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

Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu

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

Abstract. Understanding and interpreting human actions is a long-standing challenge and a critical indicator of perception in artificial intelligence. However, a few imperative components of daily human activities are largely missed in prior literature, including the goal-directed actions, concurrent multi-tasks, and collaborations among multi-agents. We introduce the LEMMA dataset to provide a single home to address these missing dimensions with meticulously designed settings, wherein the number of tasks and agents varies to highlight different learning objectives. We densely annotate the atomic-actions with human-object interactions to provide ground-truths of the compositionality, scheduling, and assignment of daily activities. We further devise challenging compositional action recognition and action/task anticipation benchmarks with baseline models to measure the capability of compositional action understanding and temporal reasoning. We hope this effort would drive the machine vision community to examine goal-directed human activities and further study the task scheduling and assignment in the real world.

Keywords: Dataset, Multi-agent Multi-task Activities, Compositional Action Recognition, Action and Task Anticipations, Multiview

1 Introduction

Activity understanding is one of the most fundamental problems in artificial intelligence and computer vision. As the most readily available learning source, videos of daily human activities could be used to train intelligent agents and, in turn, to assist humans. However, compared to recent progress in learning from static images [2,23,22,41], current machine vision's ability to understand activities from videos still falls short. Admittedly, activity understanding is inherently more challenging, which requires reason about the complex structures in activities along the additional temporal dimension; but we argue there are more profound reasons that we must look back to the origin of activity understanding.

The study and analysis of human motion perception are rooted in the field of neuroscience [54]. Using a dot-representation of human motions, Johansson [26] adopted a method to produce proximal patterns (*i.e.*, the moving light display experiment), which demonstrated that human perception of activities does not
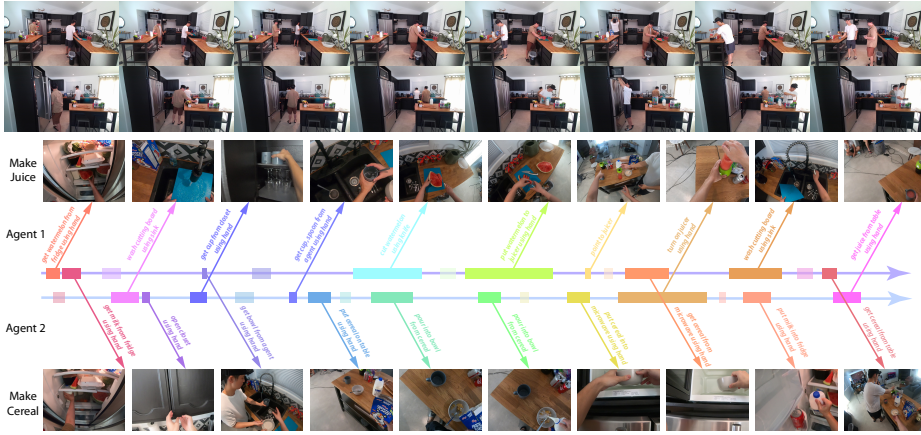
Fig. 1: Illustrations of the proposed multi-view dataset with annotations. From top to bottom: frames captured from the third-person primary view, frames captured from the third-person side view, annotated segments of each agent executing tasks, and corresponding frames captured from the first-person view.

tightly couple with *pixel-based features*; human subjects can still perceive the semantics of activities from *sparse* representations of motions. Evidence from developmental psychology, the classic Heider-Simmel experiment, further suggests that we perceive human activities from as *goal-directed* behaviors [57,5,15,11]; it is the underlying intent, rather than the surface pixels or behavior, that matters when we observe motions [4]. Such a **goal-directed [31] perspective** of activity understanding has been largely left untouched in computer vision.

Daily human activities are intrinsically multi-tasked [36,44]; understanding activity naturally demands a learning system to interpret concurrent interactions. As agents' decision-making processes are deeply affected by their unique social values, task scheduling is significantly affected by interactions (*e.g.*, cooperation, competition, subordination) among multi-agents [28]. These observations implicate that the machine vision system must objectively understand how a given task should be decomposed into atomic-actions, how multi-tasks should be executed and coordinated in parallel among multi-agents, and take the perspective from human agents to understand why the observed human activities are optimal solutions. Such a **decompositional, multi-task, multi-agent, diagnostic-driven, social perspective** of activity understanding is critical for an intelligent agent to understand human behavior and team with humans collaboratively; yet it is broadly missing in activity understanding literature.

The semantics of human actions are intrinsically ambiguous when described in natural language. For instance, although both "opening the fridge" and "opening a book" use the action verb "open," their semantics of the actions are utterly different. In this paper, we take the stance of Grice's influential work on language act [20]—technical tools for reasoning about rational action should elucidate lin-

guistic phenomena [17]. Specifically, the compositional relations between the verbs and nouns could reveal the functionality of the object and the patterns of human-object interactions, which subsequently facilitate the understanding of the observed human activities and the language that describes them. Though the previous work [19] attempted to address this issue, more general and flexible **compositional relations for describing human actions interacting with objects** are requisite for a goal-directed activity understanding.

Motivated by these deficiencies in prior work, we introduce the LEMMA dataset to explore the essence of complex human activities in a goal-directed, multi-agent, multi-task setting with ground-truth labels of compositional atomic-actions and their associated tasks. By quantifying the scenarios to up to two multi-step tasks with two agents, we strive to address human multi-task and multi-agent interactions in four scenarios: single-agent single-task (1×1), single-agent multi-task (1×2), multi-agent single-task (2×1), and multi-agent multi-task (2×2). Task instructions are only given to one agent in the 2×1 setting to resemble the robot-helping scenario, hoping that the learned perception models could be applied in robotic tasks (especially in HRI) in the near future.

Both the third-person views (TPVs) and the first-person views (FPVs) were recorded to account for different perspectives of the same activities; see Fig. 1. We densely annotate atomic-actions (in the form of compositional verb-noun pairs) and tasks of each atomic-action, to facilitate the learning of multi-agent multi-task task scheduling and assignment; see more details in Section 2.

1.1 Related Work

In this section, we review and compare prior indoor activity datasets on the basis of tasks and captured video contents; see a detailed summary in Table 1.

Crowd-sourced from online videos and movie sharing platforms, typical large-scale video datasets [49,27,8,9,14] focus on **video-level summarization and classification**. Although activity classes exhibit a large inter-class variability, spanning from outdoor sports activities to indoor household activities, they generally lack sequential, goal-directed activities. Notably, they suffer from a major

Table 1: Comparisons between LEMMA and relevant indoor activity datasets.

