# Action Scene Graphs for Long-Form Understanding of Egocentric Videos

Ivan Rodin*[1] Antonino Furnari*[1] Kyle Min*[2] Subarna Tripathi[2] Giovanni Maria Farinella[1]

[1]University of Catania [2]Intel Labs

{ivan.rodin,antonino.furnari,giovanni.farinella}@unict.it {kyle.min,subarna.tripathi}@intel.com

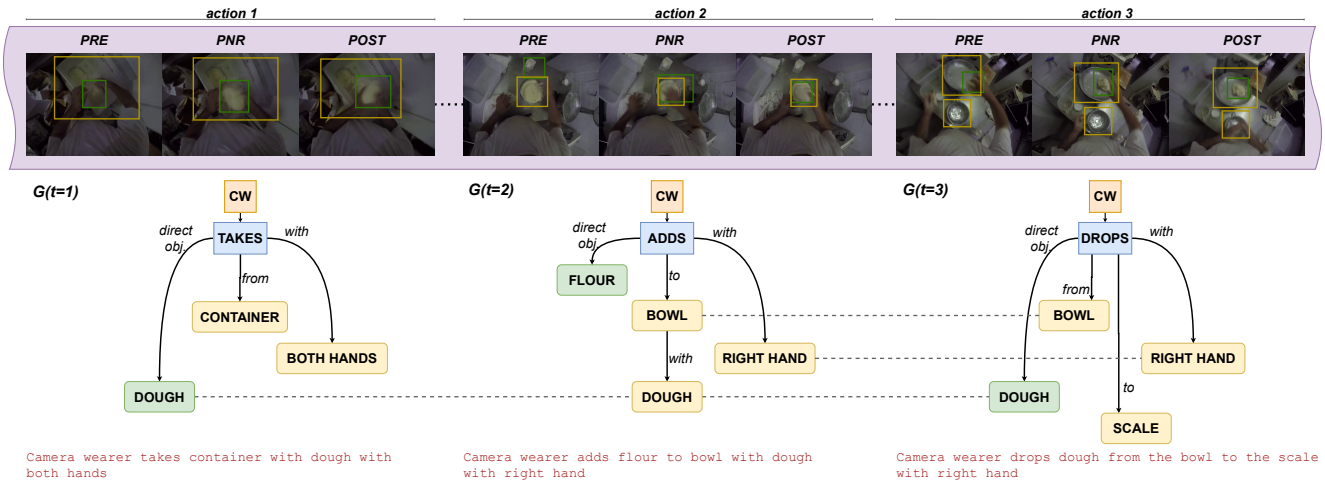


Figure 1. Egocentric Action Scene Graphs are temporal dynamic graphs ($G(t)$) capturing the action verbs (nodes in blue), direct or active objects (nodes in green), and other objects (nodes in yellow) involved in the activity performed by a camera wearer (the orange CW node). Edges between nodes represent relationship between the verb and the objects or between object pairs. The graph evolves through time providing a long-from representation of the egocentric video (dashed lines). Objects of interaction are grounded with bounding boxes.

## Abstract

*We present Egocentric Action Scene Graphs (EASGs), a new representation for long-form understanding of egocentric videos. EASGs extend standard manually-annotated representations of egocentric videos, such as verb-noun action labels, by providing a temporally evolving graph-based description of the actions performed by the camera wearer, including hands and objects involved in actions, their relationships, and how actions unfold in time. Through a novel annotation procedure, we extend the Ego4D dataset adding manually labeled Egocentric Action Scene Graphs which offer a rich set of annotations for long-from egocentric video understanding. We hence define the EASG generation task and provide a baseline approach, establishing preliminary benchmarks. Experiments on two downstream tasks, action anticipation and activity summarization, highlight the effectiveness of EASGs for long-form egocentric video understanding. We will release the dataset and code to replicate experiments and annotations*[1].

*These authors contributed equally to this work.

