A Visually-grounded First-person Dialogue Dataset with Verbal and Non-verbal Responses

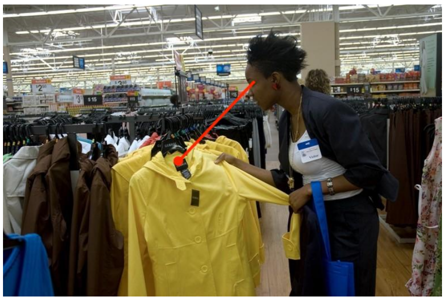
Hisashi Kamezawa[1] **Noriki Nishida**[2] **Nobuyuki Shimizu**[3] **Takashi Miyazaki**[3] **Hideki Nakayama**[1]
[1] The University of Tokyo
[2] RIKEN Center for Advanced Intelligence Project (AIP)
[3] Yahoo Japan Corporation
hisas@nlab.ci.i.u-tokyo.ac.jp
noriki.nishida@riken.jp
{nobushim, takmiyaz}@yahoo-corp.jp
nakayama@ci.i.u-tokyo.ac.jp

Abstract

In real-world dialogue, first-person visual information about where the other speakers are and what they are paying attention to is crucial to understand their intentions. Non-verbal responses also play an important role in social interactions. In this paper, we propose a visually-grounded first-person dialogue (VFD) dataset with verbal and non-verbal responses. The VFD dataset provides manually annotated (1) first-person images of agents, (2) utterances of human speakers, (3) eye-gaze locations of the speakers, and (4) the agents' verbal and non-verbal responses. We present experimental results obtained using the proposed VFD dataset and recent neural network models (e.g., BERT, ResNet). The results demonstrate that first-person vision helps neural network models correctly understand human intentions, and the production of non-verbal responses is a challenging task like that of verbal responses. Our dataset is publicly available[1].

1 Introduction

In recent years, visually-grounded dialogue systems have attracted increasing attention (Zhu et al., 2016; Ben-Youssef et al., 2017; Liao et al., 2018; Kottur et al., 2018). For example, Huber et al. (2018) developed an image-grounded conversational agent that uses visual sentiment, facial expression, and scene features, and Mostafazadeh et al. (2017) constructed the publicly available IGC dataset, which comprises image-grounded conversations.



U: これのＬはないのかしら
V: 同じ服がたくさんあるからどれかはLじゃないかな
N: 同じ服のサイズをチェックする

U: I wonder if there is an L for this.
V: We have a lot of the same clothes, so I'm guessing one of them is an L.
N: Check out the same clothing size.

Figure 1: Example of proposed VFD dataset. "U", "V", and "N" denote a human utterance, the agent's *verbal* response, and the agent's *non-verbal* response (i.e., action), respectively. All utterances and responses are represented in Japanese. English translations are added below for easier understanding. The red line links the eyes to the gaze location.

Although these studies and resources have been shown to be useful, there are currently two limitations. First, in image-grounded dialogue tasks, human speakers do not appear in the agents' vision because images are used as the topic of conversation, and the speakers are required to discuss the input image. However, in real-world dialogue scenarios, first-person visual information about where the human speaker is and what they are paying attention to is crucial for agents to understand human intentions. To understand this, we show an example in Figure 1. Without the first-person image, it is difficult for the agent to recognize that the pronoun "this" in the human utterance (U) refers to the article of yellow clothing rather than any other products (e.g., brown clothes).

Another important limitation is that, although

[1] https://randd.yahoo.co.jp/en/softwaredata

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 3299–3310,
November 16–20, 2020. ©2020 Association for Computational Linguistics

Dataset	Type	Perspective	Response	Size
VisDial (Das et al., 2017)	Task oriented	Third-person	Verbal	120K
MMD (Saha et al., 2018)	Task oriented	Third-person	Verbal	150K
TalkTheWalk (de Vries et al., 2018)	Task oriented	Third-person	Verbal	10K
AVSD (Alamri et al., 2019)	Task oriented	Third-person	Verbal	11K
IGC (Mostafazadeh et al., 2017)	Task & Non-task oriented	Third-person	Verbal	4K
SDG (Hu et al., 2016)	Non-task oriented	Third-person	Verbal & Non-verbal	50
VFD (ours)	Task & Non-task oriented	First-person	Verbal & Non-verbal	308K

Table 1: Comparison of existing visually-grounded dialogue datasets in terms of dialogue types (task-oriented or non-task-oriented), visual perspectives, response types, and the dataset size.

previous studies considered non-verbal input information (e.g., human facial expressions), they did not consider the agents' non-verbal responses (i.e., actions). Non-verbal responses often play an important role in dialogue systems. For example, a museum tour-guide robot should use non-verbal gestures to explain things to the audience better. Even in ordinary conversation, non-verbal responses such as "making a smile" or "helping to lift luggage" are often crucial for social interactions in conjunction with verbal responses.

Thus, we propose a visually-grounded first-person dialogue (VFD) dataset with verbal and non-verbal responses. As shown in Figure 1, the VFD dataset comprises (1) *first-person images* of agents, (2) *utterances* of human speakers, (3) *eye-gaze locations* of the speakers, and (4) the agents' *verbal and non-verbal responses* to the utterances. Here, human utterances and agents' verbal and non-verbal responses were manually annotated for first-person images (with eye-gaze locations) in the GazeFollow dataset (Recasens et al., 2015) using crowdsourcing with carefully-designed settings, resulting in 308K verbal and 81K non-verbal dialogues. This paper also presents experimental results obtained using the VFD dataset and recent neural network models, e.g., BERT (Devlin et al., 2019) and ResNet (He et al., 2016).

Our primary contributions are summarized as follows. (1) We present a new multimodal dialogue dataset that contains visually-grounded first-person dialogues with human speakers' eye-gaze locations. (2) We provide the manually-annotated non-verbal responses of agents, which are often crucial for social communication in the real world. (3) Our experimental results demonstrate that first-person vision helps recent neural network models understand human intentions accurately and that the production of non-verbal responses is a challenging task like that of verbal responses.

2 Related Work

Table 1 summarizes the related visually-grounded dialogue datasets.

