Helping Hands: An Object-Aware Ego-Centric Video Recognition Model Supplementary Materials

1. Discussion: Text-Region Alignment

One of the most related work to our paper is [6], Yao et al. train dual encoders to align image patches and textual words. Fine-grained pre-training helps the model to achieve better results on image classification and image-text retrieval. GLoRIA [3] is another similar work on medical image recognition, where they show region-word matching is a more label-efficient pre-training method compared to image-sentence matching on retrieval, classification and segmentation. While our work focuses on egocentric videos and utilizes detections from hand-object detector to supervise the training for alignment. This is because scenes in egocentric videos are often crowded and objects are prone to be heavily occluded. Boxes from off-the-shelf detectors are easy to obtain and can largely ease the training process. Furthermore, as an important factor in first-person videos, hands are not often mentioned in the narrations; explicit supervision helps the model to focus on the motion during training. Similarily, another line of work [1, 7] pre-trains a vision-language model to predict object boxes, but relies on manually labeled ground-truth. While our model can be trained with imperfect supervision. Other works [5, 11] train models to do pixel-text or region-text alignment for open-vocabulary detection or segmentation.

2. Implementations

2.1. Training

Given a video clip, we uniformly sample 4 frames from the clip, and resize the image to 224×224 without cropping, color jittering is applied as data augmentation. We use the 3.8M video clips from EgoClip for training. Each clip is paired with its original narration from EgoClip and the rephrased ones from LaViLA [10]. The additional pseudo-labelled video clips from LaViLA are not used.

2.2. Evaluation

EgoMCQ. EgoMCQ dataset is a multiple choice question dataset built on Ego4D. Given one narration as question, the model is tasked to find the paired video clip from 5 candidates. It has 39k questions in total, which are categorized into 'inter-video' and 'intra-video' multiple-choice questions. There are 24k questions in the "inter-video" split, where the candidates are from different videos. The "intra-video" split has 15K questions, where the candidates are from the same video. The average temporal gap between the intra-video candidates is 34.2 seconds. We sample 4 frames uniformly from each clip and resize them to 224×224 as input in evaluation.

EpicKitchens-MIR. Epic-Kitchens Multiple Instance Retrieval is a dataset from Epic-Kitchens 100 for video-text and text-video retrieval. Given a query video/caption, the task is to rank the instances from the other modality such that higher-ranked instances are more semantically relevant to the query. We use the val split for zero-shot transfer evaluation, which contains 9668 video-caption pairs. The captions are in the format of 'verb + noun', with totally 78 verb classes and 211 noun classes. In evaluation, We sample 16 frames uniformly across the clip, and resize frames to 224×224 as input. Mean Average Precision (mAP) and normalized Discounted Cumulative Gain (nDCG) are used as evaluation metrics.

EGTEA. EGTEA contains 28 hours of cooking activities from 86 unique sessions of 32 subjects. We evaluate the model on action classification and use top-1 accuracy and mean-class accuracy as metrics. The descriptions of 106 action classes are encoded into text embeddings using the text encoder. We compute the similarity score between every video embedding and the 106 text embeddings, and take the text embedding with the highest similarity score as the predicted class. Evaluation is done on its first test split with 2022 instances. We uniformly sample 10 clips from the full span of one video instance, each has 16 frames with a temporal stride of 2. We resize the frames to 224×224 as input to the model. For each video instance, we predict logits for 10 clips and then max-pool the logit as the final prediction.

EgoNLQ. Given a video clip and a query expressed in natural language, the task is to localize the temporal window within all the video history where the answer to the question is evident. We evaluate the model on the val split covering 45-hour videos, with 0.3k clips and 3.9k queries.

We follow [4] and extract all the video and text embeddings using our model, and input them to VSLNet [8] for fine-tuning on EgoNLQ. The evaluation metrics are based on the overlap of top-1 or top-5 predicted temporal windows with the ground-truth at IoU thresholds of 0.3 and 0.5.

EgoMQ. In this task is a natural language grounding task, where activities are used as queries to find responses consisting of all temporal windows where the activity occurs in a video. There are 13.6 training instances from 1.5k clips and 4.3k validation instance from 0.5k clips. We extract all the video features using our model as input, and train a VSGN [9] to perform the task. We report mAP and recall as evaluation metrics.