[1]The code is available at https://github.com/fpv-iplab/EASG

## 1. Introduction

Wearable devices allow to capture video of human activities from an egocentric perspective. A proper analysis of such video can enable a detailed understanding of how humans interact with the environment, how they manipulate objects with their hands and tools, and, ultimately, what are their goals and intentions. Easily covering sequences of activities performed by the camera wearer in different physical locations, egocentric video is by its own nature *long-form* [27]. Hence, typical applications of egocentric vision systems require algorithms able to represent and process video over temporal spans that last in the order of minutes or hours. Examples of such applications are action anticipation [2, 8, 20], video summarization [5], and episodic memory retrieval [8]. Despite the relevance of such applications in the panorama of egocentric vision [17], progress in this area has been hindered by the lack of a comprehensive and long-form representation of videos that algorithms can rely on, with popular high-level human-gathered representations being in the form of textual narrations [2], verb-noun action labels [6], temporal bounds for action segments [2, 6, 12], object bounding boxes [16], object state changes [8], and

hand-object interaction states [4, 21], all short-range representations describing temporal spans lasting few seconds.

We present Egocentric Action Scene Graphs (EASGs), a novel graph-based representation for capturing actions performed by a camera wearer in egocentric videos. The proposed representation builds on the literature of scene graphs [9, 10, 18] and extend traditional *verb-noun* action labels to a structured temporal dynamic graph format that encodes the objects involved, hands of the camera wearer, action verbs, and their relationships throughout the video.

We augment the Ego4D dataset [8] with manually annotated EASG labels, gathered through a novel multi-stage annotation and validation procedure. Following the scene graph literature [9, 29], we benchmark the EASG generation task, providing baseline results and demonstrating the feasibility of automatically generating these representations. Initial experiments show the effectiveness of EASGs in tasks like action anticipation and activity summarization.

## 2. Egocentric Action Scene Graphs

Egocentric Action Scene Graphs (EASGs) provide annotations for a video clip in the form of a dynamic graph. We formalize an EASG as a time-varying directed graph $G(t) = (V(t), E(t))$, where $V(t)$ is the set of nodes at time $t$ and $E(t)$ is the set of edges between such nodes. Each temporal realization of the graph $G(t)$ corresponds to an egocentric action spanning over a set of three frames defined as in [8]: the *precondition* (PRE), the *point of no return* (PNR) and the *postcondition* (POST) frames. The graph $G(t)$ is hence effectively associated to three frames: $\mathcal{F}(t) = \{PRE_t, PNR_t, POST_t\}$. $G(t)$ has two fixed nodes: the camera wearer node $v_{cw}(t)$ representing the camera wearer, and the verb node $v_{verb}(t)$, describing the action performed by the camera wearer at time $t$. Each graph $G(t)$ also contains a set of object nodes $V_{obj}(t)$ encoding the objects involved in the actions. In this formulation, the camera wearer's hands will appear as object nodes.

Apart for the camera wear node, each other node is associated to one or more attributes through a function $att$. The verb node is associated to a *verb class attribute*: $att(v_{verb}(t)) = verb$. Noun nodes $v_i(t)$ are associated to a *noun class attribute* $noun$ and to three bounding box attributes grounding the noun to the $PRE(t)$, $PNR(t)$ and $POST(t)$ frames associated to the action taking place at time $t$: $att(v_i(t)) = (noun, box_{PRE}, box_{PNR}, box_{POST})$. Additionally, we provide the object segmentation masks from the initial bounding box groundings (Fig. 2).

The edges in the graph describe the relationships between nodes. Relations between verb and object nodes can be of a *direct object* kind (e.g., puts – *dobj* –package), or a preposition (i.e., puts – *in* – fridge), while relationships between object nodes are characterized by the prepositions only (i.e., package –*with* – carrot). Objects $v_i(t)$ which are in a *direct object* relation with the verb node $v_{verb}(t)$ are also referred to as "direct objects", while all other objects are referred to as "indirect objects".

Figure 2. An example of segmentation masks for the PNR frame of the graph *CW takes dough from the container with both hands, from Fig.1*

## 3. Ego4D-EASG Dataset

We build our EASG dataset, *Ego4D-EASG*, by annotating a subset of 552 Ego4D [8] clips containing labels for the State Change Object Detection benchmark (SCOD).We labeled an independent EASG $G_i(t)$ for each clip $C_i$. Each temporal realization of the graph, $G_i(t)$ is seeded from the annotation tuple $a_t^i = (a_t^{i,PRE}, a_t^{i,PNR}, a_t^{i,POST})$.