Several multimodal dialogue datasets have investigated task-oriented situations. For example, MMD dataset (Saha et al., 2018) contains dialogues between shoppers of fashion products and sales agents. TalkTheWalk dataset (de Vries et al., 2018) aims to guide tourists to their destinations. In VisDial dataset (Das et al., 2017) and AVSD dataset (Alamri et al., 2019), an agent must answer questions about an input image (or video) given dialogue history. Unlike these datasets, which can only work in some limited scenarios, we aim to cover both task-oriented and non-task-oriented dialogue systems.

As shown in Table 1, IGC dataset (Mostafazadeh et al., 2016), like our VFD dataset, assumes both task-oriented and non-task oriented situations. However, in IGC, images are used as a conversation topic, and the human speakers do not appear in the agents' vision. In contrast, VFD dataset contains dialogues based on "first-person" images (and eye-gaze information), which are useful for figuring where the human speaker is and what he or she is focusing on.

Like our VFD dataset, SDG dataset (Hu et al., 2016) contains dialogues with non-verbal actions. However, SDG focuses on gestures (or body languages), e.g., "making a cup shape with the right hand", which are categorized into 271 gesture classes. In contrast, VFD dataset represents non-verbal responses as text (typically sentences) to cover a wider range of gestures, e.g., "Check out the same clothing size", "Buy one of the pumpkins a girl has", etc.

In addition, our VFD dataset is large in comparison to other datasets. It is twice the size of the MMD dataset and approximately 75 times the size of the IGC dataset. IGC dataset is small because it provides only validation and test sets.

3 VFD Dataset

3.1 Task Definition

In this paper, we define the visually-grounded first-person dialogue as to produce an utterance or take action given a human utterance and the agent's first-person vision.

Formally, the input to the system can be represented as a tuple of a human utterance u and the agent's first-person vision v. The first-person vision v is assumed to be used to understand human intentions. Thus, v can be factorized into first-person image i and more explicit visual hints for the human intentions g, i.e., $v = (i, g)$. We use eye-gaze locations for the explicit hints g. For the input triplet (u, i, g), an agent is assumed to produce a verbal response r_v and non-verbal response (i.e., actions) r_n. Here, we use textual descriptions to represent non-verbal responses, as shown in Figure 1.

The VFD dataset can be interpreted as a collection of quintuples, i.e., $\{(u, i, g, r_v, r_n)\}$. We describe how we collected these five elements in the following.

3.2 Dataset Construction

First-person Images & Eye-gaze Locations. We used the 34,775 first-person images with eye-gaze annotations in the GazeFollow dataset (Recasens et al., 2015). Here, eye-gaze locations are represented as coordinates $(x, y)_{\text{eye}}$ and $(x, y)_{\text{gaze}}$. In Figure 1, the eye location and the gaze location are linked by a red line.

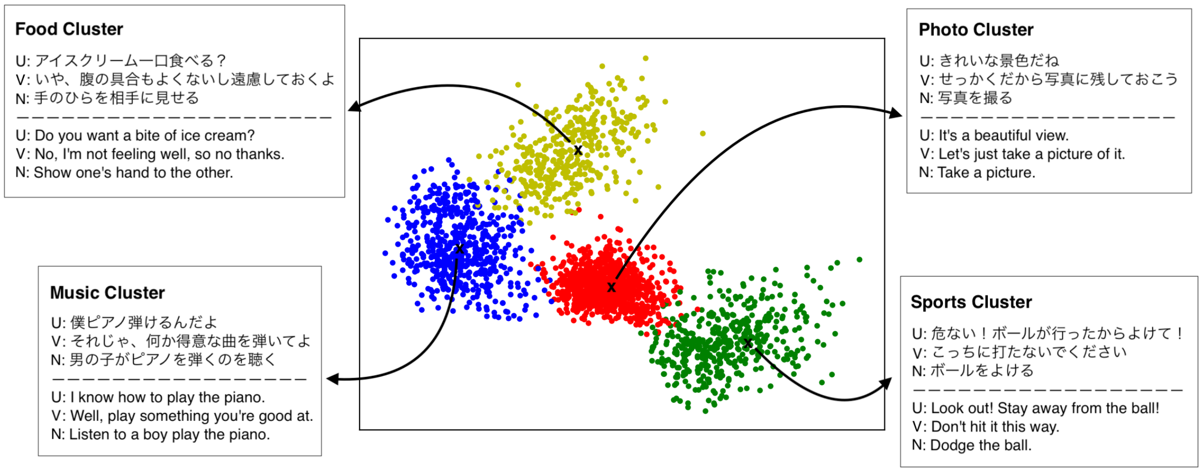
Human Utterances. Following Le Minh et al. (2018), who collected English utterances for first-person images using Amazon Mechanical Turk (AMT), we first translated their English instructions into Japanese. Then, we used a crowdsourcing platform similar to AMT called Yahoo! Crowdsourcing, operated by Yahoo Japan Corporation. It can be safely assumed that Yahoo! Crowdsourcing participants will be proficient in Japanese because such proficiency is required to sign up, navigate the user interface, and participate in the microtask market. In the annotation instructions, we showed an image with a single person marked with a red dot and asked the participants to imagine this person is speaking. We then asked the participants to submit what they think the speaker is likely saying.

The following notes were included in the instructions to avoid unexpected or trivial annotations. Note 1: "Never use the same lines again. Please write a different sentence every time." Note 2: "Do not put a commentary from a third-party perspective." Note 3: "Please do not write something people would not usually say in this situation. Please avoid lines that contain abuse and prejudice, words likely to cause a quarrel, and over-familiar tone. Please do not assume that the talking person has an extreme personality. As it is not a comedy, you do not have to write a funny line."

Verbal and Non-verbal Responses. The participants were shown the images and the utterances collected in the previous step. Then, they were asked to enter what to say (i.e., a verbal response) and what to do (i.e., a corresponding non-verbal action). To focus on dialogues requiring visual grounding, we also asked the participants the following question: "Whenever possible, please try to use some additional information found in the image to frame your response, so that your response is not entirely predictable from the utterance." We also asked the participants to enter a special dummy response "x" if it is inappropriate to respond. For a single utterance, five participants were asked to enter a response and an action.

After conducting this pilot task, we examined the results and selected promising participants (comprising a whitelist) for future task requests. Only participants on the whitelist could perform the next task. We also used the whitelist from our previous study for text entry tasks. We repeated this selection process until the final whitelist included approximately 1,600 participants. Here, approximately 200-250 of these participants regularly participated in the actual VQA collection task. Note that we allocated tasks in small batches over the course of a few months to prevent participants from working long hours.

