A Gaze-grounded Visual Question Answering Dataset for Clarifying Ambiguous Japanese Questions

Shun Inadumi[1,2], **Seiya Kawano**[2,1], **Akishige Yuguchi**[3,2]
Yasutomo Kawanishi[2,1], **Koichiro Yoshino**[2,1]
[1] Nara Institute of Science and Technology, Nara, Japan
[2] Guardian Robot Project, RIKEN, Kyoto, Japan [3] Tokyo University of Science, Tokyo, Japan
inazumi.shun.in6@naist.ac.jp, akishige.yuguchi@rs.tus.ac.jp
{seiya.kawano,yasutomo.kawanishi,koichiro.yoshino}@riken.jp

Abstract

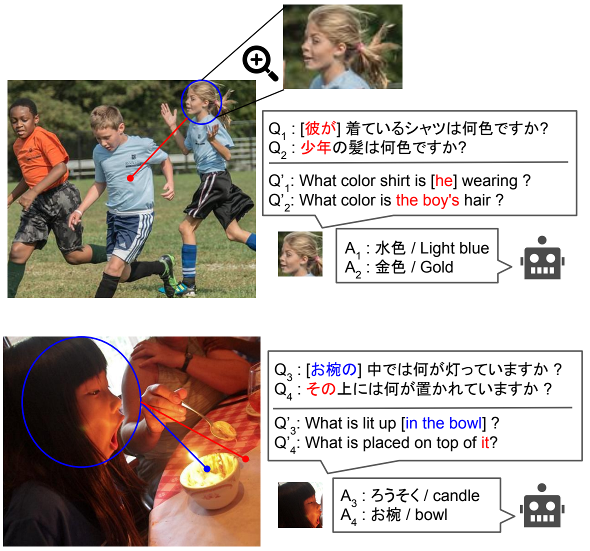
Situated conversations, which refer to visual information as visual question answering (VQA), often contain ambiguities caused by reliance on directive information. This problem is exacerbated because some languages, such as Japanese, often omit subjective or objective terms. Such ambiguities in questions are often clarified by the contexts in conversational situations, such as joint attention with a user or user gaze information. In this study, we propose the Gaze-grounded VQA dataset (GazeVQA) that clarifies ambiguous questions using gaze information by focusing on a clarification process complemented by gaze information. We also propose a method that utilizes gaze target estimation results to improve the accuracy of GazeVQA tasks. Our experimental results showed that the proposed method improved the performance in some cases of a VQA system on GazeVQA and identified some typical problems of GazeVQA tasks that need to be improved.

Keywords: Visual Question Answering, Gaze Information, Object Grounding, Human-Robot Interaction

1. Introduction

The development of interactive systems that can collaborate with humans by taking into account real-world information is one ultimate goal of vision-and-language research. Such systems should understand the given visual information to respond users based on their results. Visual question answering (VQA) (Antol et al., 2015; Goyal et al., 2017; Shimizu et al., 2018) and visual dialog (Das et al., 2017; Agarwal et al., 2020) have been proposed to achieve this goal.

VQA tasks generally assume a situation where the intention of questions is clear in visual contexts, and systems can uniquely answer them. However, in actual interaction with humans, human utterances contain various ambiguities (Taniguchi et al., 2019; Sugiyama et al., 2007). A typical problem is exemplified by directives. To properly understand questions that contain directives, the directive's destination must be grounded in the real world. For example, "Could you pass it to me?" might have numerous interpretations because of the directive, "it." Some languages, such as Japanese, feature the ellipsis of such topical terms as subject and object in addition to the occurrence of indicative words (Seki et al., 2002; Sasano et al., 2008). Referring to real-world information is one key idea to resolve the ambiguity caused by directives and ellipses. For example, speaker's gaze (Emery, 2000), speaker's pointing (Nakamura et al., 2023), and joint attention (Rocca et al., 2018) are important cues for clarifying the target of ellipses and directives.



Q_1 : [彼が] 着ているシャツは何色ですか?
Q_2 : 少年の髪は何色ですか?
Q'_1: What color shirt is [he] wearing ?
Q'_2: What color is the boy's hair ?

A_1 : 水色 / Light blue
A_2 : 金色 / Gold

Q_3 : [お椀の] 中では何が灯っていますか ?
Q_4 : その上には何が置かれていますか ?
Q'_3: What is lit up [in the bowl] ?
Q'_4: What is placed on top of it?

A_3 : ろうそく / candle
A_4 : お椀 / bowl

Figure 1: Examples of questions and answers for GazeVQA proposed in this research: Square brackets denote omitted gaze target names. Multiple target points are assigned that correspond to source points.

In this research, we address the problem of ambiguity in human questions, especially when it refers to gaze information. We propose a Gaze-grounded VQA dataset (GazeVQA) that includes Japanese questions and gaze information[1]. GazeVQA assumes a situation where a speaker in

[1] Our dataset is publicly available at https://github.com/riken-grp/GazeVQA.

LREC-COLING 2024, pages 558–571
20-25 May, 2024.

an image asks to a system an ambiguous question that may contain directives or abbreviations, and a system answers it taking into account the speaker's gaze information. For example, since there are two boys in the upper image in Figure 1, the following question from the girl is ambiguous: ”What color is the boy's hair?” However, if the system knows the girl's gaze information, it can clarify the ambiguity of the question and answer the question. We collected questions and answers by focusing on speaker's gaze targets by crowdsourcing on the MS-COCO (COCO) subset derived object recognition image dataset (Lin et al., 2014) in Gazefollow (Recasens et al., 2015). We collected questions that are difficult to answer without information about speaker's gaze targets and required that the workers not mention the names of the gaze target objects when they created their questions. As a result, GazeVQA contains 17,276 QA pairs for 10,760 images, of which 1,680 were used as the test-set. To ensure diverse answers, we assigned ten answers to each question in the GazeVQA test-set. Our primary contribution is the construction of GazeVQA.

In addition, we propose a model that accurately answers ambiguous questions using gaze information. Existing vision-and-language models can take a target image and a question about it as input and generate an answer (Cho et al., 2021; Mokady et al., 2021). In this research, we investigate whether models can improve QA accuracy using areas highlighted by gaze information. A study on segmentation using text and images as prompts (Lüddecke and Ecker, 2022) is related to our idea. Inspired by this work, we added an adapter consisting of linear layers (Dumoulin et al., 2018) to a baseline (Mokady et al., 2021) consisting of a pre-trained image encoder (Radford et al., 2021) and a text decoder (Radford et al., 2019). We proposed a method for integrating a regions of interest (RoI) that represent gaze targets into the whole image with adapters. We used an existing gaze target estimation model for the estimated a RoI (Chong et al., 2020).