### 3.1. Egocentric Action Scene Graph Annotation

The data annotation is performed in two stages: 1) the graph annotation stage, and 2) the graph validation stage. In the first stage we initialize the graph with the verb-noun annotations from Ego4D dataset, and ask Amazon Mechanical Turk workers to add relevant information about the actions (new nodes, edges and groundings). In the second stage, we aggregate the annotations from multiple workers and ask the annotators to remove contradictions from data.

After the graphs are annotated, we perform temporal recollection: in this stage we reason globally on the dynamic graph $G_i(t), t = 1 \ldots, T$ and re-assign node indices to ensure that object nodes representing the same object instance are assigned the same index across time. To achieve this, we leverage the EgoSTARK model introduced in the [24].

Finally, we perform the object segmentation using Segment Anything Model (SAM) [11].

### 3.2. Dataset Statistics and Comparison with Other Scene Graph Datasets

Table 1 reports statistics on the proposed Ego4D-EASG dataset and compares it with existing video scene graph datasets. The proposed dataset is the only one designed for long-form egocentric video understanding and it fea-

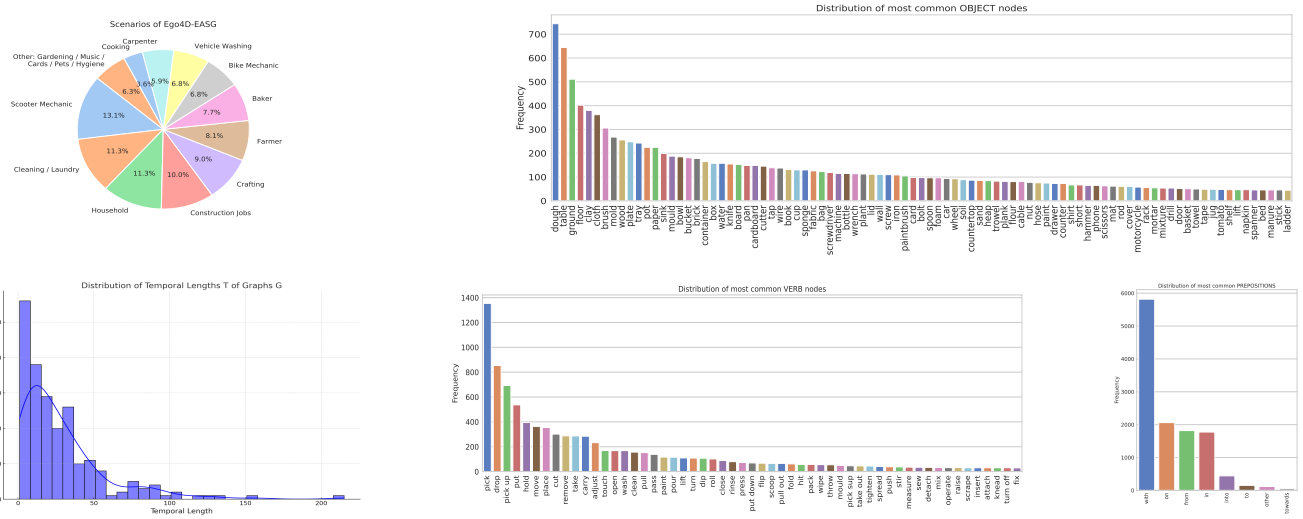


Figure 3. Left-to-right, top-to-bottom: Distributions of clips across scenarios, object nodes, temporal lengths $T$ of graphs $G$, verb nodes, and relation categories (excluding *action* and *direct object* relations). Data is distributed across different scenarios related to egocentric perception, long-tailed objects, verb distributions, and prepositions. The distribution of temporal length of graphs shows the long-form nature of our annotations, with most graphs having a length of up to 50 timesteps.