Despite the above measures, however, the resulting 327,884 data instances contained noisy or trivial responses. To eliminate such undesirable responses automatically, we created a list of erroneous patterns manually via visual inspection, and responses matching the patterns were removed. The dummy responses "x" were also removed. Finally, the total number of verbal and non-verbal responses were 308,793 and 81,867, respectively. The gap between the number of verbal and non-verbal responses is due to the fact that non-verbal responses contained more dummy responses than verbal responses.



Food Cluster

U: アイスクリーム一口食べる？
V: いや、腹の具合もよくないし遠慮しておくよ
N: 手のひらを相手に見せる

U: Do you want a bite of ice cream?
V: No, I'm not feeling well, so no thanks.
N: Show one's hand to the other.

Photo Cluster

U: きれいな景色だね
V: せっかくだから写真に残しておこう
N: 写真を撮る

U: It's a beautiful view.
V: Let's just take a picture of it.
N: Take a picture.

Music Cluster

U: 僕ピアノ弾けるんだよ
V: それじゃ、何か得意な曲を弾いてよ
N: 男の子がピアノを弾くのを聴く

U: I know how to play the piano.
V: Well, play something you're good at.
N: Listen to a boy play the piano.

Sports Cluster

U: 危ない！ボールが行ったからよけて！
V: こっちに打たないでください
N: ボールをよける

U: Look out! Stay away from the ball!
V: Don't hit it this way.
N: Dodge the ball.

Figure 2: Dialogue topics in VFD dataset. Each dialogue is represented using BERT-based vectors and colored according to the associated cluster, i.e., *food*, *photo*, *music*, or *sports*. The dialogue topics are widely distributed.

Quality Evaluation. To assess the quality of the resulting dataset, we qualitatively inspected 1,000 randomly-sampled data instances. Of those 1,000 samples, there was only one sample that was clearly as bad as spam. In addition, the percentage of slightly inappropriate samples was only 2% of the total. Therefore, we considered the quality of the VFD dataset to be sufficient for our purposes.

Among those 2% noisy samples, we found the following erroneous patterns: For utterances, some were for the person who took the photo rather than the person appearing in the photo. One utterance was very comedic. For the images, there were two images without people, e.g., a mannequin or food. In addition, there was one image that did not show the speaker's face and one image that shows many people. These errors mainly stem from the original GazeFollow dataset (Recasens et al., 2015). For verbal or non-verbal responses, one response ignored the human utterance. In addition, some responses ignored the images or were not from the robot's perspective, and some responses were offensive to the speakers. Some non-verbal responses were not actionable, e.g., ”Nice Shot!” and ”That's tough.”

We did not remove these noisy samples in the current version because it was difficult to remove them all automatically, and the noisy samples represent only 2% of the total.

3.3 Dataset Analysis

We perform a more detailed analysis of the VFD dataset.

		Response	
	Utterance	Verbal	Non-verbal
Text length	7.6	6.8	3.5
Unique words	13,352	23,880	7,711

Table 2: Linguistic statistics of utterances, and verbal and non-verbal responses in VFD dataset. Verbal responses tend to be diverse, and non-verbal responses tend to be much simpler.

We explore the topical diversity of the dataset. Specifically, we use a Japanese BERT model pre-trained on Japanese Wikipedia from HuggingFace's Transformers library (Wolf et al., 2019) and project each word in dialogue text (i.e., utterance, verbal response, and non-verbal response) to 768-dimensional vectors. Then, we average the word embeddings to obtain a vector representation of the dialogue text (utterance + two responses). Finally, we use agglomerative clustering (Karypis et al., 2000) to obtain 70 clusters for the dialogues. We select 4 of the 70 clusters and visualize them by principal component analysis (PCA), as shown in Figure 2. These 4 clusters, i.e., *food*, *photo*, *music*, and *sports*, represent typical dialogue topics in the VFD dataset. Figure 2 shows that the dialogue topics are widely distributed in the VFD dataset.

We also calculate the linguistic statistics of the texts. Here, we use MeCab morphological analyzer (Kudo et al., 2004) to tokenize the dialogue text into tokens. Table 2 summarizes the results. The average numbers of tokens (or text length) in the utterances and verbal and non-verbal responses are 7.6, 6.8, and 3.5, respectively. The number of uniques words (i.e., vocabulary size) in the ut-

U: 大きいね
V: 一つ買っていこうか？
N: 女の子が持っているかぼちゃを一つ買う

U: It's a big one.
V: Do you want me to buy you one?
N: Buy one of the pumpkins a girl has.

U1: Place near my house is getting ready for Halloween a little early.
V1: Don't you think Halloween should be year-round, though?
U2: That'd be fun since it's my favorite holiday!
V2: It's my favorite holiday as well!
U3: I never got around to carving a pumpkin last year even though I bought one.
V3: Well, it's a good thing that they are starting to sell them early this year!

Figure 3: Comparison of VFD dataset (left) and IGC dataset (Mostafazadeh et al., 2017) (right). U, V, and N denote an utterance, a verbal response, and a non-verbal response, respectively.

terances and verbal and non-verbal responses are 13,352, 23,880, and 7,711, respectively. These facts imply that verbal responses tend to be diverse, which is desirable for training well-generalized machine learning models. In contrast, the textual description of the non-verbal responses is much simpler than the utterances and verbal responses, which is desirable when building a model to perform actual actions from a textual description of a given non-verbal response.

3.4 Comparison

Here, we emphasize the characteristics of the VFD dataset by comparing it to the IGC dataset (Mostafazadeh et al., 2016), which is most similar to the VFD dataset. Figure 3 compares two examples each from the VFD dataset (left) and IGC dataset (right). In the IGC dataset, an image is used as a topic of conversation, and the human speaker does not appear in the agent's vision. In contrast, our VFD dataset uses an image as taken from the agent's first-person camera as a dynamic visual environment. In addition, the VFD dataset contains manually-annotated non-verbal responses and human eye-gaze locations.

