RefEgo: Referring Expression Comprehension Dataset from First-Person Perception of Ego4D

Shuhei Kurita[1*] Naoki Katsura[2,1] Eri Onami[3,1]
[1]RIKEN [2]University of Tsukuba [3]Nara Institute of Science and Technology
shuhei.kurita@riken.jp, n-a-katsura@mercari.com, onami.eri.ob6@is.naist.jp

Abstract

Grounding textual expressions on scene objects from first-person views is a truly demanding capability in developing agents that are aware of their surroundings and behave following intuitive text instructions. Such capability is of necessity for glass-devices or autonomous robots to localize referred objects in the real-world. In the conventional referring expression comprehension tasks of images, however, datasets are mostly constructed based on the web-crawled data and don't reflect diverse real-world structures on the task of grounding textual expressions in diverse objects in the real world. Recently, a massive-scale egocentric video dataset of Ego4D was proposed. Ego4D covers around the world diverse real-world scenes including numerous indoor and outdoor situations such as shopping, cooking, walking, talking, manufacturing, etc. Based on egocentric videos of Ego4D, we constructed a broad coverage of the video-based referring expression comprehension dataset: RefEgo. Our dataset includes more than 12k video clips and 41 hours for video-based referring expression comprehension annotation. In experiments, we combine the state-of-the-art 2D referring expression comprehension models with the object tracking algorithm, achieving the video-wise referred object tracking even in difficult conditions: the referred object becomes out-of-frame in the middle of the video or multiple similar objects are presented in the video.

1. Introduction

It is a truly demanding task to identify surrounding objects in real world scenes from video clips of egocentric viewpoints with free-form language supervisions. Such a task is necessary for glass-devices or autonomous robots that help with daily-life tasks and communicate with us in language because they need to understand the intuitive expressions of languages and ground them into the surrounding world. It is an ultimate goal for the referring expression comprehension (REC) or shortly "visual grounding" task because it maps the referred entities in text to the corresponding objects identified and tracked from the observed sequence of images.

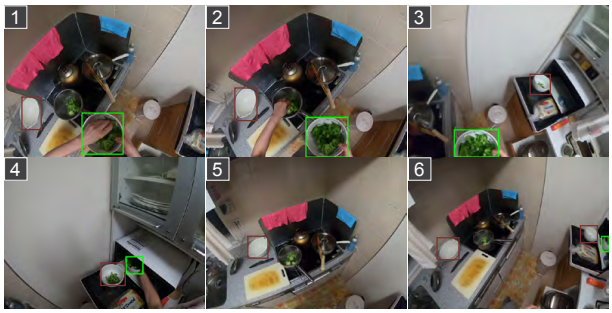


Figure 1. Sample frames of RefEgo for "the large white bowl with broccoli inside that is used to load the pan of broccoli." The referred object of the bowl in green and other bowls in brown.

Extensive efforts are being made in 2D image reference expression comprehension [12, 21, 33, 19]. Recent semi-supervised approaches contribute to the open-vocabulary object detection from 2D images [11, 35]. However, compared to these extensive studies on 2D image referring expression comprehension, we notice that comparably less efforts are taken in video-based referring expression comprehension [16, 13, 24]. Video clips in such datasets are mostly collected in the Internet and aren't suitable for the real-world daily-task understandings. The number of video clips are also limited. Ideally, video clips for such tasks are collected in embedded form in our daily lives and cover variety domains such as walking streets, shopping, chatting with others, staying in indoor, cleaning laundries, or cooking foods, when we pursue general purpose models in our daily scenes. However, it was nearly prohibitive to create datasets on such tasks because of the lack of the collection of real-world setting egocentric videos.

Recently, Ego4D [9], a massive-scale collection of egocentric video and annotation is proposed. Ego4D videos are gathered by 931 unique participants in 74 locations world-

*Corresponding author.

Video REC	Base Dataset	# Clips	# Object Annotations	# Objects	# Categories
Lingual OTB99 [16]	OTB100 [18]	99	58,733	99	-
Lingual ImageNet Videos [16]	ImageNet VID [23]	100	23,855	100	25
Video Object Segmentation	**Base Dataset**	**# Clips**	**# Object Annotations**	**# Objects**	**# Categories**
ReferDAVIS-16 [13]	DAVIS [20]	50	3,440	50	-
ReferDAVIS-17 [13]	DAVIS [20]	90	13,540	205	-
Refer-Youtube-VOS [24]	Youtube-VOS [30]	3,252	133,886	6,048	78
RefEgo (ours)	Ego4D [9] (First-person video)	12,038	226,319	12,038	505

Table 1. Comparison of the video-based referring expression comprehension datasets.

REC dataset	# Images	# Object Annotations
RefCOCO [33]	19,994	50,000
RefCOCO+ [33]	19,992	49,856
RefCOCOg [19]	26,711	54,822
RefEgo (ours)	226,319	226,319

Table 2. Comparison of the RefEgo dataset to 2D REC datasets. RefCOCO/+/g datasets are based on MSCOCO [5] images while ours is based on real-world egocentric video of Ego4D.

wide. They are captured by a head-mounted camera device, intending to capture various daily-life activities in first person vision, covering hundreds of daily life activities including various locations: in-store, office-space, houses, wirehouse, street and so on. Based on Ego4D videos, we constructed the novel in a margin larger RefEgo video-based referring expression comprehension dataset with the help of object detection and human annotation, aiming to ground intuitive language expressions on various contexts in real-world first person perception.

