Supplementary Material for LEMMA: A Multi-view Dataset for LEarning Multi-agent Multi-task Activities

Baoxiong Jia[0000−0002−4968−3290], Yixin Chen[0000−0002−8176−0241], Siyuan Huang[0000−0003−1524−7148], Yixin Zhu[0000−0001−7024−1545], and Song-Chun Zhu[0000−0002−1925−5973]

UCLA Center for Vision, Cognition, Learning, and Autonomy (VCLA)
{baoxiongjia, ethanchen, huangsiyuan, yixin.zhu}@ucla.edu
sczhu@stat.ucla.edu

1 Annotation Details

In this section, we describe additional details of the annotation process, including the design of the verb and noun classes, as well as the verb patterns.

We build our action verb taxonomy and collect a compact dictionary of action verbs for annotation based on the previous dataset that captures goal-directed activities in the kitchen [5,3,1] and living room [8,6]. Specifically, we start from the action verb vocabulary summarized in the EPIC-KITCHENS dataset [1] to cover common actions conducted in the kitchen and living room. We then

Table 1: Verb vocabulary, corresponding verb patterns, and annotated examples.

Verb	Template	Example
blend	blend targets with tools	blend coffee with spoon
clean	clean targets with tools	clean cup with sink
close	close targets	close drawer
cook	cook targets in location with tools	cook meat in pan with fork
cut	cut targets on location with tools	cut watermelon on cutting-board with knife
drink	drink targets with tools	drink milk with cup
eat	eat targets with tools	eat meat with fork
fill	fill targets with tools	fill juicer with sink
get	get targets from location with tools	get spoon, fork from drawer using hand
open	open targets	open closet
play	play targets with tools	play game-console with controller
point-to	point to targets	point to kettle
pour	pour targets into location with tools	pour coffee into cup with spoon
put	put targets to location with tools	put meat to pan with fork
sit-on	sit on location	sit on sofa
sweep	sweep targets with tools	sweep floor with vacuum
switch	switch targets with tools	switch TV with remote
throw	throw targets into location	throw wrapping into trash-can
turn-off	turn off targets with tools	turn off TV with remote
turn-on	turn on targets with tools	turn on microwave with hand
watch	watch targets	watch TV
wash	wash targets	wash cup, spoon
work-on	work on targets	work on cup-noodles

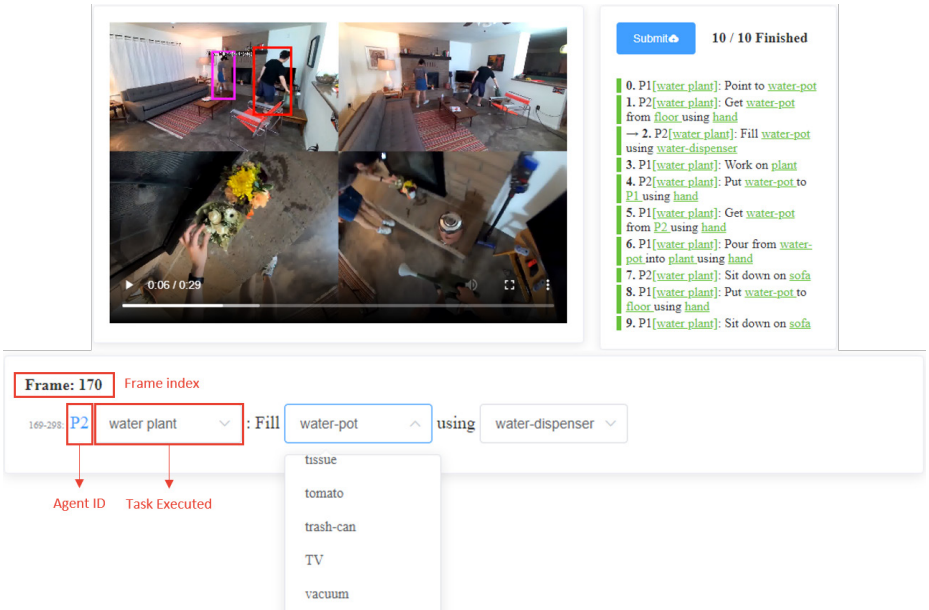

Fig. 1: A visualization of the annotation tool developed for annotating the nouns and governing tasks.

reduce the vocabulary size by eliminating unrelated action verbs, such as "pet-down," "walk," and "decide-if." We further add action verbs (*e.g.*, "point to") to incorporate human-human interactions in multi-agent collaboration scenarios. We provide this action verb vocabulary to the AMT workers for the first-round annotation. After resolving the ambiguities in language descriptions, 24 action verbs remain in the final action verb vocabulary. For the noun vocabulary, we enumerate all possible objects that could potentially be interacted in the 15 daily tasks, resulting in a noun vocabulary with a size of 64. For each action verb, we design the action verb patterns with three semantic positions for interacting objects, target/source location, and tools. The action verb vocabulary and the action verb patterns are summarized in Table 1.