4 Experiments

4.1 Task Setting

In this section, we perform experiments with the task of selecting a verbal response and a non-verbal response from candidate response sets given a human utterance, a first-person image, and eye-gaze locations. Although it is possible to train a response generator using VFD dataset, the selection task was chosen for ease of evaluation and simplicity. It is worth noting that, in our experiments, the eye-gaze locations are given to the input as an oracle during validation and testing. In the real world, this information can be given by automatic gaze-estimation techniques (Chong et al., 2018; Wei et al., 2018) developed in computer vision.

4.2 Data

For the verbal response selection task, VFD dataset is split into training, validation, and test sets each containing 569K, 12K, and 12K dialogues. For the non-verbal response selection task, the training, validation, and test sets consist of 151K, 3K, and 3K dialogues. The images are completely separated across the training/validation/test sets. For the training data, we sample negative responses randomly from the training set and fix them throughout the epochs. For the validation and test data, we perform the same negative sampling across the models for a fair comparison. The data splits and the negative samples used for validation and testing will be provided along with the VFD dataset.

4.3 Metrics

Following Lowe et al. (2015), we use Recall@k (denoted $R_n@k$) for response-selection evaluation. Here, the model selects the k most likely responses from n available candidates. Note that only one response among the n candidates is true, and the others are sampled randomly from the same set. The prediction is correct if the true response is among the top k list. We report $R_{10}@1$, $R_{10}@2$, $R_{10}@5$, and $R_2@1$.

4.4 Baseline Models

Figure 4 shows the architecture of the baseline models. We follow the same ranking strategy of Lowe et al. (2015) to develop the baseline neural network models for our selection-based dialogue task. That is, the response-selection problem in our experiments is to find a verbal (or non-verbal) response with the highest score for an input triplet $x = (u, i, g)$, i.e.,

$$r^* = \underset{r \in \mathcal{C}}{\text{argmax}} \ \text{Score}(x, r), \tag{1}$$

where $\text{Score}(x, r) \in \mathbb{R}$ denotes a real-valued score of the response r for the input utterance u, input image i, and input eye-gaze locations g.

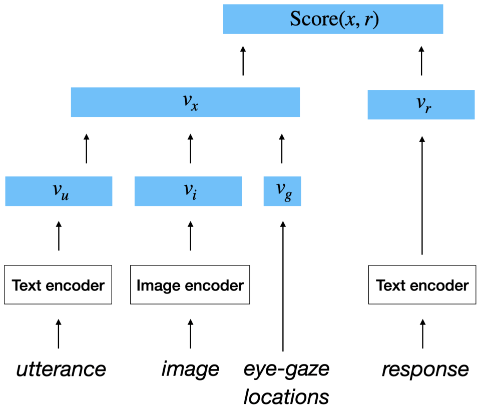


Figure 4: Overview of the baseline architecture for the response scoring. Given an input triplet of (utterance u, image i, eye-gaze locations g) and a candidate response r, the baseline model calculates the matching score between the input and the response.

We define the scoring function in Eq.(1) as follows:

$$\text{Score}(x, r) = \boldsymbol{v}_x^\top \boldsymbol{W} \boldsymbol{v}_r + \boldsymbol{b}, \tag{2}$$

where $\boldsymbol{v}_x$ and $\boldsymbol{v}_r$ denote the feature vectors for $x = (u, i, g)$ and r. $\boldsymbol{W}$ and $\boldsymbol{b}$ are a weight matrix and a bias vector, respectively.

We first apply two neural encoders, f_u and f_i, to extract feature vectors from the input utterance u and the input image i:

$$\boldsymbol{v}_u = f_u(u),\ \boldsymbol{v}_i = f_i(i). \tag{3}$$

We also represent the coordinates of eye-gaze locations as a four-dimensional vector, $\boldsymbol{v}_g \in \mathbb{R}^4$. We concatenate these feature vectors to get $\boldsymbol{v}_x$:

$$\boldsymbol{v}_x = [\boldsymbol{v}_u;\ \boldsymbol{v}_i;\ \boldsymbol{v}_g], \tag{4}$$

where $[\ \cdot\ ;\ \cdot\ ]$ denotes concatenation of vectors. The feature vector of a candidate response r is also calculated using a different text encoder f_r:

$$\boldsymbol{v}_r = f_r(r). \tag{5}$$

For training, we minimize the binary cross-entropy loss by applying a sigmoid function to the predicted scores.

In the following subsection, we describe the text encoders (i.e., f_u and f_r) and the image encoder (i.e., f_i) we used in our experiments.

Text Encoder: We employ two neural network variants for encoding utterances and responses: Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) and Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019). With the LSTM model, we use the last hidden state as the utterance or response features. With the BERT model, we insert a [CLS] token before and a [SEP] token after the utterance (or response) and use the hidden state of [CLS] tokens in the last layer of BERT as the feature vector. It is worth noting that we develop two different text encoders for f_u and f_r, which are optimized during the training.

Image Encoder: We employ two neural network models for image encoding: VGGNet (Simonyan and Zisserman, 2015) and ResNet (He et al., 2016), which are used widely for image classification and have proven to be effective methods. We use the 16-layer VGGNet and replace the last linear layer named `fc6` with a learnable linear layer whose output dimensionality is 4096. We use the 4096-dimensional vector as the image features. We also use the 50-layer ResNet. We use the last fully connected layer as the image features.

4.5 Other Settings

We used the Adam (Kingma and Ba, 2015) optimizer for training. The learning rate was fixed at 0.0001, and the mini-batch size was fixed at 64. The training was terminated when validation accuracy drops more than 1.5 points compared to the highest validation accuracy. The training typically converged in approximately 3 days for the verbal response selection task and 1 day for the non-verbal response selection task on an Nvidia GeForce GTX 1080 GPU. For the LSTM-based text encoding, we used MeCab (Kudo et al., 2004) for tokenization and used fastText (Bojanowski et al., 2017) for word embeddings, which were pre-trained on Japanese Common-Crawl and Wikipedia articles. The word-embedding and LSTM dimensions were set to 300 and 100, respectively. For the BERT-based encoding, we used a BERT model named “bert-base-japanese-whole-word-masking” from Hugging Face’s (Wolf et al., 2019) library, which was pre-trained on Japanese Wikipedia using Whole-Word-Masking. For data augmentation, we applied random cropping, random horizontal flipping, and normalization transformations to the original images during training. The baseline models were trained separately for verbal and non-verbal response selection tasks.

