [21], BM25 [28] was used. Both visual features and linguistic features were then stacked together then passed through the encoder-decoder architecture proposed by [22].

6. RESULT ANALYSIS

According to Table 4 and Subsection 5.4, top-3 solutions all used prefix mechanism to preprocess questions before pass them into the model. While the VL-UIT although they take advantage of various pretrained models, their results are not impressive in spite of high-cost resources and training time. Moreover, none of top-3 solutions carefully conducts ablation study for the impact of prefix technique on their results, hence we analyse results solely based on their submitted answers.



(a) Questions (b) Answers

Figure 6: Results of the top-5 methods in English partition on the private test set

6.1. Linguistic complexity

Nguyen et al. [29] indicated that the linguistic complexity of a sentence can be measured through its length (in terms of number of tokens). In more detail, Nguyen et al. [29] proposed LCS and LLS algorithms to measure the linguistic complexity of a sentence and showed that they are proportional to the length of that sentence. Accordingly, we mainly analyze the results of methods coming from participating teams based on the length of questions and answers in each language. The categorization detail of this sentence-length analysis is elaborated specifically in the Appendix. We sequentially report the results of the top-5 models in terms of scores and visualization in quantitative analysis and qualitative analysis, respectively, to support our statements as well as indicate what the research community should pay attention to when constructing the novel open-ended VQA dataset.

From Figure 6, we have the top-5 models share the same characteristics on the English part of the UIT-EVJVQA dataset. Particularly, when we observe the results based on the question length, we can see that all submitted models have the same behavior when they give a higher performance on short and medium questions, while they yield a few drawbacks on long and very long questions.

Turn the attention to the results based on answer length, we have different behaviors. The top-5 models give better results on medium and long answers while yielding worse results on short and very long answers. Coming from Figure 4b, we can see that most English answers are short answers, which means models have more short answers for learning hence logically, they should have better performance on short answers rather than medium and long answers. To answer this weird insight, we showed all answers given by the top-5 models, and we found out that all top-5 models tend to give medium answers, even for questions having short answers. The most exciting thing here is how models give medium or lengthy answers: they repeat some tokens from the questions, and this is the primary way our annotators give medium and lengthy answers to questions.

For questions with medium or long answers, answers from top-5 models mostly repeat some tokens from questions, and the gold medium or gold answers also repeat some tokens from questions. Thanks to these matched tokens with questions, $F1$ scores, and average BLEU scores of the answers from top-5 models are pretty high, while the crucial tokens, which determine whether or not the information in these answers is correct or not, usually have a length of 2 or 3 tokens; in case of totally wrong, they still do not affect the overall scores significantly. To gain a better understanding, some samples are shown in the Appendix for an intuitive explanation.

Coming into the Vietnamese part of the UIT-EVJVQA dataset results, we have quite different

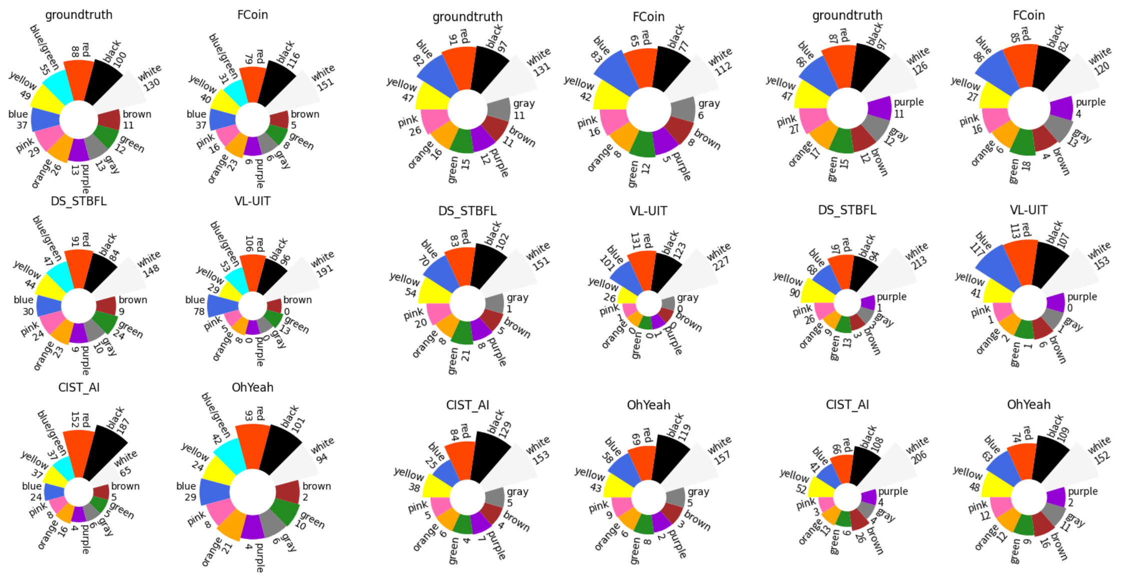
Figure 7: Results of the top-5 methods in Vietnamese partition on the private test set

