MOMA-LRG: Language-Refined Graphs for Multi-Object Multi-Actor Activity Parsing
Supplementary Material

Zelun Luo, Zane Durante, Linden Li, Wanze Xie, Ruochen Liu, Emily Jin, Zhuoyi Huang, Lun Yu Li, Jiajun Wu, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei
Stanford University
{alanzluo, durante, lindenli, wanzexie, ruochenl, emilyjin, zhuoyih, tinally}@stanford.edu
{jiajunwu, jniebles, eadeli, feifeili}@cs.stanford.edu

A Implementation Details

VLM Evaluation To evaluate two VLMs (Frozen in Time [1] and VideoCLIP [13]), we use a hybrid approach that leverages both prototypical networks [11] and the video-language similarity metrics learned by both models. Below, we show an ablation study where we use only the video prototype networks. We show the performance of using only language similarity in the few-shot case to demonstrate the effects of sample removal, and we also show the effects of our hybrid weighting scheme, where we weight the language embeddings five times more than the video embeddings when constructing the hybrid prototype (as opposed to equal weighting during the regular hybrid approach). Although we perform our ablation study with Frozen-in-Time, and use the same weighting scheme and prototype strategy for VideoCLIP as well.

Table 1: Frozen-in-Time Evaluation ablation study. For this study, we show activity and sub-activity classification accuracy in the 5-shot case. We visualize whether a given method uses language, video, or both to create its prototype embeddings.

	5-shot Video Classification			
	Video	Language	Activity	Sub-activity
Video prototypes	✓	–	87.9	25.2
Language prototypes	–	✓	91.1	18.7
Hybrid prototypes	✓	✓	89.9	25.7
Weighted hybrid prototypes	✓	✓	**92.5**	**26.2**

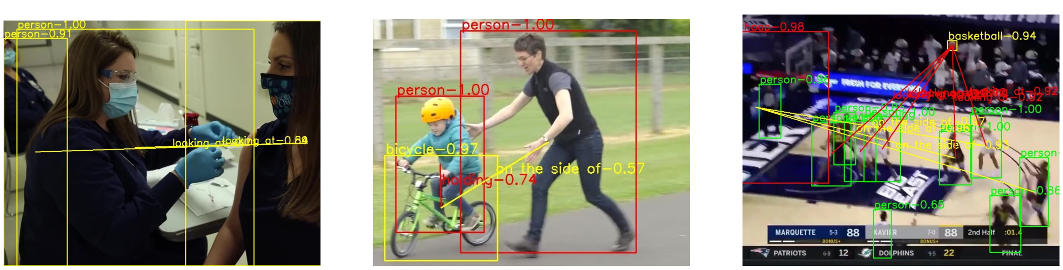

Figure 1: Example outputs of scene graph detection on the MOMA-LRG test set. As input, our model is given a static frame and outputs the objects, bounding boxes, and relationships occurring during the activity.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

Table 2: Entity detection and tracking results.

	Entity Detection						Entity Tracking			
	AP	AP50	AP75	APs	APm	APl	HOTA	DetA	AssA	LocA
Actor	38.3567	58.1256	41.2369	7.8053	19.4897	40.0392	38.859	35.669	45.191	73.963
Object	14.1730	24.5994	13.7196	5.4347	10.7436	15.8665	35.686	23.734	57.127	74.924

Table 3: A comparison of MOMA-LRG's vocabulary with related video datasets. MOMA-LRG's hierarchy unifies several definitions together (src: source, trg: target, atr: actor, obj: object, c: classified, g: grounded, t: tracked).

	Unary predicate			Binary predicate				
Dataset	Name	src_atr	src_obj	Name	src_atr	src_obj	trg_atr	trg_obj
AVA [4]/AVA-Kinetics [6]	Pose	g,t	-	Person-person/object interaction	g,t	-	-	-
Action Genome [5]	-	-	-	Relationship	g	-	-	c,g
FineGym [10]	Sub-action	-	-	-	-	-	-	-
Home Action Genome [9]	-	-	-	Relationship	g	-	-	c,g
MultiSports [7]	Action	g,t	-	Action	-	-	-	-
Something V2 [3]	-	-	-	Human-object interaction	-	-	-	c
DALY [12]	Action	g,t	c,g	-	-	-	-	-
MEVA [2]	Activity	g,t	g,t	Activity	g,t	-	g,t	g,t
TITAN [8]	Individual Atomic Actions Vehicle State/Action	c,g,t	c,g,t	Communicative Contextual/Transportive	c,g,t	-	c,g,t	c,g,t
MOMA-LRG	Attribute	c,g,t	c,g,t	Relationship	c,g,t	c,g,t	c,g,t	c,g,t

