Let Me Help You! Neuro-Symbolic Short-Context Action Anticipation

Sarthak Bhagat, Samuel Li, Joseph Campbell, Yaqi Xie, Katia Sycara, Simon Stepputtis

***Abstract*— In an era where robots become available to the general public, the applicability of assistive robotics extends across numerous aspects of daily life, including in-home robotics. This work presents a novel approach for such systems, leveraging long-horizon action anticipation from short-observation contexts. In an assistive cooking task, we demonstrate that predicting human intention leads to effective collaboration between humans and robots. Compared to prior approaches, our method halves the required observation time of human behavior before accurate future predictions can be made, thus, allowing for quick and effective task support from short contexts. To provide sufficient context in such scenarios, our proposed method analyzes the human user and their interaction with surrounding scene objects by imbuing the system with additional domain knowledge, encoding the scene object's affordances. We integrate this knowledge into a transformer-based action anticipation architecture, which alters the attention mechanism between different visual features by either boosting or attenuating the attention between them. Through this approach, we achieve an up to 9% improvement on two common action anticipation benchmarks, namely *50Salads* and *Breakfast*. After predicting a sequence of future actions, our system selects an appropriate assistive action that is subsequently executed on a robot for a joint salad preparation task between a human and a robot. Videos and dataset available on the website: `https://sarthak268.github.io/NeSCA/`.**

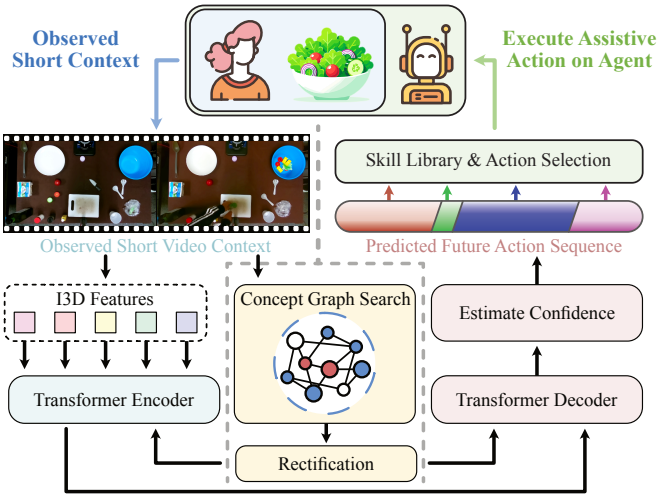

Fig. 1: NeSCA: Given a short video segment (blue), our system anticipates future actions and their respective confidences (color gradients) utilizing our proposed neuro-symbolic attention approach to re-focus attention between visual features. Finally, if sufficiently confident about the prediction, a robot executes assistive actions (green).

I. Introduction

Action anticipation is a crucial step in the development of intelligent agents [1] for the task of human-robot collaboration (HRC). For example, accurately anticipating a human's future action allows a robot assistant to provide improved support, moving past reactive toward proactive agents. Prior work in action anticipation has mainly focused on short-term or next-action anticipation [2], [3], [4], [5], [6], [7], [8], [9]; however, to enable proactive agent behavior, multiple future actions must be predicted for long horizons as the immediate next action may not always be the most appropriate assistive action. For example, taking over a task that the human is already doing or about to start may interfere with the user's immediate actions, necessitating the prediction of multiple future actions. Selecting one of multiple future actions depends on various factors, including if such a task can be executed in parallel by the robot and if its likelihood of occurring is sufficiently high given the current observation. In an assistive task, predicting action sequences requires a quick understanding of the user's current behavior, necessitating making decisions given short observation contexts of task-relevant behavior. However, making accurate predictions of future actions given only a short horizon of relevant observations is challenging due to its inherent lack of context. To this end, we propose **NeSCA**, **Ne**uro-**S**ymbolic Short-**C**ontext **A**ction Anticipation, which imbues a neural action anticipation pipeline with additional symbolic domain knowledge in the form of a Knowledge Graph (KG). NeSCA utilizes domain knowledge to connect scene objects to their relevant affordances [10], [11] through a structured prior. For example, with the knowledge that a *tomato* has the affordance being *cuttable* and knowing about the presence of a *knife* that can be used for cutting, NeSCA can boost the attention between these concepts to increase the likelihood of the human's intent of *cutting tomatoes* in the future while simultaneously attenuating the attention between other unrelated features (see Fig. 1). Imbuing a neural network that can effectively comprehend complex inputs like videos with symbolic knowledge can greatly enhance the performance of downstream tasks [12], i.e., subsequent action anticipation and user assistance. Empirically, we find that utilizing the knowledge graph that connects objects to their affordances reduces the required task-relevant observation by $\approx 50\%$ when predicting future actions as compared to current state-of-the-art baselines.

To process high-dimensional inputs like videos, transformers [13] have proven to be efficient at comprehending sequential data and lend themselves well to action

[1]The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. {sarthakb, swli, jcampbell, yaqix, sycara, sstepput}@andrew.cmu.edu

anticipation from videos [14], but remain largely black-box end-to-end approaches. On the other hand, structured domain knowledge remains interpretable and has previously been investigated in the image domain [15], demonstrating improved performance for image classification [12]. In this work, we seek to integrate neural video comprehension with external symbolic domain knowledge pertaining to the objects in the scene, linking them to their respective affordances. Given a previously unseen video sequence, we extract the relevant scene objects via a neural object detector and employ graph search through our KG to assign relevant affordances to them. To achieve the integration of the video understanding and extracted domain knowledge, we propose to imbue the attention mechanism of the transformer with an addition rectification matrix that influences how queries and keys interact with each other. Intuitively speaking, the learned knowledge-conditioned rectification matrix boosts or attenuates the attention between various video features, thus, aiding the prediction of future actions. A particular benefit of this approach is that our proposed method significantly improves performance when only short-horizon contexts are given – a key aspect for effective human-robot collaboration that prior works in action anticipation [16], [17], [18], [6], [19], [14] only addressed to a limited extent. Before utilizing our action anticipation approach for human-robot collaboration, we demonstrate its efficacy on two common long-term action-anticipation benchmarks, namely the *50Salads* [20] and *Breakfast* [21] datasets, and show superior performance as compared to current state-of-the-art methods.