In experiments, we pre-trained a baseline and the proposed models with a Japanese caption dataset (Yoshikawa et al., 2017) and a Japanese VQA dataset (Shimizu et al., 2018) and fine-tuned them on our GazeVQA dataset. In the experimental conditions, we compared the results with and without gaze information in the adapter (ground-truth RoI and estimated RoI). Our experimental results found that using gaze information improved the GazeVQA's performance in some cases. Our second contribution is a proposal of a model that integrates gaze information.

2. Related Work

2.1. Visual Question Answering with Contextual Information

Visual Question Answering (VQA) is a task where the system derives answers to questions about images (Antol et al., 2015; Goyal et al., 2017; Shimizu et al., 2018). Since this study targets questions that are ambiguous without gaze information, we intentionally collected questions that did not include the names of gaze objects.

Previous works proposed VQA datasets that contain a variety of contextual information in addition to images and questions. A visual dialog provides accurate answers to ambiguous questions that arise during dialogues (Das et al., 2017; Agarwal et al., 2020). To answer questions, the previous dialog history is used as a supplement. VQA-HAT (Das et al., 2016) and VQA-MHUG (Sood et al., 2021) improved the accuracy of VQA tasks using a saliency map. By incorporating the answerer's subjective gaze information generated while solving the question with the VQA model, both works grounded fine-grained visual and linguistic representations. Point and Ask (Mani et al., 2020) employed pointing information to answer ambiguous questions that contain directives. They used paintings to ground directives in questions and the objects in images. In our study, we exploit gaze information to answer ambiguous questions.

There are two differences between these previous works and our work on the GazeVQA. First, we use the questioner's gaze information from the images as additional contextual information. Second, GazeVQA questions contain not only directives but also Japanese subject and object ellipsis.

Some research that uses gaze information is based on the gazing information of a user looking at an image (Ilaslan et al., 2023). In this research, we assume applications such as robots that need to understand the situation from a third person view by extracting the questioner from the image.

2.2. Gaze Target Estimation

Gaze target estimation predicts a person's gaze target from a head image. Gazefollow (Recasens et al., 2015) is a gaze target estimation dataset that is annotated with the sources and target points as gaze information. Gazefollow covers people collected from various image datasets, including COCO (Lin et al., 2014). Another work maps gaze information to objects for images taken in retail environments (Tomas et al., 2021). In this research, we constructed GazeVQA based on Gazefollow since no special environments are assumed. Here the gaze destinations in Gazefollow do not nec-

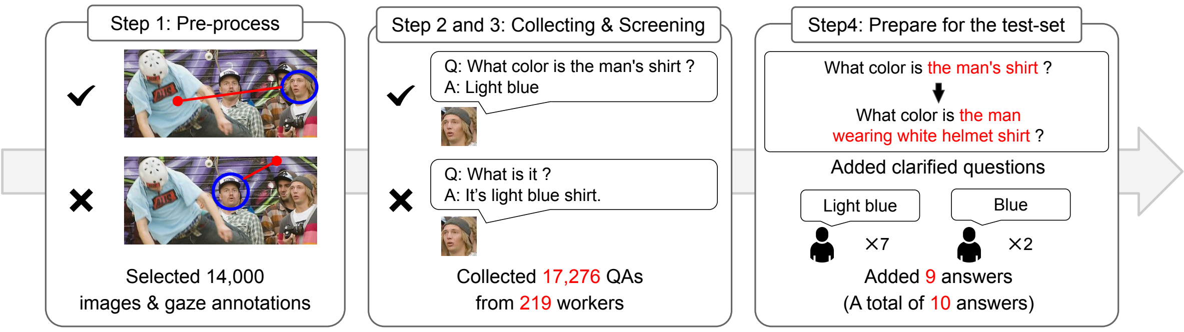

Step 1: Pre-process
✓
✗
Selected 14,000
images & gaze annotations
Step 2 and 3: Collecting & Screening
Q: What color is the man's shirt ?
A: Light blue
Q: What is it ?
A: It's light blue shirt.
Collected 17,276 QAs
from 219 workers
Step4: Prepare for the test-set
What color is the man's shirt ?
What color is the man
wearing white helmet shirt ?
Added clarified questions
Light blue
Blue
×7
×2
Added 9 answers
(A total of 10 answers)

Figure 2: Data collection process of our Gaze-grounded VQA dataset

essarily refer to gaze targets; nor do they specify their names. Therefore, we collected questions and answers about gaze objects based on the object annotations in the COCO subset of Gazefollow. For the actual gaze target estimation, we used the head image of the person associated with the gaze source (Chong et al., 2018).

3. Gaze-grounded VQA Dataset

In this section, we describe our proposed Gaze-grounded VQA dataset (GazeVQA). As in the case of VQA, GazeVQA's task is to answer questions about the given image. However, the questions in GazeVQA contain ambiguities and require consideration of the gaze information from the person in the image. We first describe GazeVQA's task settings and then explain the data collection process. We also describe GazeVQA's statistics.

3.1. Task Setting

We consider a case where questions without contextual information are given by a speaker from the system's first-person view. Models must clarify any ambiguity using the estimated region of interest (RoI) that represents gaze target. The main task is defined as follows:

GazeVQA task: Given question q, corresponding image I, and a RoI I_s, the task outputs answer $\boldsymbol{y}$.

Our data also contain the ground-truth RoI, which is the COCO bounding box; however, we assume that this RoI is not given in a real task. In this case, the system also needs to solve the following gaze target estimation task, which is defined as follows to obtain I_s:

Gaze target estimation task: Given image I and speaker's head image I_h, the task outputs I_s.

3.2. Data Collection

Figure 2 shows the process of constructing GazeVQA. We collected questions and answers for images by crowdsourcing[2]. We used images in the COCO subset of Gazefollow to acquire gold labels of the gaze sources and destinations. The specific procedure is described below.

Step 1: Selection of images and gaze information: We selected 14,000 pairs of image and gaze information and excluded the following cases: those in which the gaze destinations do not point to objects and those in which the gaze destinations point outside of the image. We used COCO's object segmentation for judgments. If the gaze destinations do not point to object segmentation, this gaze information is removed.

Step 2: Collecting questions and answers: We collected 26,296 questions and answers through crowdsourcing. Workers wrote questions and answers about gaze targets based on images with gaze information and object labels in COCO. However, if the gaze targets could not be confirmed due to image blur, we asked them not to create questions and answers for such targets.

