Appendix - ACE:
Action Concept Enhancement of Video-Language Models in Procedural Videos

Reza Ghoddoosian Nakul Agarwal Isht Dwivedi Behzad Darisuh
Honda Research Institute, USA
{reza_ghoddoosian, nakul_agarwal, idwivedi, bdariush}@honda-ri.com

1. Overview

In this appendix, we provide important implementation details of our method as well as the list of unseen synonyms used during Synonym Robustness Test (SRT). We also compare our model with LaVila [7], as an egocentric-focued baseline, on egocentric videos of the GTEA [3] dataset.

2. Implementation Details

We use a 12-layer TimeSformer [2] video encoder pretrained on Howto100M via ProcVLR [8] and the original 12-layer CLIP text encoder (ViT-B/16). The video input to our encoder comprises of 30 frames over 3 seconds in ATA, 25 frames over 2 seconds in IKEA, and 15 frames over 1 second in GTEA videos. These temporal windows are centered at the mid point of each action segment during training. For evaluation, we follow the practice of sampling three temporal clips of 224x224 crops per video and report the average [4]. GPT-4 [1] generates synonyms for our method. The number of first and second order children in synonym trees are 2,9 and 11 for the IKEA, GTEA and ATA datasets, respectively. Furthermore, batch size is set to 16, temperature τ is adjusted to 0.02, and SGD optimizes the model for up to 15 epochs on ATA, and 12 epochs for IKEA and GTEA datasets. Our parameters are the same for our VLM and the one used in our direct baseline ProcVLM [8].

Importantly, during training at each iteration, we ensure that actions sharing the same root verb are assigned the same randomly-selected verb synonym to compute L_{rand}. However, during testing, actions with the same root verb may have different synonyms chosen in each run, as synonyms for each action are sampled independently at test time (refer to Table 2-4).

Furthermore, we prompt GPT-4 as follows to generate M synonyms for action X of the ATA dataset as an example: *”what are M synonyms of the action (X) during toy assembly? please follow the constraints below:*
1- list each synonym in a new line without any numbering and period or commas.
2- make sure the resulting sentences semantically and contextually make sense given the assembly context.
3- start each line with a verb and all in small letters.
4- use the same object and sentence structure as the query.
5- if the verb is a phrasal verb like ’put down’, then place the whole phrasal verb in the begging of the sentence.
6- if the query is a phrasal verb, then I encourage you to output phrasal verbs too, specially if the phrasal verb indicates some spatial information about the scene. ”

2.1. Comparison with LaVila

LaVila is a video-language model that is pretrained on augmented captions in long-term and untrimmed videos. Having said that, LaVila's pretrained checkpoints are only available based on the egocentric videos of Ego4D [5]. On the other hand, ACE is pretrained on execontric videos of Howto100M. Therefore, a direct comparison between LaVila and ACE on GTEA's egocentric videos is not entirely fair, as LaVila has a pretraining advantage. Nonetheless, we compare our method with LaVila in Table 1. While LaVila outperforms ACE on default labels, it struggles with action synonym understanding, where ACE demonstrates more robust performance, despite being pretrained on third-person videos.

3. Synonym Robustness Test (SRT) Labels

In order to make our SRT results comparable with future work, we provide the GPT-generated synonyms for the novel actions across our three benchmark datasets (Table 2-4). The default/root labels are indicated in bold in each table. In few instances, for a given run, the connection between the AI-generated synonyms and their underlying action concept might be loose, yet such synonyms describe their associated concepts more accurately relative to the labels of other actions in the same run. Besides, the coarser labels evaluate the action concept understanding of VLMs at various degrees of granularity. A robust method should have a high mean a low standard deviation when tested against these 10 sets of action synonyms.

Table 1. Action classification *acc* of the egocentric GTEA videos.

		GTEA Dataset *[6 base and 4 novel classes]*			
		Default Labels			SRT
Method	Pretrained on	Seen	Unseen	HM	Unseen
LaVila [7]	Ego4D [5]	**96.0**	**71.4**	**81.9**	37.5±21.0
Ours	Howto100M [6]	85.1	67.2	75.1	**45.0±16.8**

Table 2. GPT-genrated action synonyms for the Synonym Robustnss Test (SRT) on the novel classes of the ATA dataset. Labels in each column correspond to the same action concept. The root labels are highlighted in bold.