Figure 8: Results of the top-5 methods in Japanese partition on the private test set in

behaviors where most models perform better on medium and long questions. When we observe the length distribution of questions in Figure 4a, most questions have a length range of around eight tokens, or most have medium length. While in English, most questions fell in the range of five and seven tokens. Therefore the top-5 models have a good performance on short and medium questions in English, while they achieved better performance on medium and long answers in Vietnamese. For Vietnamese answers, as indicated in Table 7, top-5 models share the same performance as the English answers: they perform better on medium and long answers than on short and very long answers. Some samples are presented in the Appendix to better demonstrate the method with English answers.

On the Japanese part of the UIT-EVJVQA dataset, the top-5 models have the same color on the English part when they give better results on medium and long questions. However, unlike their performance in Vietnamese and English, where they achieved better scores for medium and long answers, the top-5 models have its results as increase as the length of answers. From Figure 4c, although the occurrence of short answers is the highest, the cumulative of medium and long answers are higher than the cumulative of short answers, which indicates the top-5 models not only have more medium and lengthy answers to learn but also tend to give medium or long answers to have the optimal loss on the training set. Some samples are shown in the Appendix for better demonstration.

Apart from the previous discussion, we can conclude on the UIT-EVJVQA dataset that most deep learning models tend to give lengthy answers to given questions with images, and the way they give lengthy answers is by repeatedly using some tokens of questions as a starting point of the answers, and the main wrong parts of these answers are at the tokens indicating vital information to answer the questions such as objects, colors, or side (see Appendix for more examples). Hence, to better understand how wrong the top-5 model gives predictions, we conduct analyses focusing on the usage of side, object, and color words in each language.



(a) Vietnamese (b) English (c) Japanese

Figure 9: Color word distribution in ground truth and predictions of each team in three languages

6.2. Side words

One of the most confusing attributes while inferring the description of objects is the use of side or direction. This is the case where an object is observed on the left or right of another object or appears on the left or right side of the scene. According to the $F1$-scores of side word usage in the generated answers in Table 5, most models failed to exactly indicate the side of objects while giving answers to the questions.

Table 5: $F1$-score for side predictions of every team for each language

Team	Vietnamese	English	Japanese
CIST AI	**0.4948**	**0.3889**	0.3922
OhYeah	0.3814	0.3235	**0.4085**
DS-STBFL	0.4811	0.3137	0.3366
FCoin	0.4021	0.3039	0.3595
VL-UIT	0.2268	0.2418	0.1471

6.3. Color words

As an attribute frequently appears in the QAs of our dataset, color is worth being observed as part of the inference performance of the models. Here we measure the $F1$-score of color word matches between predicted and gold answers for every team. The color words have been defined in the guideline in Subsection 4.2.1. However, there are some minor exceptions due to the crowdsourcing process.

Table 6: $F1$-score for color predictions of every team for each language

Team	Vietnamese	English	Japanese
CIST AI	0.3230	0.3434	0.3450
OhYeah	0.2954	0.3326	0.3229
DS-STBFL	**0.5161**	**0.4933**	**0.3934**
FCoin	0.3619	0.3142	0.3234
VL-UIT	0.2716	0.2420	0.2355