Having demonstrated the efficacy of NeSCA, we showcase a joint salad creation task in a real-world tabletop scenario that leverages the sequence of predicted future human actions. Given a set of predicted actions, the system selects an appropriate action for the robot to execute while the user keeps working on their current task given a set of selection criteria (see Figure 1). Among others, these criteria mainly include checking whether the anticipated action's preconditions are already satisfied and if the action is predicted with sufficient confidence. When such an action is identified, the robot executes the action to support the user. With our approach, we achieve a 50.1% accuracy in selecting and executing an appropriate assistive action while also reducing the required length of context to half compared to the current state-of-the-art to achieve a similar success rate.

In summary, our contributions are as follows:

- We propose a novel approach utilizing knowledge graphs to augment the attention mechanism for transformer-based action anticipation, which we refer to as NeSCA.
- Through extensive experiments, we demonstrate that our proposed method outperforms current state-of-the-art methods for action anticipation on two challenging benchmarks, *50Salads* and *Breakfast*.
- We show how our proposed method can be utilized for effective HRC that anticipates tasks and subsequently supports human users in the creation of a salad in a real-world tabletop manipulation setting.

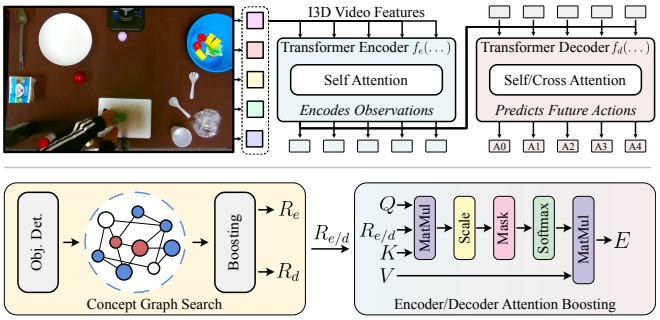

Fig. 2: NeSCA utilizes a transformer architecture for action anticipation (top); however, in parallel, Concept Graph Search (bottom) is utilized to obtain the set of active concepts, including related affordances, in the scene. These concepts are further used to refocus the attention in the transformer toward the relevant visual features.

II. Related Work

Action anticipation is a field of research that is currently gaining a lot of attention due to its usefulness in areas such as autonomous driving and human-robot interaction [22]. In this study, we introduce a new approach that makes use of structured domain knowledge to predict long-term action sequences based solely on short video contexts.

Knowledge Graphs for Computer Vision. The emergence of the utilization of structured domain knowledge, in the form of knowledge graphs, in vision models is gaining traction as it grounds their predictions by establishing a comprehensive understanding of entities and their interconnected relationships, thereby, enhancing overall model interpretability and performance. [15] introduced a knowledge graph as a structured prior for image classification and proposed the Graph Search Neural Network, demonstrating its performance improvement by integrating knowledge graphs into the vision classification pipeline. Further, [12] extended it to include the augmentation of novel concepts, encompassing visual objects and compound concepts such as affordances, attributes, and scenes. In this work, we extend the idea by refining the propagation framework from [12] to identify relevant object affordances along with the relevant tools that can be used to afford it in the desired manner. To the best of our knowledge, our approach is the first one to utilize the information about the affordances of the objects in the scene to perform action anticipation. Our methodology represents a novel approach to leveraging information regarding the affordances of objects within a scene for action anticipation.

Action Anticipation. The task of action anticipation from videos [23] revolves around predicting future actions based on a specific segment of the video. With recent advancements in foundational vision models and the availability of large-scale human-centric datasets [24], this domain has gained significant attention. Many recent approaches have been developed to predict a single future action within a short time frame, typically spanning a few seconds [2], [3], [4], [5], [6], [25], [7], [8], [9]. However, a notable emerging trend

is long-term action anticipation, which emphasizes predicting a sequence of future actions occurring in the distant future from a lengthy video [16], [17], [18], [6], [19], [14]. While much attention has been paid to predicting long-term actions with ample video context, limited research has addressed using short video contexts to predict long-term future action sequences. Our work addresses this particularly challenging task: Action anticipation for long-horizon predictions given only a short observation context.

Human-Robot Collaboration. Several approaches have performed human-robot collaboration by anticipating what actions might be useful in the current setup, by either utilizing gaze information from the user [26] or performing action prediction [27]. While these works can infer current actions, they fall short in capturing the temporal aspect of visual inputs to make predictions not only about ongoing actions but also anticipate future actions. Other methods have been proposed utilizing a human-in-the-loop approach to improve the learned policy in an online manner [28]; however, enabling these interactions can be expensive, and therefore, offline finetuning approaches have been identified as an effective solution to deploy robots in real-world scenarios [29]. This work integrates the advantages of predicting future actions and offline fine-tuning with a finite curated dataset in a novel environment to enhance the robot's prediction of useful actions considering the human's actions.

