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

A. Dataset Details

We constructed the RefEgo video-based referring expression comprehension dataset based on the world-wide various-domains first-person videos of Ego4D. Our dataset covers 5,012 videos of all Ego4D [2] videos. The total length of video clips are more than 41 hours. We summarize the statistics of our dataset in Table 5. Each video clip has two referring expression annotations. In the validation and test set, we also collect additional classification labels for the referred object in each video clip: the *object-class uniqueness* and *referred object movement*. *Object-class uniqueness* is the label whether there are any other objects of the same class with the referred object or not. *Referred object movement* is the label whether the referred object is moving/moved in the video clip or not. These additional labels are used for detailed analyses in Sec. B.

# Annotated video clips	12,038
# Sampled Ego4D videos	5,012
Annotated FPS	2
Total video length (sec)	147,765
Ave. video length (sec)	12.3
Ave. ref. exp. length (words)	13.4

Table 5. Detailed statistics for RefEgo dataset.

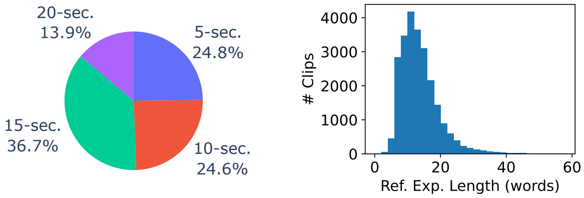


Figure 7. **Left**: Ratio of video clip length. **Right**: Length of referring expression.

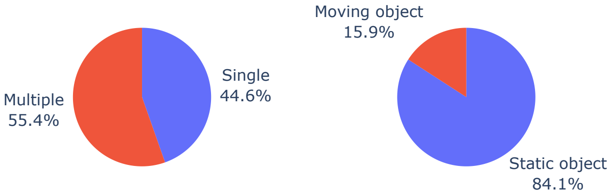


Figure 8. **Left**: Ratio of the single object of the same-class and the multiple object of the same-class in image frames. **Right**: Ratio of the static referred object and the moving referred object.

A.1. Video clips in RefEgo

In the RefEgo dataset, we firstly tried to cover as much as videos in Ego4D in order to follow a variety of topics. However, we soon notice that some videos are not suitable for the referring expression comprehension task because they have a considerably small amount of detected objects in image frames. The variety of object classes is also limited. Therefore we chose video clips in terms of the variety of objects. Firstly we extracted two images per second from all images and applied the object detector Detic. Then we chose the video clips where many classes of objects are in them according to the Detic results. We also manually removed several videos that do not include many objects and hence are not suitable for the referring expression comprehension task (e.g., farm fieldwork or boarding on a leisure boat). For the remaining videos, we sampled video clips that include as many objects as possible and they are in motion. We notice there are some videos with very few motion of the view points (e.g., watching a movie). We therefore sampled video clips in motion using both absolute difference of images and difference of the detected objects. The length of the sampled video clips are chosen in $\{5, 10, 15, 20\}$ seconds. We present the ratio of each clip length and the distribution of the referring expression length in Fig. 7. The video clip ratio of *object-class uniqueness* and *referred object movement* in the RefEgo validation and test set is presented in Fig. 8. Frequently seen object categories inferred by Detic are presented in Fig. 9 and frequently seen words in captions in Fig. 10.

A.2. Dataset Annotation

The annotation process via Amazon Mechanical Turk (MTurk) is performed by following two steps: (i) object tracking and first referring expression attachment, (ii) tracked object check and second referring expression attachment. In the first step, we collected the tracking data for a single object in the video clip. In the second step, we ask workers to check whether the same object is tracked for the results of the step-(i). We also manually check the failed video clips and remove some of those clips if they are not suitable for our task. Such filtered video clips are re-annotated through step-(i) again.

(i) object tracking and first referring expression attachment In object tracking annotation, our approach is the selection and correction from the boundary box candidates. As written in the previous section, we automatically extracted object boundary boxes in each image with Detic. We then present MTurk workers all images in the sampled video clips with the extracted boundary boxes for each image. The all extracted boundary boxes serve as candidates for the single tracked object. We recommend that workers view all images in the sampled clip first and then select a single boundary box for each image. Workers can reshape bounding boxes to fit the target object if no boundary boxes are presented on the target object in some images.