We also visualize the color word distribution in the prediction of every team for each language. It is worth noting that in the Vietnamese QAs, there is a substantial quantity of "xanh" color words which are unclear to be whether green or blue, but rather depend on the context of the visual scene. As we can see in Figure 9a, Figure 9b, and Figure 9c, the ground truth color words are not evenly distributed because some colors such as white, black, and red are used more frequently in the dataset, while the instances of brown, gray, and purple are scant. This skewness is then emphasized through the overall inference of models. In Japanese, most submitted models clearly express bias as they intensively describe objects in white color, while less-appeared colors are poorly used. This kind of behavior is similar to the prior-language phenomenon pointed out in the previous work [18] when being given questions about colors, top-5 models tend to use most-appearance colors despite the colors in images, and this phenomenon is also noticeable in other languages for some teams. For instance, the submitted model of team VL-UIT even ignores many of the less-appeared colors. As an attribute is regularly used to describe an object, color also plays an important role in distinguishing between objects, which is an essential factor in visual question answering. Therefore, more efforts must be conducted to degrade the skewness in the color distribution in a way that helps the model infer better and more precisely describe objects out of images. We suggest future works should define a fixed range of colors and pay attention while asking and using colors to answer in order that we can ignore such prior-language phenomena on the VQA dataset.

6.4. Objects

We retrieve the distribution of objects in the UIT-EVJVQA dataset and investigate the behavior of top-5 models on perceiving these objects. We end up excluding the effect of the prior-language phenomenon associated with object words in the dataset vocabulary. Hence, we proceed with using pre-trained image models, such as Faster-RCNN and Cascade RCNN, to detect the objects in images as an object-specific way of approximately simulating the grid-based visual feature extraction of top-5 solutions. In this way, we find that there exists an incorrect image understanding of pre-trained image models trained on images captured outside of Vietnam which indirectly affects the overall performance of VQA models. In conclusion, available pre-trained image models are not relevant to scenes captured in Vietnam, and the Computer Vision (CV) community in Vietnam should research and develop a more appropriate pre-trained image model, especially for images taken in Vietnam so that we can effectively tackle recent trending tasks where multi-modeling task such as VQA is one of them.

7. CONCLUSION AND FUTURE WORK

The VLSP2022-EVJVQA Challenge on multilingual image-based question answering has been organized at the VLSP 2022. Even though 36 teams had legally signed up to get the training dataset, only eight teams submitted their results. Because several teams enrolled for many challenges at the VLSP 2022, the other teams might not have enough time to explore VQA models. The highest performances was 0.4392 in F1-score and 0.4009 in BLEU on the private test set. The multilingual VQA systems proposed by the top 2 teams used ViT for the pre-trained vision model and mT5 for the pre-trained language model. EVJVQA is a challenging dataset including the training set, the development set (public test set), and the test set (private test set) that motivates NLP and CV researchers to further explore the multilingual models or systems in visual question answering.

To increase performance in multilingual visual question answering, we intend to increase the amount and quality of annotated questions in the future. In addition, we also make human-adversarial questions based on findings proposed by the research work [30].

ACKNOWLEDGMENTS

We would like to thank the efforts of annotators who have contributed to the construction of a high-quality resource for the natural language processing research community. The VLSP 2022 was supported by organizations: Aimesoft, VBee, INT2, and DAGORAS, and educational organizations: VNU-HCM University of Information Technology, VNU University of Science, VNU University of Engineering and Technology, Hanoi University of Science and Technology, Vietnam Lexicography Centre, University of Science and Technology of Hanoi, ThuyLoi University, and VAST Institute of Information Technology. This work is partially supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2020.DA14.

REFERENCES

[1] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6700–6709.

[2] D. Gurari, Q. Li, A. J. Stangl, *et al.*, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3608–3617.

[3] A. Singh, V. Natarajan, M. Shah, *et al.*, "Towards vqa models that can read," in *Proceedings of The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8317–8326.

[4] S. Changpinyo, L. Xue, I. Szpektor, *et al.*, "Towards multi-lingual visual question answering," *arXiv preprint arXiv:2209.05401*, 2022.

[5] S. Antol, A. Agrawal, J. Lu, *et al.*, "VQA: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.

[6] A. K. Gupta, "Survey of visual question answering: Datasets and techniques," *arXiv preprint arXiv:1705.03865*, 2017.

[7] L. Qi, S. Lv, H. Li, *et al.*, "Dureadervis: A: A chinese dataset for open-domain document visual question answering," in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 1338–1351.

[8] N. Shimizu, N. Rong, and T. Miyazaki, "Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1918–1928.

[9] K. Q. Tran, A. T. Nguyen, A. T.-H. Le, and K. V. Nguyen, "ViVQA: Vietnamese visual question answering," in *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, Shanghai, China: Association for Computational Lingustics, Nov. 2021, pp. 683–691. [Online]. Available: https://aclanthology.org/2021.paclic-1.72.