Our RefEgo dataset exhibits unique characteristics that make it challenging to localize the textually referred objects. It is based on egocentric videos and hence includes frequent motions in video clips. The referred objects are often surrounded by other similar objects of the same class. The referred object may appear at the edge of the image frame or even goes out-of-frame in some frames, requiring models to discriminate images that contain and don't contain the referred object. Fig. 1 presents the selective frames from a single video clip with a referred expression of "the large white bowl with broccoli inside that is used to load the pan of broccoli." There are several other bowls in image frames and the referred object goes out-of-frame in the fifth frame. In this case, the models are expected to predict that the referred object is not presented in the image, illustrating the challenging characteristics of the proposed RefEgo.

We prepared valuable baseline models from multiple approaches for RefEgo. We applied the state-of-the-art REC models of MDETR [11] and OFA [28] for RefEgo. We introduced MDETR models trained with all images including image frames with no annotated referred objects, and observed it performs better at discriminating images without referred objects than other models. We also introduce MDETR with a special binary head to discriminate images without the referred object presented. We finally apply the object tracking of ByteTrack [34] for combining multiple detection results, allowing the models to spatio-temporal localization of the referred object.

2. Related Work

2.1. Referring expression comprehension

Grounding textual expressions in objects is a key component in vision-and-language studies. It includes several formalism depending on the data type of visual information reflecting the spatial and temporal diversity of the real-world data: a single image, sequence of images or video clips, and 3D reconstructed data. Referring expression comprehension in image, or simply *visual grounding*, is a task to localize objects in an image from open vocabulary texts [12, 21, 33, 19, 17, 29]. This is one of the most active research field in the vision and language field and many advanced approaches are proposed in these years [32, 15, 11, 28]. However, as these studies are limited to a given single image and hence these REC models have limited knowledge of the referred object and its surrounding environments. 3D-scene based approaches for referring expression are also another major branch in real-worlds scene grounding [4, 27, 22, 7, 1]. Although there are numerous benefits in 2D and 3D referring expression comprehension datasets, these spatial datasets lack of the temporal localization in the real-world.

2.2. Video-based REC

Language-based object tracking Video based localization is a both temporal and spatial localization of objects in video frames. These datasets are often provided as a language-annotation extension to the existing dataset and changes the core concept to determine what object to track. Lingual OTB99 and ImageNet Videos [16] are language-based object tracking datasets that are based on existing object tracking dataset [18, 23]. Here language annotation is attached for first image to specify the object to track in later frames. For limited domain sets, person category annotation

is performed [31]. VID-Sentence dataset [6] also annotated in a part of the ImageNet VID dataset. Co-grounding network [25] and DCNet [2] is proposed for these video-based REC dataset. It is also notable that while multi-object tracking datasets [14] tend to cover limited object classes, the object tracking dataset of TAO [8] includes 894 objects of 345 free-form text classes in a part of their dataset.

Language-based video object segmentation Video object segmentation (VOS) is the segmentation task of a target object in a video clip [13, 24]. These tasks are constructed on the existing VOS datasets [20, 30]. In conventional setting, the target object is specified by a pixel-accurate mask, while the language-based localization annotation was proposed [13, 24] for language-based specification instead of the pixel masks.

Our dataset is based on first-person videos and have specific characteristics discussed in Sec. 3.1. We provide further detailed comparisons with language-based object tracking and video segmentation datasets in Table 1 and the scale comparisons with the conventional 2D REC datasets in Table 2. Our dataset is in a magnitude larger than the existing REC datasets in terms of the number of the total object annotations and the unique objects tracked.

2.3. Ego4D episodic memory benchmark

The Ego4D dataset provides the episodic memory benchmark that aims to query videos and localize the answer of the query. They provide the *visual query*-based VQ2D and VQ3D tasks, the *natural language queries*-based NLQ task for determining the temporal window of the video history where the query answer is evident, and the *moments queries* of MQ task for localizing all instances of the given activity name. Among these tasks, the NLQ task is based on the flexible natural language query. However, there are steep differences between NLQ and our task: in the NLQ task, models localize the temporal window of the occurrence (e.g., "What did I put in the drawer?"), while in our dataset, models localize the directly referred object (e.g., "the cushion on the right end of the sofa") in spatial and temporal manner. As it requires explicit spatial localization and tracking for all temporal frames where the referred object is presented, the video clips of our dataset become shorter than NLQ. Our focus is on creating a comprehensive dataset for first-person video-based referring expression comprehension, which aims to not only concentrate on the different aspects from the Ego4D episodic memory benchmarks, but also contribute them by providing rich annotated data for natural language-based object localization and tracking.