For the target object class, we sampled one from the frequent object classes in the video clip depending on the LVIS class of Detic. We present workers this target object class and ask them to find one of the objects and track it. Note that we allow workers to track objects that are not in the target object class if they are unavailable or difficult to track in images. We infer the tracked object labels based on the most frequent object labels after the entire annotation.

(ii) tracked object check and second referring expression attachment We present workers images where a single referred object with a bounding box is attached for each image and ask them to check whether the same object is tracked through the images. **Here, to make sure the same objects are tracked in the dataset, object tracking annotations that do not pass the same object tracking test in step-(ii) are removed from the dataset or re-annotated in (i) step.** 5.8% annotations are marked as they do not track the same object in the video clips and hence removed or re-annotated. We re-annotate or remove these video clips. Note that this figure doesn't include the results of "null workers" in MTurk because results of "null workers" are removed and reassigned to other workers immediately upon they are found. We also ask workers to compose an additional referring expression in addition to the auxiliary information labels of the object stationary and uniqueness in all instances in the validation and test sets. Therefore, each video clip has two referred expression annotations.

After the step (ii), we quickly reviewed overall annotations and manually edited them if necessary. We confirmed that we gathered two different referring expressions for each tracked object. We present a sample annotation website on MTurk to adjust and select the bounding boxes in Fig. 11.

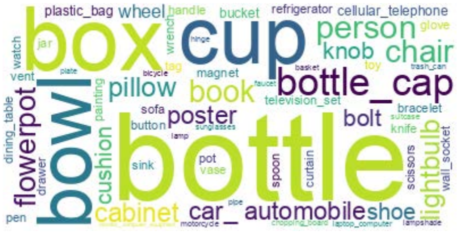


Figure 9. Frequently seen object categories in RefEgo by Detic (LVIS).

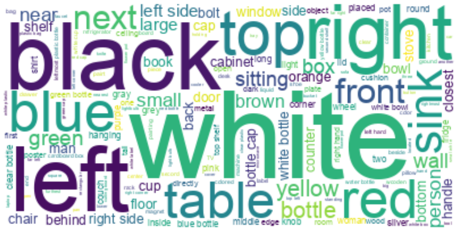


Figure 10. Frequently seen words in referential expression of RefEgo.

Model	Val. mIoU	Val. mAP@50	Test mIoU	Test mAP@50
CG-SL-Att [4]	33.2	**38.0**	32.9	**38.0**
DCNet [1]	**34.2**	37.0	**33.2**	36.5

Table 6. RefEgo val. and test sets (Images w/ targets).

B. Video-based REC baselines

We also derive the scores with existing video-based REC models of Co-grounding network [4] and DCNet [1]. It is considerable that these models do not concentrate on the discrimination of images without target objects and therefore we train these models with images w/ targets. Results are evaluated in image-based metrics of mIoU and mAP@50 in Table 6. These models use alignments between frames that are not suitable when we apply these models for extracted frames that include the target objects. It is notable that OFA and MDETR have a strong object detection ability from pretraining and hence achieve a good performance even in the detection setting.

C. Object tracking implementation details

For ByteTrack [5], we used GIoU [3] for this similarity criteria to enable robust matching in motion videos. We set thresholds of the high and low scores for the detection boxes to 0.1 and -0.5, respectively. For object matching between adjacent image frames, the matching candidate object bounding boxes are allowed only when the REC confidence score (or objectness scores of MDETR) are greater than 0.9 and GIoU is greater than 0.9. In addition, we apply NMS

Track high threshold	0.1
Track low threshold	-0.5
Confidence Score threshold	0.9
GIoU Match threshold	0.9

Table 7. ByteTrack parameters.

to predictions in each frame to reduce overlapped bounding boxes. We summarize the hyper-parameters of ByteTrack in Table 7.

We notice that the confidence scores from REC models are sufficiently reliable even when objects that are not well-tracked. Therefore, we use a simple heuristic to merge the object tracking results with the original REC results. We first assign the REC confidence score for the new confidence score of the candidate object bounding boxes. Then, following the object tracking results, we obtained the averaged REC confidence score for the time-sequence of the bounding boxes that are sequenced by ByteTrack. If an averaged confidence score is higher than the original score for some bounding boxes, we updated the confidence score with the averaged one. By doing so, we can maintain the REC confidence score when it is sufficiently high while we can update the low REC confidence score for the bounding boxes when the bounding boxes for the same object in adjacent frames are sufficiently high.