[10] D. Gupta, P. Lenka, A. Ekbal, and P. Bhattacharyya, "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*, 2020, pp. 900–913.

[11] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[12] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, pp. 2383–2392.

[13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of The Association for Computational Linguistics*, 2002, pp. 311–318.

[14] K. Q. Tran, A. T. Nguyen, A. T.-H. Le, and K. Van Nguyen, "ViVQA: Vietnamese visual question answering," in *Proceedings of The 35th Pacific Asia Conference on Language, Information and Computation*, 2021, pp. 546–554.

[15] M. Ren, R. Kiros, and R. Zemel, "Exploring models and data for image question answering," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[16] T. Vu, D. Q. Nguyen, D. Q. Nguyen, M. Dras, and M. Johnson, "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*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 56–60. DOI: 10.18653/v1/N18-5012. [Online]. Available: https://aclanthology.org/N18-5012.

[17] D. Teney, P. Anderson, X. He, and A. Van Den Hengel, "Tips and tricks for visual question answering: Learnings from the 2017 challenge," in *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4223–4232.

[18] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "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*, 2017, pp. 6904–6913.

[19] S. Changpinyo, L. Xue, I. Szpektor, *et al.*, *Towards multi-lingual visual question answering*, 2022. DOI: 10.48550/ARXIV.2209.05401. [Online]. Available: https://arxiv.org/abs/2209.05401.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[22] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[23] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[24] X. Chen, X. Wang, S. Changpinyo, *et al.*, "Pali: A jointly-scaled multilingual language-image model," *arXiv preprint arXiv:2209.06794*, 2022.

[25] P. Wang, A. Yang, R. Men, *et al.*, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *International Conference on Machine Learning*, PMLR, 2022, pp. 23 318–23 340.

[26] H. Bao, L. Dong, and F. Wei, "Beit: Bert pre-training of image transformers," *arXiv preprint arXiv:2106.08254*, 2021.

[27] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[28] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Found. Trends Inf. Retr.*, vol. 3, no. 4, pp. 333–389, Apr. 2009, ISSN: 1554-0669. DOI: 10.1561/1500000019. [Online]. Available: https://doi.org/10.1561/1500000019.

[29] N. H. Nguyen, D. T. Vo, K. Van Nguyen, and N. L.-T. Nguyen, "OpenViVQA: Task, dataset, and multimodal fusion models for visual question answering in Vietnamese," *arXiv preprint arXiv:2305.04183*, 2023.

[30] S. Sheng, A. Singh, V. Goswami, *et al.*, "Human-adversarial visual question answering," *Advances in Neural Information Processing Systems*, vol. 34, pp. 20 346–20 359, 2021.

Received on March 08, 2023
Accepted on August 22, 2023

APPENDIX

A Example model demonstration

In this section, we provide extensively samples (Figure 10, 11, 12, 13, and 14) from top-5 models on the three languages for better demonstrating our observation in Section 6.

what is the girl walking holding on her hand?
Human: she is holding a disk of food
Top 1: she is holding a handbag
Top 2: the girl is holding a bouquet
Top 3: the girl is holding a milk
Top 4: she is holding a camera
Top 5: the girl is holding a bouquet

how many girls in white shirts are standing at the cashier?
Human: two girls in white shirts are standing at the cashier
Top 1: there are two girls in white shirts standing at the cashier
Top 2: there are two girls in white shirts standing at the cashier.
Top 3: there are two girls in white shirts
Top 4: there are two girls in white shirts
Top 5: there is one man standing in the stall

how many people are on the advertisement board of the supermarket?
Human: there are three people on the advertisement board.
Top 1: there are two people standing on the advertisement board
Top 2: there are two people on the advertisement board of the supermarket
Top 3: there are three people on the wall is a person
Top 4: five people
Top 5: there are two people on the left of the woman

what stall are the children standing at?
Human: the children are standing at the children's toy stall
Top 1: the children are standing at the vegetable stall
Top 2: standing at the lantern stall
Top 3: the children are standing at the dairy stall
Top 4: the broom stall
Top 5: standing at the checkout stall