3. RefEgo Dataset

3.1. Task

Our task is to localize a textually referred object in a sequence of images from an egocentric video. Given a single phrase of a referring expression for one target object referred, the model predicts bounding boxes for the referred object in a sequence of images drawn from the Ego4D videos. As the first-person videos include high viewpoint motions, the referred objects sometimes locate out of frames for some images. Therefore, we introduce the task of *discriminating images that do not include the referred object* from other images that include the referred object when the models are given a sequence of images. This is in addition to the conventional object localization task by *predicting a single bounding box of the referred object for images that include the referred object*. We summarize the special conditions that make our RefEgo dataset plausible in real-world experiments as follows.

Frequent viewpoint motion Ego4D videos are captured by wearable cameras and hence experience a number of viewpoint motions. The frequent viewpoint motion makes it challenging to identify the same objects in distinct frames. Here, we assume the REC models can help the object tracking methods as the tracking-by-detection approach. If REC models are accurate enough, it can successfully ground the single target object in images onto the unique referred expressions, therefore they simultaneously solve the visual grounding and object tracking problems, independent of the frequent viewpoint motion.

Detection of the no-referred-object images In existing referring expression comprehension tasks, the target object always appears in the given image. This is, however, unrealistic and uncommonly happens when we develop some glass-devices or autonomous robots that move around in scenes, capture images and search for the referred objects. Therefore our annotation includes images where the referred object does not appear in them in the video video clips. This imposes the referring expression comprehension models on practical and ambitious experimental settings: the models are required to discriminate images that include the referred objects from other images in the sequence of images of the video clip. Our dataset includes 295,530 images in total and 226,319 images (76.6% of the total images) contains the annotated bounding box of the referred target object in them. We also confirmed that at least four images in a single video clip have a target object annotation. The lack of the target object from some frames is common in egocentric videos and becomes one of the greatest challenges for the conventional image-based REC models because such models are trained with images that surely include the target objects.

Multiple similar objects in scenes In previous referring expression comprehension datasets, it sometimes happens

Figure 2. The process of attaching the bounding box for the blue pillow placed far distant in a Ego4D video clip. The annotated referring expression is "a square bright blue pillow on the chair in front of the ottoman."

that some object classes are unique in one image and hence easy to localize the referred object just from the object class. We consider this is due to the lack of the diverseness of the same-class objects in the same scene: if there is only one mug in some room, it is easy to select a mug from other objects just from an object class name. However, it is much more rigorous to localize a single mug cup on a shelf full of mug cups from referring expressions. We therefore choose the objects when there are multiple same-class of objects for them in the video clip for further annotation. We select temporal scenes of video clips that include many objects for annotation. The average number of detected objects in images is 22.1, and the average number of the same class objects present in the images with the target object is 3.5 for all splits [1].

3.2. Dataset creation

Video clip extraction from Ego4D We firstly extracted images from all of the original Ego4d videos and then applied the object detection model of Detic [35] trained with the LVIS [10] dataset to automatically detect object bounding boxes for the extracted images. Based on the object detection results, we sampled video clips that include many detected objects in frames. We did this because detecting and localizing objects from images with multiple objects can be challenging. We also choose video clips that have motions, avoiding video clips with less motion. We also avoided stereo videos for annotations because they have special depth treatments and hence require additional annotation costs in later steps. The frames of the sampled video clips consist of 10 to 40 images. For preserving a wide variety of Ego4D dataset, we ensured the extracted video clips cover the most of activity themes annotated in Ego4D. The detected bounding boxes are also used to provide bounding box candidates of possible target objects for later human annotation on Amazon Mechanical Turk (MTurk). We chose 2 frames per second image extraction for the annotation, considering the annotation cost for longer video clips and the ability of tracking the same object in video clips of the existing object tracking datasets, such as TAO [8] where frames are sampled in 1 frame per second.

Human annotation We used Amazon Mechanical Turk (MTurk) for collecting annotations to massive scale extracted video clips of 12,038. We asked MTurk workers for choosing the same object in frames, editing the detected bounding box of the tracked object, and writing the referred expressions for the target object. For this purpose, we developed an interactive visualization website that presents images that include candidates of bounding boxes from Detic drawn from the sampled video clips. Workers are asked to select the same object in images and write down a referring expression to specify the object. They are also asked to edit the bounding box of the target object by clicking the edges of them to fit the object. Workers are asked to compose a referential expression that is enough detailed to localize the target object by observing all frames in the video clips. We further collected supplementary annotations of the referred object. The further details of the video clip selection and annotation process are in S.M.

[1] We used Detic for this statistics.

Split	# Clips	# Images	# Images with BBox
Train	9,172	225,500	173,183
Val.	1,549	38,470	29,322
Test	1,317	31,560	23,814

Table 3. Dataset statistics.

3.3. Dataset statistics

We finally gathered annotations on 12,038 video clips. The total length of the video clip is 147,765 seconds and the averaged length is 12.3 seconds. Each clip has two different referring expression writings for one annotated object. The Ego4D dataset has its own video clips for episodic memory, hands and objects and audio-visual diarization & social tasks. We assumed our annotation serves supplementary roles for these existing tasks. Therefore, for dataset splitting, we followed these existing splits as much as possible, namely, the Forecasting + Hands & Objects (FHO) splitting. For some video clips without FHO, we follow Episodic Memory(EM) splitting. The remaining videos are for the training set. Table 3 presents the statistics for each split. We make sure that clips sampled from the same video are assigned to the same split. The further detailed dataset statistics, including human accuracy of the REC task, and construction details of the annotations are in the S.M.