Dataset	Task Annotation	Multi-agent	Multi-task	Multi-view	Samples	Frames	Action Classes	Action Segments	Actions per Video	Modality	Year
MPII Cooking [42]	✓	✗	✗	✗	273	2.9M	88	14,105	51.7	RGB	2012
ADL [39]	✗	✗	✓	✗	20	1.0M	32	436	13.6	RGB	2012
50Salads [50]	✓	✗	✗	✗	50	0.5M	17	966	19.3	RGB-D	2013
CAD-120 [29]	✗	✗	✗	✗	120	0.1M	10	1,175	9.8	RGB-D	2013
Breakfast [30]	✓	✗	✗	✓	433	3.0M	50	3,078	7.1	RGB	2014
Watch-n-Patch [59]	✓	✗	✗	✗	458	0.1M	21	2978	6.5	RGB-D	2015
Charades [48]	✗	✗	✓	✗	9,848	7.4M	157	67,000	6.8	RGB	2016
Something-Something [19]	✗	✗	✗	✗	108,499	-	174	108,499	1.0	RGB	2017
EGTEA GAZE+ [34]	✓	✗	✗	✗	86	2.4M	106	10,325	120.1	RGB	2018
EPIC-KITCHENS [12]	✗	✗	✓	✗	432	11.5M	149	39,596	91.7	RGB	2018
LEMMA (proposed)	✓	✓	✓	✓	324	4.6M	641	11,781	36.4	RGB-D	2020

drawback [16]; activities are highly correlated to the general scene and object context, possessing a strong dataset bias for activity understanding.

Some datasets tackle the **human atomic-actions** using short clips or limited tasks, with a focus on the semantics of action verbs and objects [19], 3D action analysis [33,25,45], and action grounding with multi-modality inputs [35]. Although such datasets are suitable for atomic-actions, they are intrinsically impaired at studying the long-term reasoning of goal-directed human activities.

Recently, **concurrent actions** have been taken into consideration. For instance, Charades [48] is a large-scale benchmark for household activities, and Charades-Ego [47] steps further with both FPVs and TPVs. However, the activities involved are mostly unrelated to specific goals due to the crowdsourced script generation process. Similarly, although Multi-THUMOS [60] and AVA [21] focus on highly paralleled activities, and some datasets look at the temporal order of activities [7,52], the unnaturally scripted activities result in the lack of meaningful goal-directed tasks exhibited in our daily life.

Conversely, **instructional video** datasets [1,50,30,29,43] tackle goal-directed multi-step tasks, mostly in cooking, repairing, and assembling activities. In spite of their relevance, they fail to account for multi-agent or multi-task problems. EPIC-KITCHENS [12] is perhaps the only exception; it records naturally paralleled task execution of agents in kitchen environments, but with no task specification or multi-agent interactions. Additionally, prior instructional video datasets have either drastic view perspective changes [61,1,51,53] or limited egocentric view with severe occlusions [39,34], hindering the activity understanding.

Another related stream of work is the learning of group-level activities in a **multi-agent** setting [24], such as detecting key actors [40], predicting future trajectories [38,32], and recognizing collective activities [10,37,46]. However, such coarse-grained multi-agent interactions leave the latent subtlety of collaboration and task assignment untouched. Although simulation-based multi-agent environments [3,55,6] can partially address such an issue, learning from noisy and real visual input in physical work is still essential for understanding collaborative planning behaviors of agents in the context of complex daily tasks.

The collected LEMMA dataset strives to address the shortcomings of the aforementioned works, capturing goal-directed, decompositional, multi-task activities with multi-agent collaborations. As shown in Table 1, the size, annotation, and actions per video of LEMMA are at a comparable scale to state-of-the-art benchmarks. We hope such a design will boost the study of human activity understanding and potentially motivate new cross-disciplinary research insights.

1.2   Contributions

This paper's contribution is three-fold. (i) We design and collect a multi-view video dataset, capturing multi-agent, multi-task activities with goal-directed daily tasks. (ii) We annotate the dataset, focusing on the compositionality of actions and the governing task for each atomic-action. (iii) We provide compositional action recognition and action/task anticipation benchmarks by consider-

ing the aforementioned features; we also compare and analyze multiple baseline models to promote future research on human activity understanding.

2   The LEMMA Dataset

This section describes the design, data collection, and data annotation process of the LEMMA dataset. The dataset is profiled by various statistics from diversified perspectives to highlight its potentials in activity understanding.[1]

2.1   Activities and Scenarios

We first build a task pool of 15 common tasks in the kitchen (*e.g.*, "make juice," "make cereal") and living room (*e.g.* "watch TV," "water plant"). On top of these tasks, we design four types of scenarios (with a different focus) to study goal-directed multi-step multi-task indoor activities in multi-agent settings.

1. **Single-agent Single-task** (1×1)**:** Each participant was first asked to perform all tasks from the task pool independently; this ensures participants are clear with the goal of each task and could schedule and assign tasks efficiently in later multi-task or multi-agent scenarios. Participants were asked to read the instructions and walk around to get familiarized with the new environments.

2. **Single-agent Multi-task** (1×2)**:** Each participant was then asked to simultaneously perform two tasks, randomly sampled from the task pool. The participants determined the order of task executions without any restrictions.

3. **Multi-agent Single-task** (2×1)**:** Two participants were asked to perform a single task cooperatively; the task is randomly selected from the task pool. To emulate human-robot teaming accurately, only one participant (primary) was provided with task instructions; the other participant (helper), with no knowledge of the task, was asked to collaborate with the primary agent to finish the task efficiently. Only nonverbal communications (*e.g.*, gestures) were allowed between two participants; this design would open up new venues on nonverbal communications and the emergence of language in real-world environments.

4. **Multi-agent Multi-task** (2×2)**:** Both participants were provided with task instructions. Since both participants were asked to accomplish two complex multi-step tasks collaboratively, this scenario has the most natural activity/task patterns and richest mechanisms for learning task scheduling and assignment.

In total, the LAMMA dataset includes 37 unique task combinations in the multi-task scenarios. Participants were explicitly instructed to perform tasks efficiently and provided with a brief task instruction with basic environment information. Except for the specification of the goal states for each task, we add no additional constraint to the order of task execution; participants perform tasks naturally and freely. Fig. 2 shows a sample instruction for the 2×1 scenario.

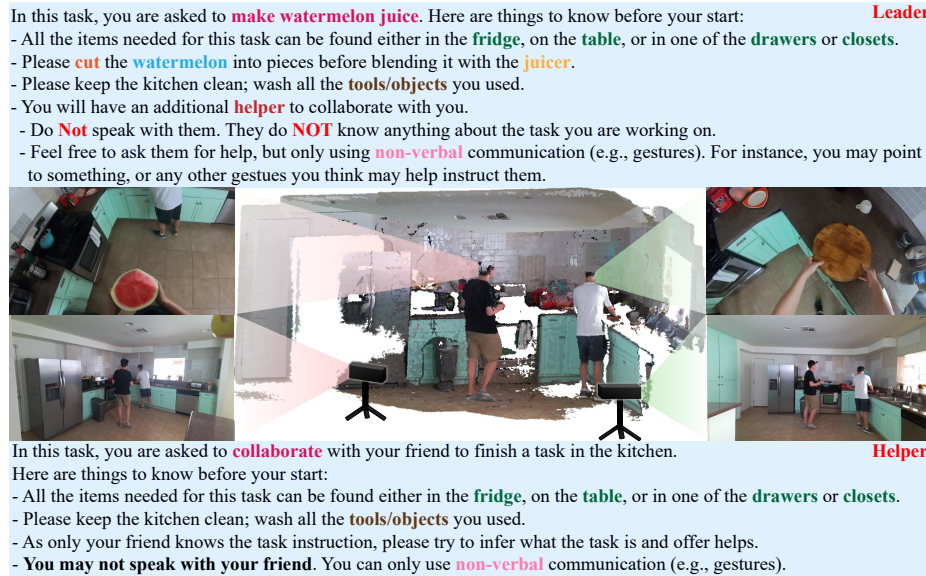[1] The dataset will be made publicly available at the following website with download links and util code: `https://sites.google.com/view/lemma-activity`.