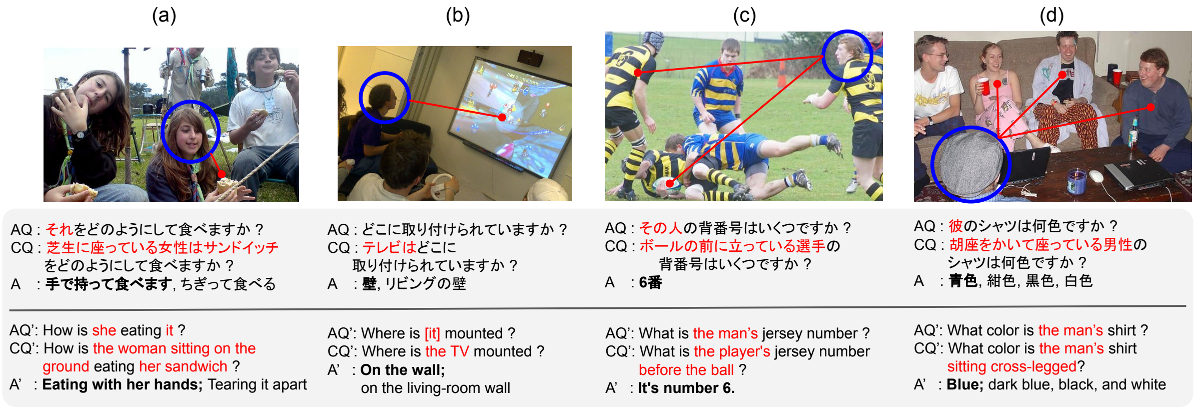
The workers were given the following instructions:

- Make the questions at least ten Japanese characters long.
- Do not include the names of the gaze target objects in the questions.
- Create questions that can be answered using only the image.

We designed the first and second instructions to create a variety of ambiguous questions that require gaze information. We designed the third instruction to exclude from GazeVQA any questions that are ambiguous outside of the image content. For example, such a question as "What will he do after this?" is not covered in this research because it requires some inference.

Step 3: Screening of questions and answers: Since the raw crowdsourcing results are noisy, we

[2] https://crowdworks.jp/



(a)
AQ : それをどのようにして食べますか？
CQ : 芝生に座っている女性はサンドイッチ
をどのようにして食べますか？
A : **手で持って食べます**, ちぎって食べる

AQ': How is she eating it ?
CQ': How is the woman sitting on the
ground eating her sandwich ?
A' : **Eating with her hands;** Tearing it apart

(b)
AQ : どこに取り付けられていますか？
CQ : テレビはどこに
取り付けられていますか？
A : **壁**, リビングの壁

AQ': Where is [it] mounted ?
CQ': Where is the TV mounted ?
A' : **On the wall;**
on the living-room wall

(c)
AQ : その人の背番号はいくつですか？
CQ : ボールの前に立っている選手の
背番号はいくつですか？
A : **6番**

AQ': What is the man's jersey number ?
CQ': What is the player's jersey number
before the ball ?
A' : **It's number 6.**

(d)
AQ : 彼のシャツは何色ですか？
CQ : 胡座をかいて座っている男性の
シャツは何色ですか？
A : **青色**, 紺色, 黒色, 白色

AQ': What color is the man's shirt ?
CQ': What color is the man's shirt
sitting cross-legged?
A' : **Blue;** dark blue, black, and white

Figure 3: Examples of GazeVQA test-set: AQ and answers in bold denote ambiguous questions and answers obtained through Step 3. CQ denotes questions clarified by annotator's work. The original questions and answers are given in Japanese. We put English translation in the bottom. Words denoted by square brackets are supplements in translations; the terms are omitted in the original Japanese questions.

need to screen them. In the process of entering questions in Step 2, we placed and used a bonus question: "Do not enter any text in this field." As a post-processing step, we manually checked the unnaturalness of the questions for workers who answered the bonus question. We excluded the annotation of 27 workers (from the original 246) whose annotated questions were repetitive or too vague. We selected 17,276 questions and answers to exclude unnatural questions. We call this question set ambiguous questions (AQs).

Step 4: Preparation for test-set: GazeVQA questions are associated with one or more of the 80 types of gaze objects. We divided the GazeVQA train/valid/test-set into 13,785/1,811/1,860 (0.8 : 0.1 : 0.1).

We expanded the test-set to ensure a variety of answer sets and assigned ten answers to the test-set questions, following a previous work (Antol et al., 2015; Goyal et al., 2017). Nine workers created additional answers for each question in the test-set. Each worker was given only gaze sources and ambiguous questions and answered without gaze destinations and names of the gaze target objects.

We also added a question to clarify each test-set, which contains the names or characteristics of the gaze targets from a single annotator. The annotator referred to the questions, the answers, and the gaze information. We called these clarified questions (CQs).

3.3. Example

Figure 3 shows a few examples included in the GazeVQA test-set. These questions suffer from ambiguities due to both directives (Fig. 3 (a)) and ellipsis peculiar to Japanese (Fig. 3 (b)). The questions are determined as answers based on the content of the questions, even if the gaze targets consist of more than two candidate objects (Fig. 3 (c)). However, some questions are too vague, where the answers are inconsistent with gaze information (Fig. 3 (d)).

3.4. Statistics and Analysis

We compared the statistics of GazeVQA and the Japanese VQA dataset (VQA-ja) (Shimizu et al., 2018) to highlight the former's characteristics.

Table 1 shows the statistics of these dataset. The percentage of unique questions in GazeVQA (46.46%) exceeds that in VQA-ja (45.21%), and its average length of questions is also slightly longer. The percentage of unique answers in GazeVQA (33.87%) is larger than that in the Japanese VQA (17.10%), and its average length of answers is also slightly longer. This is because the GazeVQA questions assumed supplemental information acquired by gaze information in addition to the question itself.

Table 2 shows the typology of question types included in GazeVQA. The percentage of "what" types is 81.85%, which is about 10% higher than the percentage of the VQA-ja (Shimizu et al., 2018). GazeVQA includes many questions that ask about the attributes of the gaze target object, such as color and shape, because the gaze target was the question's subject. The percentage of "where" types about the location of objects and "how" types about the number of objects was 12.04% in total. GazeVQA also contains other types of questions, including "when" types that ask about time and "who" types that ask about a person.

Table 1: Statistics on GazeVQA and Japanese VQA (VQA-ja) (Shimizu et al., 2018)

	GazeVQA	VQA-ja
Images	10,760	99,208
Question and answers	17,276	793,664
Unique questions	8,628	358,844
Unique answers	5,853	135,743
Avg. question length	15.37	14.82
Avg. answer length	4.92	4.56

Table 2: Typology of question types for GazeVQA

Types (Keywords in Japanese)	#Counts
What (*nani*, *dono*, *donna*)	14,141
is/are/do/does	7,215
color	3,626
condition	1,240
kind	903
shape	703
others	454
Where (*doko*)	1,085
How (*dore*, *ikutsu*)	996
Which (*dochira*)	295
Others (*itsu*, *dare*, *naze*)	875

Table 3 shows the frequency of arguments in the predicate-argument structure[3] in the GazeVQA test-set. Ambiguous questions in the GazeVQA test-set often result in the ellipsis of nominative and accusative cases, related to subjects and objects in questions, compared with clarified questions. This result suggests that GazeVQA contains questions in which the nominative and accusative cases are omitted, which is often in Japanese. For example, Fig. 3(b) is a typical example of the ellipsis of the nominative case.