	Encoders		Verbal Response				Non-verbal Response			
Input	Text	Image	R_{10}@1	R_{10}@2	R_{10}@5	R_2@1	R_{10}@1	R_{10}@2	R_{10}@5	R_2@1
U	LSTM	-	50.0	**69.2**	**91.1**	**84.8**	35.6	56.3	84.9	78.6
U	BERT	-	**50.1**	67.4	89.7	84.3	**42.3**	**60.1**	**86.2**	**80.6**
U+I	LSTM	VGGNet	49.1	68.9	92.1	85.1	41.5	61.5	89.7	82.0
U+I	LSTM	ResNet	49.4	69.9	**92.4**	85.3	40.0	61.4	89.7	81.6
U+I	BERT	VGGNet	**52.7**	**71.1**	91.9	**86.1**	**44.8**	**65.7**	**89.7**	**82.6**
U+I	BERT	ResNet	52.5	71.1	91.9	86.0	43.4	64.5	89.1	82.1
U+I+G	LSTM	VGGNet	50.2	69.5	92.0	85.2	39.6	61.2	89.1	81.8
U+I+G	LSTM	ResNet	49.1	69.3	92.1	85.1	39.6	61.0	89.8	81.7
U+I+G	BERT	VGGNet	**53.6**	**72.1**	92.5	**86.6**	**46.2**	**66.3**	**90.7**	**82.9**
U+I+G	BERT	ResNet	53.2	71.8	**92.6**	86.5	43.7	65.7	89.7	82.2

Table 3: Comparison results of the baseline models in verbal and non-verbal response selection tasks. U, I, and G denote that we use utterances, images, and eye-gaze locations for inputs, respectively. First-person images and eye-gaze locations improve the performance for almost all encoder combinations.

4.6 Quantitative Results

We report the evaluation scores of the baseline models in the verbal and non-verbal response selection tasks. We summarize the results in Table 3. U, I, and G denote that we use utterances, images, and eye-gaze locations for inputs, respectively.

For almost all encoder combinations (e.g., BERT × VGGNet), first-person images improve the verbal and non-verbal response-selection performance by up to 5.6 points (See U vs. U+I). In addition, especially when using BERT, eye-gaze locations always improve the performance further by up to 1.4 points (See U+I vs. U+I+G). These results indicate that the eye-gaze information from the agents' first-person perspective is effective in understanding the human intentions.

Overall, the BERT scores are higher than the LSTM scores for all input variations: U, U+I, U+I+G. This is consistent with results in other NLP tasks. As for image encoders, VGGNet achieves higher scores than ResNet, which is often observed in multimodal tasks (Wang et al., 2017; Ouyang et al., 2017; Yudistira and Kurita, 2017). BERT × VGGNet using all the input modalities achieves the highest R_{10}@1 score of 53.6%.

Interestingly, the best R_{10}@1 score for non-verbal response selection is about 7 points worse than the score for verbal-response selection. This fact indicates that producing non-verbal responses is more difficult than producing conventional verbal responses and there is room for improvement.

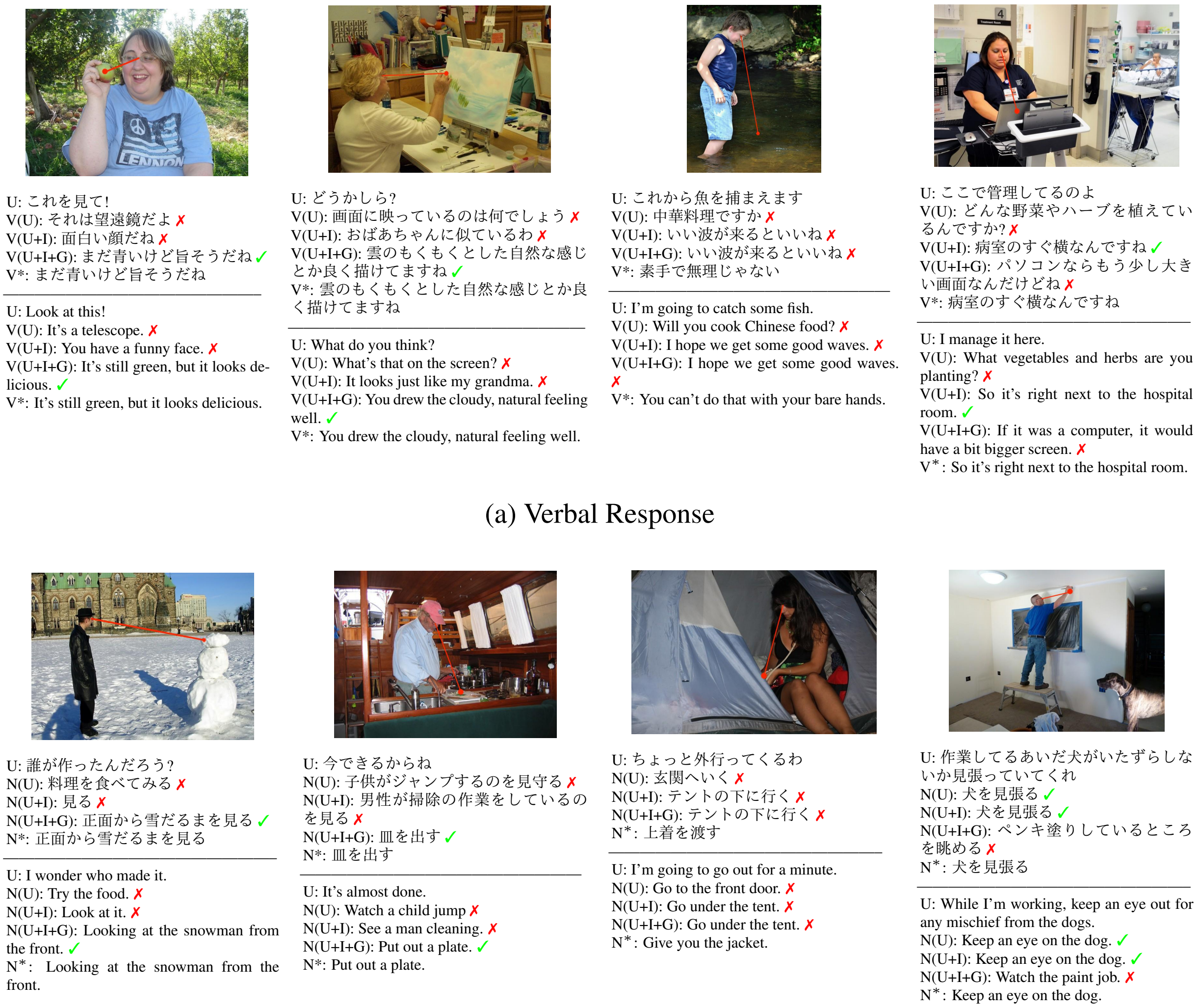
4.7 Qualitative Analysis

Here, we inspect the verbal and non-verbal responses selected by the baseline model, BERT × VGGNet. Figure 5 (a) shows the selected verbal responses. The selected non-verbal responses are shown in Figure 5 (b). The other examples can also be found in the supplemental material.