In this task, you are asked to **make watermelon juice**. Here are things to know before your start: **Leader**
- All the items needed for this task can be found either in the **fridge**, on the **table**, or in one of the **drawers** or **closets**.
- Please **cut** the **watermelon** into pieces before blending it with the **juicer**.
- Please keep the kitchen clean; wash all the **tools/objects** you used.
- You will have an additional **helper** to collaborate with you.
  - Do **Not** speak with them. They do **NOT** know anything about the task you are working on.
  - Feel free to ask them for help, but only using **non-verbal** communication (e.g., gestures). For instance, you may point to something, or any other gestues you think may help instruct them.

In this task, you are asked to **collaborate** with your friend to finish a task in the kitchen.   **Helper**
Here are things to know before your start:
- All the items needed for this task can be found either in the **fridge**, on the **table**, or in one of the **drawers** or **closets**.
- Please keep the kitchen clean; wash all the **tools/objects** you used.
- As only your friend knows the task instruction, please try to infer what the task is and offer helps.
- **You may not speak with your friend**. You can only use **non-verbal** communication (e.g., gestures).

Fig. 2: An exemplar task instruction of making juice for two agents in a **Multi-agent Single-task** (2×1) scenario. Middle: Point clouds, TPVs, and FPVs.

2.2   Data Collection

We recorded the data in 7 different Airbnb houses, performed by 8 individuals in 14 unique kitchens/living rooms. To provide different views of performing the daily activities and avoid occlusion in narrow spaces, we set up two Kinect Azure cameras to capture the RGB-D videos of the global scene and human bodies. In addition, each participant was instructed to wear a head-mounted GoPro camera to capture detailed agent-specific actions in an egocentric view. In post-processing, we synchronize the camera recordings of all views at a frame rate of 24 FPS. Fig. 2 shows an example of a scene with a point cloud merged from two Kinects and four RGB views from both Kinects and GoPros. Combining TPVs and FPVs captures most of the details of performing daily activities, provides sufficient data for understanding human activities, and benefits future research in embodied vision. The additional depth information and 3D human skeletons captured by Kinects can also be adopted for future 3D understanding tasks.

2.3   Ground-truth Annotation

We used the Amazon Mechanical Turk (AMT) to annotate both human bounding boxes and action information in the synchronized recordings. Specifically, action information includes the temporal localization of segments, semantic labels, and the governing task of each atomic-action. The semantic labels of atomic-actions are composed of verbs and nouns, representing flexible compositional relations to describe human actions. Additional details are provided below.

(a) Frequency of annotated noun classes across all frames

(b) Frequency of recorded tasks (c) Frequency of annotated verb   (d) Action wordle
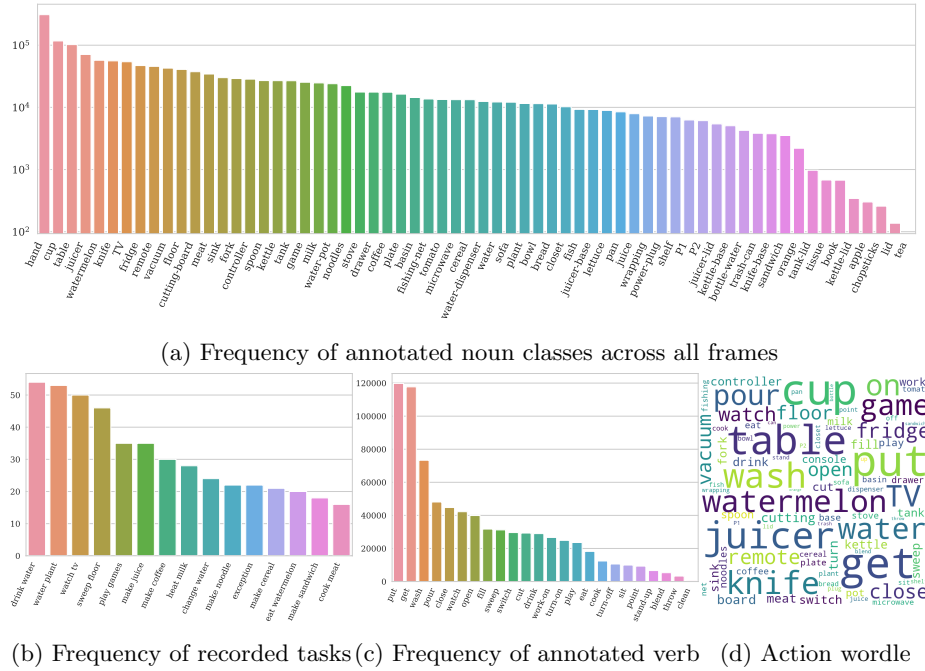
Fig. 3: Statistics of the LEMMA dataset.

Bounding Boxes and Segments: Bounding boxes of humans are annotated on the primary view of TPVs. Skeletons captured by Kinects are used to provide initial estimations of bounding boxes. Next, we use Vatic [56] to adjust bounding boxes and annotate the segments of atomic-actions. The segments of atomic-actions are defined by verbs without corresponding nouns, for example, "put __ to __ using __," "pour into __ from __." Each video was first annotated by two AMT workers; task-irrelevant actions (*e.g.*, "walking," "holding") are ignored. We then compute the Intersection over Union (IoU) of both bounding boxes and temporal segments. A third AMT worker is asked to fine-tune the annotations if the IoU of bounding boxes or segments annotated is lower than 0.5.

Atomic-actions and Activities: Given the verbs of the atomic-action segments, two AMT workers were asked to fill in the blanks of the verb patterns and annotate the governing tasks in multi-task scenarios with a self-developed interactive annotation tool (see *supplementary material*). We allow concurrent actions for each agent with multiple nouns for the same verb; for example, "get spoon, cup from table using hand." As there might exist ambiguities in describing the atomic-actions with natural languages, such as the possible annotations of "wash cup using water" *vs.* "wash cup using sink," we manually go through all the annotations and resolve the ambiguous action annotations following a uniform criterion. Examples of annotation results are shown in *supplementary.*