Bounding boxes and verbs. We use Vatic [7] to annotate the human bounding boxes and verb patterns with blanks to be filled in. The human bounding boxes are annotated as rectangles, and the verb patterns are treated as attributes for each human bounding box.

Nouns. We further develop an interactive annotation tool to fill nouns into the blanks for each semantic position of each action verb. We also use this tool to annotate the governing task of each compositional action. Each blank can be filled by choosing from a set of given options, as shown in Fig. 1. During the annotation process, the synchronized video from egocentric views and TPVs are

merged into the same window and presented to the AMT worker. We visualize the bounding box of agents with their ID "P1" and "P2" to help AMT workers find the correspondences. A full snapshot of the annotation interface is shown in Fig. 1. After filling all blanks, we manually go through all the annotations and resolve the ambiguous action annotations by eliminating and merging the nouns with occurrence frequencies of less than 50. We show annotation results in Fig. 5.

2 Implementation Details

Compositional Action Recognition Below, we detail the designs and implementations of the two proposed models, "branching" and "sequential," for the compositional action recognition task. We build both models on top of the backbone 3D CNN model and use a multi-branch network to train verbs, nouns, and their correspondences. We start from the "sequential" model as the "branching" model is a variant of the "sequential" model; see an illustration in Fig. 2.

For the verb branch, we propose 3 verb candidates for each segment and extract verb visual features for verb recognition. Specifically, the verb visual features $f_{\text{verb}} = \{F^{(i)}_{\text{verb}}(f_{\text{vis}})\}$ are generated using three different linear projections $\{F^{(i)}_{\text{verb}}\}_{i=1,2,3}$ applied onto the feature f_{vis} extracted by 3D CNN. We sort ground-truth action labels according to their index in the verb vocabulary and use cross-entropy loss $\mathcal{L}_{\text{verb}}$ as the supervision for verb recognition.

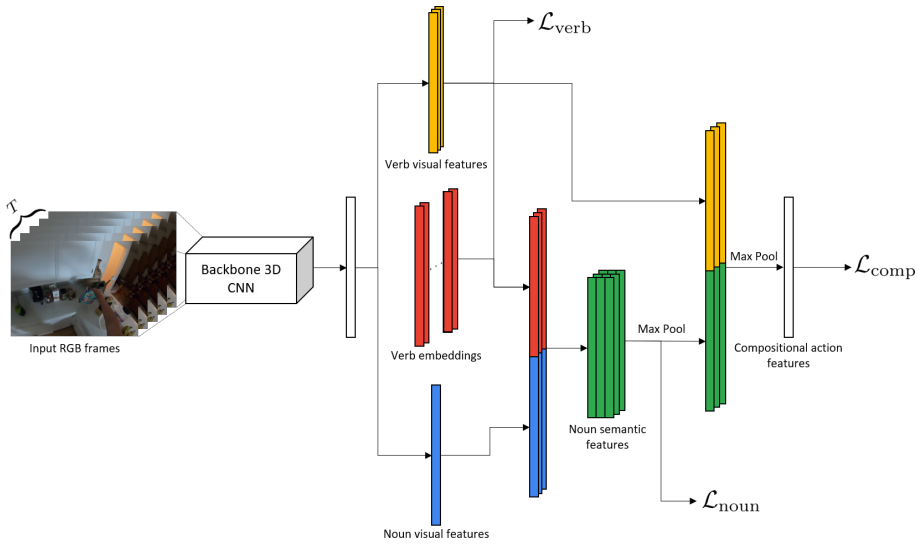

Fig. 2: An illustration of the proposed "sequential" model, which predicts verbs, nouns, and compositional actions jointly.

For the noun branch, we utilize the embeddings of each verb as additional features by GloVe [4]. The embedding of each verb is passed into a linear projection layer and concatenated with the extracted visual features to generate noun feature vectors $f_{\text{noun-vis}}$. Next, we use three different linear projections $\{F^{(i)}_{\text{noun}}\}_{i=1,2,3}$ to generate features for each of the noun visual feature vectors and obtain noun semantic features $f_{\text{noun-sem}} = \{[F^{(i)_{\text{noun}}}(f^{(j)}_{\text{noun-vis}})]_{i=1,2,3}\}_{j=1,2,3}$. As we generate ground-truth labels following the same scheme, we use binary cross-entropy loss $\mathcal{L}_{\text{noun}}$ as the supervision for recognizing nouns at their correct semantic positions using $f_{\text{noun-sem}}$. During training, the embeddings of the ground-truth verbs are fed into the network. During testing, we use the embedding of the predicted top-3 verbs.

