Knowledge-Guided Short-Context Action Anticipation in Human-Centric Videos

Sarthak Bhagat, Simon Stepputtis, Joseph Campbell, Katia Sycara
The Robotics Institute, Carnegie Mellon University
{sarthakb, sstepput, jacambe, sycara}@andrew.cmu.edu

Abstract

This work focuses on anticipating long-term human actions, particularly using short video segments, which can speed up editing workflows through improved suggestions while fostering creativity by suggesting narratives. To this end, we imbue a transformer network with a symbolic knowledge graph for action anticipation in video segments by boosting certain aspects of the transformer's attention mechanism at run-time. Demonstrated on two benchmark datasets, Breakfast and 50Salads, our approach outperforms current state-of-the-art methods for long-term action anticipation using short video context by up to 9%.

1. Introduction

The ability to predict which actions may happen after a particular video segment ends has multiple use cases in video understanding, including video production and editing [20, 41, 6, 38]. This work focuses on anticipating actions from short video segments and provides potential avenues to enhance the editing process. In particular, the ability to extract actions from a video segment can be utilized in two manners: 1) It allows for intelligent clip suggestions for future editing, namely the ability to suggest videos given what will likely happen next, and 2) it provides information on what *generally would happen*, which allows editors to refine their composition to either confirm or contradict a viewer's expectation. As a particular challenge, our work addresses video understanding in the context of short video segments that only span seconds to predict which actions will happen for how long after the end of the segment [3].

This work presents a novel approach to action anticipation from video segments by combining symbolic domain knowledge with the video comprehension capabilities of transformer-based architectures [33]. While such transformers often require a long context window to comprehend the underlying scene [42], extracting the information about relevant objects and associating them with the possible set of actions that could be taken with them enables us to make inferences about future actions even with little context. We propose to augment the attention mechanism [37] of transformer-based action-anticipation architectures by utilizing symbolic domain knowledge to boost or suppress the attention given to various features presented in the video. This allows us to predict future actions accurately, particularly from short-horizon observations – a key aspect that prior works [11, 1, 22, 34, 2, 17] in action anticipation fail to cater to.

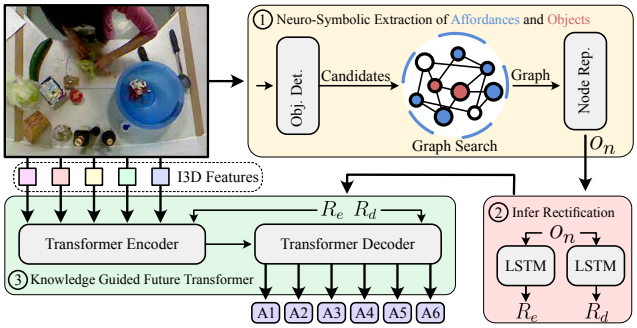

Figure 1. Overview of our proposed Knowledge Guided Action Anticipation approach.

In our work, we utilize Knowledge Graphs (KG) to capture the relationship between entities present in the video and link them to their respective affordances and the potential tools that could be used to afford them in a particular way. Prior work [40, 26, 23, 15] has introduced efficient methods of identifying such relationships, which can subsequently be utilized to identify the potential for certain actions. For example, in a kitchen scene, seeing a tomato and a knife can enable the potential action of cutting the tomato as a tomato has the affordance of *can be cut* while a knife affords the ability to *cut*. Intuitively, we utilize such knowledge to alter the attention mechanism of transformer-based action-anticipation methods to re-focus what features the transformer pays attention to when predicting the list and duration of future actions.

We demonstrate our method on two common long-term action-anticipation benchmarks, namely the 50Salads [36] and Breakfast [27] datasets, and show superior performance as compared to current state-of-the-art methods.

2. Related Works

Concept Learning. The emerging field of relevant concept extraction from visual inputs involves identifying and extracting relevant visual and non-visual entities from the input data [21, 25, 24, 7, 30]. [29] 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, [5] 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 [5] to identify relevant object affordances. We then utilize these concepts to predict a possible set of actions in a video based on a short previous context.

Action Anticipation. The task of action anticipation from videos [19] 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 [9, 10], 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 [13, 14, 31, 35, 34, 12, 16, 32]. 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 [11, 1, 22, 34, 2, 17]. 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. Particularly in this challenging domain, a deep understanding of the scene objects and their contextual utilization can be helpful.

3. Knowledge Guided Action Anticipation

In this section, we introduce the two main components of our approach: 1) a transformer-based architecture to extract future actions based on [17] (see Section 3.2) and 2) a neuro-symbolic architecture to extract affordances from video frames (see Section 3.3). Subsequently, in Section 3.4, we introduce how the domain knowledge from our neuro-symbolic pipeline can be utilized to re-focus the attention of the video transformer by introducing a correction matrix to its attention mechanism.

3.1. Problem Setup

Given an observed video segment $\mathbf{V}_O$, we create a policy $\mathbf{a} = \pi(\mathbf{V}_O)$ that predicts a sequence of actions $\mathbf{a}$ happening after the end of $\mathbf{V}_O$. The video sequence $\mathbf{V}_O \in \mathcal{R}^{H\times W\times C\times N}$ 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 frames N.

We train our model from a dataset $\mathcal{D} = [\mathbf{s}_n, \ldots, \mathbf{s}_N]$ where each sample $\mathbf{s}_n = [\mathbf{V}_n, a_n]$ contains the video-frame $\mathbf{V_n}$ and action-label $\mathbf{a}_n$, where n is the frame index. After training, we provide the trained policy π with a new, previously unseen video sequence showing α percent of the full video, tasked with predicting the most likely action for each frame in the following β percent of the remaining video.