(a) verbs (y-axis) and tasks (x-axis)   (b) verbs (y-axis) and next verbs (x-axis)
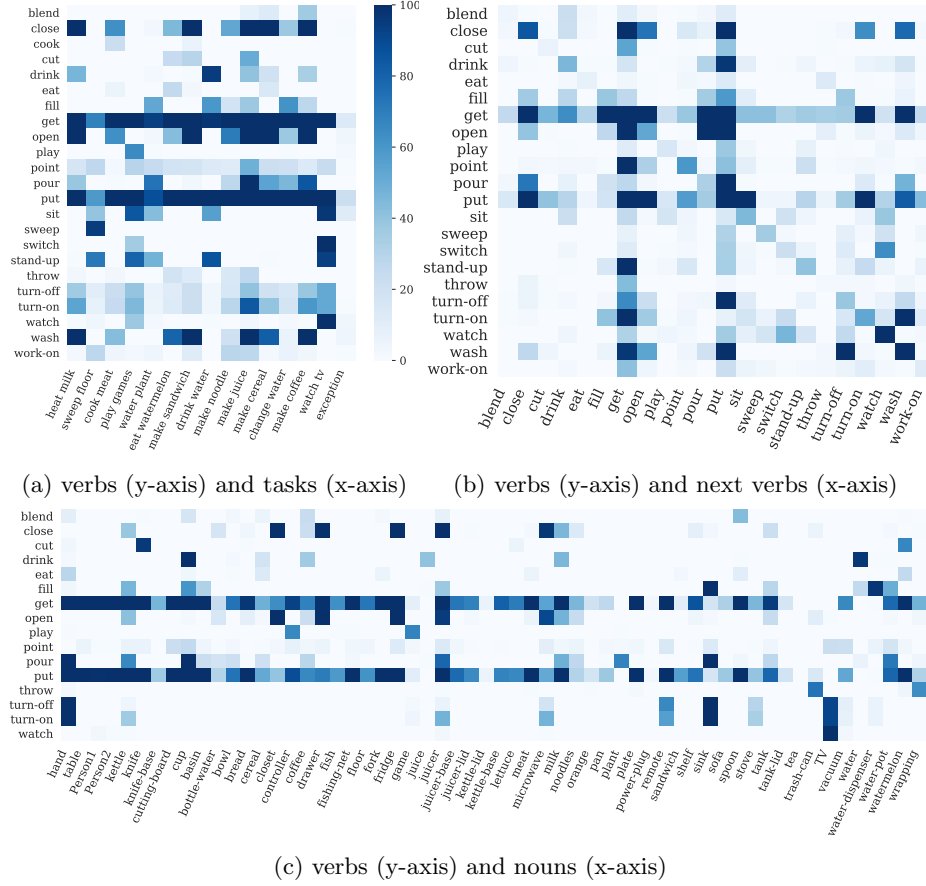
(c) verbs (y-axis) and nouns (x-axis)

Fig. 4: The co-occurrence statistics for verbs, nouns, and tasks in LEMMA.

2.4   Dataset Statistics

In total, we recorded 324 activities, generating 324×2 TPV videos (from both Kinects) and 445 FPV videos. Among them, 136 activities were performed in kitchens and the remaining 188 in the living rooms. The collected LEMMA dataset consists of 127 1×1 activities, 76 1×2 activities, 66 2×1 activities, and 55 2×2 activities. The frequency of the recorded tasks is shown in Fig. 3b. The total duration of all the activities is 10.1 hours, with an average duration of 2 minutes per video and the longest activity of 7 minutes.

We retrieved a total of 4.6 million images during post-processing, including 2.9 million RGB images captured by both GoPros and Kinects and 1.7 million depth images captured by Kinects. We annotated 0.9 million RGB frames captured by the primary view Kinect and gathered 0.8 million annotated frames with one or more actions performed by each of the agents (if multiple).

After resolving annotation ambiguities, we collected 24 verb classes and 64 noun classes, resulting in 862 compositional atomic-action labels, of which 641 appear more than 50 times. We show the frequencies of annotated verbs and nouns in Figs. 3a and 3c; both distributions roughly follow the Zipf's law.

Co-occurrence relations among annotated verbs, nouns, and tasks are shown in Fig. 4. As we can see from Figs. 4a and 4c, verbs like "get" and "put" co-occur with various nouns in almost all of the tasks, which aligns with our intuition that moving objects around consists a large portion of our daily activities. Interactive actions between participants are captured by verbs (*e.g.*, "point-to") and nouns (*e.g.*, "P1," short for "participant 1") in the form of annotations like "get knife from P1 using hand" or "point-to sink."

3   Benchmarks

Aligned with our motivations, two general goals are constructed to evaluate indoor human activity understanding on the collected LEMMA dataset: (i) recognize atomic-actions and their semantics; and (ii) understand the goal-directed activities and monitor multiple concurrent tasks, especially in multi-agent scenarios. Specifically, we define two challenging benchmarks to test the capability of understanding complex goal-directed activities for computer vision algorithms.

3.1   Compositional Action Recognition

Human indoor activities are composed of fine-grained action segments with rich semantics. As mentioned by Goyal *et al.* [19], interactions with objects are highly purposive. From the simplest verb of "put," we can generate a plethora of combinations of objects and target places, such as "put cup onto table," "put fork into drawer." Situations could become even more challenging when objects were used as tools; for example, "put meat into pan using fork."

Motivated by the above observation, we propose the compositional action recognition benchmark on the collected LEMMA dataset with each object attributed to a specific semantic position in the action label. Specifically, we build 24 compositional action templates; see Fig. 5a for some examples. In these action templates, each noun could denote an interacting object, a target or a source location, or a tool used by a human agent to perform certain actions.

The proposed compositional action recognition benchmark is challenging; it requires computational models to correctly detect the ongoing concurrent action verbs as well as the nouns at their correct semantic positions. We evaluate model performances by metrics on compositional action recognition in both FPVs and TPVs. Specifically, the model is asked to predict (i) multiple labels in verb recognition for concurrent actions (*e.g.*, "watch tv" and "drink with cup" at the same time), and (ii) multiple labels in noun recognition for each semantic position given verbs, representing the interactions with multiple objects using the same action (*e.g.*, "wash spoon, cup using sink"). Fig. 5b shows the schematics of the evaluation process. For training and testing on TPVs, we provide ground-truth bounding boxes of humans as additional information on spatial localization.

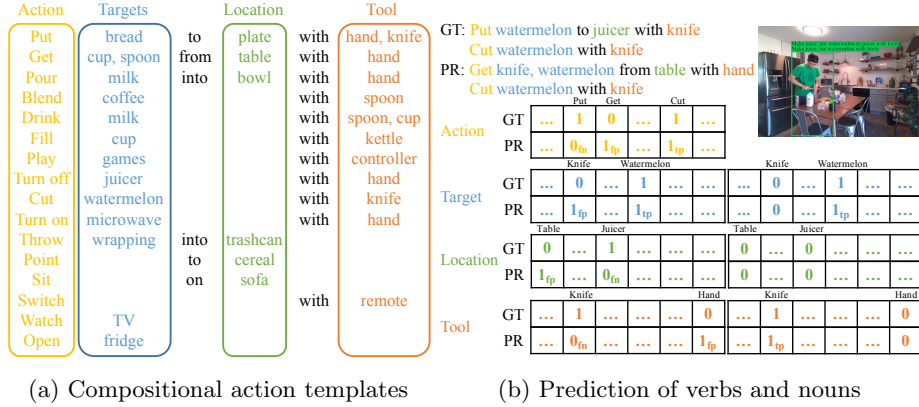(a) Compositional action templates   (b) Prediction of verbs and nouns

Fig. 5: Compositional action recognition benchmark on LEMMA. (a) Examples of Compositional action templates. Yellow denotes verbs. Blue, green, and brown denote nouns for an interacting object, target/source location, and tool, respectively. (b) Examples of predictions of the verbs and nouns in compositional action recognition. Verbs and nouns are evaluated through multi-label classification.