4. Methodology

In this section, we first describe ClipCap (Mokady et al., 2021), which is our baseline for the GazeVQA task, and next describe our proposed model, "ClipCap + Adapter," which adds adapters (Dumoulin et al., 2018) to ClipCap. Finally, we explain the procedure for obtaining the region of interest (RoI) of the gaze targets in the gaze target estimation task.

4.1. Baseline Model: ClipCap

ClipCap is a vision-and-language model consisting of an image encoder and a text decoder. We used ClipCap as the baseline for the GazeVQA task because there is no representative VQA model pre-trained in Japanese.

Table 3: Frequency of predicate term relationships in test-set of GazeVQA questions: Note that "nom.", "acc." and "dat." denote numbers of nominative, accusative and dative cases, AQ and CQ refer to caption of Fig. 3.

Types	*ga* (nom.)	*wo* (acc.)	*ni* (dat.)
AQ	2,044	1,028	440
CQ	2,912	1,584	569

Image encoder: Given a RGB image $I \in \mathbb{R}^{W \times H \times 3}$, the baseline image encoder outputs image series $\boldsymbol{r} = \{r_1, \ldots, r_n\}$ that can be input to the text decoder. Here n is the length of the image series, and element r_i in $\boldsymbol{r}$ has the same dimensions as the token embedding in question $\boldsymbol{q}$.

Given image I, CLIP image encoder (Radford et al., 2021) outputs image series $\boldsymbol{p} = \{p_1, \ldots, p_n\}$ using a single linear layer f:

$$\{p_1, \ldots, p_n\} = f(CLIP(I)). \quad (1)$$

Given image series $\boldsymbol{p}$, the multi-layer transformer blocks (Vaswani et al., 2017) the F output $\boldsymbol{r}$:

$$\{r_1, \ldots, r_n\} = F(\{p_1, \ldots, p_n\}). \quad (2)$$

We call these transformer blocks a mapping network, following the previous work (Mokady et al., 2021).

Text decoder: Given question tokens $\boldsymbol{q} = q_1, \ldots, q_m$ and image series $\boldsymbol{r}$, the autoregressive text decoder generates answer tokens $\boldsymbol{y}$. The following is the input series of the text decoder:

$$\{r_1, \ldots, r_n, [SEP1], q_1, \ldots, q_m, [SEP2]\}, \quad (3)$$

where $[SEP1]$ and $[SEP2]$ are "Question:" and "Answer:" and represent the decoder prompts.

4.2. Proposed Model: ClipCap + Adapter

Figure 4 shows the structure of our proposed model. We added adapters to a mapping network (Dumoulin et al., 2018), inspired by work on object segmentation using text and objects as queries (Lüddecke and Ecker, 2022). Adapters merge image I and RoI I_s, and the mapping network outputs an image series that takes into account a gaze target. Each mapping network's transformer block has an adapter (Fig. 4, right). The CLIP image encoder constructs two image series: one for image $\boldsymbol{p}$ and another for the RoI of the gaze target $\boldsymbol{s} = \{s_1, \ldots, s_n\}$ from I and I_s, similar to the baseline image encoder. Given $\boldsymbol{p}$ and $\boldsymbol{s}$, the

[3] We calculated this frequency through an integrated Japanese text analyzer (Ueda et al., 2023).

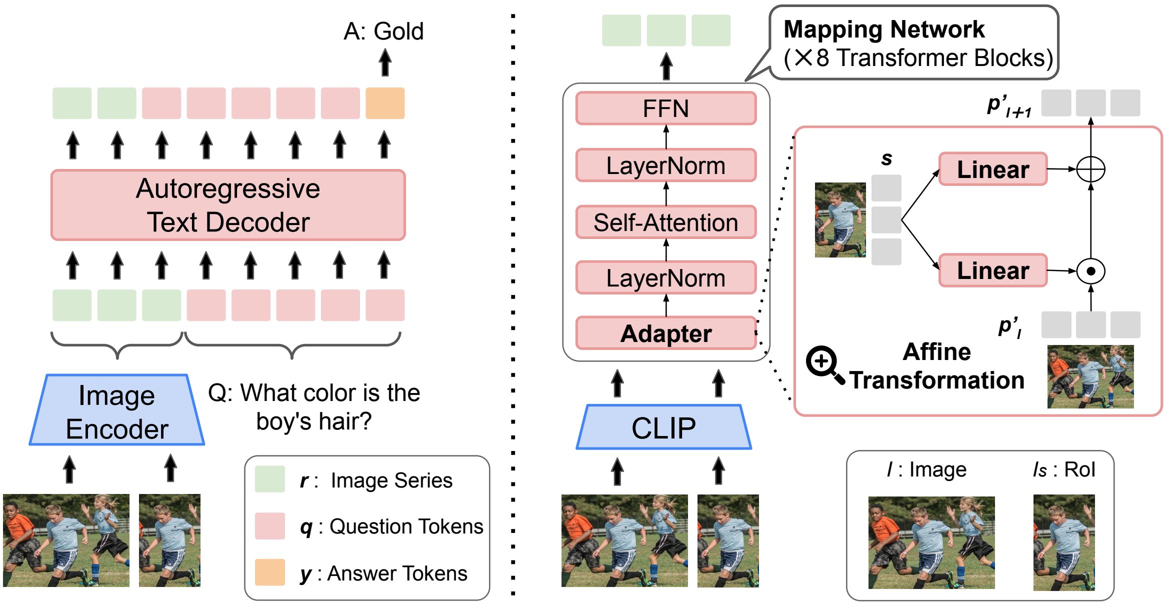


Figure 4: **Left:** Overview of proposed system **Right:** Details of Image Encoder architecture

adapter computes the element-wise affine transformation and outputs a mixture of features $\boldsymbol{p}'$ from I and I_s:

$$\boldsymbol{p}'_{l+1} = g(\boldsymbol{s}) \odot \boldsymbol{p}'_l \oplus h(\boldsymbol{s}), \tag{4}$$

where g and h denote a linear layer and $\boldsymbol{p}'_l$ denotes the input of the transformer block of the mapping network in the l-th layer. Note that the input of the first transformer block is $\boldsymbol{p}'_l = \boldsymbol{p}$.

4.3. Process of Gaze Target Estimation

We obtain RoI I_s, which is an input to the adapter, from a head image of gaze source I_h. Given image I and head image I_h, the gaze target estimation model (Chong et al., 2020) outputs a gaze heatmap H. We binarize a threshold value of 0 for H and obtain I_s, which is a bounding box corresponding to the gaze target (Ardizzone et al., 2013). We consider I to be I_s since it is difficult to get I_s from H if every element of H is 0.

5. Experiments

5.1. Experimental Setup

Dataset: We used 123,287 images and 616,435 captions from the Japanese image caption dataset (Yoshikawa et al., 2017) (STAIR) and 99,208 images and 793,664 question-answer pairs from the Japanese VQA dataset (Shimizu et al., 2018) (VQA-ja) as pre-training for the models. We fine-tuned them using the GazeVQA train-set.