Figure 10: Results of top-5 models on medium answers in English of the UIT-EVJVQA dataset

what does a flag every house has on the front look like?
Human: every house has a red flag with a yellow star in front
Top 1: the flag has a red background and a yellow star
Top 2: a red background and a yellow star
Top 3: a red shirt and a flag
Top 4: the red flag
Top 5: the lanterns

which hand is the girl using to holding the hand of the man?
Human: the girl is using left hand to holding the hand of the man
Top 1: the girl is using her right hand to holding the hand of the man
Top 2: the girl is using her right hand to holding the hand of the man
Top 3: the girl is using right hand to take the hand of the man
Top 4: the girl is holding the hand of the man
Top 5: the girl is holding the left hand of the girl in white ao dai

where is the red charging light hung?
Human: the red charging light is hung on the parasol on the left
Top 1: the red charging light is hung at the checkout of the supermarket
Top 2: the red charging light is hung on the left side of the road
Top 3: the red chair is on the ceiling of the ceiling
Top 4: above a grocery store
Top 5: the market is very crowded with the market

how many motorbikes are parked in front of the rice store?
Human: there are two motorbikes parked in front of the rice store
Top 1: there is one motorbike parked in front of the rice store
Top 2: here is one motorbike parked in front of the rice store
Top 3: there are two motorbikes parked in front of the rice store
Top 4: there is one motorbike parked in front of the rice store
Top 5: there are two motorbikes parked in front of the road

Figure 11: Results of top-5 models on long answers in English of the UIT-EVJVQA dataset

lá cờ trên cánh cổng có màu gì? *(what is the color of the flag on the gate?)*
Human: lá cờ trên cánh cổng có màu đỏ *(the flag on the gate is red)*
Top 1: lá cờ có màu đỏ *(the flag is red)*
Top 2: lá cờ có màu đỏ *(the flag is red)*
Top 3: lá cờ có màu đỏ *(the flag is red)*
Top 4: lá cờ có màu đỏ *(the flag is red)*
Top 5: lá cờ có màu đỏ *(the flag is red)*

cô gái mang khăn choàng cổ đang mặc áo màu gì? *(what is the color of the outfit the that girl having a scarf is wearing?)*
Human: cô ấy mặc áo màu đỏ *(she is in red)*
Top 1: cô gái mang khăn choàng cổ đang mặc áo màu đen *(the girl wearing a scarf is in black)*
Top 2: cô gái mang khăn choàng cổ đang mặc áo mà *(the girl wearing a scarf is wearing a shirt that)*
Top 3: cô gái mang khăn màu đỏ *(she is wearing a red scarf)*
Top 4: màu hồng *(pink one)*
Top 5: cô ấy mang áo màu đỏ *(she is wearing a red shirt)*

ngôi nhà nằm ở phía bên nào của con đường? *(which side of the road was the house built?)*
Human: ngôi nhà nằm ở phía bên trái của con đường *(the house was built in the left of the road)*
Top 1: ngôi nhà nằm bên trái con đường *(the house was built in the left of the road)*
Top 2: ngôi nhà nằm bên trái con đường *(the house was built in the left of the house)*
Top 3: ngôi nhà nằm bên phải con đường *(the house was built in the right of the road)*
Top 4: ngôi nhà nằm bên trái con đường *(the house was built in the left of the road)*
Top 5: con chó ở bên phải con đường *(the dog is in the right of the road)*

người ta sơn màu gì cho cây trụ của bảng tên đường? *(what is the color of the pillars of the street sign?)*
Human: người ta sơn màu đỏ và trắng *(the pillars is in red and white)*
Top 1: cây trụ của bảng tên đường được sơn màu đỏ *(the pillars is in red)*
Top 2: người ta sơn màu đỏ cho cây trụ của bảng *(they painted the pillars with red)*
Top 3: người ta sơn màu trắng cho cây trụ của bảng tên đường *(they painted the pillars of the street sign with white)*
Top 4: sơn màu trắng *(painted with white)*
Top 5: màu xanh lá cây *(in green)*

Figure 12: Results of top-5 models on medium answers in Vietnamese of the UIT-EVJVQA dataset

B Detailed result analysis implementation

Sentence length: To alleviate the result analysis based on sentence length, we define the four different ranges for the length of questions and answers. In particular, we define the *short questions* (respective, *short answer*) are questions whose length is shorter or equal to 5 tokens, *medium questions* (respective, *medium answer*) questions whose length is between 6 to 10 tokens, *long questions* (respective, *long answer*) are questions whose length is between 11 to 15 tokens, and finally, *very long questions* (respective, *very long answer*) are questions whose length is greater than 15 tokens. Tokens of questions and answers are defined in the same manner when we do statistics in Section 4.4.