D. Human performance on RefEgo

To compare the current model performance with those by the human experts, we provide the human performance for the test set of the RefEgo. For this human expert test, we first sampled video clips in the test set and presented them in the same website including the object detection results used in the annotation process (ii) to two expert workers. Similar to the annotation process, we asked the expert workers to select bounding boxes from the auto-detected ones and modify them if they are not fitted to the tracked objects following one of the annotated referring expressions. In Table 8, we present the human performance compared with the best model prediction, confirming the great performance gap between them. The all image frames metrics of mSTIoU, mIoU+n and mAP@50+n have larger margins to the human performance than the metrics for REC of mIoU and mAP@50. This suggests that the video clip-wise object localization from the referred expression is a much more difficult task than the simple REC task in 2D images.

E. Additional qualitative analyses

We provide the further qualitative for MDETR models here. Figure 12 is the comparison of the bounding boxes by the annotated referred object, MDETR†, MDETR‡ and MDETR‡ (all).

References

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

[2] 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.

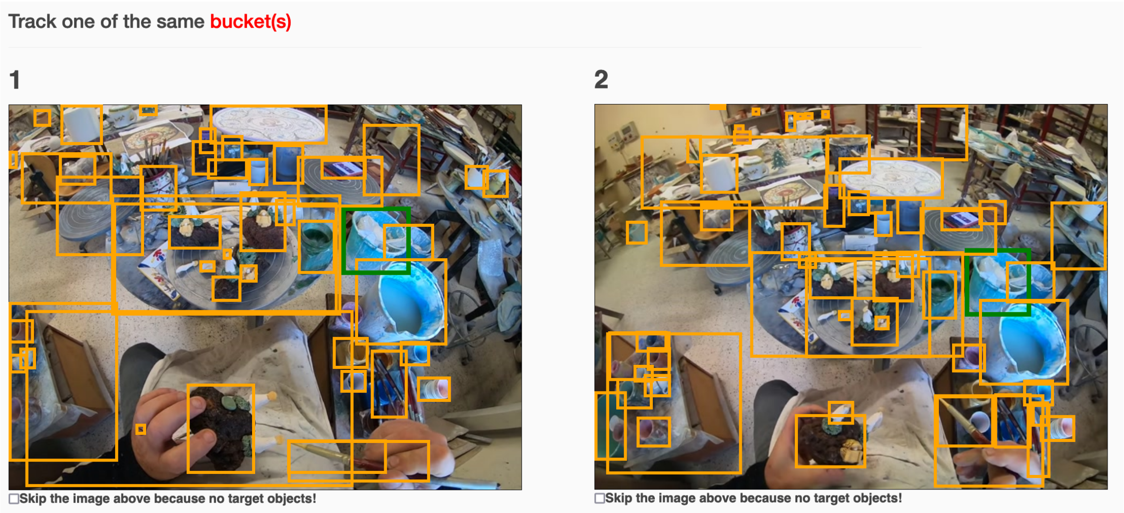
[3] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union. June 2019.

[4] 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.

[5] 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.

	RefEgo Test				
	All images			Images w/ targets	
Model	mSTIoU	mIoU+n	mAP@50+n	mIoU	mAP@50
OFA†	15.4	28.9	29.3	27.8	27.1
OFA‡	31.7	44.4	47.5	**51.0**	**56.2**
MDETR+BH‡ (all)	36.9	**45.7**	**51.1**	45.7	53.0
+Object tracking	**37.6**	45.4	51.0	46.0	53.4
Human	77.1	82.9	85.4	78.8	82.1
Δ	+39.5	+37.2	+34.3	+27.8	+25.9

Table 8. Comparisons with human performance. Δ represents the difference between the human performance and the best model prediction (**bold**).



Track one of the same bucket(s)

1

□Skip the image above because no target objects!

2

□Skip the image above because no target objects!

Figure 11. MTurk annotation website.