B Dataset Statistics

Please see Figures 4-9 for the detailed dataset statistics. Specifically,

- 148 hours of videos
- 1,412 activity instances from 20 activity classes ranging from 31s to 600s and with an average duration of 241s.
- 15,842 sub-activity instances from 91 sub-activity classes ranging from 3s to 31s and with an average duration of 9s.
- 161,265 atomic action interaction instances.
- 636,194 image-level actor instances and 104,564 video-level actor instances from 26 classes.
- 349,034 image-level object instances and 47,494 video-level object instances from 225 classes.
- 1,037,319 relationship instances from 52 classes.
- 704,230 attribute instances from 13 classes.

C Dataset Access

Along with the MOMA-LRG dataset, we release a dataset toolkit [1] that allows easy access and will facilitate replication of the results in our work and future works as well. This code base quickly processes the dataset and allows for easy integration of MOMA-LRG within any existing framework.

The layout of the MOMA-LRG dataset directory is described in Figure 9.

[1]https://github.com/d1ngn1gefe1/moma

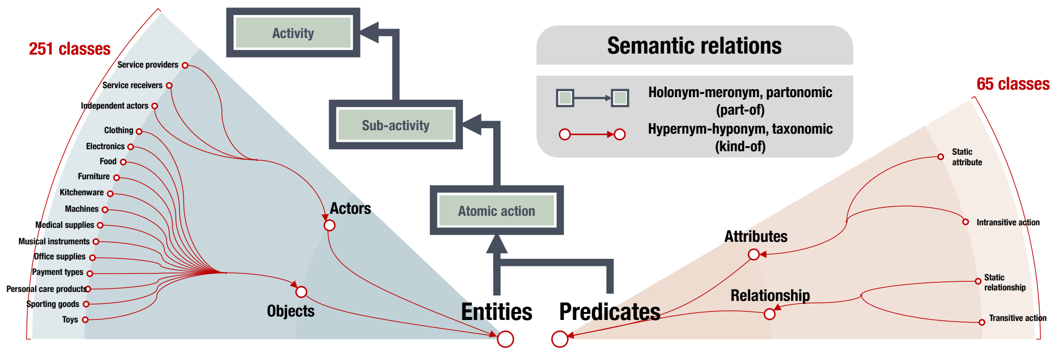

Figure 2: Partonomic and taxonomic hierarchies of MOMA-LRG. MOMA-LRG breaks down activities into sub-activities, which are in turn described by atomic actions. Atomic actions are broken down into entities (actors and objects), whose interactions with each other are described by predicates that either attributes (unary, involving one entity) or relationship (binary, involving two entities).

```
$ tree dir_moma
.
├── anns
│   ├── anns.json
│   ├── split_std.json
│   ├── split_fs.json
│   ├── clips.json
│   └── taxonomy
└── videos
    ├── all
    ├── raw
    ├── activity_fr
    ├── activity
    ├── sub_activity_fr
    ├── sub_activity
    ├── interaction
    ├── interaction_frames
    └── interaction_video
```

Figure 9: The dataset directory layout.

D Main Manuscript Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]