3.2   Action and Task Anticipation

As emphasized throughout the paper, the most significant factor of human activities is the goal-directed, teleological stand. An in-depth understanding of goal-directed tasks demands a predictive ability of latent goals, action preferences, and potential outcomes. To tackle these challenges, we propose the action and task anticipation benchmark on the collected LEMMA dataset. Specifically, we evaluate model performances for the anticipation (*i.e.*, predictions for the next action segment) of action and task with both FPV and TPV videos.

This benchmark provides both the training and testing data in all four scenarios of activities to study the goal-directed multi-task multi-agent problem. As there is an innate discrepancy of prediction difficulties among these four scenarios, we gradually increase the overall prediction difficulty, akin to a curriculum learning process, by setting the percentage of training videos to be $3/4$, $1/4$, $1/4$, and $1/4$ for 1×1, 1×2, 2×1 and 2×2 scenarios, respectively. Intuitively, with sufficient clean demonstrations of tasks in 1×1 scenario, interpreting tasks in more complex settings (*i.e.*, 1×2, 2×1, and 2×2) should be easier, thus requiring less learning samples; such a design encourages the model to generalize. The model performance is evaluated individually for each scenario.

4   Experiments

In this section, we conduct experiments on the two proposed benchmarks with details on evaluation metrics, experimental settings, and baseline results. We further discuss the results to highlight the underlying challenges of each task.

4.1   Compositional Action Recognition

Experimental Setup: We randomly split all the video samples into training and test sets with a ratio of 3:1, resulting in 243 recorded activities for training and the remaining 81 for testing. Due to the multi-agent setup, each activity may have multiple FPVs; 333 (out of 445) FPV videos are split into training. In TPVs, the recordings of the primary view with the ground-truth human bounding box annotations are given for both training and testing videos. Results are evaluated on two separate sources of inputs: FPVs and TPVs.

Evaluation Metrics: Model performances are evaluated separately for verbs, nouns, and compositional action recognition. Verb and compositional action recognition are treated as multi-label classifications with 25 verb classes and 863 compositional action classes (including a "null" action). After generating multi-hot labels for each semantic position in the presented verb, noun recognition is evaluated as multi-label classification (64 object classes). Average precision, recall, and F1-score for all predictions are reported on testing sets. During the evaluation, we sample image frames at 5 FPS and evaluate on these frames.

Methods: We adopt two recent 3D-CNN networks, I3D [9] and SlowFast Network [13], as the baseline models. The baseline models predict the compositional action directly. Considering compositionality of verbs and nouns, we propose two variants of the baseline models: (i) a multi-branch network (branching model) that builds on the bottleneck layer of the backbone models to leverage both verb and noun supervision, and (ii) a multi-step inference model (sequential model), wherein verbs are first inferred with a beam search and then fed into object inference with their verb embeddings for joint learning.

Implementation Details: The training procedure utilizes all annotated segments in the training set. Additionally, we re-scale all the images with the short side to 256 pixels. To feed data into 3D-CNN models, 4 frames are first sampled for each action segment as center frames, and an additional 8 frames are then uniformly sampled around center frames with a window length of 32. We train each model on 8 Titan RTX GPUs on a single computing node for 50 epochs (20k iterations) with a batch size of 96. We use warm-up strategy and perform large mini-batch batch normalization, as suggested in [18]. The learning rate is initially set to 0.0125 for each parallel branch and decays with a cosine annealing. Other settings of the backbone models are the same as in [13]. For the proposed sequential model, we use the beam search with a size of 5 for action inference. We extract bounding box features of humans with ROIAlign [22] for frames in TPVs. More implementation details are provided in *supplementary material*.

Results and Discussion: Table 2 shows quantitative results of predicting verbs, nouns, and compositional actions for the compositional action recognition task. For FPVs, rather than directly predicting the compositional actions (baseline models), predicting the verbs and nouns with their semantic positions

Table 2: Comparisons of compositional action recognition on LEMMA.

View Type	Method	Verb			Noun			Compositional Action		
		Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1
FPV	I3D	17.09	43.89	24.60	3.42	16.15	5.72	11.07	39.49	17.30
	Slowfast	22.27	56.42	31.94	4.31	20.60	7.13	18.68	**50.65**	27.3
	I3D sequential	25.04	**57.00**	34.80	**19.36**	**75.29**	**30.80**	18.00	50.04	26.47
	Slowfast sequential	24.30	49.71	32.64	17.95	59.11	27.54	26.80	38.41	31.57
	I3D branching	25.73	55.62	**35.8**	18.63	69.76	29.41	22.29	48.46	30.53
	Slowfast branching	**26.16**	56.33	35.73	18.18	73.46	29.15	**27.97**	48.87	**35.58**
TPV	I3D	14.18	36.34	20.40	2.29	11.05	3.79	6.85	**23.82**	10.64
	Slowfast	14.28	**37.38**	20.66	2.32	11.14	3.83	**7.76**	23.25	**16.31**
	I3D sequential	16.17	30.17	21.05	7.79	**25.41**	11.93	2.23	12.67	3.79
	Slowfast sequential	15.31	28.84	20.00	6.37	22.39	9.92	3.27	9.16	4.82
	I3D branching	12.92	32.09	18.43	12.75	17.70	14.82	4.67	20.76	7.6
	Slowfast branching	**16.64**	33.40	**22.21**	**17.29**	18.36	**17.81**	6.52	21.55	10.01

boosts the performance on all metrics, indicating that understanding the compositional structures of human actions indeed supports the prediction. We also observe that the results of compositional action recognition in the sequential models are slightly lower than the branching model due to the aggregated error brought in by a relatively low precision ($\sim 25\%$) of the verb recognition.

In comparison, the results of compositional action recognition in TPVs are significantly lower than those in the FPVs due to severe occlusion. It also shows that predicting the composition of verbs and nouns makes no significant improvement compared with predicting compositional action directly. Such a result implies that current models could not capture the details of compositions between verbs and nouns from TPVs. Taken together, the results indicate that fusion among the representations of visual embodiment between TPVs and FPVs might be a crucial ingredient to tackle this problem in the future.

Fig. 6 shows qualitative results for the composed action recognition task.

4.2   Action and Task Anticipations

Experimental Setup: We split the training and test sets with ratios $3 : 1$, $1 : 3$, $1 : 3$, $1 : 3$ for the four scenarios 1×1, 1×2, 2×1, 2×2, respectively. Such a spit results in training set with (96, 19, 16, 13) activities and a test set with (31, 57, 50, 42) activities in four scenarios. During training and testing, the computational models have access to both FPVs and TPVs, together with the ground-truth human bounding boxes annotations of the TPV primary view.