III. Knowledge-Guided Action Anticipation

This section introduces the details of our proposed method, NeSCA, as well as its application to Human-Robot Collaboration (HRC). At its core, NeSCA, consists of two core components: (1) a neuro-symbolic graph-search approach that extracts relevant scene concepts (i.e., objects and their affordances, see Sec. III-A); and (2) a modified attention mechanism informed by our extracted concepts, allowing us to anticipate future actions from short observation context (see Sec. III-B and III-C). With this set of predicted actions, we demonstrate the utility of NeSCA in an HRC task, utilizing our fast action anticipation from short observed contexts, which allows us to effectively assist a user (see Sec. III-D).

Problem Statement. NeSCA (see Fig. 2) addresses the problem of predicting a sequence of future actions $\mathbf{a}$ from a *short* video observation $\mathbf{V}_O$ that can subsequently be used to provide effective assistance to a human user. In our setting, we observe α-percent of a video and predict β-percent of future actions with respect to the average total video length. Given an observation, we learn a function $\mathbf{a} = f_\theta(\mathbf{V}_O)$ that predicts a sequence of actions $\mathbf{a}$ happening after the end of $\mathbf{V}_O$. The video sequence $\mathbf{V}_O \in \mathbb{R}^{H\times W\times C\times N}$ is represented as a four-dimensional matrix describing the height H, width W, and channels C of each video frame $\mathbf{V}_i$, and the number of observed frames N_O. The action sequence $\mathbf{a} = [(a_0,d_0,c_0),...,(a_n,d_n,c_n)]$ contains a list of N_P tuples describing the action a, its duration d, and confidence c.

To actuate the robot, we propose a policy $a_r = \pi(f_\theta(\mathbf{V}_O),\mathbb{S})$ utilizing the predicted set of future actions.

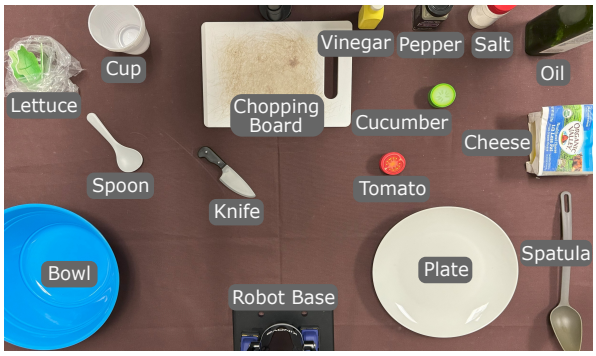

Fig. 3: Our *Dummy Kitchen* setup and available objects for creating salads in an HRC task involve cutting/peeling vegetables, preparing dressing, and mixing/serving the salad.

Given a skill library $\mathbb{S}$ and a list of future actions $f(\mathbf{V})$, π identifies a suitable action a_r for a robot to execute in our HRC task.

Training Procedure. We train our model from a dataset $\mathbb{D} = [\mathbf{s}_n,\ldots,\mathbf{s}_N]$ where each sample $\mathbf{s}_n = [\mathbf{V}_n,\mathbf{a}_n]$ contains the video-frame $\mathbf{V}_\mathbf{n}^\mathbf{i}$ and action-label $\mathbf{a}_n^i$, where i is the frame index of video n. After training, we provide the trained action anticipation model $f_\theta(\ldots)$ with a new, previously unseen video sequence, showing α percent (with respect to time) of the full video, tasking the policy with predicting the most likely action for each frame in the following β percent of the remaining video.

A. Extracting Domain Knowledge

We provide a hand-crafted KG as the source of our domain knowledge, establishing connections between various concepts. In the following, we refer to objects and affordances as concepts. Each node in our KG is initialized by utilizing Grounded-DINO [30]. To utilize this knowledge during inference, given a short sequence of video frames, we extend our prior work [12] to the domain of videos. Intuitively, this approach utilizes a neural object detector to extract a set of initial concepts and subsequently utilizes them as a starting point for a graph search through the knowledge graph $\mathbb{K}$. We create the graph $\mathbb{K}$ consisting of two types of nodes: object nodes (e.g., *salt*, *knife*, *bowl*) and affordance nodes (e.g., *graspable*, *pourable*, *cuttable*). For example, a *tomato* has a connection to *cuttable*, which, in turn, connects to *knife*.

In the first step, we extract a set of relevant concepts from the video frames $\mathbf{V_0} \in \mathbf{V_O}$, using open-vocabulary object detection as proposed in [30]. These initial concepts $\mathbb{C}_O$ are then utilized as a starting point for our iterative Concept Graph Search (CGS), forming the initial set of active concepts in our KG. CGS has two main components: a) the Propagation Network, which generates frame-conditioned representations (based on $\mathbf{V}_0$) for all candidate concepts connected to the active ones using Graph Attention Network v2 [31], and b) the Importance Network, responsible for computing a scalar importance value for each candidate node, given $\mathbf{V}_0$. Concepts above a predefined importance threshold

Fig. 4: Example of our assistive HRC system: Shortly after the user starts to prepare the dressing, the robot identifies the intention and correctly assists the user in creating the dressing by adding further ingredients.

are incorporated into the set of active concepts. This process is repeated for T iterations, alternating the Propagation and Importance Networks. After expanding all relevant concepts $\mathbb{C}_F$ through T iterations, we generate a latent representation $\mathbf{C}_n$ for each active concept. Intuitively, CGS allows us to extract the relevant concepts concerning the observed video and utilize them as additional domain knowledge during the action anticipation (see Sec. III-B).

B. Action Anticipation with Domain Knowledge

This section introduces our main contribution, detailing how domain knowledge $\mathbf{C}_n$ can be utilized for action anticipation. In particular, our architecture is motivated by [14]; however, we alter the attention mechanism of the encoder and decoder to allow for the integration of additional domain knowledge, thus, improving the contextual reasoning capabilities of the action-anticipation pipeline.