Implementation details: We used a ResNet-based $RN \times 4$ (Tan and Le, 2019) as the CLIP image encoder and processed images I and regions of interest I_s in a manner that resembles CLIP normalization[4]. The input of the CLIP image encoder is a resized image with 224 dimensions (height and width); the output is a 640-dimensional vector. We composed a mapping network of eight layers of transformer blocks and set length n of the image series ($\boldsymbol{p}$, $\boldsymbol{s}$, and $\boldsymbol{r}$) to 10. We used GPT-2 as our text decoder (Radford et al., 2021), which was pre-trained on a Japanese corpus[5].

For a batch size of 32, we trained 10 epochs for STAIR, VQA-ja, and GazeVQA. The optimizer was AdamW (Loshchilov and Hutter, 2019), with a learning rate of 2e-5 in pre-training and 1e-4 in fine-tuning. We used a beam search with beam width of 10 for the GazeVQA evaluation.

Training target: We next describe the results of training the parameters of the mapping network and the text decoder due to the limited data available in Japanese. Our model has about 426M training parameters: 410M baseline training parameters and 16M adapter parameters. We also report the results of training the mapping network or only the adapters with GazeVQA to explicitly update the adapter weights. There are 74M baseline training parameters, and our model has 90M training parameters only when the mapping network is trained.

[4] https://github.com/openai/CLIP

[5] https://huggingface.co/rinna/japanese-gpt2-medium

Table 4: Evaluation results of baseline and proposed models with GazeVQA test-set: $|\theta|$ is number of trainable parameters for each model.

Models	$\|\theta\|$	Acc	Bs
Fine-tuned Text Decoder & Mapping Network			
ClipCap	410	**36.80**	**81.75**
ClipCap + Adapter (I)	426	34.78	81.39
ClipCap + Adapter (I_s)	426	34.15	81.28
ClipCap + Adapter (GT)	426	34.72	81.33
Fine-tuned Mapping Network			
ClipCap	74	35.83	81.21
ClipCap + Adapter (I)	90	**38.45**	**81.74**
ClipCap + Adapter (I_s)	90	38.11	81.71
ClipCap + Adapter (GT)	90	38.01	81.70
Fine-tuned Adapter Only			
ClipCap + Adapter (I)	16	40.06	81.91
ClipCap + Adapter (I_s)	16	39.03	81.92
ClipCap + Adapter (GT)	16	**40.09**	**82.01**

Evaluation metrics: We evaluated the model with VQA score Acc that takes into account the diversity of the answers in the VQA task (Antol et al., 2015; Goyal et al., 2017). We also evaluated the model using a BERT score, Bs (Zhang et al., 2020), which takes into account the variability of the responses. We used a multilingual BERT sentence vector for our evaluation and calculated the similarity of the vectors between the predicted answer and each element in the gold answer set[6]. Bs is the arithmetic mean of all these similarities.

5.2. Quantitative Evaluation

Table 4 shows the evaluation results of the proposed model and the baseline. Table 5 shows the ablation study results for the baseline inputs. Here all the scores (Acc and Bs) are the averages of five training and evaluation iterations of the GazeVQA. We denote the image as I, the RoI obtained from the gaze target estimation as I_s, and the gold RoI as GT, which is a COCO bounding box associated with the question, with respect to the model inputs.

Our model vs. baseline: We compared our proposed model (ClipCap + Adapter (I_s)) with the baseline (ClipCap) with the RoI I_s input to the adapter. Appendix A shows detailed evaluation results for each question type described according to the classification in Table 2.

As shown in Table 4, our model underperformed the baseline when the mapping network and the text decoder are trained with GazeVQA. However, it outperformed the baseline performance when only the mapping network and the adapters were trained with GazeVQA. In particular, the VQA score of our model trained only with adapters is 39.03, which is about four points higher than the baseline trained with the mapping network and the text decoder. Our model can generate accurate answers to ambiguous questions with about 16M parameter updates, compared to the baseline, which requires a full tuning both text decoder and a mapping network.

Table 5: Ablation evaluation results of baseline: I_s and GT denote that baseline only uses a limited region of image I pointed by their bounding boxes.

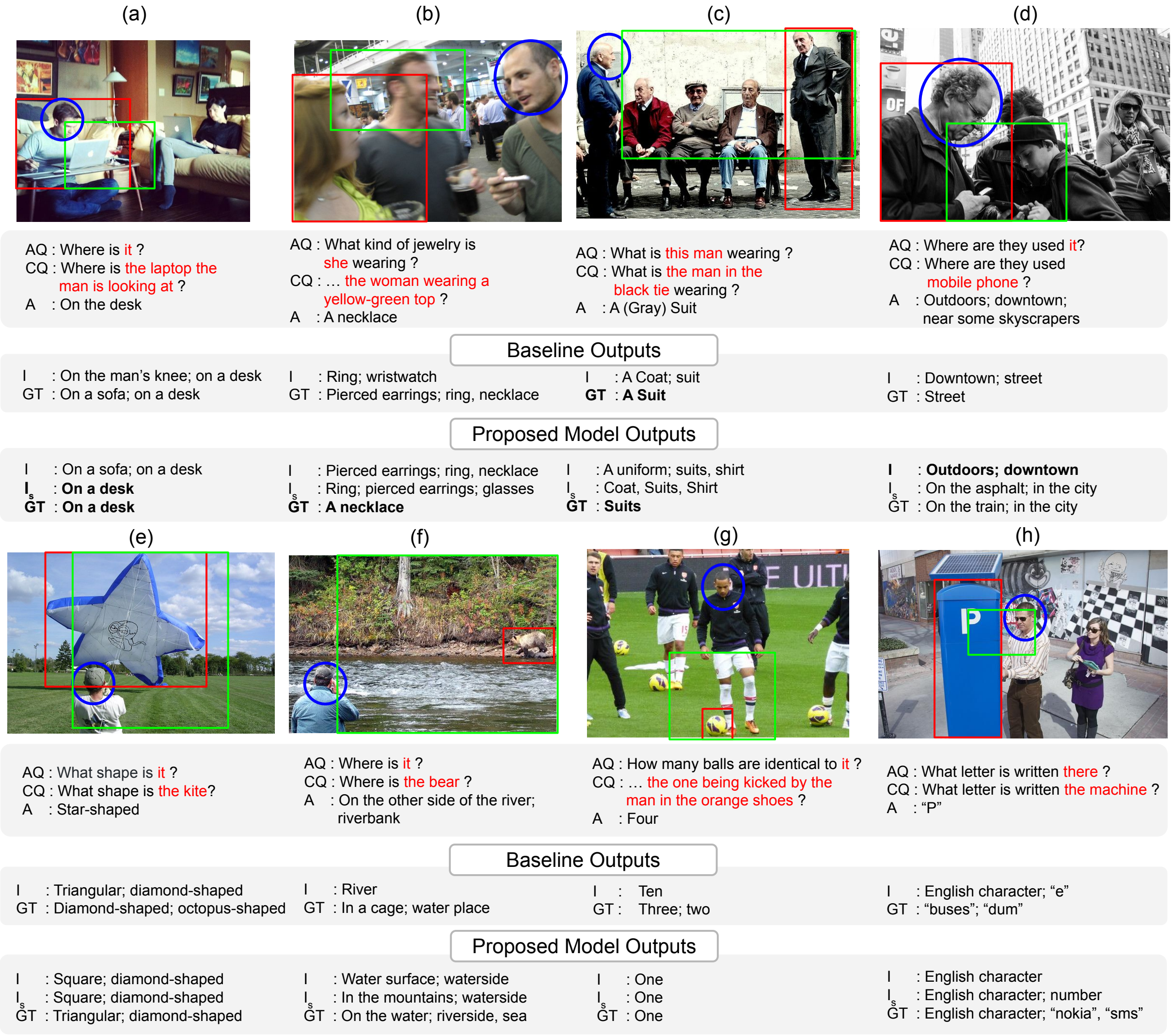
Models	Acc	Bs
ClipCap	36.80	81.75
w/o image series	16.10	78.48
w/o question tokens	3.66	65.93
ClipCap (I_s)	34.53	81.28
ClipCap (GT)	34.27	81.26