4. Model

4.1. Referring expression comprehension models

We first apply the conventional image-based referring expression comprehension models of MDETR [11] and OFA [28] for a sequence of images from video clips.

MDETR MDETR is an end-to-end text-modulated detector based on the DETR [3], state-of-the-art detection frameworks. MDETR archived high performance on the REC benchmark, such as RefCOCO/+/g [33, 19]. It uses the soft token prediction to ground parts of textual expressions and detected regions in images through N learnable embeddings, called object queries, given to the MDETR decoder. Each bounding box prediction is also paired with a special token that represents that the bounding box is not grounded to the given textual phrase. For prediction of each token t_n^* of the referred expression with L tokens, the MDETR decoder derives the probability s_n^i that i-th token is paired to n-th object query as $s_n^i = \frac{\exp t_n^i}{\sum_{j=1}^{L+1} \exp t_n^j}$. Here i is a whole number and $1 \leq i \leq L+1$. s_n^{L+1} is the special token prediction for "no object", of the n-th object query. We use $(1 - s_n^{L+1})$ as the confidence score for the bounding box prediction from the n-th object query.

MDETR with all images In contrast to existing REC datasets, some image frames in video clips don't include the referred objects in the RefEgo dataset. When we train MDETR only with images that include the referred object, MDETR tends to predict bounding boxes with high confidence even if there are no referred objects in the images. To assign low confidence scores for no referred object images, we trained a MDETR model with all images by treating all predicted bounding boxes as negative samples for images without referred objects during training.

MDETR+BH For better confidence scores to discriminate images without referred objects, we also expand the existing MDETR and add an additional object query to the existing prediction heads of MDETR. This additional object query is combined with a binary head (BH) to perform the binary classification trained with the binary cross-entropy loss by determining whether referred objects are in images or not.

OFA OFA [28] is the unifying architecture for various vision and language tasks, e.g., image captioning, visual question answering and visual grounding. Unlike existing object detection models, OFA predicts a bounding box by directly determining the region position in $\langle x1, y1, x2, y2 \rangle$-order with an autoregressive language model prediction. This nature, however, makes it difficult to obtain the "confidence score" for the predicted bounding box. We use the prediction probability of the sequence of $\langle x1, y1, x2, y2 \rangle$ tokens as "confidence score" for the confidence of the prediction. Similar to the original MDETR, we used extracted images that include the bounding box annotations for training.

No-referred-object images detection In RefEgo, models are required to discriminate images that don't contain the referred object. For this purpose, we use the confidence scores of the predicted bounding boxes by simply assigning a threshold on it to determine whether images include the referred object or not.

4.2. Object tracking

Unlike conventional referring expression comprehension on 2D images, RefEgo is a video-based REC dataset where an object is localized through the video frames. Images in video frames often become worse because of motion blur and occlusion. Therefore it is difficult for REC models to consistently find the referred object in all image frames in sequence. In addition, because of the temporal movement of objects in videos, REC models often detect totally different objects in some frames, resulting in inconsistent object tracking in video frames. To reduce inconsistent localization across video frames, We took the tracking-by-detection approach with attaching a tracking algorithm to the image-based REC models. We applied ByteTrack [34] for the results from MDETR prediction. ByteTrack is a state-of-the-art tracking algorithm that tracks objects based on the overlap between adjacent frames. Because both the camera and objects can be in motion, ByteTrack calculates the overlap after predicting positions in the next frame with the Kalman filter. We first extracted 30 frame per second im-

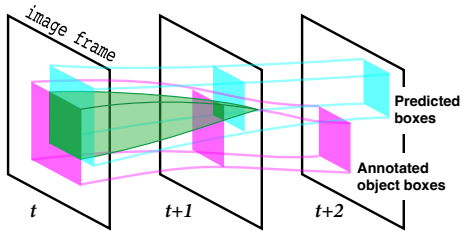


Figure 3. Image frames and STIoU overview. STIoU is defined with the intersection area (green) over the sum of the predicted (purple) and annotated (cyan) object boxes through time frames.

ages from the RefEgo video clips and obtained the MDETR predictions on them. We used ByteTrack for these MDETR prediction results and obtained the candidates of multiple tracked objects. We then introduced a simple heuristics to score the sequences of bounding boxes of the tracked objects with the confidence score from MDETR, and update the predicted bounding boxes from MDETR if the confidence score of the tracked bounding boxes are higher than the original MDETR confidence score. We consider this heuristics-based approach serves as the baseline of the object tracking over referring expression comprehension models. We didn't apply object tracking for OFA results because OFA always predicts a single object in each frame. Thus there are no chances for object tracking to select better bounding boxes. Please see S.M. for further details of the implementation and hyper parameters of the object tracking over referring expression comprehension models.