However, before we detail the novel attention mechanism in Section III-C, we briefly outline the standard transformer-based part of our pipeline, consisting of an encoder $f_e(\dots)$ and decoder $f_d(\dots)$ (see Fig 2). The encoder $\mathbf{e}_{enc} = f_e(\mathbf{V}_O)$ utilizes I3D [32] features $\mathbf{x}_n^{I3D}$ of the observed video $\mathbf{V}_O$ and produces a set of embeddings $\mathbf{e}_{enc} \in \mathbb{R}^{n \times D}$ for each video frame n and encoding dimension D. The encoder processes visual features extracted from the observed segment of a video $\mathbf{V}_O$ by employing multi-head self-attention. The resulting output is then provided to a classifier, $\mathbf{a}_O = f_{obs}(\mathbf{e}_{enc})$, determining the actions corresponding to the observed part of the video segment.

The decoder employs the embeddings of the observed sequence $\mathbf{e}_{enc}$ generated by $f_e(\dots)$ along with learnable tokens referred to as actions queries, initialized with zero vectors. Similarly to the encoder, the decoder $\mathbf{e}_{dec} = f_d(\mathbf{e}_{enc}, \chi)$ produces a set of embeddings $\mathbf{e}_{dec} \in \mathbb{R}^{p \times D}$ where p is the upper bound of the future actions that can be predicted and $\chi \in \mathbb{R}^{p \times D}$ are the p action queries. Subsequently, we utilize two separate, fully connected networks for predicting the future actions $\mathbf{a}_{pred}$ and their durations $\mathbf{d}_{pred}$ respectively.

$$\mathbf{d}_{pred} = f_{dur}(\mathbf{q}_n^{L_d}) \quad \text{and} \quad \mathbf{a}_{pred} = f_{act}(\mathbf{q}_n^{L_d}) \tag{1}$$

Finally, we retrieve confidence $\mathbf{c}_{pred}$ for each predicted action. We quantify the certainty of the model's prediction using negative entropy of the predicted distribution of actions, i.e., $\mathbf{c}_{pred} = \sigma(\mathbf{a}_{pred}) \log(\sigma(\mathbf{a}_{pred}))$, where $\sigma(.)$ represents the softmax function.

C. Knowledge-Guided Attention Mechanism

So far, we have discussed how relevant domain knowledge is retrieved from a symbolic KG, as well as the general action anticipation pipeline. In this section, we describe our main contribution: Altering the multi-head attention layers of the encoder $f_e(\dots)$ and $f_d(\dots)$ to improve contextual prediction by leveraging our extracted domain knowledge $\mathbf{C}_n$. Intuitively, the extracted domain knowledge establishes a connection between the objects in the scene and their respective affordances, improving the predicted actions' relevance by boosting or attenuating the attention between different features. To this end, we introduce a rectification matrix $\mathbf{R}$ inside the multi-head attention equation. We obtain a separate knowledge-guided rectification matrix for our encoder and decoder, namely $\mathbf{R}_e$ and $\mathbf{R}_d$, with which we modify the attention mechanism:

$$\text{KG-Attn}_{e/d}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\Big(\frac{\mathbf{Q}\mathbf{R}_{e/d}\mathbf{K}^T}{\sqrt{d_k}}\Big)\mathbf{V} \tag{2}$$

This modification allows our model to prioritize the features associated with objects having relevant affordances, giving them higher importance than those not present in the scene. The rectification matrix is presented as a diagonal matrix for which we retrieve the diagonal by predicting it from $\mathbf{C}_n$. Particularly, we a utilize $\mathbf{R}_{e/d} = f_{e/d}^R(\mathbf{C}_n)$ to predict each diagonal, where $f_{e/d}^R(...)$ is implemented as an LSTM. Note that $f_e^R(\dots)$ and $f_d^R(\dots)$ are separate networks that do not share parameters amongst themselves.

D. Human-Robot Collaboration using Anticipated Actions

Having access to a sequence of likely future actions as well as their durations and confidences, we define a policy $a_r = \pi(f_\theta(\mathbf{V}_O), \mathbb{S})$ that chooses an appropriate assistive robot action a_r from a set of possible skills $\mathbb{S}$ (see Fig 4). Selecting the appropriate action $a_r \in \mathbb{S}$ is a challenging task as, upon selection of an action, the robot is committed to performing it. This commitment requires time and utilizes objects in the environment that could have been used otherwise by the human partner.

Thus, we define four selection criteria to choose an appropriate action or not to choose an action at all and continue to observe the user. First, the cumulative duration of actions $d_s = \sum_{i=0}^{i=r-1}(d_i)$ for any action candidate a_r, where $0 \leq r \leq |\mathbf{a}|$ must be larger than the average length d_r of action candidate a_r. This constraint ensures that the human collaborator would not have done or needed to do the chosen task before the robot can complete it. Secondly, we ensure that all objects