Evaluation Metrics: Model performances are evaluated individually (per agent) for the action and task anticipations task. Specifically, both action and task anticipations are evaluated as multi-label classifications with 863 compositional action classes (including a "null" action) and 15 task classes. Average precision, recall, and F1-score are reported individually for each of the four scenarios on the testing sets. Similar to the protocol used in the above compositional action recognition task, we re-sample image frames at 5 FPS and evaluate these sub-sampled frames during the testing phase.
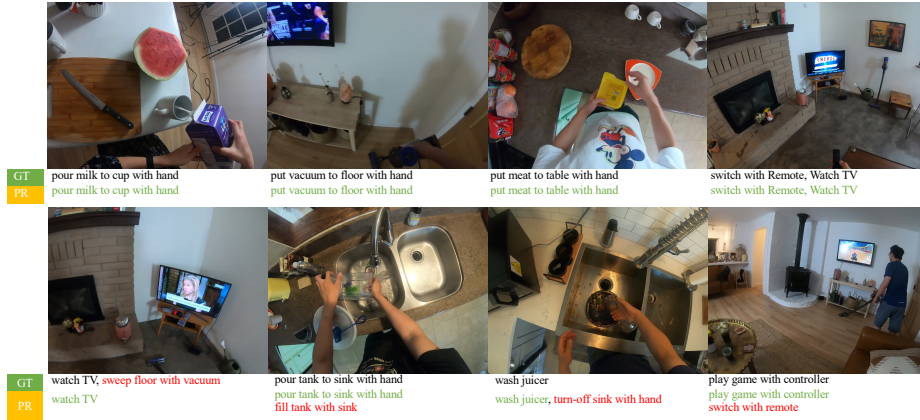
Fig. 6: Qualitative results of compositional action recognition on LEMMA. From top to bottom, we show correct predictions and failure examples. Red marks wrong verb or noun predictions, green indicates correct verb or noun predictions.

Methods: We leverage the visual features extracted by the pre-trained SlowFast model in compositional action recognition for baseline models. Specifically, we compare two backbone models: (i) using segment-level recognition feature (SF) directly by adding an MLP on top of the features, and (ii) using long-term feature bank (LFB) with max pooling [58]. For activities with multi-agent interactions, we use the other agent's FPV features together with their own's to capture the joint task execution progress for learning and inference; these variants are denoted as M-SF (FPV) and M-LFB (FPV) For comparison, we also use the concatenation of the FPV feature and primary TPV feature as the input; the corresponding models are denoted as M-SF (TPV) and M-LFB (TPV).

Implementation Details: For the LFB model, we use a history window size of 10 and aggregate the features using max-pooling, as described in [58]. For the multi-agent variants, we use max-pooling to fuse features of two views and process them with a different branch as another temporal inference module. We train models on a single Titan Xp GPU for 50 epochs with a learning rate of 0.001. See *supplementary material* for more details on network architectures.

Results and Discussion: Table 3 shows quantitative results of action and task anticipation. The proposed multi-agent variants (M-) of baseline models perform the best among all models. For single-agent activities (1×1, 1×2), we have the following crucial observations. First, models that consider temporal relations between frames generally perform better than the models using segment features. Second, adding additional TPV features to single-agent activities slightly helps interpret the task being executed and therefore promotes anticipation. This result matches the intuition that computational models having access to both FPVs and TPVs would perceive more holistic scene information. We also

Table 3: Comparisons of the action and task anticipations on LEMMA.

Scenario	Method	1 × 1			1 × 2			2 × 1			2 × 2		
		Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1	Avg.Prec	Avg.Rec	Avg.F1
Compositional action	SF	23.42	22.25	22.82	20.13	20.06	20.10	18.89	19.22	19.05	18.31	16.67	17.45
	LFB	23.03	28.67	25.54	20.48	25.4	22.67	18.31	22.30	20.11	18.53	20.97	19.68
	M-SF (TPV)	**24.22**	28.05	25.99	20.10	24.48	22.08	19.15	16.71	17.85	19.64	15.18	17.12
	M-LFB (TPV)	23.54	**37.81**	**29.01**	21.10	**31.86**	25.39	19.67	21.03	20.33	**20.11**	20.30	**20.15**
	M-SF (FPV)	23.30	25.41	24.31	**21.34**	23.18	22.22	**19.70**	17.46	18.51	19.82	15.8	17.58
	M-LFB (FPV)	23.26	31.07	26.60	20.78	27.40	23.63	19.42	**21.73**	**20.51**	19.49	20.12	19.8
Task	SF	50.53	79.08	61.66	48.07	67.78	56.25	39.05	57.43	46.49	44.88	62.09	52.1
	LFB	57.57	**84.31**	68.42	52.12	68.94	59.36	38.40	53.08	44.56	48.17	64.61	55.19
	M-SF (TPV)	58.61	79.96	67.05	55.45	67.24	60.78	**45.73**	58.98	**51.51**	49.66	64.47	56.10
	M-LFB (TPV)	**60.27**	82.19	**69.54**	**56.2**	**72.46**	**63.30**	43.94	61.41	51.23	48.85	**67.48**	**56.67**
	M-SF (FPV)	51.12	79.18	62.13	48.42	69.04	56.92	41.00	58.11	48.08	46.04	65.97	54.24
	M-LFB (FPV)	55.56	82.83	66.51	52.22	70.01	59.82	41.33	**64.49**	50.38	46.65	69.59	55.86

find that the performances of task anticipation in the 1×1 single-task scenario are better than the one in the 1×2 multi-task scenario, matching what we would expect from more complicated task execution patterns.

For multi-agent activities (2×1, 2×2), we observe that the aggregation of FPV and TPV features generally performs better. It supports our hypothesis that observing the other agents' actions helps the computational models to "understand" task scheduling and assignment. We also observe that, models' performances in 2×1 activities are slightly worse than in 2×2 activities. We hypothesize that task plans in the 2×2 scenarios change less frequently, with a clear task assignment coordinates the individual tasks. In comparison, in the 2×1 scenarios, the sequential ordering of the task requires more frequent communications between agents to coordinate. Such a performance gap calls for better modeling of multi-agent task assignments. Due to the page limit, we show qualitative results of action and task anticipation in the *supplementary material*.

5   Conclusions

In this paper, we introduce the LEMMA dataset with a focus on natural multi-agent multi-task daily activities. Dense annotations are provided on both compositional action and task for learning and inference on four different activity scenarios with increasing difficulty. Additionally, we propose two challenging tasks on LEMMA to measure existing models' competence in action understanding and temporal reasoning: (i) compositional action recognition, and (ii) action/task anticipations. We hope this effort would attract the computer vision community to look into natural and realistic goal-directed human activities and further study the task scheduling and assignment in real-world scenarios.