Factors contributing to GazeVQA task accuracy: We compared our proposed model with a baseline trained only on the mapping network. As shown in Table 4, our model with image I as input to the adapter (ClipCap+ Adapter (I)) outperformed the baseline, and there is no difference in our model with RoI I_s and GT as input: ClipCap+ Adapter (I_s) and ClipCap+ Adapter (GT). This result suggests that the increase in training parameters due to the addition of adapters is one reason for the improved accuracy of the GazeVQA task. This result also suggests that using RoI I_s, which is a model for gaze target estimation, may reduce the accuracy when the estimation is incorrect. Our qualitative evaluation in Section 5.3 discusses these results.

Characteristics of our dataset: We identified the elements needed to resolve ambiguous questions in GazeVQA through an ablation study on the baseline. As shown in Table 5, the performance of the baseline, which excludes question tokens or image series from the input, is significantly worsened. Models need to jointly understand the images/questions to solve GazeVQA tasks.

The performance of the baseline with regions of interest (I_s and GT) as input to the image encoder falls below the baseline with image I as input. This result suggests that keeping some information outside the gaze targets, rather than completely removing such information, improves the accuracy of the GazeVQA task.

[6] https://huggingface.co/bert-base-multilingual-cased



(a)

AQ : Where is it ?
CQ : Where is the laptop the man is looking at ?
A : On the desk

Baseline Outputs
I : On the man’s knee; on a desk
GT : On a sofa; on a desk

Proposed Model Outputs
I : On a sofa; on a desk
I_s : On a desk
GT : On a desk

(b)

AQ : What kind of jewelry is she wearing ?
CQ : … the woman wearing a yellow-green top ?
A : A necklace

Baseline Outputs
I : Ring; wristwatch
GT : Pierced earrings; ring, necklace

Proposed Model Outputs
I : Pierced earrings; ring, necklace
I_s : Ring; pierced earrings; glasses
GT : A necklace

(c)

AQ : What is this man wearing ?
CQ : What is the man in the black tie wearing ?
A : A (Gray) Suit

Baseline Outputs
I : A Coat; suit
GT : A Suit

Proposed Model Outputs
I : A uniform; suits, shirt
I_s : Coat, Suits, Shirt
GT : Suits

(d)

AQ : Where are they used it?
CQ : Where are they used mobile phone ?
A : Outdoors; downtown; near some skyscrapers

Baseline Outputs
I : Downtown; street
GT : Street

Proposed Model Outputs
I : Outdoors; downtown
I_s : On the asphalt; in the city
GT : On the train; in the city

(e)

AQ : What shape is it ?
CQ : What shape is the kite?
A : Star-shaped

Baseline Outputs
I : Triangular; diamond-shaped
GT : Diamond-shaped; octopus-shaped

Proposed Model Outputs
I : Square; diamond-shaped
I_s : Square; diamond-shaped
GT : Triangular; diamond-shaped

(f)

AQ : Where is it ?
CQ : Where is the bear ?
A : On the other side of the river; riverbank

Baseline Outputs
I : River
GT : In a cage; water place

Proposed Model Outputs
I : Water surface; waterside
I_s : In the mountains; waterside
GT : On the water; riverside, sea

(g)

AQ : How many balls are identical to it ?
CQ : … the one being kicked by the man in the orange shoes ?
A : Four

Baseline Outputs
I : Ten
GT : Three; two

Proposed Model Outputs
I : One
I_s : One
GT : One

(h)

AQ : What letter is written there ?
CQ : What letter is written the machine ?
A : “P”

Baseline Outputs
I : English character; “e”
GT : “buses”; “dum”

Proposed Model Outputs
I : English character
I_s : English character; number
GT : English character; “nokia”, “sms”

Figure 5: Outputs for baseline and proposed models: AQ, CQ, and A are respectively ambiguous questions, clarified questions, and examples of correct answers. Bolded results are models that scored best among five attempts. GT and I_s are denoted by red and green boxes.

5.3. Qualitative Evaluation

Figure 5 shows examples of the actual outputs for the baseline and our models. We examined the impact of the differences in the inputs to the adapters on the results of our proposed model. First, our model with the RoI GT input to the adapter tended to provide unique answers to ambiguous questions about the attributes of the gaze targets, such as the object’s shape and name. As shown in Figures 5 (a) and (b), this tendency is more pronounced when the RoI contains visual features that contribute to providing an accurate answer. In other words, our model outputs inconsistent answers when the gaze target estimation model cannot narrow down the objects at the gaze target (Fig. 5 (c)). Finally, our model with image I input to the adapter tends to give accurate answers to questions that require an understanding of the image (Fig. 5 (d)).

6. Discussion and Limitations

We proposed GazeVQA to achieve a system that can understand the ambiguities in human utterances using a speaker’s gaze information. The visual features contained in GazeVQA were a single image, and the gaze destinations and sources were within its frame. However, since the visual features captured by an actual system, such as a robot, are dynamic, they contain uncertainty. This situation makes it difficult for the system to recognize speakers and disambiguation cues. We believe we should fully use gaze information and such modalities as pointing (Nakamura et al., 2023) and the dialog context before utterances (Das et al., 2017; Yu et al., 2019) to account for visual uncertainty.