Model	Frame-wise (↑ larger is better) / Action Sequence (↓ smaller is better)								Next Action (↑)	
	5-10	5-20	5-30	5-50	10-10	10-20	10-30	10-50	5	10
50Salads										
KG Baseline [12]	6.92 / 4.88	6.21 / 6.20	6.01 / 7.44	5.58 / 7.92	7.13 / 4.50	6.48 / 5.98	6.07 / 7.37	5.78 / 7.88	8.0	9.0
Video-Llama [33]	- / 6.44	- / 7.20	- / 7.90	- / 9.12	- / 6.12	- / 6.80	- / 7.86	- / 9.02	6.0	7.0
CNN [16]	7.42 / 3.22	6.97 / 5.07	6.67 / 5.86	6.40 / 6.11	8.50 / 3.33	7.80 / 4.87	7.45 / 5.20	6.92 / 6.60	10.0	28.0
RNN [16]	7.98 / 3.00	6.90 / 5.46	6.48 / 6.30	6.42 / 6.16	8.78 / 2.94	7.92 / 4.83	7.57 / 5.20	7.26 / 6.52	12.0	30.0
FUTR [14]	8.90 / 2.98	7.46 / 4.52	7.29 / 5.40	8.63 / 6.80	15.17 / 2.74	11.34 / 4.04	11.31 / 4.98	11.36 / 6.04	12.0	36.0
NeSCA ($T=0$)	7.95 / 3.08	7.86 / 4.42	6.15 / 5.20	7.10 / 6.58	24.0 / 2.60	16.90 / 3.72	11.17 / 4.98	11.30 / 6.80	10.0	34.0
NeSCA ($T=1$)	**17.86 / 2.84**	**16.25 / 4.22**	**10.84 / 5.14**	**9.38 / 6.70**	**23.15 / 2.54**	**17.28** / 3.78	**16.62** / 4.76	**13.61 / 5.74**	**14.0**	**42.0**
NeSCA ($T=2$)	13.67 / 2.90	9.60 / 4.40	8.62 / 5.32	8.51 / 6.60	22.86 / 2.56	16.86 / **3.71**	14.70 / **4.52**	12.66 / 5.75	12.0	38.0
Breakfast										
KG Baseline [12]	5.44 / 8.22	4.95 / 9.10	4.22 / 9.66	3.98 / 10.02	6.02 / 7.90	5.15 / 8.77	4.86 / 9.21	4.51 / 9.78	7.22	12.31
Video-Llama [33]	- / 11.20	- / 12.24	- / 13.62	- / 13.82	- / 11.08	- / 12.04	- / 12.98	- / 13.22	5.39	9.80
CNN [16]	5.76 / 6.98	5.52 / 7.22	5.45 / 7.98	4.80 / 8.43	7.84 / 6.48	6.62 / 6.95	6.02 / 7.44	5.17 / 8.13	11.45	18.90
RNN [16]	6.16 / 6.76	5.60 / 7.05	5.53 / 7.69	4.96 / 8.09	7.67 / 6.67	6.73 / 6.90	6.15 / 7.44	5.22 / 8.12	12.02	19.96
FUTR [14]	9.54 / 1.63	7.24 / 2.07	6.42 / 2.40	5.58 / 3.02	14.70 / 1.41	12.55 / 1.76	12.10 / **2.06**	11.71 / 2.62	23.97	30.05
NeSCA ($T=0$)	9.69 / 1.65	7.20 / 2.04	6.55 / 2.40	5.62 / 3.06	15.30 / 1.43	13.23 / 1.82	12.24 / 2.22	11.65 / 2.68	20.55	25.47
NeSCA ($T=1$)	**9.91 / 1.60**	**7.95 / 2.02**	**6.86 / 2.34**	**5.88 / 2.98**	**15.53** / 1.41	**13.52** / 1.76	**13.07** / 2.09	**11.94** / 2.63	**25.25**	26.45
NeSCA ($T=2$)	9.75 / 1.62	7.60 / 2.05	6.70 / 2.38	5.76 / 3.00	15.52 / **1.36**	13.46 / **1.72**	12.68 / 2.15	11.84 / **2.60**	23.32	**30.35**

TABLE I: NeSCA performance compared to the current state-of-the-art in long-term action anticipation for different horizons of $\alpha - \beta$ (top row). The numbers in boldface and underlined indicate the highest and the second-highest accuracy, respectively.

needed for a chosen action a_r, as defined in our skill library $\mathbb{S}$, are observed in our set of active concepts $\mathbf{C}_n$ and that all objects have the appropriate affordances. For example, if we consider the action of cutting a tomato, the robot requires a knife, cutting board, and tomato, but also that the tomato has the affordance of being *cuttable* (i.e., is not already in a diced state, which would not afford the ability to be cut it further). Thirdly, we verify whether the prerequisites for the specific task have already been fulfilled; for instance, the action of *placing tomato in bowl* necessitates that the *cut tomato* action precedes it. Lastly, we consider the confidences $\mathbf{c}_{pred}$ for the candidate action a_r. Specifically, we only consider actions for which the estimated confidence is above a pre-defined threshold to ensure that the robot only executes the most likely actions.

With these four constraints, we define policy $\pi(\dots)$ that, given the predicted action sequence for horizon β over an observed time-horizon α, selects a single action a_r that should be executed by the robot. However, note that if no such action that satisfies all four constraints can be found, policy π will return a no-op action. In such cases, the policy will continue attempting to identify an appropriate action as further video frames are available. Similarly, when the robot is currently committed to executing a previously selected action a_r, the robot will ignore action choices made by policy π until the prior action is completed.

