2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Understanding Contexts Inside Robot and Human Manipulation Tasks through Vision-Language Model and Ontology System in Video Streams

Chen Jiang†, Masood Dehghan† and Martin Jagersand†

***Abstract*— Manipulation tasks in daily life, such as pouring water, unfold through human intentions. Being able to process contextual knowledge from these Activities of Daily Living (ADLs) over time can help us understand manipulation intentions, which are essential for an intelligent robot to transition smoothly between various manipulation actions. In this paper, to model the intended concepts of manipulation, we present a vision dataset under a strictly constrained knowledge domain for both robot and human manipulations, where manipulation concepts and relations are stored by an ontology system in a taxonomic manner. Furthermore, we propose a scheme to generate a combination of visual attentions and an evolving knowledge graph filled with commonsense knowledge. Our scheme works with real-world camera streams and fuses an attention-based Vision-Language model with the ontology system. The experimental results demonstrate that the proposed scheme can successfully represent the evolution of an intended object manipulation procedure for both robots and humans. The proposed scheme allows the robot to mimic human-like intentional behaviors by watching real-time videos. We aim to develop this scheme further for real-world robot intelligence in Human-Robot Interaction.**

I. Introduction

Recent advances in fusing computer vision with linguistic knowledge enables researchers to model human-like commonsense knowledge using semantic contexts for intelligent robots. Studies both in robot vision and Natural Language Processing (NLP) [1]–[5] provide promising tools for robots to better understand human tasks and assist humans in their daily life. Still, intelligent robots are far from perfect. Intelligent robots face challenges in: (1) interpreting sensor inputs of vision and force contact interactions through modeling and learning from daily life knowledge; and (2) performing intelligent actions that take into account the surrounding physical environment as well as human actions and intentions. However, it is challenging to extract contextual knowledge directly from daily life in ways similar to human thinking.

Understanding contexts semantically is important for robotics because humans express intention during the process of action execution. For example, in a pouring manipulation task, the context involves a sequence of actions executed over time, including, grasping an object that contains liquid, pouring the liquid into an empty container, and releasing the currently held pouring object. From studies in action recognition [6]–[8], we know that the execution of manipulation tasks requires hand-eye coordination, and humans intentionally focus their visual attention onto relevant regions of objects. Commonsense knowledge in a pouring task can be summarized by humans as constraint-like rules (e.g. "grasping an object full of liquid is needed before pouring", "destination of pouring should usually be an empty container that I can grasp safely", etc.). There is no doubt that attention and commonsense knowledge on-scene can serve as strong prior knowledge to logically deduce human intention. To plan for future manipulation actions, intelligent robots will need to infer from contextual knowledge in real-time. Therefore, it is essential to develop techniques to semantically process visual information and interpret manipulation actions on-scene for both robot and human.

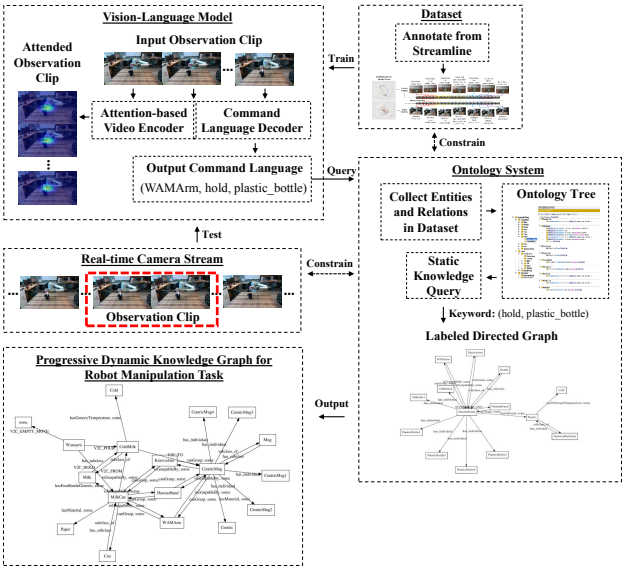

Fig. 1. Overview of our scheme. Under our ontology system for commonsense knowledge, our proposed scheme is able to utilize a Vision-Language model and an ontology system to extract spatial attentions consecutively over a camera stream, and interpret a manipulation task into an evolving dynamic knowledge graph filled with details of manipulation actions, objects, and their relations.