5. Experiments

5.1. Evaluation metric

In 2D referring expression comprehension, the mean intersection-over-union overlap (**mIoU**) is commonly used for the quality of the target object selection. Following the widely used metric for referring expression comprehension [19, 33], we count IoU > 0.5 cases for positive cases and otherwise negative cases for **AP@50**. **AP@50** is not sensitive to the details of the shape of the predicted bounding boxes but the selection of the objects from other similar objects. These metrics, however, are applicable only when the target object is presented in the images. We apply the traditional **mIoU** and **AP@50** for images that include the annotated target object bounding box, ignoring image frames that don't contain the referred objects.

As the video-based object tracking task, the target object can be invisible or out-of-frame in the sequence of image frames. This is a major difference of this task compared with the existing object tracking datasets. We therefore expand the existing mean IoU metric for the video-based evaluation including frames that don't contain the target object: mean Spatio-Temporal IoU (mSTIoU) and mean IoU with negative prediction (mIoU+n).

Suppose we have a single video clip that consists of $\mathcal{N}$ frames in the frame-per-second evaluated. For i-th image frame in $\mathcal{N}$ frames include an annotated bounding box t_i where its area size is given by $|t_i|$. Among $\mathcal{N}$ frames, $\mathcal{M}$ frames include the target object and hence $|t_i| > 0$ while $|t_i| = 0$ for the remaining $\overline{\mathcal{M}} \cap \mathcal{N}$ frames (here $0 \leq |\mathcal{M}| \leq |\mathcal{N}|$). For each image frame, models predict a bounding box p_i that may become size-0 ($|p_i| \geq 0$), suggesting that the image frame doesn't contain the referred object in it for $|p_i| = 0$ case. The conventional mIoU only for images that include the target object is:

$$\mathrm{mIoU} = \frac{1}{|\mathcal{M}|} \sum_{\mathcal{M}} \frac{|p_i \cap t_i|}{|p_i \cup t_i|}. \tag{1}$$

This is an image-wise metric, ignoring images that don't contain the target object ($\overline{\mathcal{M}} \cap \mathcal{N}$).

We introduce the Spatio-Temporal IoU (STIoU) as the multi-frame summation of the intersection and union of IoUs over a single video clip of $\mathcal{N}$ frames:

$$\mathrm{STIoU} = \frac{\sum_{\mathcal{N}} |p_i \cap t_i|}{\sum_{\mathcal{N}} |p_i \cup t_i|}. \tag{2}$$

STIoU satisfies $0 \leq \mathrm{STIoU} \leq 1$ where $\mathrm{STIoU} = 1$ for the exact math in all frames while $\mathrm{STIoU} = 0$ for complete mismatch of all annotated and predicted bounding boxes. When $|t_i| = 0$, STIoU is not penalized if $|p_i| = 0$ while it decreases due to the larger denominator of $|p_i \cup t_i| = |p_i|$ if $|p_i| > 0$. We use mean STIoU (mSTIoU) where the STIoU is calculated in each video clip and then we take the mean of STIoU inside the validation and test splits.

For Eq. 5.1, we cannot replace $\mathcal{M}$ with $\mathcal{N}$ without any assumptions because $|p_i \cup t_i|$ can become 0 for $|t_i| = 0$ case. However, for $|t_i| = 0$ case, $|p_i \cap t_i|$ is always 0 for any $|p_i|$. Therefore, we extend the conventional IoU metric for $\mathcal{N}$ under an assumption that for $|t_i| \to 0$ and $|p_i| \to 0$ case, $\frac{|p_i \cap t_i|}{|p_i \cup t_i|} \to 1$. We call this **IoU+n** as IoU with the negative prediciton. **IoU+n** can be expressed in the following simple form:

$$\mathrm{IoU+n} = \begin{cases} \frac{|p_i \cap t_i|}{|p_i \cup t_i|} & (|p_i| > 0 \ \text{ or } \ |t_i| > 0) \\ 1 & (|p_i| = 0 \ \text{ and } \ |t_i| = 0) \end{cases} \tag{3}$$

Unlike simple IoU, IoU+n includes the detection of images without the target object. Similar to IoU, we take mean of IoU+n for images and video-clips (mIoU+n), and similar to AP@50, we also introduce **AP@50+n** where we count the IoU + n > 0.5 cases. Fig. 3 visualize the STIoU metric as an IoU extension for time sequences.