IV. Experiments

In this section, we evaluate NeSCA on two common benchmarks for action anticipation – *50Salads* and *Breakfast* – and demonstrate how action anticipation can be used for human-robot collaboration in a real-world task. Our benchmarks (see Sec. IV-A) extensively evaluate the ability to utilize short video contexts while predicting long-horizon future actions. In our real-world setup (see Sec. IV-B), we utilize the ability to correctly anticipate actions to facilitate the collaborative creation of a salad.

a) Datasets: We evaluate the effectiveness of NeSCA using two publicly available benchmark datasets for action anticipation for in-home environments, particularly kitchen scenarios, as well as one real-world robotics dataset: 1) The *50Salads* dataset [20] with its five splits, densely annotated with 17 fine-grained action labels and three high-level activities; 2) The *Breakfast* dataset [21] with four splits, categorizing each frame into one of 10 breakfast-related activities using a comprehensive set of 48 fine-grained action labels; and 3) a dataset of 20 videos collected from our dummy kitchen setup (see Fig 3). Among these dummy kitchen videos, we designate half of them for fine-tuning the model, while the remaining half are reserved for assessing the performance of the fine-tuned model.

b) Metrics: To evaluate the efficacy of our approach, we calculate the *Mean over Classes* (MoC) accuracy. This metric is computed by comparing the predicted actions to the ground-truth actions for all future frames within the horizon window defined by β, making it the most comprehensive metric as it captures action sequence and action durations. To quantify the ability of our model to identify the sequence of the next actions without considering their durations, we employ a metric that computes the minimum number of addition, deletion, or substitution operations required to exactly match the predicted to the ground truth action sequence. While neglecting action durations, this metric captures the semantic understanding of the task composition. Derived from this metric, we also employ immediate single next-action prediction as a metric. Finally, in our real-world setup, we utilize the accuracy of completing an action, i.e., anticipating the right action and executing it, as our primary evaluation metric.

A. Action Anticipation Benchmark

a) Action Anticipation Performance: We evaluate NeSCA by comparing the performance on all metrics averaged across all splits against long-term action anticipation

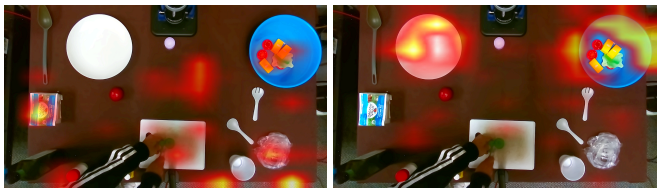

Fig. 5: Attention to visual features relevant to our task, as attended to by FUTR (Left) and NeSCA (Right). With our re-focusing approach, attention is heightened for areas having objects relevant to tasks after the current *cutting lettuce*.

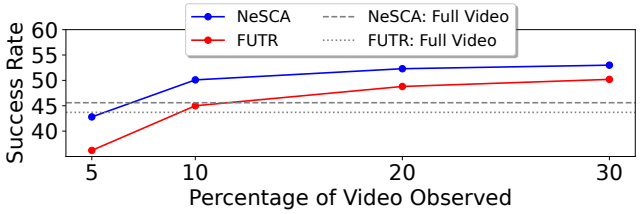

Fig. 6: Success rate of NeSCA in our kitchen setup with varying context lengths. The observed percentage is reported with respect to the average length of finetuning videos.

baselines [16], [14], depicted in Table I. [16] uses action labels extracted from the action segmentation model, while our work and the most recent state-of-the-art [14] use visual features from the observed video segments. In addition, we conduct a comparative analysis with two additional baseline methods. The first is a KG-only approach proposed by [12], which aims to extract all objects along with their associated affordances in each p^{th} frame of the video. This method incorporates a decay mechanism with a rate of γ to account for the diminishing importance of active nodes over time as the video moves on. The second baseline is a multimodal fusion model utilizing large language models [33], where we begin by providing a comprehensive explanation of the entire scenario and subsequently prompt it to produce predictions for future actions from a predefined list of possibilities. As can be seen in Table I, NeSCA outperforms the current state-of-the-art in long-term action anticipation using short context in all the metrics on the *50Salads* dataset and on nine out of the ten metrics we used on the *Breakfast* dataset. On the MoC metric, NeSCA outperforms the baseline by up to 9% on *50Salads* and 1% on *Breakfast*.

As our method relies on a fixed number of iterations T during CGS, we also evaluated varying numbers with $0 \leq T \leq 2$. The most favorable outcome was observed when T was set to 1. In the case of $T = 0$, no graph propagation was performed, and the model relied solely on objects detected by our object detector. As a result, its performance resembled that of [14], which lacks information about associated object affordances. On the other hand, when $T = 2$, the model's consideration expanded beyond the context relevant to the video. Empirically, we chose a propagation of $T = 1$ for all our experiments.

Approach	Finetuning	Confidence	Success		MoC	
			$\alpha = 5\%$	$\alpha = 10\%$	$\alpha = 5\%$	$\alpha = 10\%$
Autoregressive			13.0	17.4	6.2	7.4
Autoregressive	✓		27.3	36.4	8.9	12.2
FUTR			16.6	20.8	6.7	9.2
NeSCA			19.2	23.1	6.9	9.9
NeSCA	✓		33.7	41.8	12.4	18.1
NeSCA (Full)			35.2	43.6	14.4	20.2
NeSCA	✓	✓	42.8	50.1	-	-

TABLE II: Performance of the action anticipation pipeline, NeSCA, for human-robot collaboration on our kitchen setup. *Success* values represent the real-time joint performance of anticipating the sequence of actions and performing the actions in the kitchen setup, while the *MoC* values represent the accuracy of framewise prediction of actions over the collected trajectories from the kitchen setup. The average length of sequences (according to which the percentages are calculated here) is 120 seconds.

b) Qualitative Evaluation: We showcase an example to compare NeSCA against [14] by evaluating the time-series segmentation of the predicted future actions. Figure 7 depicts an example from our kitchen setup where the model observes two actions in the $\alpha = 5\%$ (≈ 6 seconds) observed segment of the video and then predicts what actions take place in the next $\beta = 30\%$ (≈ 36 seconds) of the video. While our model accurately identifies the sequence of all four ground-truth future actions and their approximate durations, the baseline approach failed to identify two out of the four actions correctly. We attribute our approach's improved performance to our model's ability to focus on the objects currently in use and objects that could be used later by extracting their associated affordances.