In this paper, we propose a scheme[1] to explore the fundamental problem of perceiving and interpreting on-scene dynamic knowledge and commonsense knowledge over time for both robots and humans using an attention-based Vision-Language model and an ontology system. The purpose is to represent the evolution of the intended manipulation procedure, and generate visual attention tracks and dynamic

†Authors are with Department of Computing Science, University of Alberta, Edmonton AB., Canada, T6G 2E8. {cjiang2, masood1, mj7}@ualberta.ca

[1]Our code and collected dataset are publicly available at: https://github.com/zonetrooper32/robot_semantics.

knowledge graphs that can be used as inputs for robotic manipulation decisions. Figure 1 shows the logic flow of this scheme. Our contributions can be summarized as follows:

- We present a scheme to capture manipulation concepts into a time-independent knowledge domain. An ontology system is constructed to store objects and relations in a taxonomic structure, which serves as our commonsense knowledge over a particular domain of manipulation tasks.
- We collect a benchmark dataset of RGB-D videos from manipulation tasks performed by both robots and humans. Under our domain knowledge constraints, we manually annotate ground truth object and action information to the video frames.
- We present a Vision-Language model based on the popular sequence-to-sequence structure [9] with spatial attention mechanism to caption manipulation knowledge for video streams. The Vision-Language model is able to implicitly learn spatial attention on the salient regions corresponding to the manipulation actions and activities at hand.
- We combine the Vision-Language model with the ontology system, allowing the model to semantically interpret the evolution of the manipulation task into a linguistic dynamic knowledge graph filled with commonsense knowledge.

II. Related Work

A. Manipulation Knowledge in Robotics

Various methods have studied the effect of introducing contextual knowledge for robotic behaviors. Multiple studies [10]–[14] have discussed generic schemes to represent collective commonsense knowledge on manipulation objects and relations for scene understanding and commonsense reasoning. The evolution of manipulation tasks is another wildly studied aspect that directly utilizes manipulation knowledge in robotics. Task evolution can be represented by structures such as semantic trees [15]–[17] or behavior trees [18], [19]. However, most of those evolution representations rely heavily on human annotations, and few of the aforementioned studies have discussed how to automatically acquire evolution representations on-scene. In our work, initiated from human knowledge in manipulation contexts, we utilize a Vision-Language model to automatically interpret task evolution and robotic actions semantically.

B. Vision and Language in Robotics

Originating from action recognition and video captioning, there have been a number of studies on introducing language in combination with vision to learn semantic actions for robotic uses. Nguyen et al. [3], [5] proposed to caption human actions into command sentences, which can be used to control robotic actions. Similar works can be observed to improve the capabilities of Vision-Language models under robotic settings for problems like Human-Robot Interaction [20]–[22], action learning and planning [1], [2], [4], [23], [24], etc. However, the evaluation of these methods usually involves: (1) sampling of a small fixed number of frames, which is not suitable when intermediate feedback is continuously requested in a real-time video stream; or (2) heavy reliance on object detection, which is only weakly associated to manipulation contexts. We highlight these points in our scheme with a Vision-Language model under a more realistic semantic context using video streams.

III. Robot Semantics

In this section, we introduce our main framework to semantically model contextual knowledge in both robot and human manipulation tasks. We first propose a general scheme to model the timing constraints typically found in robot vision, then present our scheme to collect, construct and model contextual knowledge for manipulation tasks in general.

A. Sampling from Streams

1) Stream, Video and Dataset: A camera stream CS_{T_0}, from start time T_0 to (potentially) infinity, observes a scene of a human or robot performing a sequence of actions and produces image frames I_{T_k}. We define a video of length $j-i+1$ in the form $V_{T_i...T_j} = \{I_{T_i}, I_{T_{i+1}}, ..., I_{T_j}\}$, initiating from start time T_i to end time T_j. Each video shows a full demonstration of a task. A dataset $D = \{V_1, V_2, ..., V_N\}$ is defined as a collection of N videos from multiple camera streams of different time periods with annotations.

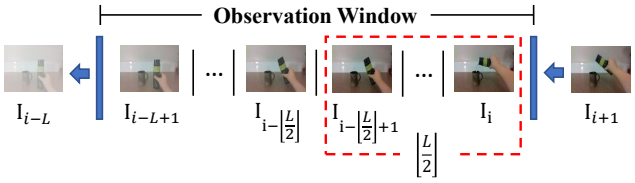

Fig. 2. Representing camera vision as a stream of data. A queue of observations is maintained and stream signals for inference when $\lfloor\frac{L}{2}\rfloor$ of the observation is queued with new information.

2) Observation: We assume weak to no long-term time dependencies, i.e. an observation is "most dependent and trust-able" within a time window. We further define this intermediate observation as a small clip of frames $C_{t_k...t_{k+L}} = \{I_{t_k}, I_{t_{k+1}}, ..., I_{t_{k+L}}\}$, initiating from the start time t_k and persisting for a time length L. Figure 2 presents a simple scheme to sample a series of overlapping clip observations from a camera stream. The goal is to always maintain a queue of maximum L length serving as our observation window, while streams of image frames continuously arrive into the observation window. A clip is collected when: (1) the observation window is filled for the first time; or (2) $\lfloor\frac{L}{2}\rfloor$ of the observation window is flushed with newer images. Hence, the camera stream CS_{T_0} is treated as an overlapping sequential composition of several L length clips.