VISOR. VISOR is a dataset built on videos from Epic-Kitchens 100 for segmenting hands and active objects in egocentric videos. We re-propose VISOR for a in-contact hand-object grounding by 1) converting the segmentation masks to bounding boxes 2) filtering out not-in-contact objects in the frames. Given a list of names (hand + in-contact objects), the model is tasked to predict a bounding box for each instance. We do evaluation on the val split with 7,747 images, 182 entity classes from 4 videos. After filtering, the model has 1.4 hands and 0.9 objects per frame on average. We resize each image to 224×224 and repeat it for 4 times along the temporal dimension to make it a 4-frame clip as input to the model. For hands, we always use the first hand query for left hand box prediction, and the second hand query for right hand box predication, as we find hand queries have learned to specify without explicit supervision. For objects, we match the text embedding of object names with the predicted object embeddings for grounding as in training.

For the baseline image detector, we also resize the shorter side of the image to 224p as input for fair comparison. The detector produces two types of output: hand boxes with 'left' and 'right' labels, and object boxes without object class. Since there is no specific grounding predicted by the detector, we conduct two types of matching in our evaluation:

- **Random matching:** The predicted object boxes are randomly assigned to ground-truth objects
- **Hungarian matching:** We compute the IoU between predicted boxes and ground-truth boxes, and apply Hungarian matching for grounding.

3. Statistics

3.1. Grounded Nouns in EgoClip

The Ego4D taxonomy dictionary [2] is a thesaurus that records meaningful nouns/verbs in Ego4D narrations, it has

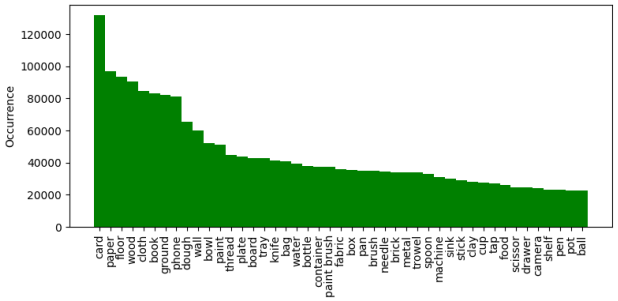

Figure 1. **The distribution of op 45 nouns in EgoClip.**

	Unseen	Seen	Overall
Occurrence	2,041	15,800	17,841
Localization Acc	52.8%	82.0%	78.7%

Table 1. **Localization accuracy on seen and unseen nouns/phrases on VISOR.**

581 noun groups with 1610 nouns. We match all the single words and two-word phrases in the narrations with nouns in the dictionary to extract the nouns from the narrations. We remove nouns that refer to the background or someone who is holding the camera, including: 'man', 'woman', 'person', 'lady', 'they', 'ground', 'camera', 'table', and 'leg'. We also remove nouns related to 'hand' because we use hand supervision from the object-hand detector instead of the narrations. As a result, we find 5,020,303 nouns from 3,847,723 narrations in training. Below, we plot a histogram of the top 45 nouns in EgoClip.

3.2. Out-of-Distribution Nouns in VISOR

We compare the 1610 nouns in the pre-training dataset EgoClip, and 411 nouns in the downstream grounding dataset VISOR. There are 250 noun words/noun phrases in VISOR that have not appeared in the pre-training. Some are new combinations with an additional adjective, e.g., small bread, hot water, aluminium foil. Some are objects that have not appeared in the pre-training, e.g., basil, scale, drainer. As results shown in table 1, the localization accuracy is 48.4% on unseen concepts and 70.9% on seen concepts. The reason that our model is able to ground some of the unseen concepts is probably: 1) Some unseen nouns/phrases have similar semantic meaning with the seen ones, hence the word embeddings can be similar. e.g., hot water and water. 2) When there is no other distractive object in the scene, all the object queries localize the same object that is in contact with the hand. In this case, the proposed box can always be matched to the object of interest no matter whether it is seen or not.

References

[1] Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*, 2022. 1

[2] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proc. CVPR*, 2022. 2

[3] Shih-Cheng Huang, Liyue Shen, Matthew P Lungren, and Serena Yeung. Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3942–3951, 2021. 1

[4] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *arXiv preprint arXiv:2206.01670*, 2022. 2

[5] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 1

[6] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1

[7] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. *arXiv preprint arXiv:2111.08276*, 2021. 1

[8] Hao Zhang, Aixin Sun, Wei Jing, Liangli Zhen, Joey Tianyi Zhou, and Rick Siow Mong Goh. Natural language video localization: A revisit in span-based question answering framework. *IEEE PAMI*, 2021. 2

[9] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667, 2021. 2

[10] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. *arXiv preprint arXiv:2212.04501*, 2022. 1

[11] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proc. CVPR*, 2022. 1