| Dataset | Dynamic | Egocentric | Sequences | Hours | Avg. Len. (seconds) | Avg. Graphs per Vid. | Obj Cls | Verb Cls | Rel Cls |
|---|---|---|---|---|---|---|---|---|---|
| VidVRD [22] | ✗ | ✗ | 1,000 | 3 | 11 | 3.9* | 35 | 25** | 132 |
| VidOR [23] | ✗ | ✗ | 10,000 | 99 | 35 | 8.8* action + 29.2* spatial | 80 | 42 | 50 |
| Action Genome [28] | ✓ | ✗ | 10,000 | 82 | 30 | 5 | 35 | - | 25 |
| PVSG [30] | ✗ | Partly (28%) | 400 | 9 | 77 | 382 | 126 | 44 | 57 |
| HOMAGE [18] | ✗ | paired ego-exo | 1,752 | 25 | 3 | 3.8 | 86 | 453 | 29 |
| Ego4D-EASG (Ours) | ✓ | ✓ | 552 | 22.3 | 197 | 27.1 | 789 | 329 | 21 |

Table 1. Comparison with existing video scene graph datasets. Our Ego4D-EASG dataset is the only one explicitly designed for long-form egocentric video understanding, featuring egocentric videos, dynamic graphs, an average sequence length of 3.1 minutes and an average number of 27.1 graphs per sequence. *measured in object-relation-object triplets. **intransitive + transitive verb predicates.

| Method | With Constraint | | | | | | | | | No Constraint | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Edge Cls* | | | *SG Cls* | | | *EASG Cls* | | | *Edge Cls* | | | *SG Cls* | | | *EASG Cls* | | |
| | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 | R@10 | R@20 | R@50 |
| Random Guess | 8.0 | 8.0 | 8.0 | 0.2 | 0.4 | 1.0 | 0.0 | 0.0 | 0.0 | 36.5 | 72.6 | 99.9 | 0.3 | 0.5 | 1.0 | 0.0 | 0.0 | 0.0 |
| Baseline (Ours) | 60.4 | 60.4 | 60.4 | 41.4 | 44.3 | 50.6 | 14.3 | 16.4 | 17.9 | 94.4 | 99.8 | 100 | 51.6 | 58.2 | 62.4 | 14.7 | 18.3 | 20.9 |

Table 2. Baseline results for three EASG generation tasks (i.e. *Edge Cls*, *SG Cls*, and *EASG Cls*) in terms of Recall@K.

tures 552 egocentric video sequences, 11.4 hours of video, comprising an average labeled sequence length of 3.1 minutes, $T = 27.1$ graphs per video in average, 789 object classes, 329 verb classes, and 21 relation classes. Compared to previous datasets, ours is the only one that includes verb nodes explicitly encoding actions. All the object and hand nodes are manually annotated with 103,027 bounding boxes. Figure 3 reports statistics on the distribution of scenarios, nouns, verbs, relations, and temporal graph lengths.

## 4. Egocentric Action Scene Graphs Generation

**Task Definition** Unlike standard scene graph generation, EASG generation aims to predict the action verbs as well as objects and their relationships. We define three EASG generation tasks as follows: (1) Edge classification (*Edge Cls*) is to predict verb-object and object-object relationships given visual features, the ground-truth action verb and object classes, (2) Scene Graph Classification (*SG Cls*) is to predict both the object classes and the edge relationships given visual features and the ground-truth action verb, and (3) Egocentric Action Scene Graph Classification

We measure the length of each sequence from the timestamp of the $G(1) : PRE$ frame to the timestamp of the $G(T) : POST$ frame.

(*EASG Cls*) is to predict all these three components, which encompass action verbs, objects, and edge relationships. We follow [9] and report results for predicate (*Edge Cls*) and scene graph (*SG Cls*) classification, and extend it with *EASG Cls* to evaluate time-evolving graphs.