B. Robot Semantics Dataset

While many datasets exist for manipulation tasks and human intentions [3], [25]–[27], few span both robot and human manipulation tasks. We collect 720p videos over a set of

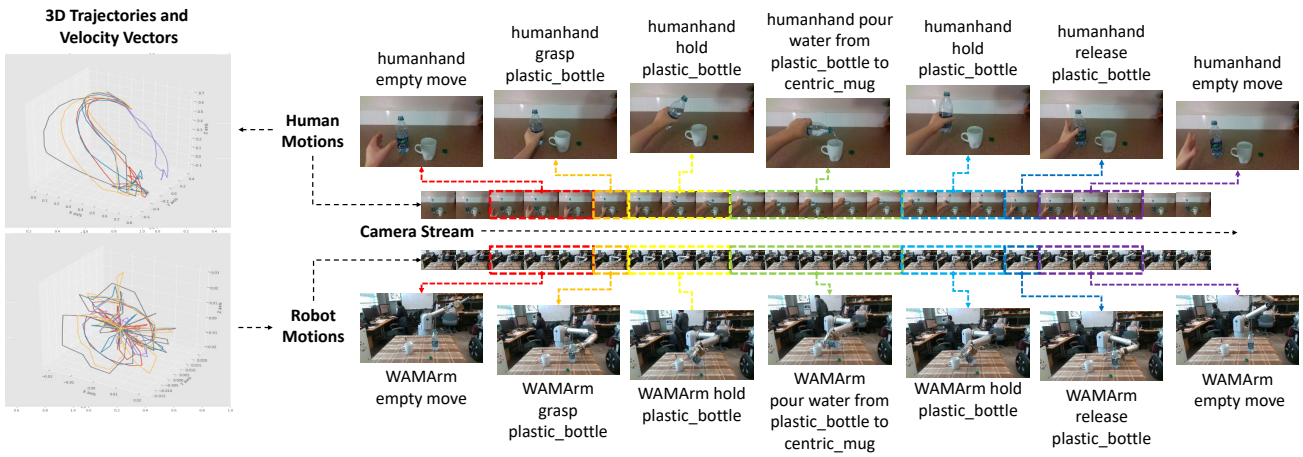

Fig. 3. Samples from our Robot Semantics dataset. Initiated from human motions, manipulation videos are collected for both human and WAM robot arms, and annotated. Left: Ten of the 3D trajectories and the velocity vectors from the robot motions.

manipulation tasks using an Intel RealSense D435i Camera. Figure 3 presents an overview over our benchmark data with frame-by-frame command language annotations and robot way-point trajectories. Our Robot Semantics Dataset consists of RGB-D videos, where each video demonstrates a complete particular manipulation task such as "pouring water to the cup". A manipulator (robot or human) executes a sequence of motions with some container objects to complete the full manipulation task. There are two types of manipulators included: human subjects and a Barrett Whole Arm Manipulator (WAM) robot.

1) Human: The camera was setup with an egocentric view in a kitchen environment. Human subjects were asked to use their left or right hand to perform a series of actions for a complete manipulation task. For manipulation tasks performed by a human, there are 94 videos - 42,681 images in total.

2) WAM: A Barrett WAM robot was used to perform the same set of manipulation tasks as the human subjects. We used the experimental protocol originating from IVOS benchmark [28]; a human operator guided the WAM to reach the target and perform the intended manipulation actions. Robotic way-point trajectories, in the form of quaternions over the 7 joints poses, were recorded during the kinesthetic teaching. The WAM robot then executed the manipulation actions by following the kinesthetic teaching. For the WAM robot arm, there are 46 videos - 69,368 images and 43 recorded trajectories in total.

C. Domain Knowledge Ontology

An ontology is a well known way to store machine-interpretable definitions of concepts in a static knowledge domain. Here, we construct an ontology to store and query for the set of explicitly defined commonsense knowledge over the concepts of manipulation. Given a set of linguistic entities $E = \{e_1, e_2, ..., e_n\}$, including the objects and manipulators presented in our dataset, for any two entities $e_i, e_j \in E$, we impose a set of binary logical constraints LC in a taxonomic and relational structure using a linguistic relation $a_k \in A$:

$$e_i \xrightarrow{a_k,\, r} e_j \in LC \tag{1}$$

where $A = \{a_1, a_2, ..., a_m\}$ is the set of relations, r represents restrictions (Quantifier, Cardinality, and hasValue). $E + A$ represents the complete linguistic vocabulary over the entire manipulation domain knowledge. In general, relations can originate from: (1) hierarchical definitions, for example, *"*$\forall_{(GlassCup,Cup)}$ *isA(GlassCup, Cup)"*, *"*$\forall_{(PlasticBottle,HardBottle)}$ *disjoint(PlasticBottle, HardBottle)"*; (2) action relation between any two entities, for example, *"*$\forall_{WAM}$ *canPour some ColdMilk"* and *"*$\forall_{PlasticBottle}$ *isGraspableBy some HumanHand"*; and (3) attributes or properties of any entity, for example, *"*$\forall_{ColdMilk}$ *hasTemperature some Cold"*, *"*$\forall_{GlassCup}$ *canHold some HotWater"*, etc. Entities are stored as classes with individuals, while relations are stored as binary object properties with restrictions.