We use max-pooling to summarize the noun semantic features and concatenate it with verb visual features. We use another layer of max pooling to generate the final compositional action feature and use binary cross-entropy loss as $\mathcal{L}_{\text{comp}}$ to provide supervision for compositional action recognition. The joint loss is

$$\mathcal{L} = \mathcal{L}_{\text{verb}} + \mathcal{L}_{\text{noun}} + \mathcal{L}_{\text{comp}}.$$

For the "branching" model, we follow the same basic scheme of the "sequential" model but remove the connection between the verb branch and the noun branch by discarding the additional verb embeddings. The remaining details of the architecture, as well as the optimizing objectives, remain the same.

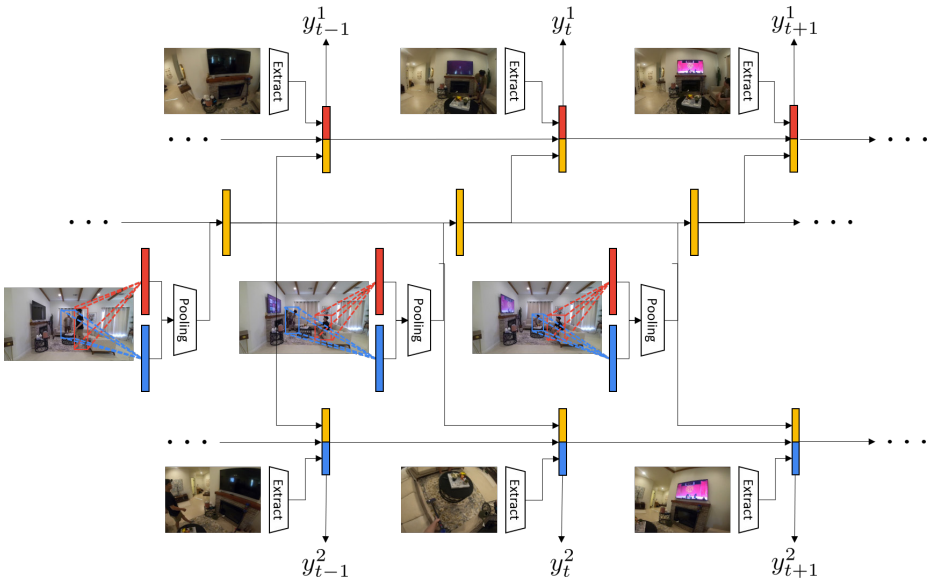

Fig. 3: An illustration for the multi-agent variants of the original sequential model with TPV features as additional features.

(a) Results of action anticipation. $\text{GT}_{\text{Action}}$ indicates the ground-truth action in the next frame.

(b) Results of task anticipation. GT_{Task} indicates the ground-truth task for actions in the next segment.

Fig. 4: Qualitative results of action and task anticipation on LEMMA.

.

Action and Task Anticipation We explain the details of the multi-agent variants of the compared sequential models. For scenarios where two agents collaborate, we incorporate the egocentric features of another agent (denoted as Ego in Table 3) or TPV features (denoted as TPV in Table 3) through a pooling mechanism, similar to [2]. We use these pooled features to incorporate global task execution information to each agent. Specifically, we concatenate the extracted global features to features extracted by the backbone 3D CNN models from the target agent's egocentric view for training and inference. For TPV, we use ROIAlign to extract visual features corresponding to each agent's bounding box. An illustration of the pipeline with TPV features as additional features is shown in Fig. 3.

3 Additional Experiment Results

We show some qualitative results for action and task anticipation.

Fig. 5: Examples of the annotated bounding boxes and compositional actions.

References

1. Damen, D., Doughty, H., Maria Farinella, G., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Scaling egocentric vision: The epic-kitchens dataset. In: Proceedings of European Conference on Computer Vision (ECCV) (2018)
2. Gupta, A., Johnson, J., Fei-Fei, L., Savarese, S., Alahi, A.: Social gan: Socially acceptable trajectories with generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
3. Kuehne, H., Arslan, A., Serre, T.: The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
4. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP) (2014)
5. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
6. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: Pigraphs: learning interaction snapshots from observations. ACM Transactions on Graphics (TOG) **35**(4), 1–12 (2016)
7. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision (IJCV) **101**(1), 184–204 (2013)
8. Wu, C., Zhang, J., Savarese, S., Saxena, A.: Watch-n-patch: Unsupervised understanding of actions and relations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)