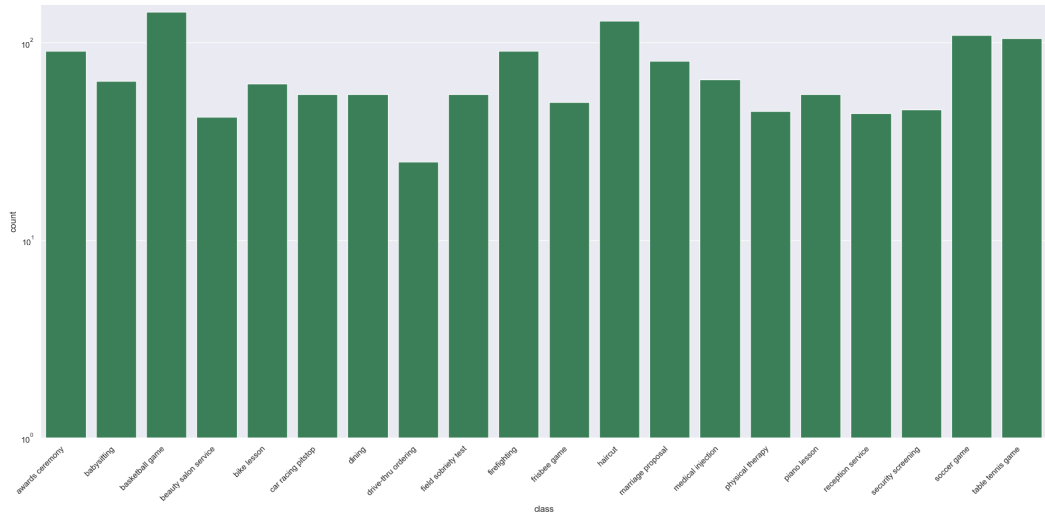

Figure 3: The class distribution of MOMA-LRG activities.

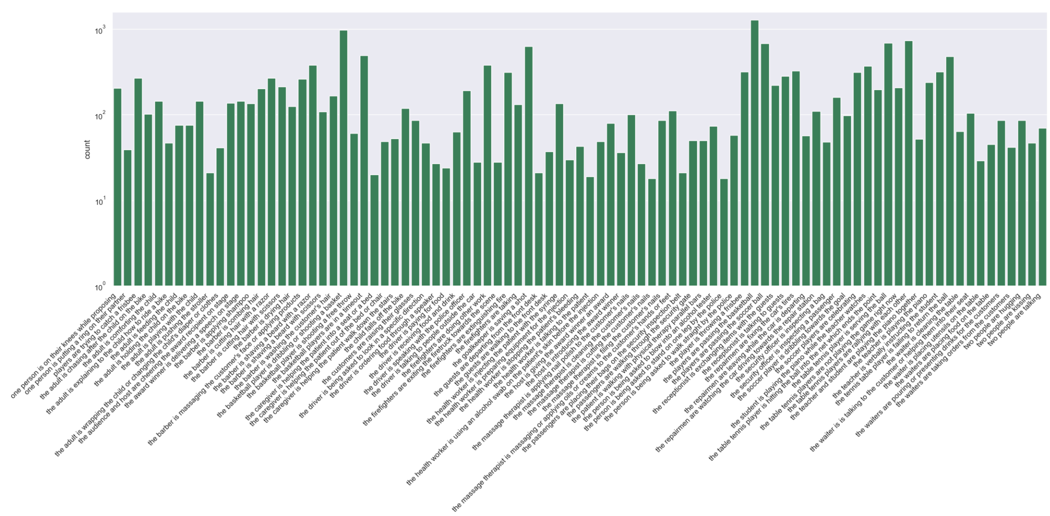

Figure 4: The class distribution of MOMA-LRG sub-activities.

   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

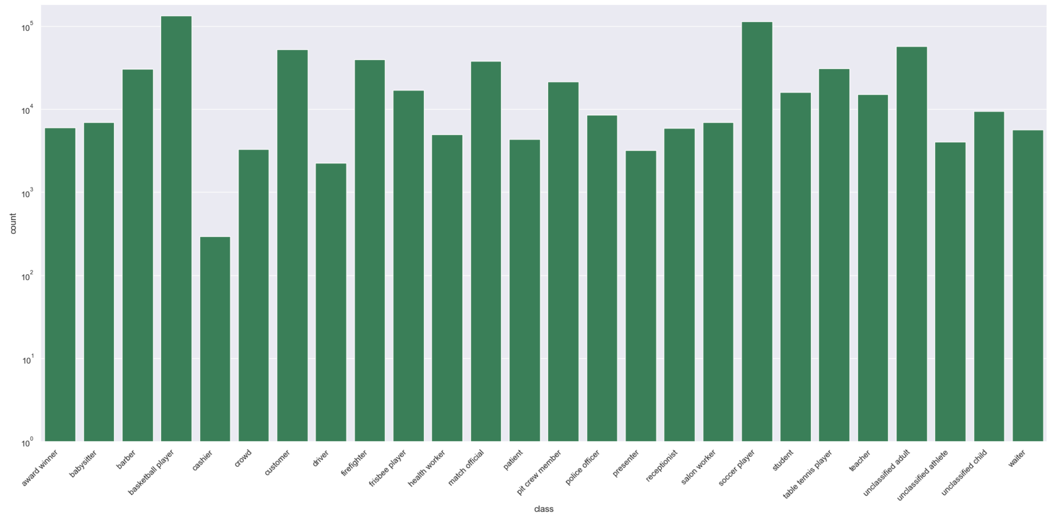

Figure 5: The class distribution of MOMA-LRG actors.