D. Defining Dynamic Knowledge

To interpret on-scene manipulation knowledge over time for visual data, we use a dynamic knowledge graph with time attributes. A dynamic knowledge graph $G_{T_i...T_j} = (N, E)$ is a Labeled Directed Graph where edges inside are composed of the binary logical constraints in LC. Any edge $e \in E$ in a knowledge graph spans LC and any node $n \in N$ spans $E + A$. Nodes can be spatially connected. Additionally, a timing constraint is applied in correspondence to the visual data. For any visual data within a time period $T_i...T_j$, including a clip, a video and a stream, a dynamic knowledge graph $G_{T_i...T_j}$ describes the set of relations that are presented during this time period.

For a compact inference representation, we define a command language, which can be seen as a basic skeleton form of the dynamic knowledge graph. A command language $S_{t_i...t_j} \subseteq G_{T_i...T_j}$ describes the most important relations

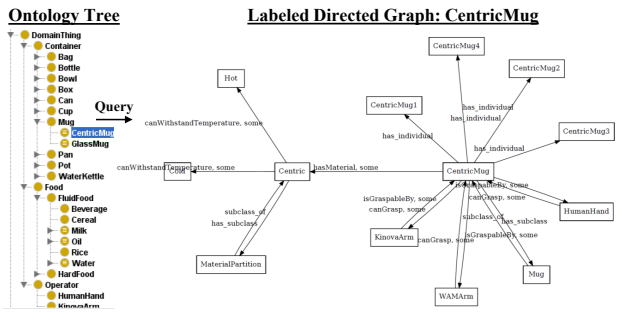

Fig. 4. From ontology system to object taxonomy. Given a keyword concept, for example "CentricMug", we can query from the ontology tree and represent the associated concepts into a Labeled Directed Graph.

composed by the manipulation actions at hand. A command language $S_{t_i...t_j}$ is represented as:

$$e_1 \xrightarrow{a_1} e_2 \xrightarrow{a_2} ... \xrightarrow{a_n} e_n \tag{2}$$

where $e_1, e_2, ..., e_n \in E$ and $a_1, a_2, ..., a_n \in A$, $t_i...t_j \subseteq T_i...T_j$, and S has a length of L. Edges inside $S_{t_i...t_j}$ are sequentially composed and restrictions can be neglected for command languages. The timing constraint also applies to the command language. For any entity e_i perceived inside a command language, the commonsense knowledge over the entity can be queried from an ontology system into an Labeled Directed Graph G_{e_i}, see Figure 4. A dynamic knowledge graph $G_{T_i...T_j}$ can be seen as the union over the command language and all queried Labeled Directed Graphs:

$$G_{T_i...T_j} = S_{t_i...t_j} \cup G_{e_1} \cup ... \cup G_{e_n} \tag{3}$$

E. Combining Vision and Language

Given an observation as a unit of clip $C_{t_1...t_L} = \{I_{t_1}, I_{t_2}, ..., I_{t-1}, I_t, ..., I_{t_L}\}$, our goal is to caption a command language and acquire the related visual attentions over this time period $t_1...t_L$. We propose to train an end-to-end attention-based sequence-to-sequence (seq2seq) model to infer for a command language at any time period of the manipulation task. Figure 5 shows the detailed architecture for our neural command parser using seq2seq structure with spatial attention and output command language.

1) Spatial Attention: Originally proposed in Xu et al. [29] for image captioning tasks, the implicitly learned attention adaptively attends to relevant salient regions inside a clip of N frames by the semantic labels assigned. The context vector z_t at timestamp t is a dynamic representation of the relevant salient part of the image feature a_{ti} of size $(L, H \times W)$. A positive scalar weight α_{ti} is generated, interpreted as the relative importance to give to a location i:

$$\begin{aligned} e_{ti} &= f_{att}(a_{ti}, h_{t-1}) \\ \alpha_{ti} &= \frac{\exp(e_{ti})}{\sum_{j=1}^{H \times W} e_{tj}} \end{aligned} \tag{4}$$

where f_{att} is a mechanism that determines the amount of attention allocated to different regions of the image feature, conditioned on the previous hidden state h_{t-1} of the encoding LSTM. a_t can be extracted by any generic CNN network and the Bahdanau attention mechanism [30] is used here. The attended visual feature is calculated simply as a weighted sum:

$$z_t = \sum_{i=1}^{H \times W} \alpha_{ti} a_{ti} \tag{5}$$

2) Command Language Generation with seq2seq: Long Short-Term Memory network is a type of Recurrent Neural Network that learns long-term dependencies from the input data. Given the attended visual feature input z_t, the hidden state h_t and the memory cell state c_t at the next timestamp t are computed as:

$$\begin{aligned} i_t &= \sigma(W_{ii} z_t + b_{ii} + W_{hi} h_{t-1} + b_{hi}) \\ f_t &= \sigma(W_{if} z_t + b_{if} + W_{hf} h_{t-1} + b_{hf}) \\ g_t &= \tanh(W_{ig} z_t + b_{ig} + W_{hg} h_{t-1} + b_{hg}) \\ o_t &= \sigma(W_{io} z_t + b_{io} + W_{ho} h_{t-1} + b_{ho}) \\ c_t &= f_t * c_{t-1} + i_t * g_t \\ h_t &= o_t * \tanh(c_t) \end{aligned} \tag{6}$$

where σ is the sigmoid function, i_t, f_t and o_t represent the input state, forget state and output state over the current timestamp t.