GazeVQA was designed for Japanese questions, and the availability of Japanese vision-and-language data is limited. For this reason, our study investigated a good training efficiency baseline (Mokady et al., 2021) and method (Dumoulin et al., 2018). However, none of the models used in this research accurately answered questions about the shape of special objects, positional relationships, number of objects, or character comprehension (Fig. 5 (e)-(h)). A system needs to understand the gaze information to identify what is the object indicated by directives or ellipsis, but the question requires information from other areas in the image; as in the case of Figure 5 (g). We believe a model structure must be used that allows for a fine-grained understanding of the correspondence between vision-and-language (Cho et al., 2021; OpenAI, 2023) to alleviate this problem. Appendix B shows evaluation results of how these models can handle GazeVQA ambiguous questions and clarified questions.

7. Conclusion

We introduced a Gaze-grounded VQA dataset (GazeVQA) to address the problem of ambiguities in human utterances in real world. Answering GazeVQA questions is challenging without the speaker's gaze information and contains ambiguities about directives and ellipsis peculiar to Japanese. Furthermore, we proposed a model that integrates the region of interest of the gaze target as gaze information in addition to images and questions. Quantitative results show that our model improves the performance over a baseline on the GazeVQA task. Qualitative results show that our model provides accurate answers to ambiguous questions about the attributes of gaze objects through gaze information. Our future work will address the difficult cases in our study by exploring model architectures and methods for integrating gaze information.

8. Acknowledgements

This research was supported by JSPS KAKENHI Grant Number 22H04873 and JST Moonshot R&D Program Grant Number JPMJMS2236.

9. Bibliographical References

Shubham Agarwal, Trung Bui, Joon-Young Lee, Ioannis Konstas, and Verena Rieser. 2020. History for visual dialog: Do we really need it? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8182–8197.

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

Edoardo Ardizzone, Alessandro Bruno, and Giuseppe Mazzola. 2013. Saliency based image cropping. In *Proceedings of the 17th International Conference Progress in Image Analysis and Processing*, volume 8156, pages 773–782.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 1931–1942.

Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. 2018. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *Proceedings of 15th the European Conference on Computer Vision*, pages 397–412.

Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. 2020. Detecting attended visual targets in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5395–5405.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2016. Human attention in visual question answering: Do humans and deep networks look at the same regions? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 932–937.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Vincent Dumoulin, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville, and Yoshua Bengio. 2018. Feature-wise transformations. *Distill*, 3(7):e11.

Nathan J. Emery. 2000. The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, 24(6):581–604.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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*, pages 6904–6913.

Muhammet Ilaslan, Chenan Song, Joya Chen, Difei Gao, Weixian Lei, Qianli Xu, Joo Lim, and Mike Shou. 2023. GazeVQA: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10462–10479.

Tsung Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the 13th European Conference on Computer Vision*, pages 740–755.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the 7th International Conference on Learning Representations*.

Timo Lüddecke and Alexander Ecker. 2022. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096.

Arjun Mani, Nobline Yoo, Will Hinthorn, and Olga Russakovsky. 2020. Point and ask: Incorporating pointing into visual question answering. arXiv:2011.13681.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. ClipCap: CLIP prefix for image captioning. arXiv:2111.09734.

Shu Nakamura, Yasutomo Kawanishi, Shohei Nobuhara, and Ko Nishino. 2023. DeePoint: Visual pointing recognition and direction estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20577–20587.

OpenAI. 2023. GPT-4 technical report. arXiv:2303.08774.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 8748–8763.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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

Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking? In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, volume 1, pages 199–207.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of the 28th Advances in Neural Information Processing Systems*, volume 28, pages 91–99.

Roberta Rocca, Mikkel Wallentin, Cordula Vesper, and Kristian Tylén. 2018. This and that back in context: Grounding demonstrative reference in manual and social affordances. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*.

Ryohei Sasano, Daisuke Kawahara, and Sadao Kurohashi. 2008. A fully-lexicalized probabilistic model for Japanese zero anaphora resolution. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 769–776.

Kazuhiro Seki, Atsushi Fujii, and Tetsuya Ishikawa. 2002. A probabilistic method for analyzing Japanese anaphora integrating zero pronoun detection and resolution. In *Proceedings of the 19th International Conference on Computational Linguistics*.

Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. 2018. 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*, pages 1918–1928.

Ekta Sood, Fabian Kögel, Florian Strohm, Prajit Dhar, and Andreas Bulling. 2021. VQA-MHUG: A gaze dataset to study multimodal neural attention in visual question answering. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 27–43.

Osamu Sugiyama, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2007. Natural deictic communication with humanoid robots. In *Proceedings of the 2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1441–1448.

Yi Lin Sung, Jaemin Cho, and Mohit Bansal. 2022. VL-Adapter: parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237.

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 6105–6114.

Tadahiro Taniguchi, Daichi Mochihashi, Takayuki Nagai, Satoru Uchida, Naoya Inoue, Ichiro Kobayashi, Tomoaki Nakamura, Yoshinobu Hagiwara, Naoto Iwahashi, and Tetsunari Inamura. 2019. Survey on frontiers of language and robotics. *Advanced Robotics*, 33(15–16):700–730.

Henri Tomas, Marcus Reyes, Raimarc Dionido, Mark Ty, Jonric Mirando, Joel Casimiro, Rowel Atienza, and Richard Guinto. 2021. Goo: A dataset for gaze object prediction in retail environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*.

Nobuhiro Ueda et al. 2023. KWJA: A unified Japanese analyzer based on foundation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 3, pages 538–548.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale Japanese image caption dataset. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 417–421.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5123–5132.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations*.

A. Evaluation by Question Types

We compared our proposed model with the baseline based on the typology of question types for GazeVQA shown in Table 2. As shown in Figure 6, our model performs well with “What is” questions about object attributes, “What condition” questions about the current state of an object, and “Which” questions that are multiple choice questions. The baseline with RoI GT (ClipCap(GT)) performs well with “What color” questions that ask for an object color.

B. Discussion on Evaluation with Clarified Questions

Figure 6 shows comparative evaluation results of ambiguous questions and clarified questions with the baseline (ClipCap) and modern vision-and-language models: VL-T5 (Cho et al., 2021; Sung et al., 2022) and GPT-4V (OpenAI, 2023). We used 300 samples from our GazeVQA test-set for evaluation and did not fine-tune any models with GazeVQA train-set.

B.1. Implementation details

VL-T5 is a vision-and-language model consisting of an image encoder (Ren et al., 2015) and the text encoder-decoder (Raffel et al., 2020). We constructed VL-T5 with the CLIP image encoder (Tan and Le, 2019) and the Japanese T5 model [7] [8], based on the implementation of Sung et al. (2022). We used the same conditions as in Section 5.1 for the VL-T5 training setup.