**Experimental Setting** We design a baseline for the novel EASG generation task consisting of task-specific fully-connected layers working on top of pre-extracted visual features. For *Edge Cls*, we use a single-layer model to predict the edge relation from the clip-level features and ROIAlign features of each object bounding box. For the clip-level features, we take the average of SlowFast [7] features (pre-extracted and provided within the Ego4D dataset [8]) for the whole clip spanning from *PRE* to *POST* frames. We extract the ROIAlign features using the Faster-RCNN [19] pre-trained for the short-term action anticipation benchmark [8]. For *SG Cls*, we add an additional fully-connected layer to predict the object classes from the ROIAlign features. For *EASG Cls*, we add another additional layer to predict the action verb from the clip-level features.

**Results** We report the results for all tasks and setups in Table 2, similar to scene graph generation datasets like Action Genome, which evaluate using Recall@K, with K=10, 20, 50. Baseline results are compared with random guess. We can observe that the scores of *EASG Cls* are significantly lower than other results, indicating that action verbs introduce another layer of difficulty to EASG understanding.

## 5. Downstream long-from video understanding tasks with Egocentric Action Scene Graphs

In this section, we report experiments showing the potential of the EASG representation in the downstream tasks of action anticipation and activity summarization. Following recent results showing the flexibility of Large Language Models (LLMs) as symbolic reasoning machines [14], we perform these experiments with open-source LLMs from the LLaMa-2 series [25]. We show that EASG offers an expressive way of modeling long-form activities, in comparison with the gold-standard verb-noun action encoding, extensively adopted in previous work [3, 8].

**Experimental Setting** We prompt the model to predict the future action from a sequence of length $T \in \{5, 20\}$. We compare two types of representations - EASG and sequences of verb-noun pairs. The input sequence of graphs can be represented as $s_{EASG} = [G(t_0), G(t_0 + 1), ..., G(t_0+T-1)]$, with $t_0+T-1 \geq 20$. Each graph $G(t)$ is represented as a string of triplets, where each triplet encapsulates the relationship between nodes (e.g., *CW - verb - wash; wash - direct object - car; wash - with - sponge*). As an output, we request to provide the future unobserved scene graph $G(t + T)$ in the same format. From the predicted graph, we extract the pair of verb and direct object node classes for evaluation. Given the uncertainty in forecasting future events, we prompt the LLM to output up to $N = 5$ predictions, a standard practice in anticipation [2, 8]. We evaluate results using top-k accuracy, with $k \in \{1, 5\}$, reported for verb, noun, and actions.

| | Seq. length $T$ | Avg. duration | Verb | | Noun | | Action | |
|---|---|---|---|---|---|---|---|---|
| | | | Top-1 | Top-5 | Top-1 | Top-5 | Top-1 | Top-5 |
| V-N | 5 | 21s | 1.52 | 3.14 | 41.55 | 51.20 | 1.02 | 2.04 |
| EASG | 5 | 21s | 2.11 | 7.23 | **42.18** | 53.08 | 2.11 | 4.86 |
| V-N | 20 | 82s | 3.32 | 7.07 | 40.12 | 52.94 | 2.23 | 4.47 |
| EASG | 20 | 82s | **5.32** | **14.81** | 41.27 | **54.33** | **3.45** | **8.20** |
| Improvement | | | +2.0 | +7.58 | +0.63 | +1.25 | +1.22 | +3.34 |

Table 3. Performance Comparison for the Action anticipation task.

| | CIDEr | ROUGE-1 | ROUGE-2 | ROUGE-L | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR |
|---|---|---|---|---|---|---|---|---|---|
| V-N | 9.42 | 31.5 | 10.3 | 29.7 | 35.7 | 18.6 | 7.6 | 3.9 | 26.09 |
| EASG | 13.79 | 33.3 | 10.7 | 31.4 | 37.3 | 19.0 | 7.8 | 4.2 | 26.30 |
| Narrations | 19.99 | 37.7 | 14.0 | 34.4 | 42.0 | 24.0 | 11.7 | 6.7 | 29.43 |

Table 4. Results of activity summarization with EASGs and verb-noun representations.

In the similar fashion we tackle long-form activity summarization task. In these experiments we take the Ego4D clip summaries as ground-truth, and evaluate the produced summaries using the CIDEr [26] metric, adopted in the image captioning literature, and standard metrics for NLG (ROUGE [13], BLEU [15], METEOR [1]).