The seq2seq model [9] is an encoder-decoder architecture where an encoding vector representation v is learned by an encoding LSTM, and a decoding LSTM learns to generate the command sentence sequence $s_1..., s_K$ conditioned on the encoding vector:

$$p(s_1..., s_K | z_{t_1}, ..., z_{t_L}) = \sum_{k=1}^{K} p(s_k | v, s_1, ..., s_{k-1}) \tag{7}$$

where $v = (h_{t_L}, c_{t_L})$ is the last hidden state and the memory cell state of the encoding LSTM, $Z = (z_{t_1}, ..., z_{t_L})$ is the sequence of attended visual features and $S = (s_1, ..., s_K)$ is the corresponding output command sequence with a maximum length of K. $p(s_k | v, s_1, ..., s_{k-1})$ is represented with a softmax over all the tokens in the command vocabulary. The command language decoder takes the concatenation of the current command embedding feature and the last encoding hidden state as $[s_{k-1}, h_{t_L}]$ and generates the next probable command token s_k. The seq2seq structure is optimized by maximizing the log likelihood objective:

$$\underset{\theta}{\operatorname{argmax}} \sum_{(Z,S)} \log p(S | Z; \theta) \tag{8}$$

where θ is the model parameters.

IV. Experiments

A. Implementation Details

1) Vision-Language Model: The implementation for Vision-Language models are done using PyTorch. For fair comparisons, all visual features are extracted by ResNet50 pretrained on ImageNet without finetuning. The weights for

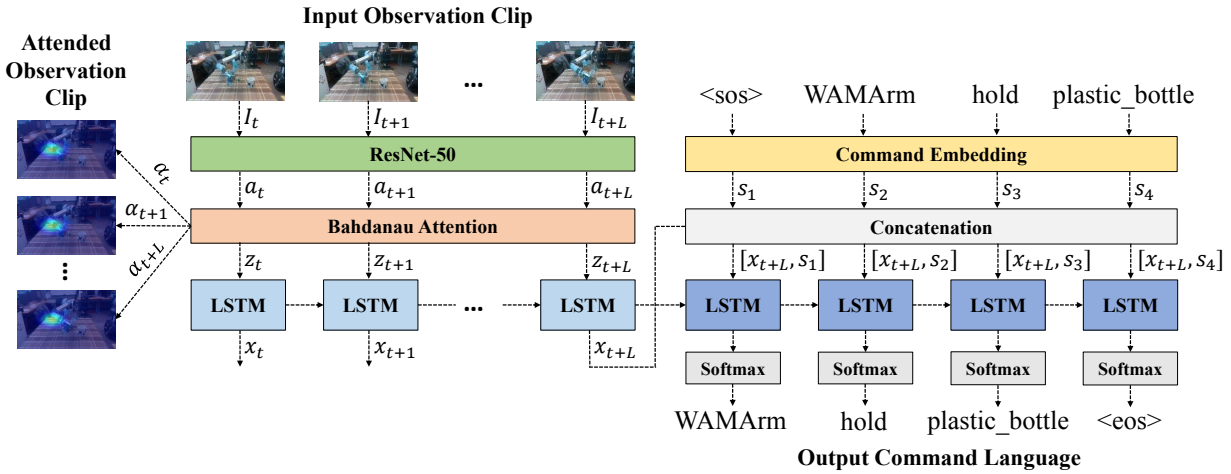

Fig. 5. Architecture to model command language using seq2seq with attention. The model first takes an observation clip of images as inputs, then outputs the command language and the related spatial attentions associated with the inputted clip.

LSTMs, attention mechanism and command word embedding are randomly initialized with a hidden unit size of 512. Training is done with Adam optimizer for 100 epochs with a learning rate of 0.0001 and a batch size of 16. The maximum command sentence length is chosen as 15.

2) Ontology System: The ontology system is jointly constructed using Protégé [31] and owlready2 [32]. HermiT reasoner is invoked to assess the reasoning correctness of the constructed ontology tree. There are 65 classes, 30 IDed individuals and 14 relations on record.

3) Dynamic Knowledge Graph: In addition to an offline evaluation setting, a real-time camera stream $CS_{T_0...T_{inf}}$ is setup to observe the manipulation scene. The proposed video stream sampling method is employed to generate observation clips over time. For any observation clip $C_{t_i...t_j}$ sampled, the Vision-Language model is inferred to generate a command language $S_{t_i...t_j}$ and the related visual attentions. For each entity e_i inside the command language, a word-based recursive tree searching algorithm is employed to query over the ontology tree $onto$, returning a Labeled Directed Graph G_{e_i} of taxonomy which are further merged with the command language into the final dynamic knowledge graph $G_{T_0...T_{inf}}$. A pseudo-algorithm for generating the Dynamic Knowledge Graph is shown in Algorithm 1.