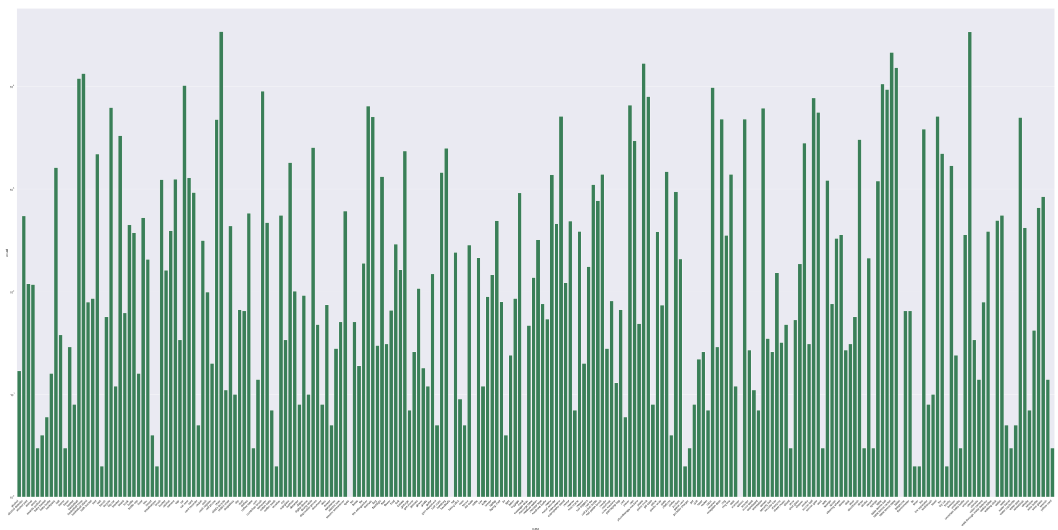

Figure 6: The class distribution of MOMA-LRG objects.

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

E Supplementary Material Checklist

1. Dataset documentation and intended uses. Recommended documentation frameworks include datasheets for datasets, dataset nutrition labels, data statements for NLP, and accountability frameworks.

Please see Section C.

2. URL to website/platform where the dataset/benchmark can be viewed and downloaded by the reviewers.

Please see Section C.

3. Author statement that they bear all responsibility in case of violation of rights, etc., and confirmation of the data license.

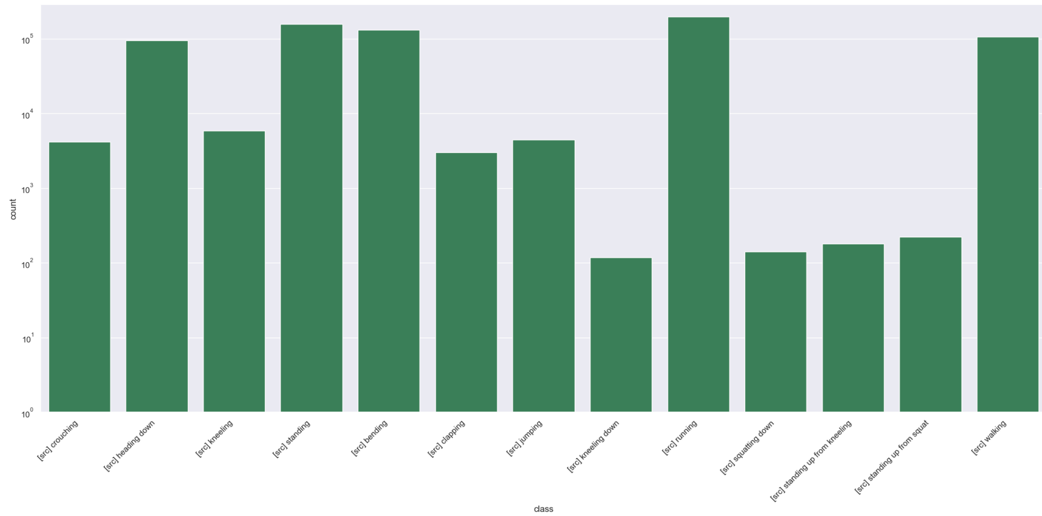

Figure 7: The class distribution of MOMA-LRG attributes.