The re-focusing of our model is demonstrated in Figure 5, highlighting the areas our approach (right) and [14] (left) focuses on. Our model directs attention to both the bowl and the plate, even in scenarios where the subject is not directly interacting with them. This capability enables our model to accurately anticipate future actions, such as *put cheese into bowl* and subsequently *serve salad onto plate*. In contrast, the baseline approach indiscriminately focuses on many objects in the scene, neglecting to discern the relevant objects based on their affordances and their potential utility in the context of ongoing and completed actions.

B. Real-World Human-Robot Collaboration

After showing the effectiveness of our NeSCA approach on two common baselines, we utilize it in an HRC scenario of preparing a salad in a joint task between a robot and a human user. To bridge the domain gap that arises due to the shift in physical attributes (for example, lighting conditions, the color of prevalent objects, etc.) of the real-world kitchen setup and the trained dataset, we finetune our trained model to a dataset comprising of both the original videos and 10 videos collected on our kitchen setup.

a) Transfer Learning on Kitchen Environment: We assess the effectiveness of our fine-tuned model on our kitchen environment depicted in Table II, which utilizes the same action space as the *50 Salads* dataset. During

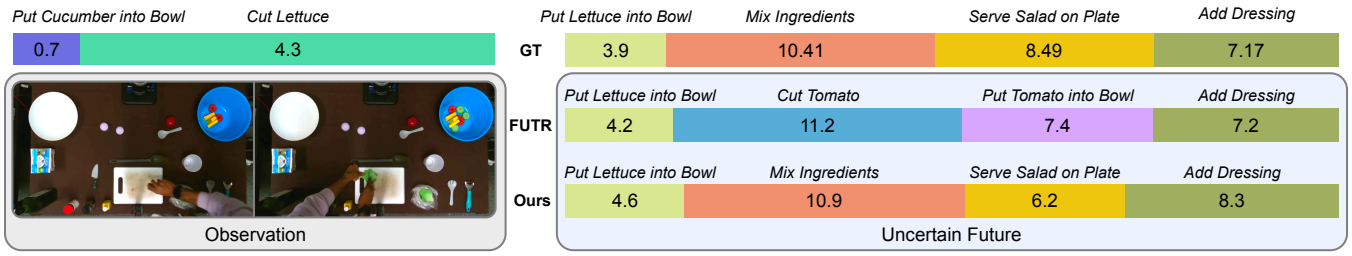

Fig. 7: Sample result of NeSCA on the real-world kitchen setup, observing 5% and predicting another 30% into the future, along with the predictions made by FUTR [14] and the ground truth action labels.

inference, we provide access to a pre-defined skill library $\mathbb{S} = \{S_0, S_1, S_2 ... S_m\}$ where each high-level skill S_i corresponds to a specific sequence of low-level control inputs, conditioned on the placement of the objects. We use a top-down RGB camera (see Figure 3) to track our objects using Scale-Invariant Feature Transform (SIFT) [34] and color feature detection. Given the skill library and video stream, our action anticipation module operates in real-time to deduce future actions and their associated confidences. When the four action criteria mentioned in Section III-D are met, the respective instruction is sent to the robot for execution.

The skills in the library are broadly categorized into three "grasp types": a top-down grasp, suitable for pick-and-place actions with items like vegetables; a sideways grasp, ideal for picking up and pouring objects such as olive oil or vinegar bottles; and an aligned grasp, designed for handling oriented tools like knives and spatulas. The aligned grasp feature is engineered to bring and hand over tools to a human collaborator. In this process, the robot brings the instructed tool near the potential area of use for easy accessibility.

For real-world experiments, we define success as the robot correctly identifying future actions and executing the respective target action. The observed % of video, denoted by α, is computed by comparing the duration of human action observed by our model with the average duration of a video in our finetuning dataset. This evaluation involves comparing its performance against several baselines, namely: (1) a non-finetuned model, (2) an autoregressive classifier that predicts the next action by considering extracted video features in addition to prior action predictions, and finally, (3) a model with the same architecture but trained from scratch on 25 videos collected in our dummy kitchen environment. While training from scratch on our dummy kitchen environment only uses 25 videos as compared to the original *50Salads* dataset, we find that providing further videos does not improve the performance of the model any further. In addition to our approach, we also compare against the best performing state-of-the-art method in long-term action anticipation, FUTR [14]. In Table II, we have observed a significant performance improvement when fine-tuning the model using a few videos from our kitchen setup. Moreover, NeSCA consistently outperforms autoregressive baselines, underscoring the significance of leveraging not only the visual-temporal features of the video but also exploiting information about objects in the scene and their associated affordances. In comparison to a model trained on the complete dataset (see NeSCA (Full) in Table II), our fine-tuned approach demonstrates a superior success rate and comparable frame-wise action prediction accuracy. Note that we have not presented the MoC values for our model with confidence estimation, since this is specifically incorporated into the model for real-time evaluation and is not applied in the assessment using our collected set of videos.

Further, we also evaluate the dependence of NeSCA and FUTR on the percentage of video observed on our kitchen setup (see Figure 6). As is expected, the performance of both approaches increases as the percentage of video increases, but the difference is much more pronounced when the context window is shorter. Further, the dashed line represents an approach that, instead of employing a sliding video window focusing on a specific fixed context, utilizes the entire video up to that point. By observing only 10% of the video, NeSCA outperforms the non-sliding window approach. This underscores the ability of NeSCA to draw meaningful inferences with a very short context window and highlights the impact of using uncertainty-based thresholding to improve the success rate in real-world scenarios.