Algorithm 1: Generate a Dynamic Knowledge Graph

Inputs: A camera stream CS_{T_0}. A Vision-Language Model $Model$. A static ontology tree $onto$.
Result: Dynamic knowledge graph $G_{T_0...T_{inf}}$ over time period $T_0...T_{inf}$.

```
initialize CS_{T_0};
initialize an empty G_{T_0...T_inf};
while True do
    C_{t_i...t_j} ← STREAM_SAMPLE(CS_{T_0});
    S_{t_i...t_j} ← Model(C_{t_i...t_j});
    UNION(S_{t_i...t_j}, G_{T_0...T_inf});
    for e_i in S_{t_i...t_j} do
        G_{e_i} ← QUERY(onto, e_i);
        UNION(G_{e_i}, G_{T_0...T_inf})
    end
end
```

B. Evaluation Settings

1) Datasets: We evaluate our seq2seq architecture designs on two datasets: (1) IIT-V2C dataset, originally proposed in Nguyen et al. [3] to process fine-grained human action understanding in the form of command languages; and (2) the proposed Robot Semantics dataset. All experimental parameters for the IIT-V2C dataset are setup as in Nguyen et al. [3]. For experiments with our Robot Semantics dataset, three specialized evaluation divisions are combined:

- **Stream**: Human operators significantly hinder the smoothness of task execution by slowing down or performing a number of task-irrelevant motions. There are 5 human videos - 6159 images for evaluation in this category.
- **Unknown**: Objects that are never presented during model training are collected into this category. There are 18 human videos - 8463 images and 15 WAM videos - 22821 images in this category.
- **Complex**: Multiple objects are presented at scene. One human uses their finger to point to some objects of interests at specific locations. The manipulator performs the action on the specified objects. 7 WAM videos - 9708 are available in this category.

For video stream sampling, an observation window size of 30 frames with an overlapping size of 15 frames is used, equivalent to a full 1 sec of observing under an FPS of 30. As a result, 4188 training clips and 3075 evaluating clips are generated.

2) Baseline Experiments: To validate the effectiveness of our architectural design, we employ ablation studies with the

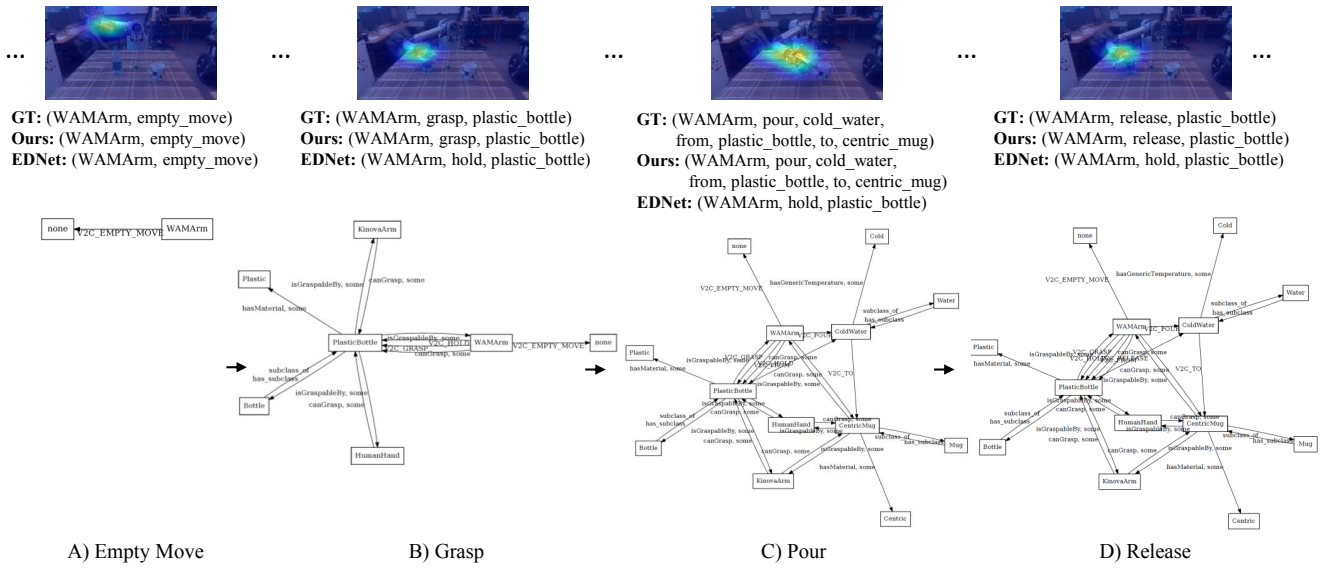

Fig. 6. Examples of the progressive dynamic knowledge graph visualized with spatial attentions and the command languages generated on our Robot Semantics dataset.

following variations:

- **no_att vs. att**, where no attention is considered vs. spatial attention is employed during the stage of visual feature encoding.
- **no_concat vs. concat**, where only the passing of the encoding vector is employed vs. the last encoded hidden state h_{t_L} is also collected and concatenated with the word embedding feature during the sequence decoding stage.