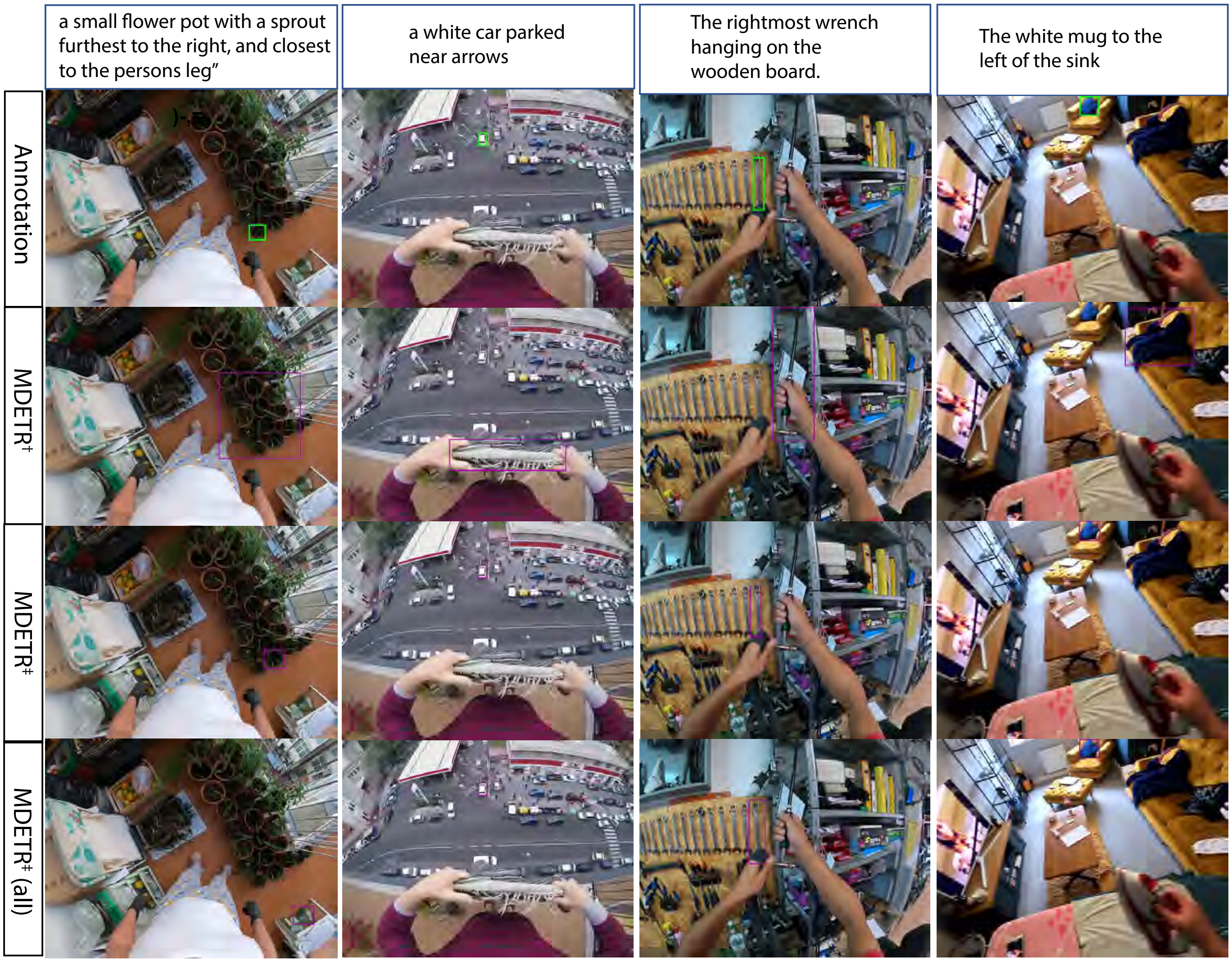

a small flower pot with a sprout furthest to the right, and closest to the persons leg”
a white car parked near arrows
The rightmost wrench hanging on the wooden board.
The white mug to the left of the sink
Annotation
MDETR†
MDETR‡
MDETR‡ (all)

Figure 12. Comparison of annotated referred object in green, MDETR†, MDETR‡ and MDETR‡ (all) bounding box predictions in purple.