C. Discussion, Limitations and Future Work

While our experiments demonstrate the value of action anticipation in human-robot collaboration, it is crucial to acknowledge that real-world human behavior is highly unpredictable. This necessitates the ability of action anticipation approaches to quickly and accurately predict actions from only short observations of task-relevant behavior. However, exploring more complex methods that incorporate additional factors such as gaze, behavior patterns, or personalized action anticipation tailored to individual differences could be promising avenues for future research. Additionally, we demonstrated that augmenting action anticipation with symbolic knowledge greatly benefits the model's performance; however, our approach relies on the availability of a hand-crafted knowledge graph that encompasses relevant scene objects and their respective affordances. To address this issue, we plan on generating relevant knowledge graphs in a data-driven manner.

V. Conclusions

Our novel knowledge-guided action anticipation approach, NeSCA, considers both objects and their affordances in the

scene, demonstrating state-of-the-art performance on two action anticipation datasets, particularly from short task-relevant observations. A key to our method's success is the integration of domain knowledge into the attention mechanism of the transformer, allowing for effective boosting or attenuation of visual features in the short context provided to the model, allowing it to make high-quality predictions faster. Given effective action anticipation through our method, we demonstrate its utility in an assistive HRC task, in which a robot successfully assists in the creation of a salad.

VI. Acknowledgements.

We would like to acknowledge the support from DARPA under grant FA8750-23-2-1015, AFOSR under grants FA9550-18-1-0251 and FA9550-18-1-0097, and ARL under grant W911NF-19-2-0146 and W911NF-2320007.

References

[1] N. F. Duarte, M. Raković, J. Tasevski, M. I. Coco, A. Billard, and J. Santos-Victor, "Action anticipation: Reading the intentions of humans and robots," *IEEE Robotics and Automation Letters*, 2018.

[2] A. Furnari and G. Farinella, "What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[3] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Predicting the future: A jointly learnt model for action anticipation," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[4] A. Miech, I. Laptev, J. Sivic, H. Wang, L. Torresani, and D. Tran, "Leveraging the present to anticipate the future in videos," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[5] F. Sener and A. Yao, "Zero-shot anticipation for instructional activities," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[6] F. Sener, D. Singhania, and A. Yao, "Temporal aggregate representations for long-range video understanding," in *ECCV 2020*.

[7] B. Fernando and S. Herath, "Anticipating human actions by correlating past with the future with jaccard similarity measures," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] R. Girdhar and K. Grauman, "Anticipative video transformer," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

[9] D. Roy and B. Fernando, "Action anticipation using pairwise human-object interactions and transformers," *IEEE Transactions on Image Processing*, 2021.

[10] Y. Zhu, A. Fathi, and L. Fei-Fei, "Reasoning about object affordances in a knowledge base representation," in *ECCV 2014*.

[11] S. Ghosh, T. Aggarwal, M. Hoai, and N. Balasubramanian, "Text-derived knowledge helps vision: A simple cross-modal distillation for video-based action anticipation," in *Findings*, 2022.

[12] S. Bhagat, S. Stepputtis, J. Campbell, and K. P. Sycara, "Sample-efficient learning of novel visual concepts," *ArXiv*, vol. abs/2306.09482, 2023.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17.

[14] D. Gong, J. Lee, M. Kim, S. J. Ha, and M. Cho, "Future transformer for long-term action anticipation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[15] K. Marino, R. Salakhutdinov, and A. Gupta, "The more you know: Using knowledge graphs for image classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Y. A. Farha, A. Richard, and J. Gall, "When will you do what? - anticipating temporal occurrences of activities," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[17] Y. Abu Farha and J. Gall, "Uncertainty-aware anticipation of activities," *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*.

[18] Q. Ke, M. Fritz, and B. Schiele, "Time-conditioned action anticipation in one shot," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[19] Y. Abu Farha, Q. Ke, B. Schiele, and J. Gall, "Long-term anticipation of activities with cycle consistency," *Pattern Recognition*, 2021.

[20] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*.

[21] H. Kuehne, A. Arslan, and T. Serre, "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.

[22] Z. Zhong, M. Martin, M. Voit, J. Gall, and J. Beyerer, "A survey on deep learning techniques for action anticipation," *ArXiv*, vol. abs/2309.17257, 2023.

[23] X. Hu, J. Dai, M. Li, C. Peng, Y. Li, and S. Du, "Online human action detection and anticipation in videos: A survey," *Neurocomputing*.

[24] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.

[25] S. Bhagat, S. Uppal, Z. Yin, and N. Lim, "Disentangling multiple features in video sequences using gaussian processes in variational autoencoders." in *European Conference on Computer Vision (ECCV)*, 2020.

[26] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: A review," *J. Hum.-Robot Interact.*, 2017.

[27] E. Aghapour and J. A. Farrell, "Human action prediction for human robot interaction," in *2016 American Control Conference (ACC)*, 2016.

[28] H. Liu, S. Nasiriany, L. Zhang, Z. Bao, and Y. Zhu, "Robot learning on the job: Human-in-the-loop autonomy and learning during deployment," *ArXiv*, vol. abs/2211.08416, 2022.

[29] Z. Li, K. Xu, L. Liu, L. Li, D. Ye, and P. Zhao, "Deploying offline reinforcement learning with human feedback," *ArXiv*, vol. abs/2303.07046, 2023.

[30] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[31] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" in *International Conference on Learning Representations*, 2022.

[32] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[33] H. Zhang, X. Li, and L. Bing, "Video-llama: An instruction-tuned audio-visual language model for video understanding," *arXiv preprint arXiv:2306.02858*, 2023.

[34] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the seventh IEEE international conference on computer vision*, 1999.