In the leftmost example of Figure 5 (a), the model cannot understand what the pronoun "this" in the human utterance refers to without the image. By using the image information (U+I), the model wrongly focuses on the human face in the image and responds, "You have a funny face." By using the eye-gaze locations (U+I+G), the model understands that the person is paying attention to the green apple and succeeds in finding the correct response.

In the second example from the left in Figure 5 (a), the human utterance, "What do you think?", is too ambiguous. By using the image information (U+I), we can see that the model wrongly focuses on the speaker, as in the previous example. The eye-gaze locations (U+I+G) allow the model to understand that the speaker asks about the painting and finds the correct response.

The right two examples in Figure 5 (a) show the failure cases. In the third example from the left, it is difficult for the model to select the correct response, "You can't do that with your bare hands", because it requires the world knowledge that fish are hard to catch without tools. In the rightmost example, the woman's gaze is on the computer, which wrongly lets the model focus on the computer instead of the next "hospital room."



U: これを見て!
V(U): それは望遠鏡だよ ✗
V(U+I): 面白い顔だね ✗
V(U+I+G): まだ青いけど旨そうだね ✓
V*: まだ青いけど旨そうだね

U: Look at this!
V(U): It's a telescope. ✗
V(U+I): You have a funny face. ✗
V(U+I+G): It's still green, but it looks delicious. ✓
V*: It's still green, but it looks delicious.

U: どうかしら?
V(U): 画面に映っているのは何でしょう ✗
V(U+I): おばあちゃんに似ているわ ✗
V(U+I+G): 雲のもくもくとした自然な感じとか良く描けてますね ✓
V*: 雲のもくもくとした自然な感じとか良く描けてますね

U: What do you think?
V(U): What's that on the screen? ✗
V(U+I): It looks just like my grandma. ✗
V(U+I+G): You drew the cloudy, natural feeling well. ✓
V*: You drew the cloudy, natural feeling well.

U: これから魚を捕まえます
V(U): 中華料理ですか ✗
V(U+I): いい波が来るといいね ✗
V(U+I+G): いい波が来るといいね ✗
V*: 素手で無理じゃない

U: I'm going to catch some fish.
V(U): Will you cook Chinese food? ✗
V(U+I): I hope we get some good waves. ✗
V(U+I+G): I hope we get some good waves. ✗
V*: You can't do that with your bare hands.

U: ここで管理してるのよ
V(U): どんな野菜やハーブを植えているんですか? ✗
V(U+I): 病室のすぐ横なんですね ✓
V(U+I+G): パソコンならもう少し大きい画面なんだけどね ✗
V*: 病室のすぐ横なんですね

U: I manage it here.
V(U): What vegetables and herbs are you planting? ✗
V(U+I): So it's right next to the hospital room. ✓
V(U+I+G): If it was a computer, it would have a bit bigger screen. ✗
V*: So it's right next to the hospital room.

(a) Verbal Response

U: 誰が作ったんだろう?
N(U): 料理を食べてみる ✗
N(U+I): 見る ✗
N(U+I+G): 正面から雪だるまを見る ✓
N*: 正面から雪だるまを見る

U: I wonder who made it.
N(U): Try the food. ✗
N(U+I): Look at it. ✗
N(U+I+G): Looking at the snowman from the front. ✓
N*: Looking at the snowman from the front.

U: 今できるからね
N(U): 子供がジャンプするのを見守る ✗
N(U+I): 男性が掃除の作業をしているのを見る ✗
N(U+I+G): 皿を出す ✓
N*: 皿を出す

U: It's almost done.
N(U): Watch a child jump ✗
N(U+I): See a man cleaning. ✗
N(U+I+G): Put out a plate. ✓
N*: Put out a plate.

U: ちょっと外行ってくるわ
N(U): 玄関へいく ✗
N(U+I): テントの下に行く ✗
N(U+I+G): テントの下に行く ✗
N*: 上着を渡す

U: I'm going to go out for a minute.
N(U): Go to the front door. ✗
N(U+I): Go under the tent. ✗
N(U+I+G): Go under the tent. ✗
N*: Give you the jacket.

U: 作業してるあいだ犬がいたずらしないか見張っていてくれ
N(U): 犬を見張る ✓
N(U+I): 犬を見張る ✓
N(U+I+G): ペンキ塗りしているところを眺める ✗
N*: 犬を見張る

U: While I'm working, keep an eye out for any mischief from the dogs.
N(U): Keep an eye on the dog. ✓
N(U+I): Keep an eye on the dog. ✓
N(U+I+G): Watch the paint job. ✗
N*: Keep an eye on the dog.

(b) Non-verbal Response

Figure 5: Verbal and non-verbal responses selected by the baseline model, BERT × VGGNet. U, V, N denote the human utterance and the selected verbal and non-verbal response, respectively. V* and N* indicates the gold-standard responses. We show the input modalities (U, I, G) used to produce the response in parentheses. We mark the correct responses by ✓, while the incorrect responses are marked ✗.

Similar phenomena can be observed for non-verbal response selection. In the left two examples in Figure 5 (b), it is hard to identify the human intentions from the utterances alone. The images (U+I) provide important contextual information, but it is still not sufficient for properly understanding the intentions of the utterances. The eye-gaze locations (U+I+G) enable the models to identify the human intentions and respond more accurately. For instance, in the second example from the left, it is hard to understand what the man is doing due to the mess in the room; however, if you look at the tip of the man's gaze, you can see that he is cutting vegetables with a kitchen knife. In such cases, eye-gaze information works particularly well when many objects are present.