C. Results and Analysis

TABLE I
QUANTITATIVE EVALUATION RESULTS WITH BASELINES ON ROBOT SEMANTICS DATASET AND IIT-V2C DATASET.

Dataset	Name	B-4	M	R	C
Ours	seq2seq-concat-att	0.581	**0.477**	**0.784**	**4.027**
	seq2seq-concat	0.576	0.471	0.771	3.905
	seq2seq	**0.596**	0.471	0.77	3.943
	EDNet [3]	0.580	0.463	0.766	3.879
	V2CNet [5]	0.574	0.459	0.763	3.792
IIT-V2C	S2VT [33]	0.159	0.183	0.382	1.431
	SCN [34]	0.190	0.195	0.399	1.561
	EDNet [3]	0.174	0.193	0.398	1.550
	V2CNet [5]	0.199	0.198	0.408	1.656
	seq2seq-concat-att	0.180	0.195	0.401	1.594
	seq2seq-concat	0.203	0.204	0.417	1.737
	seq2seq	0.203	**0.209**	**0.427**	**1.765**

We report the standard machine translation and language generation metrics with the coco-evaluation code [35]: BLEU-4, METEOR, CIDEr, and ROUGE-L.

1) Results for Vision-Language Model: Table I shows the mean over 5 experiment scores for our Robot Semantics dataset and the best experimental scores on the IIT-V2C dataset. The seq2seq-concat-att performs strongly against others, in particular, it is superior at extracting explainable visual attentions and corresponding labels of semantic meanings. This is important because human beings can attend and gaze into meaningful regions when performing manipulation tasks. Consequentially, the attention mechanism suffers when useless salient information is introduced into the video clips, as in the experiments on IIT-V2C dataset where a synthetic mean ImageNet frame needs to be padded for most video clips. Another benefit in applying attention to real-time robot manipulation is that, under our stream sampling with overlap, the manipulation concepts are consistently being attended to and traced. It can also be observed that the seq2seq structures outperform simple CNN-LSTM architectures like EDNet [3], and complex Temporal Convolutional Network (TCN) based architecture like V2CNet [5], achieving state-of-art performance on both datasets. This indicates that the sequential modeling strategy is more viable when dealing with a real-time camera stream. Additionally, we show an example in 6c where EDNet fails to distinguish between the "holding" and "pouring" actions. However, with the help of spatial attention, our seq2seq-concat-att successfully captions the command language while attending to the regions of a pouring water stream.

2) Results for Dynamic Knowledge: We demonstrate the dynamic evolution of the knowledge graph over time for a pouring action video in Figure 6, along with the generated spatial attentions and the predicted command languages. Under the inputs from a real-time camera stream, our proposed scheme is able to dynamically visualize the evolution of a robot pouring task with the intended manipulation actions. The generated spatial attentions successfully focus on the regions where concepts of manipulation present themselves, while the manipulation procedure is summarized into the command languages. With the ontology system, command languages are completed into a constantly evolving dynamic knowledge graph filled with commonsense knowledge over

presented manipulation entities. The combination of visual attention and the evolving dynamic knowledge graph fundamentally reflects the intended manipulation knowledge over the robot pouring task, which can be directly integrated into robot decision making and action execution.

V. Conclusions

In this paper, we propose a scheme to caption a combination of visual attentions and an evolving dynamic knowledge graph filled with commonsense knowledge. The scheme fuses a Vision-Language model with an ontology system and works with a real-time camera stream. In future work, there are a number of things still to be explored. Visual attention is open for further interpretation, where we plan to explore the connections between language attention and human eye gaze. Our dynamic knowledge graph can be further expanded with graph neural networks to allow for direct integration with robotic trajectory learning during real-time action planning. Ultimately, our future research will be focused on combining manipulation contexts with intelligent robot action controllers.

References

[1] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos, "Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web," in *AAAI*, 2015.

[2] S. Yang, W. Zhang, W. Lu, H. Wang, and Y. L, "Learning actions from human demonstration video for robotic manipulation," in *International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1805–1811.

[3] A. Nguyen, D. Kanoulas, L. Muratore, D. G. Caldwell, and N. G. Tsagarakis, "Translating videos to commands for robotic manipulation with deep recurrent neural networks," in *International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1–9.

[4] H. Zhang and S. Nikolaidis, "Robot learning and execution of collaborative manipulation plans from youtube videos," *arXiv preprint arXiv:1911.10686*, 2019.

[5] A. Nguyen, T.-T. Do, I. Reid, D. G. Caldwell, and N. G. Tsagarakis, "V2cnet: A deep learning framework to translate videos to commands for robotic manipulation," *arXiv preprint arXiv:1903.10869*, 2019.

[6] S. Sudhakaran and O. Lanz, "Attention is all we need: Nailing down object-centric attention for egocentric activity recognition," in *British Machine Vision Conference, (BMVC)*, 2018, p. 229.

[7] M. Lu, Z.-N. Li, Y. Wang, and G. Pan, "Deep attention network for egocentric action recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3703–3713, 2019.