References

1. Alayrac, J.B., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., Lacoste-Julien, S.: Unsupervised learning from narrated instruction videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: Proceedings of International Conference on Computer Vision (ICCV) (2015)
3. Baker, B., Kanitscheider, I., Markov, T., Wu, Y., Powell, G., McGrew, B., Mordatch, I.: Emergent tool use from multi-agent autocurricula. In: International Conference on Learning Representations (ICLR) (2020)
4. Baldwin, D.A., Baird, J.A.: Discerning intentions in dynamic human action. Trends in Cognitive Sciences **5**(4), 171–178 (2001)
5. Baldwin, D.A., Baird, J.A., Saylor, M.M., Clark, M.A.: Infants parse dynamic action. Child development **72**(3), 708–717 (2001)
6. Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., Farhi, D., Fischer, Q., Hashme, S., Hesse, C., et al.: Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680 (2019)
7. Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Weakly supervised action labeling in videos under ordering constraints. In: Proceedings of European Conference on Computer Vision (ECCV) (2014)
8. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
10. Choi, W., Shahid, K., Savarese, S.: What are they doing?: Collective activity classification using spatio-temporal relationship among people. In: International Conference on Computer Vision Workshops (ICCV Workshops) (2009)
11. Csibra, G., Gergely, G.: 'obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. Acta psychologica (2007)
12. 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)
13. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
14. Fouhey, D.F., Kuo, W.c., Efros, A.A., Malik, J.: From lifestyle vlogs to everyday interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
15. Gergely, G., Bekkering, H., Király, I.: Rational imitation in preverbal infants. Nature **415**(6873), 755–755 (2002)
16. Girdhar, R., Ramanan, D.: Cater: A diagnostic dataset for compositional actions and temporal reasoning. International Conference on Learning Representations (ICLR) (2020)
17. Goodman, N.D., Frank, M.C.: Pragmatic language interpretation as probabilistic inference. Trends in cognitive sciences **20**(11), 818–829 (2016)

18. Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., He, K.: Accurate, large minibatch sgd: Training imagenet in 1 hour. arXiv preprint arXiv:1706.02677 (2017)
19. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The "something something" video database for learning and evaluating visual common sense. In: Proceedings of International Conference on Computer Vision (ICCV) (2017)
20. Grice, H.P.: Logic and conversation. In: Speech acts, pp. 41–58. Brill (1975)
21. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., 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 (CVPR) (2018)
22. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of International Conference on Computer Vision (ICCV) (2017)
23. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
24. Ibrahim, M.S., Muralidharan, S., Deng, Z., Vahdat, A., Mori, G.: A hierarchical deep temporal model for group activity recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
25. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **36**(7), 1325–1339 (2013)
26. Johansson, G.: Visual perception of biological motion and a model for its analysis. Perception & psychophysics **14**(2), 201–211 (1973)
27. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
28. Kleiman-Weiner, M., Ho, M.K., Austerweil, J.L., Littman, M.L., Tenenbaum, J.B.: Coordinate to cooperate or compete: abstract goals and joint intentions in social interaction. In: Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci) (2016)
29. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. International Journal of Robotics Research (IJRR) **32**(8), 951–970 (2013)
30. 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)
31. Land, M., Mennie, N., Rusted, J.: The roles of vision and eye movements in the control of activities of daily living. Perception **28**(11), 1311–1328 (1999)
32. Lerner, A., Chrysanthou, Y., Lischinski, D.: Crowds by example. In: Proceedings of Computer Graphics Forum (2007)
33. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
34. Li, Y., Liu, M., Rehg, J.M.: In the eye of beholder: Joint learning of gaze and actions in first person video. In: Proceedings of European Conference on Computer Vision (ECCV) (2018)

35. Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S.A., Yan, T., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., et al.: Moments in time dataset: one million videos for event understanding. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2019)
36. Monsell, S.: Task switching. Trends in cognitive sciences **7**(3), 134–140 (2003)
37. Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.C., Lee, J.T., Mukherjee, S., Aggarwal, J., Lee, H., Davis, L., et al.: A large-scale benchmark dataset for event recognition in surveillance video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
38. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: Proceedings of International Conference on Computer Vision (ICCV) (2009)
39. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
40. Ramanathan, V., Huang, J., Abu-El-Haija, S., Gorban, A., Murphy, K., Fei-Fei, L.: Detecting events and key actors in multi-person videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
41. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2015)
42. 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)
43. Rohrbach, M., Rohrbach, A., Regneri, M., Amin, S., Andriluka, M., Pinkal, M., Schiele, B.: Recognizing fine-grained and composite activities using hand-centric features and script data. International Journal of Computer Vision (IJCV) **119**(3), 346–373 (2016)
44. Rubinstein, J.S., Meyer, D.E., Evans, J.E.: Executive control of cognitive processes in task switching. Journal of experimental psychology: human perception and performance **27**(4), 763 (2001)
45. 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)
46. Shu, T., Xie, D., Rothrock, B., Todorovic, S., Chun Zhu, S.: Joint inference of groups, events and human roles in aerial videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
47. Sigurdsson, G.A., Gupta, A., Schmid, C., Farhadi, A., Alahari, K.: Charades-ego: A large-scale dataset of paired third and first person videos. arXiv preprint arXiv:1804.09626 (2018)
48. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: Proceedings of European Conference on Computer Vision (ECCV) (2016)
49. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
50. Stein, S., McKenna, S.J.: Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (2013)
51. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: Coin: A large-scale dataset for comprehensive instructional video analysis. In: Pro-

ceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

52. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

53. Toyer, S., Cherian, A., Han, T., Gould, S.: Human pose forecasting via deep markov models. In: International Conference on Digital Image Computing: Techniques and Applications (DICTA) (2017)

54. Turaga, P., Chellappa, R., Subrahmanian, V.S., Udrea, O.: Machine recognition of human activities: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **18**(11), 1473–1488 (2008)

55. Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., et al.: Grandmaster level in starcraft ii using multi-agent reinforcement learning. Nature **575**(7782), 350–354 (2019)

56. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision (IJCV) **101**(1), 184–204 (2013)

57. Woodward, A.L.: Infants selectively encode the goal object of an actor's reach. Cognition **69**(1), 1–34 (1998)

58. Wu, C.Y., Feichtenhofer, C., Fan, H., He, K., Krahenbuhl, P., Girshick, R.: Long-term feature banks for detailed video understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

59. 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)

60. Yeung, S., Russakovsky, O., Jin, N., Andriluka, M., Mori, G., Fei-Fei, L.: Every moment counts: Dense detailed labeling of actions in complex videos. International Journal of Computer Vision (IJCV) **126**(2-4), 375–389 (2018)

61. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: Proceedings of AAAI Conference on Artificial Intelligence (AAAI) (2018)

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

Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu

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 [42,30,12] and living room [59,45]. Specifically, we start from the action verb vocabulary summarized in the EPIC-KITCHENS dataset [12] to cover common actions conducted in the kitchen and living room. We then 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