Side words: To measure the correctness of experimental methods in side words prediction, we observe the matching of words that depict sides between the generated answers and the gold answers

có bao nhiêu hình minh hoạ món ăn được dán trên biển hiệu nhà hàng? *(how many food illustrating pictures were stuck on the signboard of the restaurant?)*
Human: có ba hình minh hoạ được dán trên biển hiệu nhà hàng *(there are three illustrating pictures stuck on the signboard of the restaurant)*
Top 1: có hai hình minh hoạ món ăn được dán trên biển hiệu nhà hàng *(there are two food illustrating pictures stuck on the signboard of the restaurant)*
Top 2: có ba hình minh hoạ món ăn *(there are three food illustrating pictures)*
Top 3: có ba hình *(there are three pictures)*
Top 4: có một hình minh hoạ món ăn được dán trên biển hiệu nhà hàng *(there was a food illustrating picture stuck on the signboard of the restaurant)*
Top 5: có hai món ăn được gắn trên bàn *(there are two food illustrating pictures stuck on the signboard of the restaurant)*

các sạp bán đồ ở hai bên đường đang trưng bày những mặt hàng gì? *(what kind of products does these stalls displaying?)*
Human: các sạp bán hàng này trưng bày chủ yếu các loại quần áo *(these stalls are mainly selling clothes)*
Top 1: các sạp bán đồ ở hai bên đường đang trưng bày các loại quần áo *(these stalls along the road are displaying clothes)*
Top 2: bán các loại quần áo và vải vóc *(selling various kinds of clothes and silks)*
Top 3: các sạp hàng ở đây bán quần áo *(these stalls are selling clothes)*
Top 4: bán các loại quần áo *(selling clothes)*
Top 5: bán các loại quần áo *(selling clothes)*

bên cạnh người khách là thứ gì? *(what is next to the customer?)*
Human: bên cạnh là một chiếc xe đẩy hàng của siêu thị *(next to the customer is a shopping cart)*
Top 1: bên cạnh người khách là một chiếc xe đẩy *(next to the customer is a shopping cart)*
Top 2: bên cạnh người khách là một siêu thị *(next to the customer is a supermarket)*
Top 3: bên cạnh người khách là một tấm bảng *(next to the customer is a signboard)*
Top 4: bên cạnh người khách là những hộp sữa *(next to the customer is bottles of milk)*
Top 5: bên trái là người phụ nữ mặc áo xanh dương *(in the left there is a woman in blue)*

hai bên con phố này người ta bày bán những mặt hàng gì? *(what are being sold in both side of this road?)*
Human: người ta bày bán chủ yếu các loại vải, quần áo *(they are selling kinds of silks and clothes)*
Top 1: người ta bày bán các loại quần áo *(they are selling kinds of clothes)*
Top 2: người ta bày bán các loại quần áo *(they are selling kinds of clothes)*
Top 3: bán các loại trái cây *(selling kinds of fruits)*
Top 4: những mặt hàng thực phẩm *(kings of foods)*
Top 5: bán các loại quần áo *(selling clothes)*

Figure 13: Results of top-5 models on long answers in Vietnamese of the UIT-EVJVQA dataset

specifically in each language. In English, side words are the words "left" and "right" in our dataset. While in Japanese, they are "左" and "右". However, they are not simply "trái" and "phải" in Vietnamese, as "trái" may not essentially be used to convey a side. Therefore, we broaden the set of side words as "bên trái", "bên phải", "tay trái" (left-hand side) and "tay phải" (right-hand side) according to our observation of the dataset. To this end, we adopt the F1-score to calculate the proportion of match in terms of the above side words between the generated answers and the gold answers.

Objects: To get the distribution of objects in the UIT-EVJVQA dataset, we used the POS method from the previous work [16] and achieved objects by collecting tokens tagged as nouns. We used the same manner as investigating the behavior of top-5 models on color words but we can not find such language-prior phenomenon on objects.