**Results** Table 3 reports the results of these experiments. Best results are always achieved by EASG-based representations. As can be noted, even short EASG sequences ($T = 5$) tend to outperform long V-N sequences ($T = 20$), highlighting the higher representation power of EASG, when compared to standard verb-noun representations. EASG representations achieve the best results for long sequences ($T = 20$). EASGs bring overall significant improvements of up to $+7.58$ with respect to the best verb-noun based prediction across the different metrics.

Results for summarization, reported in Table 4 indicate strong improvement in CIDEr score over $s_{vn}$ inputs, showing that models which process EASG inputs capturing detailed object-action relationships, will generate more specific, informative sentences that align well with reference descriptions.

## 6. Conclusion

Our paper reports four key contributions: Egocentric Action Scene Graphs (EASG) as a novel representation for understanding long-form egocentric videos; A procedure for the collection of such graphs and extended the Ego4D dataset with manually annotated EASG labels: Initial baseline results for EASG generation; The validation of the effectiveness of the EASG representation in two downstream tasks, aimed at long-form egocentric video understanding. We believe that these contributions mark a step forward in long-form egocentric video understanding.

## References


[1] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. 4

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 1, 4

[3] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020. 4

[4] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022. 2

[5] Ana Garcia Del Molino, Cheston Tan, Joo-Hwee Lim, and Ah-Hwee Tan. Summarization of egocentric videos: A comprehensive survey. *IEEE Transactions on Human-Machine Systems*, 47(1):65–76, 2016. 1

[6] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR 2011*, pages 3281–3288. IEEE, 2011. 1

[7] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 4

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2, 4

[9] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10236–10247, 2020. 2, 4

[10] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 2

[11] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[12] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 619–635, 2018. 1

[13] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 4

[14] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. *arXiv preprint arXiv:2307.04721*, 2023. 4

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 4

[16] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2847–2854. IEEE, 2012. 1

[17] Chiara Plizzari, Gabriele Goletto, Antonino Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Dima Damen, and Tatiana Tommasi. An outlook into the future of egocentric vision. *arXiv preprint arXiv:2308.07123*, 2023. 1

[18] Nishant Rai, Haofeng Chen, Jingwei Ji, Rishi Desai, Kazuki Kozuka, Shun Ishizaka, Ehsan Adeli, and Juan Carlos Niebles. Home action genome: Cooperative compositional action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11184–11193, 2021. 2, 3

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 4

[20] Ivan Rodin, Antonino Furnari, Dimitrios Mavroeidis, and Giovanni Maria Farinella. Predicting the future from first person (egocentric) vision: A survey. *Computer Vision and Image Understanding*, 2021. 1

[21] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9869–9878, 2020. 2

[22] Xindi Shang, Tongwei Ren, Jingfan Guo, Hanwang Zhang, and Tat-Seng Chua. Video visual relation detection. In *Proceedings of the 25th ACM International Conference on Multimedia*, page 1300–1308, New York, NY, USA, 2017. Association for Computing Machinery. 3

[23] Xindi Shang, Donglin Di, Junbin Xiao, Yu Cao, Xun Yang, and Tat-Seng Chua. Annotating objects and relations in user-generated videos. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 279–287. ACM, 2019. 3

[24] Hao Tang, Kevin J Liang, Kristen Grauman, Matt Feiszli, and Weiyao Wang. Egotracks: A long-term egocentric visual object tracking dataset. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[25] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 4

[26] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 4

[27] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. 1

[28] Yiming Wu, Omar El Farouk Bourahla, Xi* Li, Fei Wu, Qi Tian, and Xue Zhou. Adaptive graph representation learning for video person re-identification. *IEEE Transactions on Image Processing*, 2020. 3

[29] Jingkang Yang, Yi Zhe Ang, Zujin Guo, Kaiyang Zhou, Wayne Zhang, and Ziwei Liu. Panoptic scene graph generation. In *European Conference on Computer Vision*, pages 178–196. Springer, 2022. 2

[30] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, and Ziwei Liu. Panoptic video scene graph generation. In *CVPR*, pages 18675–18685, 2023. 3