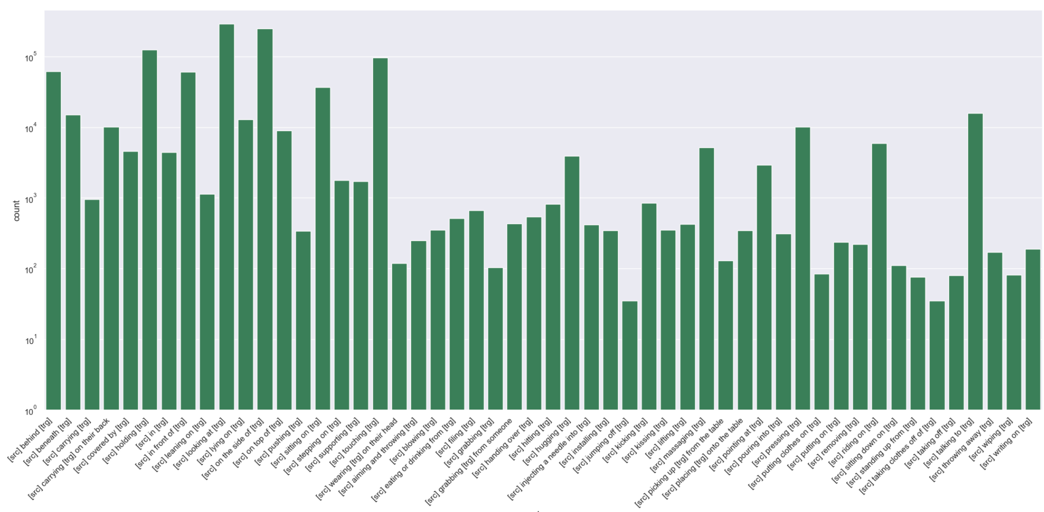

Figure 8: The class distribution of MOMA-LRG relationships.

We hereby confirm the data license and bear all responsibility in case of violation of rights.

4. Hosting, licensing, and maintenance plan. The choice of hosting platform is yours, as long as you ensure access to the data (possibly through a curated interface) and will provide the necessary maintenance.

The videos in our dataset were collected from YouTube. We provide the dataset annotations, a script to crawl raw videos, and a script to preprocess the videos into the desired format.

5. Links to access the dataset and its metadata. This can be hidden upon submission if the dataset is not yet publicly available but must be added in the camera-ready version. In select cases, e.g when the data can only be released at a later date, this can be added afterward (up to a year after the submission deadline). Simulation environments should link to open source code repositories.

Please see Section C.

6. The dataset itself should ideally use an open and widely used data format. Provide a detailed explanation on how the dataset can be read. For simulation environments, use existing frameworks or explain how they can be used.

The dataset files are in either .jpg, .png, or .mp4, all of which are widely used data format. We have provided a dataset toolkit and API (please see Section C) which allows for easy access to the dataset.

7. Long-term preservation: It must be clear that the dataset will be available for a long time, either by uploading to a data repository or by explaining how the authors themselves will ensure this.

We have a team that is dedicated to this project, being in charge of the maintenance, QA, and future extension of the MOMA-LRG dataset.

8. Explicit license: Authors must choose a license, ideally a CC license for datasets, or an open source license for code (e.g. RL environments). An overview of licenses can be found here: https://paperswithcode.com/datasets/license.

We plan to license our dataset under a CC BY license to allow for broad use of our work across research and industry. Specifically, we will use Attribution 4.0 International (CC BY 4.0).

9. Add structured metadata to a dataset's meta-data page using Web standards (like schema.org and DCAT): This allows it to be discovered and organized by anyone. A guide can be found here: https://developers.google.com/search/docs/data-types/dataset. If you use an existing data repository, this is often done automatically.

Please see Section C.

10. Highly recommended: a persistent dereferenceable identifier (e.g. a DOI minted by a data repository or a prefix on identifiers.org) for datasets, or a code repository (e.g. GitHub, GitLab,...) for code. If this is not possible or useful, please explain why.

Please see Section C.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

[2] Kellie Corona, Katie Osterdahl, Roderic Collins, and Anthony Hoogs. Meva: A large-scale multiview, multimodal video dataset for activity detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1060–1068, January 2021.

[3] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017.

[4] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018.

[5] 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.

[6] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. The ava-kinetics localized human actions video dataset. *CoRR*, abs/2005.00214, 2020.

[7] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13536–13545, 2021.

[8] Srikanth Malla, Behzad Dariush, and Chiho Choi. TITAN: future forecast using action priors. *CoRR*, abs/2003.13886, 2020.

[9] 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.

[10] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2616–2625, 2020.

[11] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

[12] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Human action localization with sparse spatial supervision. *arXiv preprint arXiv:1605.05197*, 2016.

[13] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, 2021.