服装を選ぶ女性はどれくらいいますか? *(how many women choose their clothes?)*
Human: 服装を選ぶ女性が2人います *(there are two women who choose clothes)*
Top 1: 服装を選ぶ女性は1人います *(there is one woman who chooses the clothes)*
Top 2: 服装を選ぶ女性は一人います *(there is one woman who chooses the clothes)*
Top 3: 赤いシャツを着た女性がいます *(there is a woman in a red shirt)*
Top 4: いない *(not in)*
Top 5: 黄色い帽子をかぶった女性がいます *(there is a woman wearing a yellow hat)*

ガイドは何色のシャツを着ていますか? *(what color shirt do the guides wear?)*
Human: ガイドは赤いシャツを着ています *(the guide wears a red shirt)*
Top 1: 赤いシャツを着ています *(wearing a red shirt)*
Top 2: 白いシャツを着ています *(wearing a white shirt)*
Top 3: 白いシャツを着ています *(wearing a white shirt)*
Top 4: 緑色 (green)
Top 5: 青いシャツを着ています *(wearing a blue shirt)*

男の人は何をしていますか? *(what is the man doing)*
Human: 彼は会議で話しています *(he is speaking at a meeting)*
Top 1: 本を読んでいます *(reading a book)*
Top 2: 男の人は看板を作っています *(man is making a sign)*
Top 3: 男のために商品をチェックしています *(checking products for men)*
Top 4: 男の人は報道陣に話しています *(man is talking to the press)*
Top 5: 男の人は電話を持っています *(man has a phone)*

配達人は何をしていますか? *(what does the delivery person do?)*
Human: 女の人に配達物を渡している *(giving a delivery to a woman)*
Top 1: 配達人は商品を紹介しています *(the delivery person introduces the product)*
Top 2: 配達人はバイクを押しています *(the delivery man is pushing a motorcycle)*
Top 3: 道路をしています *(doing the road)*
Top 4: 配達人はバイクに乗っています *(the delivery man is riding a motorcycle)*
Top 5: 商品を紹介しています *(Introducing products)*

このカップルは、一緒に座るためにどの位置を選択しますか? *(which position is this couple sitting together?)*
Human: 彼らはカフェの前の道端に座っています *(they are sitting by the roadside in front of the cafe)*
Top 1: 彼らは一緒に座るために、木の下の椅子を選択します *(they choose a chair under the tree to sit together)*
Top 2: 通りの真ん中 *(middle of the street)*
Top 3: このエリアはスーパーマーケットに座っています *(this area sits in a supermarket)*
Top 4: 通りの真ん中 *(middle of the street)*
Top 5: 女の子は自転車の前に座っています *(girl sits in front of a bicycle)*

女の子が手に持っているのは何ですか? *(what is the girl holding in her hand?)*
Human: 女の子はカメラを手に持っています *(the girl holds the camera in her hand)*
Top 1: 女の子はバッグを持っています *(girl has a bag)*
Top 2: 女の子は手に持っているのは一つのカメラです *(the girl has a camera in her hand)*
Top 3: 女の子が手に持っているのは、椅子です *(the girl is holding a chair)*
Top 4: お香の束 *(bundle of incense)*
Top 5: 女の子は左手に持っています *(the girl holds in her left hand)*

この店はどのような機会に装飾品を販売していますか? *(on what occasions does this store sell decorations?)*
Human: このお店は旧正月の飾りを販売しています*(this shop sells chinese new year decorations)*
Top 1: クリスマス *(Christmas)*
Top 2: 旧正月 *(lunar new year)*
Top 3: この店は、あらゆる種類の物資を販売しています *(this store sells all kinds of goods)*
Top 4: クリスマス *(christmas)*
Top 5: この店はクリスマスの飾りを売ります*(this store sells christmas decorations)*

チェックのズボンの女性はどこに立っていますか? *(where is the lady in the checked trousers standing?)*
Human: 彼女は黒い服を着た2人の間に立っています *(she stands between two people dressed in black)*
Top 1: 彼女は花畑の隣に立っています*(she is standing next to the flower garden)*
Top 2: チェックのズボンの女性は赤いシャツの女性の後ろに立っています*(a woman in checked trousers stands behind a woman in a red shirt)*
Top 3: スーパーのズボンの女性は、黒いシャツを着ています*(woman in super pants wears a black shirt)*
Top 4: 黒い服を着た女 性の後ろに立っています *(standing behind a woman dressed in black)*
Top 5: 彼女はクリスマスの飾りの飾りの飾りの店の中に立っています *(she is standing inside a store of christmas ornaments ornaments ornaments)*

Figure 14: Results of top-5 models on long answers and very long answers in Japanese of the UIT-EVJVQA dataset