GPT-4V is a large-scale vision-and-language model trained on large amounts of image-text data. We evaluated the GazeVQA test-set using GPT-4V in the 3-shot setting; each example was constructed from questions and answers and gaze targets included in the GazeVQA train-set. Tables 7 and 8 show prompts given to GPT4V for inferring an answer from either an ambiguous question and

[7] https://huggingface.co/retrieva-jp/t5-small-short

[8] https://huggingface.co/retrieva-jp/t5-base-short

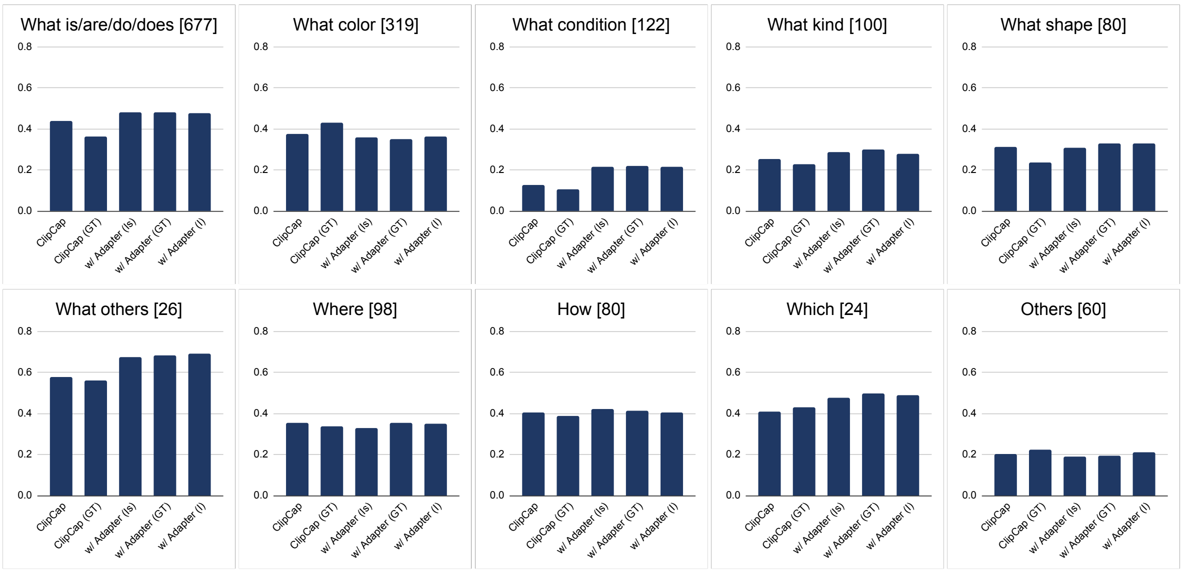


Figure 6: Evaluation results of baseline and proposed model by question types for GazeVQA test-set: Square brackets denote the number of questions.

Table 6: Comparison results between ambiguous questions (AQ) and clarified questions (CQ) using vision-and-language models

Models	Types	Acc	Bs
ClipCap	AQ	21.55	78.19
ClipCap	CQ	**25.11**	**79.41**
VL-T5 small	AQ	**32.66**	80.20
VL-T5 small	CQ	31.66	**80.27**
VL-T5 base	AQ	32.33	80.16
VL-T5 base	CQ	**34.11**	**80.77**
GPT-4V 3-shot	AQ	34.11	79.99
GPT-4V 3-shot	CQ	**39.33**	**80.17**

bounding boxes of gaze targets GT or a clarified question.

B.2. Results

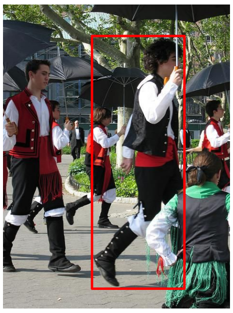
Figures 4 and 6 suggest that our model fine-tuned with GazeVQA outperforms GPT4V when ambiguous questions are used. On the other hand, Figure 6 shows that GPT-4V outperforms other models when clarified questions are used as input instead of ambiguous questions. These results indicate that large vision-and-language models such as GPT4V are highly capable, but they are not sufficient in situations such as GazeVQA task, where the question contains ambiguity and needs to be supplemented with contextual information.

Table 7: Prompts used to evaluate GazeVQA ambiguous questions: gaze targets GT are denoted by red boxes.

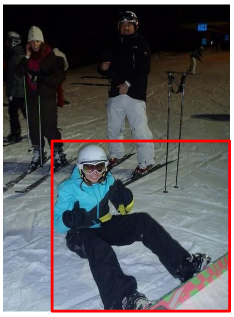
Instruction

Instruction: Given an ambiguous Japanese question that includes ellipsis or directives, an image, and bounding boxes (format:$[x_1, y_1, w, h]$), you answer the question in Japanese. Note1: Each question is answerable if you consider the bounding boxes corresponding to the ellipsis or directives of the question. Note2: Each answer will end with a noun.

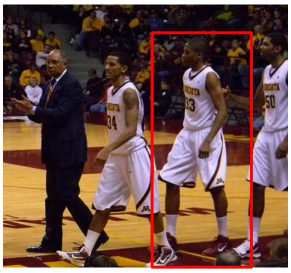
Visual input examples



{Example1}



{Example2}



{Example3}

Text input example1

The question is: What color is he wearing?
The image is: {Example1}
The bounding boxes are: [194.16,69.03,194.15,524.95]
Answer the question with a single phrase in Japanese: Black

Text input example2

The question is: What is she wearing on her head?
The image is: {Example2}
The bounding boxes are: [158.2,339.42,291.96,293.39]
Answer the question with a single phrase in Japanese: Helmet

Text input example3

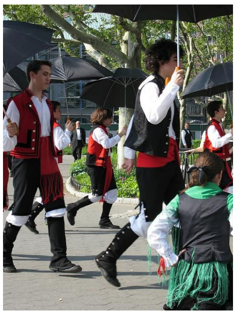
The question is: What is his number?
The image is: {Example3}
The bounding boxes are: [440.31,156.82,117.88,310.4]
Answer the question with a single phrase in Japanese: Number 33

Table 8: Prompts used to evaluate GazeVQA clarified questions.

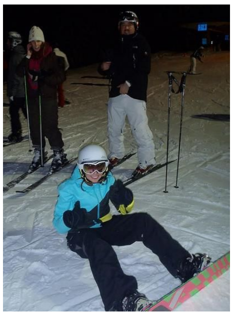
Instruction

Instruction: Given a Japanese question and an image, you answer the question in Japanese. Note: Each answer will end with a noun.

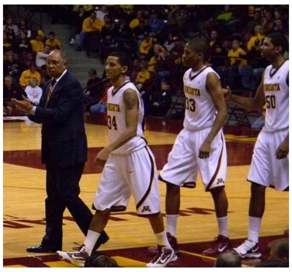
Visual input examples



{Example1}



{Example2}



{Example3}

Text input example1

The question is: What color is the man on the right with the black umbrella wearing?
The image is: {Example1}
Answer the question with a single phrase in Japanese: Black

Text input example2

The question is: What is the woman in the blue jacket wearing on her head?
The image is: {Example2}
Answer the question with a single phrase in Japanese: Helmet

Text input example3

The question is: What is his number of the second man from the right?
The image is: {Example3}
Answer the question with a single phrase in Japanese: Number 33