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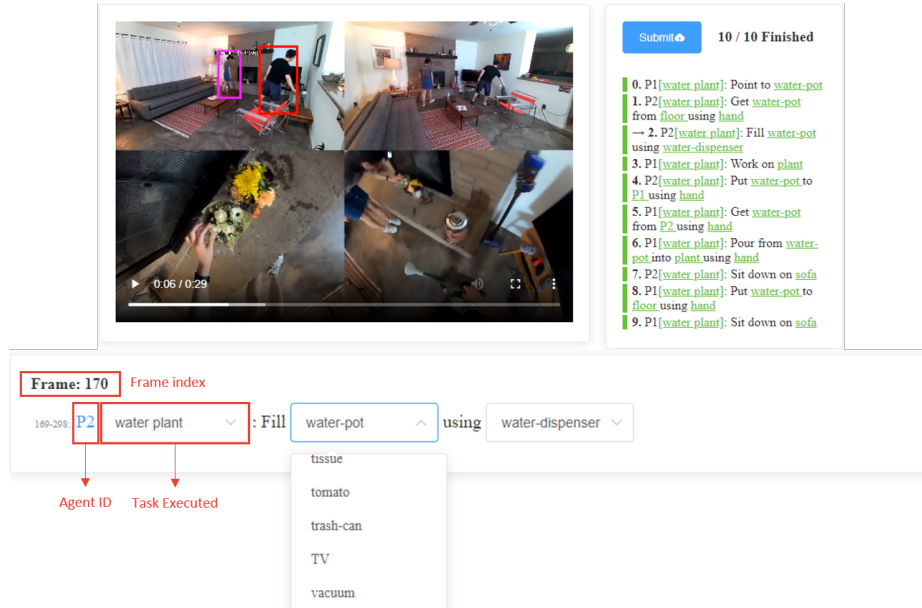
Fig. 1: A visualization of the annotation tool developed for annotating the nouns and governing tasks.

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 [56] 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 $f_{\mathrm{verb}} = \{F_{\mathrm{verb}}^{(i)}(f_{\mathrm{vis}})\}$ are generated using three different linear projections $\{F_{\mathrm{verb}}^{(i)}\}_{i=1,2,3}$ 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 $\mathcal{L}_{\mathrm{verb}}$ as the supervision for verb recognition.

For the noun branch, we utilize the embeddings of each verb as additional features by GloVe [**?**]. The embedding of each verb is passed into a linear projection layer and concatenated with the extracted visual features to generate
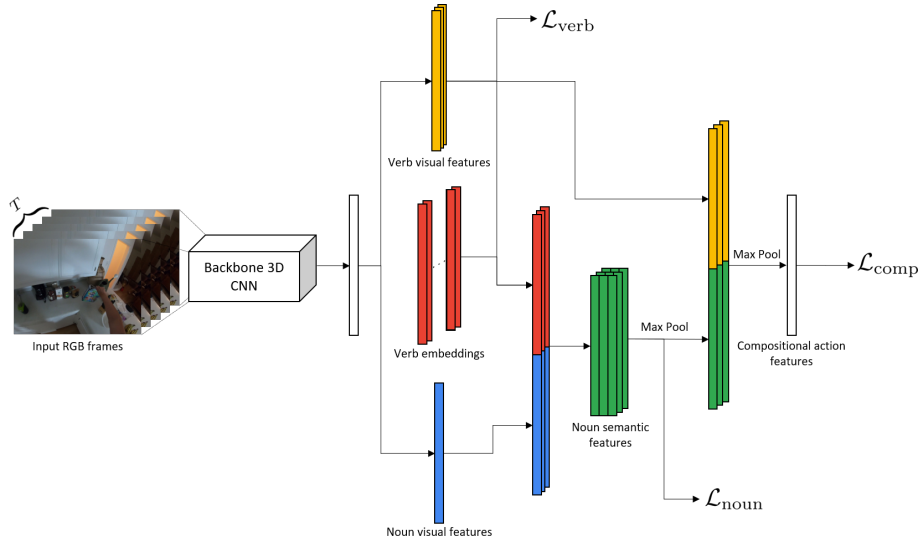

Fig. 2: An illustration of the proposed "sequential" model, which predicts verbs, nouns, and compositional actions jointly.

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.

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
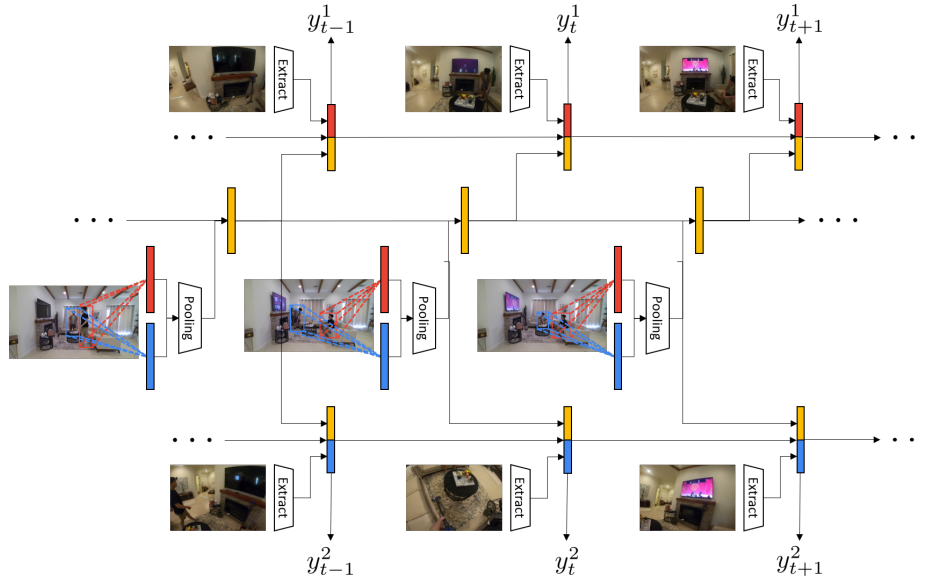

Fig. 3: An illustration for the multi-agent variants of the original sequential model with TPV features as additional features.

as Ego in Table 3) or TPV features (denoted as TPV in Table 3) through a pooling mechanism, similar to [?]. 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.

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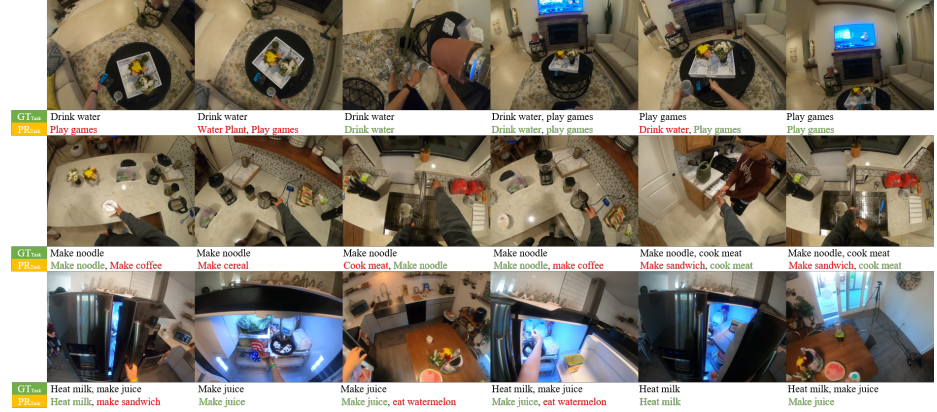
(b) Results of task anticipation. GT_{Task} indicates the ground-truth task for actions in the next segment.

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

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