[8] Z. Li, Y. Huang, M. Cai, and Y. Sato, "Manipulation-skill assessment from videos with spatial attention network," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019.

[9] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *NIPS*, 2014.

[10] S. Aditya, Y. Yang, C. Baral, C. Fermuller, and Y. Aloimonos, "Visual commonsense for scene understanding using perception, semantic parsing and reasoning," in *AAAI Spring Symposium Series*, 2015.

[11] C. Ye, Y. Yang, R. Mao, C. Fermüller, and Y. Aloimonos, "What can i do around here? deep functional scene understanding for cognitive robots," in *International Conference on Robotics and Automation (ICRA)*, 2017, pp. 4604–4611.

[12] S. Aditya, Y. Yang, C. Baral, Y. Aloimonos, and C. Fermüller, "Image understanding using vision and reasoning through scene description graph," *Computer Vision and Image Understanding*, vol. 173, pp. 33–45, 2018.

[13] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoğlu, and G. Bartels, "Know rob 2.0—a 2nd generation knowledge processing framework for cognition-enabled robotic agents," in *International Conference on Robotics and Automation (ICRA)*, 2018, pp. 512–519.

[14] S. Aditya, Y. Yang, and C. Baral, "Integrating knowledge and reasoning in image understanding," *arXiv preprint arXiv:1906.09954*, 2019.

[15] Y. Yang, A. Guha, C. Fermüller, and Y. Aloimonos, "Manipulation action tree bank: A knowledge resource for humanoids," in *International Conference on Humanoid Robots*, 2014, pp. 987–992.

[16] H. Zhang, X. Lan, S. Bai, L. Wan, X. Zhou, and N. Zheng, "A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes," *International Conference on Intelligent Robots and Systems (IROS)*, pp. 6435–6442, 2018.

[17] R. Fox, R. Berenstein, I. Stoica, and K. Goldberg, "Multi-task hierarchical imitation learning for home automation," in *International Conference on Automation Science and Engineering (CASE)*, 2019, pp. 1–8.

[18] K. French, S. Wu, T. Pan, Z. Zhou, and O. C. Jenkins, "Learning behavior trees from demonstration," *International Conference on Robotics and Automation (ICRA)*, pp. 7791–7797, 2019.

[19] M. Colledanchise, D. Almeida, and P. Ögren, "Towards blended reactive planning and acting using behavior trees," *International Conference on Robotics and Automation (ICRA)*, pp. 8839–8845, 2016.

[20] J. Hatori, Y. Kikuchi, S. Kobayashi, K. Takahashi, Y. Tsuboi, Y. Unno, W. Ko, and J. Tan, "Interactively picking real-world objects with unconstrained spoken language instructions," *International Conference on Robotics and Automation (ICRA)*, pp. 3774–3781, 2017.

[21] M. Shridhar and D. Hsu, "Interactive visual grounding of referring expressions for human-robot interaction," in *Proceedings of Robotics: Science and Systems*, 2018.

[22] M. Dehghan, Z. Zhang, M. Siam, J. Jin, L. Petrich, and M. Jagersand, "Online object and task learning via human robot interaction," in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 2132–2138.

[23] C. Paxton, Y. Bisk, J. Thomason, A. Byravan, and D. Fox, "Prospection: Interpretable plans from language by predicting the future," *International Conference on Robotics and Automation (ICRA)*, pp. 6942–6948, 2019.

[24] Q. Zhang, J.-H. Chen, D. Liang, H. Liu, X. Zhou, Z. Ye, and W. Liu, "An object attribute guided framework for robot learning manipulations from human demonstration videos," *International Conference on Intelligent Robots and Systems (IROS)*, pp. 6113–6119, 2019.

[25] F. Mahdisoltani, G. Berger, W. Gharbieh, D. Fleet, and R. Memisevic, "On the effectiveness of task granularity for transfer learning," *arXiv preprint arXiv:1804.09235*, 2018.

[26] D. Damen, H. Doughty, G. Maria Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price *et al.*, "Scaling egocentric vision: The epic-kitchens dataset," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 720–736.

[27] Y. Li, M. Liu, and J. M. Rehg, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 619–635.

[28] M. Siam, C. Jiang, S. Lu, L. Petrich, M. Gamal, M. Elhoseiny, and M. Jagersand, "Video object segmentation using teacher-student adaptation in a human robot interaction (hri) setting," in *International Conference on Robotics and Automation (ICRA)*, 2019, pp. 50–56.

[29] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[31] M. A. Musen, "The protégé project: a look back and a look forward," *AI matters*, vol. 1, no. 4, pp. 4–12, 2015.

[32] J.-B. Lamy, "Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies," *Artificial intelligence in medicine*, vol. 80, pp. 11–28, 2017.

[33] S. Venugopalan, M. Rohrbach, J. Donahue, R. J. Mooney, T. Darrell, and K. Saenko, "Sequence to sequence-video to text," 2015, pp. 4534–4542.

[34] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7206–7215.

[35] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.