Here, the introduced STIoU and IoU+n are video-wised evaluation metrics. STIoU is penalized for predicting finite-sized bounding boxes for images without target objects while IoU+n is rewarded for predicting images without the

Model	RefEgo Val					RefEgo Test				
	All images			Images w/ targets		All images			Images w/ targets	
	mSTIoU	mIoU+n	mAP@50+n	mIoU	mAP@50	mSTIoU	mIoU+n	mAP@50+n	mIoU	mAP@50
From Pretrained										
OFA	33.9	44.8	47.8	**53.2**	**58.4**	32.9	44.9	47.8	**51.9**	**56.8**
MDETR	36.2	42.1	47.9	46.0	53.3	35.4	42.4	48.0	45.0	52.3
+Object tracking	36.3	42.2	47.8	46.1	53.4	35.5	42.4	48.1	45.2	52.4
MDETR (all)	37.2	45.6	51.2	45.0	52.3	36.1	45.0	50.5	44.0	51.1
+Object tracking	37.5	45.5	51.2	45.3	52.6	36.5	45.1	50.7	44.1	51.3
MDETR+BH (all)	37.5	**46.3**	**52.0**	45.2	52.6	36.5	45.6	51.0	45.4	52.7
+Object tracking	**37.9**	46.1	51.9	45.4	52.9	**36.9**	**45.7**	**51.1**	45.7	53.0
From RefCOCOg										
OFA†	16.9	30.2	30.8	30.0	29.6	15.4	28.9	29.3	27.8	27.1
OFA‡	32.7	44.5	47.7	**52.3**	**57.7**	31.7	44.4	47.5	**51.0**	**56.2**
MDETR†	17.4	27.4	28.3	25.1	25.2	15.4	25.6	26.4	22.9	22.8
+Object tracking	17.5	27.3	28.3	25.2	25.2	15.5	25.6	26.3	23.0	22.9
MDETR‡	36.6	42.0	47.6	46.6	53.6	35.8	41.4	46.8	45.8	52.5
+Object tracking	36.7	41.9	47.5	46.7	53.7	35.9	41.3	46.7	45.9	52.7
MDETR‡ (all)	37.9	45.3	50.9	46.4	53.5	37.2	45.0	50.4	45.7	52.6
+Object tracking	38.2	45.3	50.9	46.6	53.8	37.5	45.0	50.4	45.9	52.9
MDETR+BH‡ (all)	37.5	**46.1**	**51.6**	46.4	53.6	36.9	**45.7**	**51.1**	45.7	53.0
+Object tracking	**38.4**	46.0	**51.6**	46.8	54.1	**37.6**	45.4	51.0	46.0	53.4

Table 4. Experimental results on RefEgo validation and test sets. (†) : the off-the-shelf RefCOCOg model performance. (‡) : models are trained with RefEgo from the off-the-shelf RefCOCOg model. Other models are trained with RefEgo from the pretrained checkpoints.

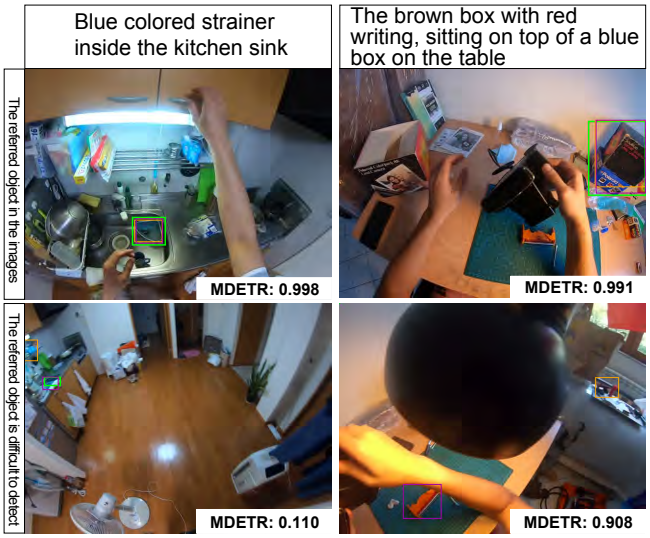
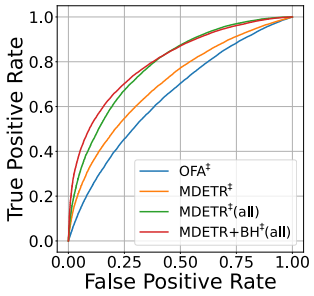


Figure 4. Example bounding box annotated in green, predicted by OFA‡ in orange and MDETR‡ (all) in purple in two images in two columns from the same video clips. When the referred object doesn't exist in the images (below), both OFA‡ and MDETR‡ models detect bounding boxes of other objects.



Model	AUC
(Chance rate)	50.0
OFA‡	64.9
MDETR‡	70.7
MDETR‡ (all)	78.8
MDETR+BH‡ (all)	80.5

Figure 5. **Left**: ROC curves for detection of the no-referred-object images. **Right**: AUC for the prediction whether the referred objects are in images or not.

targets. Among these, **STIoU is the prime video-wised metric of the video-based referring expression comprehension** because it is based on the entire video-clip coverage of the referred object. It is notable that IoU+n is too sensitive to the corner cases where the small fragment of the referred object is presented at the edge of the image frame and detecting such fragments becomes a subtle but critical problem for IoU+n. STIoU is robust to such corner cases because STIoU isn't affected too much by small bounding boxes of the predictions and annotations.

5.2. Video-based REC in RefEgo

We used the state-of-the-art pretrained REC models of OFA-Large [28] and MDETR [11] for our experiments. For MDETR, EfficientNet-B3 [26] is used the visual backbone network in experiments. We prepared OFA, MDETR and MDETR (all) models. OFA and MDETR models use extracted image frames that include the annotated bounding boxes while MDETR (all) uses all image frames including images that do not contain the referred object bounding box as described in Sec. 4. We trained OFA and MDETR models with RefEgo from the pretrained checkpoints of these models. We also prepare OFA‡ and MDETR‡ models from the off-the-shelf models of OFA† and MDETR† trained with RefCOCOg [19].