We also show the failure examples for non-verbal response selection. We consider that the third example from the left is difficult because the agent has to be thoughtful just like preparing a jacket. In the rightmost example, the man is asking someone to keep an eye on the dog; however, he is not looking at it, so it appears that his gaze has a negative effect.

In summary, we found that first-person images and eye-gaze information are effective in the following cases: (1) when the utterance is ambiguous, e.g., when it contains indicative pronouns like ”this”, and (2) when there are many objects in the image, and it is difficult to identify what the speaker is talking about. These are very common in everyday conversation. Thus, we consider that it would be effective and beneficial to develop social robots that interact with first-person visual information, including gaze, in real-world applications.

5 Conclusion

In this paper, we have presented the VFD dataset with verbal and non-verbal responses. We manually annotated 308K human utterances and 308K verbal and 81K non-verbal responses of agents, which are grounded in the agents' first-person images with human eye-gaze locations. We confirmed the validity of the first-person view in the experiments for the response selection tasks; however, this task (especially, non-verbal response production) remains challenging, and improvements are required.

References

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Atef Ben-Youssef, Chloé Clavel, Slim Essid, Miriam Bilac, Marine Chamoux, and Angelica Lim. 2017. Ue-hri: a new dataset for the study of user engagement in spontaneous human-robot interactions. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 464–472.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

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 the European Conference on Computer Vision*, pages 383–398.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Zhichao Hu, Michelle Dick, Chung-Ning Chang, Kevin Bowden, Michael Neff, Jean E Fox Tree, and Marilyn Walker. 2016. A corpus of gesture-annotated dialogues for monologue-to-dialogue generation from personal narratives. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 3447–3454.

Bernd Huber, Daniel McDuff, Chris Brockett, Michel Galley, and Bill Dolan. 2018. Emotional dialogue generation using image-grounded language models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–12.

Michael Steinbach George Karypis, Vipin Kumar, and Michael Steinbach. 2000. A comparison of document clustering techniques. In *Proceedings of the Text Mining Workshop on Knowledge Discovery and Data Mining*.

Diederik P Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic gradient descent. In *Proceedings of the International Conference on Learning Representations*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision*, pages 153–169.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237.

Thao Le Minh, Nobuyuki Shimizu, Takashi Miyazaki, and Koichi Shinoda. 2018. Deep learning based multi-modal addressee recognition in visual scenes with utterances. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 1546–1553.

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 801–809.

Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. 2015. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*, pages 462–472.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1802–1813.

Xi Ouyang, Shigenori Kawaai, Ester Gue Hua Goh, Shengmei Shen, Wan Ding, Huaiping Ming, and Dong-Yan Huang. 2017. Audio-visual emotion recognition using deep transfer learning and multiple temporal models. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pages 577–582.

Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. 2015. Where are they looking? In *Proceedings of the Advances in Neural Information Processing Systems*, pages 199–207.

Amrita Saha, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations*.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.

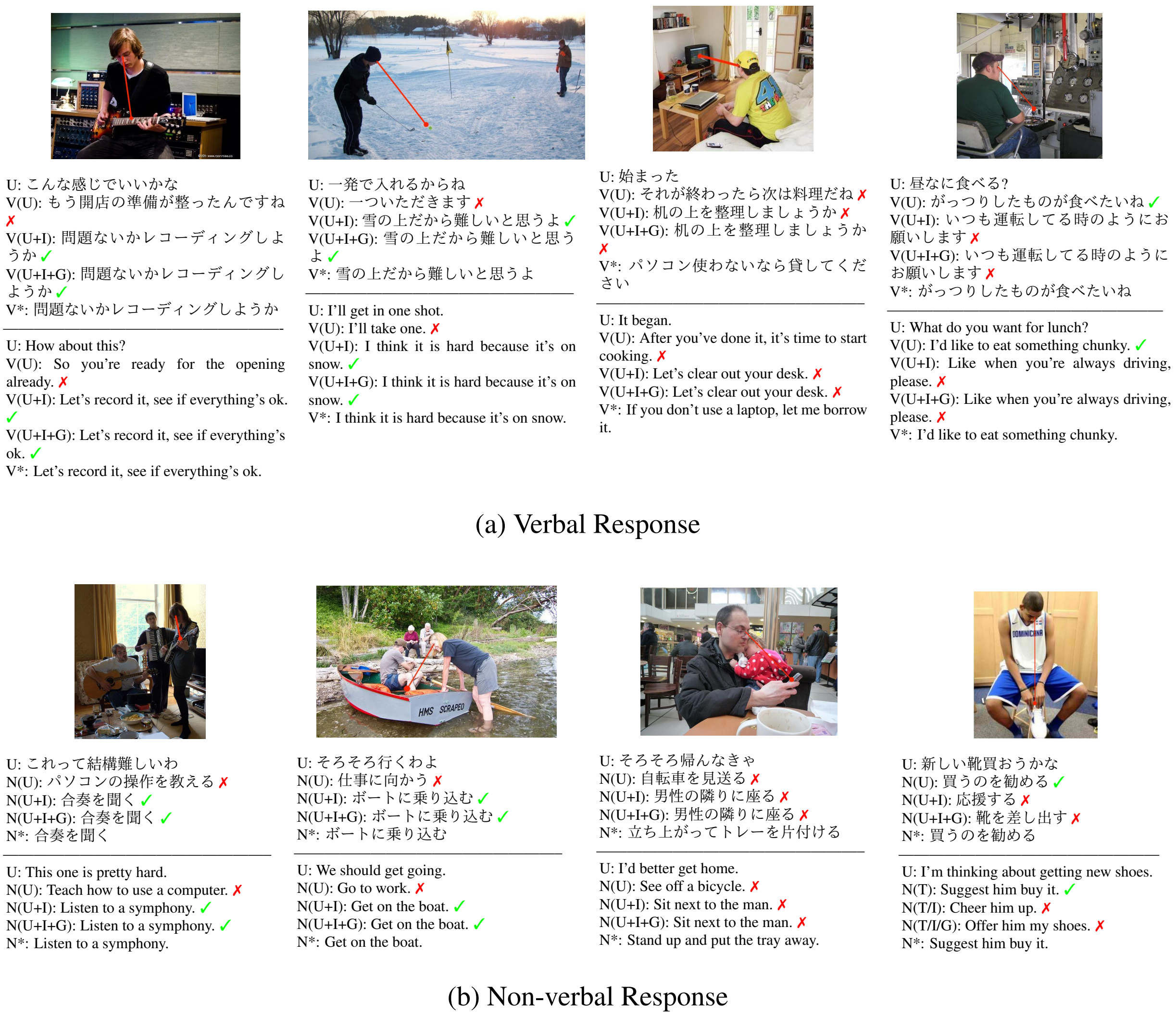
Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. 2017. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1182.