Run	Unseen Action Labels - ATA				
1	**’drop item’**	**’balance part’**	**’pick up item’**	**’spin block’**	**’hammer pin’**
2	’leave item’	’stabilize part’	’clutch item’	’wheel block’	’whack pin’
3	’set down item’	’center part’	’retrieve item’	’turn block’	’nail pin’
4	’lower item’	’align part’	’grab item’	’twirl block’	’pound pin’
5	’let fall item’	’steady part’	’catch item’	’circle block’	’bash pin’
6	’deposit item’	’level part’	’hold item’	’swivel block’	’beat pin’
7	’release item’	’adjust part’	’lift item’	’rotate block’	’drive pin’
8	’place item’	’calibrate part’	’take item’	’revolve block’	’strike pin’
9	’lay down item’	’match part’	’grasp item’	’twist block’	’thunk pin’
10	’ditch item’	’weigh part’	’harness item’	’wind block’	’smash pin’

Table 3. GPT-genrated action synonyms for the Synonym Robustnss Test (SRT) on the novel classes of the IKEA dataset. Labels in each column correspond to the same action concept. The root labels are highlighted in bold. Please zoom in for a better view.

Run	Unseen Action Labels - IKEA									
1	**’lay down shelf’**	**’lay down side panel’**	**’lay down front panel’**	**’push table top’**	**’position drawer right side up’**	**’lay down table top’**	**’slide bottom panel’**	**’push table’**	**’lay down bottom panel’**	**’lay down back panel’**
2	’set down shelf’	’place side panel’	’set down front panel’	’press table top’	’place drawer right side up’	’put down table top’	’insert bottom panel’	’press table’	’set down bottom panel’	’place down back panel’
3	’position down shelf’	’rest side panel’	’position down front panel’	’shift table top’	’align drawer right side up’	’position down table top’	’push bottom panel’	’shove table’	’position down bottom panel’	’put down back panel’
4	’put down shelf’	’position side panel’	’place down front panel’	’slide table top’	’set drawer right side up’	’set down table top’	’guide bottom panel’	’slide table’	’put down bottom panel’	’set down back panel’
5	’position down shelf’	’rest side panel’	’place down front panel’	’push table top’	’orient drawer right side up’	’set down table top’	’guide bottom panel’	’press table’	’place down bottom panel’	’set down back panel’
6	’set down shelf’	’rest side panel’	’lay down front panel’	’nudge table top’	’orient drawer right side up’	’position down table top’	’move bottom panel’	’shove table’	’position down bottom panel’	’set down back panel’
7	’place down shelf’	’place side panel’	’put down front panel’	’nudge table top’	’orient drawer right side up’	’put down table top’	’move bottom panel’	’nudge table’	’put down bottom panel’	’place down back panel’
8	’put down shelf’	’lay down side panel’	’set down front panel’	’nudge table top’	’align drawer right side up’	’place down table top’	’guide bottom panel’	’press table’	’place down bottom panel’	’position down back panel’
9	’position down shelf’	’rest side panel’	’put down front panel’	’nudge table top’	’align drawer right side up’	’set down table top’	’slide bottom panel’	’shove table’	’position down bottom panel’	’place down back panel’
10	’position down shelf’	’set down side panel’	’lay down front panel’	’slide table top’	’align drawer right side up’	’set down table top’	’guide bottom panel’	’push table’	’lay down bottom panel’	’put down back panel’

Table 4. GPT-genrated action synonyms for the Synonym Robustnss Test (SRT) on the novel classes of the GTEA dataset. Labels in each column correspond to the same action concept. The root labels are highlighted in bold. Given that we only use verbs for the GTEA datasets, the term “ingredient” is added to verbs as a placeholder to comply with the verb+object template of the actions.

Run	Unseen Action Labels - GTEA			
1	**’shake ingredient’**	**’fold ingredient’**	**’stir ingredient’**	**’spread ingredient’**
2	’toss ingredient’	’place together ingredient’	’mix ingredient’	’smear ingredient’
3	’rattle ingredient’	’combine ingredient’	’beat ingredient’	’smooth ingredient’
4	’jostle ingredient’	’tuck ingredient’	’whisk ingredient’	’garnish ingredient’
5	’tremble ingredient’	’incorporate ingredient’	’whip ingredient’	’slather ingredient’
6	’sway ingredient’	’integrate ingredient’	’swirl ingredient’	’apply ingredient’
7	’vibrate ingredient’	’mingle ingredient’	’rotate ingredient’	’cover ingredient’
8	’rock ingredient’	’merge ingredient’	’agitate ingredient’	’layer ingredient’
9	’quiver ingredient’	’interlace ingredient’	’incorporate ingredient’	’lather ingredient’
10	’wobble ingredient’	’fuse ingredient’	’dissolve ingredient’	’daub ingredient’

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 1

[2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 1

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

[4] 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. 1

[5] 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

[6] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, 2019. 2

[7] Yue Zhao, Ishan Misra, Philipp Krähenbühl, and Rohit Girdhar. Learning video representations from large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6586–6597, 2023. 1, 2

[8] Yiwu Zhong, Licheng Yu, Yang Bai, Shangwen Li, Xueting Yan, and Yin Li. Learning procedure-aware video representation from instructional videos and their narrations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14825–14835, 2023. 1