Figure 6. Qualitative analyses for three video clips. Selective images from left-top (past) to right bottom (future). The green, dotted red and dotted cyan bounding boxes are annotated, predicted by ByteTrack and predicted by REC with the top-1 confidence score, respectively. The texts at the left of images are the referring expressions.

Table 4 presents the performance of the REC models in RefEgo validation and test sets. The current state-of-the-art referring expression comprehension models somehow successfully localize objects in images when the target object is surely in the given image as the OFA models exhibit strong performances in AP@50 of the widely used REC metrics in the validation set. The MDETR‡ (all) and MDETR+BH‡ (all) models, however, achieve better performance in the all images metrics of mSTIoU, mIoU and AP@50+n. This suggests that the confidence scores of these models are more effective in determining the presence or absence of the referred object in the image compared to other models, assuming that this is because the MDETR‡ (all), trained with all images, learns the images that do not include the referred object through contrastive learning. Overall, OFA models are good at predicting accurate bounding boxes when the referred objects are in the images while MDETR models are good at predicting both the bounding boxes and discriminating images without the referred objects.

We also applied ByteTrack-based object tracking for MDETR models as described in Sec. 4.2. In Table 4, we observed the object tracking slightly improves performance especially in mSTIoU and mIoU while it is less effective for mIoU+n. We noticed that the performance with mIoU+n sometimes slightly degrades with object tracking because of the lower discrimination of images without the referred objects. We will take a close look in the quantitative analyses on the comparison between REC and object tracking results.

5.3. Discriminating images without referred objects

It is a difficult task to discriminate images without the referred objects from image frames of the video clips. When the referred objects are out-of-images or not visible, the models have a tendency to predict bounding boxes on objects that appear similar to the referenced expression but are not the correct ones. Fig. 4 presents the predictions of bounding boxes for both models for cases where the target object is in the image (top) and the referred object is difficult to detect or isn't visible in the images due to occlusion caused by other objects.(bottom). We further investigate how accurately REC models discriminate images without referred objects by metrics used for binary classification evaluations under various thresholds on the top-1 bounding box confidence scores. Fig. 5 presents AUC (Area Under the ROC Curve) for ROC. MDETR‡ (all) outperforms OFA‡ and MDETR‡ models in AUC, suggesting better performance in determining images that include the referred object. The binary head of MDETR+BH‡ (all) further contributes discriminating images without referred objects.

Model	RefEgo Val					RefEgo Test				
	All images			Images w/ targets		All images			Images w/ targets	
	mSTIoU	mIoU+n	mAP@50+n	mIoU	mAP@50	mSTIoU	mIoU+n	mAP@50+n	mIoU	mAP@50
Single object of same-class (easy)										
OFA‡	36.5	48.0	51.4	58.2	63.9	35.6	48.2	51.6	56.6	62.4
MDETR+BH‡ (all)	43.9	51.9	58.1	52.3	60.3	42.4	51.3	57.2	51.3	59.2
+object tracking	44.9	51.9	58.2	52.8	60.9	43.1	51.1	57.2	51.5	59.4
Multiple objects of same-class (hard)										
OFA‡	29.3	41.4	44.5	47.2	52.3	28.8	41.5	44.5	46.7	51.6
MDETR+BH‡ (all)	32.0	41.2	46.0	41.3	47.8	32.8	40.9	45.8	41.7	48.3
+object tracking	32.8	41.5	46.6	41.6	48.4	33.6	41.2	46.5	42.0	49.0
Static object (easy)										
OFA‡	32.9	44.7	48.0	52.6	58.2	32.9	45.5	48.8	51.9	57.4
MDETR+BH‡ (all)	37.8	46.6	52.2	46.9	54.3	37.8	46.6	52.2	46.8	54.2
+object tracking	38.6	46.4	52.2	47.3	54.9	38.5	46.4	52.1	47.0	54.6
Moving object (hard)										
OFA‡	31.3	43.1	46.0	50.4	54.9	25.9	38.8	41.3	45.9	50.2
MDETR+BH‡ (all)	35.9	43.5	48.0	43.8	49.6	32.4	40.7	45.6	40.6	46.7
+object tracking	36.9	43.6	48.2	44.1	50.0	33.1	40.6	45.7	41.1	47.4

Table 5. The performance difference due to the the *object-class uniqueness* and *referred object movement* during prediction. **Top**: single (easy) and multiple (hard) objects of the same-class in image frames. **Bottom**: static (easy) and moving (hard) referred object.

5.4. Qualitative analysis

Figure 6 shows the results of MDETR+BH‡ (all) and its object tracking counterparts. In the first two rows of the images, the referred object of the green bowl often goes at the edge of the frames or even out-of-frames in the video clip, making it a challenging scenario for conventional object tracking methods. However, we found that the ByteTrack-based object tracking mode, with the assistance of the REC results, was able to successfully combine the tracked object of the green bowl in many frames. In the images in the third and fourth rows, the REC model makes incorrect bounding box predictions in the middle of these two video clips. However, the object tracking method successfully continues to track the same object. These video clips include multiple objects of the same class, which may cause confusion for the REC models.