Ping Wei, Yang Liu, Tianmin Shu, Nanning Zheng, and Song-Chun Zhu. 2018. Where and why are they looking? jointly inferring human attention and intentions in complex tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6801–6809.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Novanto Yudistira and Takio Kurita. 2017. Gated spatio and temporal convolutional neural network for activity recognition: towards gated multimodal deep learning. *EURASIP Journal on Image and Video Processing*, 2017.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer vision and Pattern Recognition*, pages 4995–5004.



U: こんな感じでいいかな
V(U): もう開店の準備が整ったんですね ✗
V(U+I): 問題ないかレコーディングしようか ✓
V(U+I+G): 問題ないかレコーディングしようか ✓
V*: 問題ないかレコーディングしようか

U: How about this?
V(U): So you're ready for the opening already. ✗
V(U+I): Let's record it, see if everything's ok. ✓
V(U+I+G): Let's record it, see if everything's ok. ✓
V*: Let's record it, see if everything's ok.

U: 一発で入れるからね
V(U): 一ついただきます ✗
V(U+I): 雪の上だから難しいと思うよ ✓
V(U+I+G): 雪の上だから難しいと思うよ ✓
V*: 雪の上だから難しいと思うよ

U: I'll get in one shot.
V(U): I'll take one. ✗
V(U+I): I think it is hard because it's on snow. ✓
V(U+I+G): I think it is hard because it's on snow. ✓
V*: I think it is hard because it's on snow.

U: 始まった
V(U): それが終わったら次は料理だね ✗
V(U+I): 机の上を整理しましょうか ✗
V(U+I+G): 机の上を整理しましょうか ✗
V*: パソコン使わないなら貸してください

U: It began.
V(U): After you've done it, it's time to start cooking. ✗
V(U+I): Let's clear out your desk. ✗
V(U+I+G): Let's clear out your desk. ✗
V*: If you don't use a laptop, let me borrow it.

U: 昼なに食べる?
V(U): がっつりしたものが食べたいね ✓
V(U+I): いつも運転してる時のようにお願いします ✗
V(U+I+G): いつも運転してる時のようにお願いします ✗
V*: がっつりしたものが食べたいね

U: What do you want for lunch?
V(U): I'd like to eat something chunky. ✓
V(U+I): Like when you're always driving, please. ✗
V(U+I+G): Like when you're always driving, please. ✗
V*: I'd like to eat something chunky.

(a) Verbal Response

U: これって結構難しいわ
N(U): パソコンの操作を教える ✗
N(U+I): 合奏を聞く ✓
N(U+I+G): 合奏を聞く ✓
N*: 合奏を聞く

U: This one is pretty hard.
N(U): Teach how to use a computer. ✗
N(U+I): Listen to a symphony. ✓
N(U+I+G): Listen to a symphony. ✓
N*: Listen to a symphony.

U: そろそろ行くわよ
N(U): 仕事に向かう ✗
N(U+I): ボートに乗り込む ✓
N(U+I+G): ボートに乗り込む ✓
N*: ボートに乗り込む

U: We should get going.
N(U): Go to work. ✗
N(U+I): Get on the boat. ✓
N(U+I+G): Get on the boat. ✓
N*: Get on the boat.

U: そろそろ帰んなきゃ
N(U): 自転車を見送る ✗
N(U+I): 男性の隣りに座る ✗
N(U+I+G): 男性の隣りに座る ✗
N*: 立ち上がってトレーを片付ける

U: I'd better get home.
N(U): See off a bicycle. ✗
N(U+I): Sit next to the man. ✗
N(U+I+G): Sit next to the man. ✗
N*: Stand up and put the tray away.

U: 新しい靴買おうかな
N(U): 買うのを勧める ✓
N(U+I): 応援する ✗
N(U+I+G): 靴を差し出す ✗
N*: 買うのを勧める

U: I'm thinking about getting new shoes.
N(T): Suggest him buy it. ✓
N(T/I): Cheer him up. ✗
N(T/I/G): Offer him my shoes. ✗
N*: Suggest him buy it.

(b) Non-verbal Response

Figure 6: Additional verbal and non-verbal responses selected by the baseline model, BERT × VGGNet. U, V, N denote the human utterance and the selected verbal and non-verbal response, respectively. V* and N* indicates the gold-standard responses. We show the input modalities (U, I, G) used to produce the response in parentheses. We mark the correct responses by ✓, while the incorrect responses are marked ✗.

A Supplemental Material

Here, we show additional examples for verbal and non-verbal responses selected by the baseline model, BERT × VGGNet.

What these examples have in common is that the intentions of the utterances are ambiguous in isolation, which is common in everyday conversation. For instance, in the leftmost example of Figure 6 (a), it is hard for machines to identify what the pronoun "this" refers to.

We show four successful examples on the left side of Figure 6 (a), (b). By using first-person perspective visual information (U+I or U+I+G), the models can understand the intentions correctly. For instance, in the leftmost example of Figure 6 (a), the model correctly understands that the speaker is asking about his playing. In the second example from the left in Figure 6 (a), the visual information allows the model to understand that the speaker is talking about the golf game. Also, in the left two examples in Figure 6 (b), the models successfully utilize the visual information to understand the human intentions.

We also show four failure examples on the right side of Figure 6 (a), (b). In the third example from the left in Figure 6 (a), it is difficult to choose the ground truth response (V*) because the human speaker is watching TV and talking about it, while the ground truth one is talking about the laptop on the desk. In the rightmost example in Figure 6 (a), the visual information is not useful because the utterance is not sufficiently related to the given image. The third example from the left in Figure 6 (b) is also difficult because the agent has to have the common knowledge that we must put away the used trays before we leave in a cafe. In the rightmost example in Figure 6 (b), we consider that the visual information wrongly lets the models take actions related to more specific information about players or shoes rather than the more general action of suggesting to buy the shoes.