5.5. Detailed performance analyses

We present detailed analyses on the *object-class uniqueness* and *referred object movement*. If there are multiple objects of the same class in one video clip, it becomes more difficult to track the identical object. Similarly, if the referred object is moved to another place during the video clip, it is challenging to precisely localize the referred object in images frames. Based on the OFA‡ and MDETR+BH‡ (all) models, we derived the detailed scores for the following four cases in the RefEgo validation and test sets in Table 5, confirming that the model performance degrades when the objects are moved or multiple similar objects exist in the scenes.

6. Conclusion

Based on the wide-variety world-wide first-person perception dataset of Ego4D, we constructed the RefEgo dataset for the real-world and egocentric video-based referring expression comprehension. This dataset is not only larger than existing REC datasets in terms of images with annotated bounding boxes, but it is also grounded on the real-world egocentric videos, making it a valuable and challenging task for precisely grounding natural languages in real-world contexts. In experiments, we combined the REC and object tracking approaches to spatio-temporally localize referred objects even in challenging conditions, such as when the referred object goes out-of-frames in the middle of the clips or the REC model makes several incorrect predictions. This approach provides a precious baseline for first-person video-based REC datasets.

Limitations We have annotated our dataset using images from the Ego4D first-person video dataset. As a result, our usage terminology and limitations for videos and images align with those of the original Ego4D dataset. Our dataset encompasses diverse video domains, including indoor and outdoor scenes. However, it is important to note that we haven't included videos from domains rarely found in the Ego4D dataset.

Acknowledgments This work was supported by JSPS KAKENHI Grant Number 22K17983 and 22KK0184, and by JST PRESTO Grant Number JPMJPR20C2.

References

[1] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[2] Meng Cao, Ji Jiang, Long Chen, and Yuexian Zou. Correspondence matters for video referring expression comprehension. *the 30th ACM International Conference on Multimedia*, 2022. 3

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 5

[4] Dave Zhenyu Chen, Angel X. Chang, and Matthias Nießner. Scanrefer: 3d object localization in RGB-Dscans using natural language. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[5] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[6] Zhenfang Chen, Lin Ma, Wenhan Luo, and Kwan-Yee Kenneth Wong. Weakly-Supervised Spatio-Temporally Grounding Natural Sentence in Video. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 3

[7] Samyak Datta, Sameer Dharur, Vincent Cartillier, Ruta Desai, Mukul Khanna, Dhruv Batra, and Devi Parikh. Episodic memory question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19119–19128, June 2022. 2

[8] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A Large-Scale Benchmark for Tracking Any Object. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 3, 4

[9] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Christian Fuegen, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the World in 3,000 Hours of Egocentric Video. In *Proceedings of the IEEE/CVF Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[10] Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4

[11] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. MDETR - modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 5, 7

[12] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 1, 2

[13] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2018. 1, 2, 3

[14] Laura Leal-Taixé, Anton Milan, Ian D. Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. *arXiv preprint arXiv:1504.01942*, 2015. 3

[15] Muchen Li and Leonid Sigal. Referring Transformer: A One-step Approach to Multi-task Visual Grounding. In *Proceedings of the Neural Information Processing Systems (NeurIPS)*, 2021. 2

[16] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees G. M. Snoek, and Arnold W. M. Smeulders. Tracking by Natural Language Specification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[17] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. CLEVR-Ref+: Diagnosing Visual Reasoning With Referring Expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[18] Yang Lu, Tianfu Wu, and Song-Chun Zhu. Online Object Tracking, Learning, and Parsing with And-Or Graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2

[19] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille, and Kevin Murphy. Generation and Comprehension of Unambiguous Object Descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 5, 6, 7

[20] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 2, 3

[21] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 1, 2

[22] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. REVERIE: Remote Embodied Visual Referring Expression in Real Indoor Environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, (3):211–252, 2015. 2

[24] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3

[25] Sijie Song, Xudong Lin, Jiaying Liu, Zongming Guo, and Shih-Fu Chang. Co-grounding networks with semantic attention for referring expression comprehension in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 3

[26] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. 7

[27] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. RIO: 3D Object Instance Re-Localization in Changing Indoor Environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[28] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022. 2, 5, 7

[29] Chenyun Wu, Zhe Lin, Scott Cohen, Trung Bui, and Subhransu Maji. PhraseCut: Language-based Image Segmentation in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[30] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian L. Price, Scott Cohen, and Thomas S. Huang. YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2, 3

[31] Masataka Yamaguchi, Kuniaki Saito, Y. Ushiku, and Tatsuya Harada. Spatio-Temporal Person Retrieval via Natural Language Queries. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[32] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[33] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. Modeling Context in Referring Expressions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 1, 2, 5, 6

[34] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2, 5

[35] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting Twenty-thousand Classes using Image-level Supervision. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 4