Figure 1 demonstrates the main components of our approach across our three-step process. In the following sections, we will detail the Future Transformer (FUTR) [17] (Section 3.2), and how we extract symbolic knowledge for a given video segment using a neuro-symbolic architecture [5] (Section 3.3). Subsequently, we will detail how transformer architecture can be enhanced by imbuing it with symbolic knowledge (Section 3.4).

3.2. Future Transformer (FUTR)

Future Transformer (FUTR) [17] is a long-term action anticipation transformer architecture that consists of a transformer encoder and a transformer decoder. The encoder is responsible for processing visual features extracted from the observed segment of a video by employing multi-head self-attention [37]. To accomplish this, the encoder, f_e, employs a stack of L_e layers, each containing multi-head attention mechanisms, layer normalization [4], and feed-forward networks, all interconnected through residual connections [18], i.e., $\mathbf{x}_n^{L_e} = f_e(\mathbf{x}_n^0)$ for $\mathbf{x}_n^0 = \mathbf{W}_f\mathbf{x}_n^f \in \mathcal{R}^D$, where $\mathbf{x}_n^f$ are the I3D [8] features for the n^{th} frame of the observed video $\mathbf{V}_O$, D represents the encoding dimension, and $\mathbf{W}_f$ is a learnable weight matrix. The resulting output is then provided to a classifier, f_{obs}, which is a fully connected layer followed by a softmax activation function, determining the actions corresponding to the observed part of the video segment: $\mathbf{a}_{obs} = f_{obs}(\mathbf{x}_n^{L_e})$.

The decoder employs a cross-attention mechanism that uses the embeddings of the observed sequence generated by the encoder. After processing through L_d layers of the decoder, architecturally similar to the encoding ones, the output, $\mathbf{q}_n^{L_d}$, is obtained, i.e., $\mathbf{q}_n^{L_d} = f_d(\mathbf{x}_n^{L_e})$. Subsequently, we utilize two separate, fully connected networks for predicting the future actions $\mathbf{a}_{pred}$ and their durations $\mathbf{d}_{pred}$ respectively.

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

Putting both these components together, long-term relationships between past and future actions are learned via this transformer framework. The loss for the transformer is comprised of the framewise cross-entropy loss for the observed action sequence combined with the L2 and cross-entropy loss over the durations and actions of the predicted action sequence, respectively. With a defined methodology for predicting future actions from video segments in place,

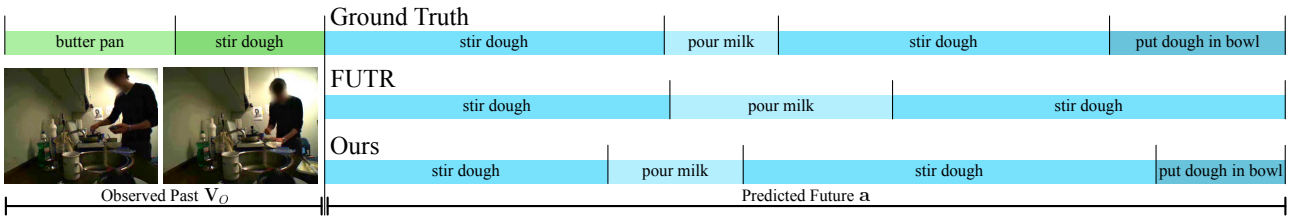

Figure 2. Example of our method on the *Breakfast* dataset, observing 10% (green hues) and predicting another 30% (blue hues).

the subsequent sections delve into the process of enhancing it with symbolic domain knowledge.

3.3. Visual Concept Extraction

To extract objects and affordances from video frames, we utilize [5], extracting these features by utilizing a domain-specific knowledge graph. Intuitively, this approach utilizes a neural object detector to extract a set of initial objects and subsequently utilizes them to initialize the knowledge graph $\mathcal{K}$. We created the graph $\mathcal{K}$ consisting of two types of nodes: object nodes (e.g., *salt*, *knife*, *bowl*) and affordance nodes (e.g., *graspable*, *pourable*, *cuttable*). The knowledge graph connects each object to its respective affordance, and each affordance is linked to its corresponding tool. For example, a *tomato* has a connection to *cuttable*, which, in turn, connects to *knife*. Then, we employ a graph-search approach to propagate information from the initial nodes to relevant connected nodes throughout an iterative process. Finally, we generate a latent representation of all relevant nodes $\mathbf{o}_n$.

Consider, $\mathbf{c}_A^k$ and $\mathbf{c}_C^k$ represent the active and the candidate nodes respectively at the k^{th} propagation step, for $k \in \{0, 1, ..., T\}$. We start by running an object detection algorithm [28], f_{OD}, to identify objects in the scene for the n^{th} frame. These objects serve as the initial nodes in our graph for the propagation process, so, $\mathbf{c}_A^0 = f_{OD}(\mathbf{V_n})$. At each step of propagation, the neighbors of the previously active nodes in $\mathcal{K}$ are considered candidate nodes. The importance network, f_I, computes the importance of each candidate node, conditioned on the visual input, and finally, the ones above a certain importance threshold, γ, are expanded.

$$\mathbf{c}_A^k = \{\mathbf{c}_C^k \mid f_I(\mathbf{c}_C^k) > \gamma\} \tag{2}$$

Through a series of T propagations, we determine the associated affordances for each object in the scene and the potential tools that can be used for those affordances. For example, if a *tomato* is detected, we associate it with the affordance of being *cuttable* and link it to the tool *knife* for cutting. This helps us understand the range of actions possible in this setting. Each identified object, affordance, and the respective tool to perform this affordance is then passed onto the context network, f_C, to obtain the vectors corresponding to each finally active node in the KG, i.e., $\mathbf{o}_n = f_C(\mathbf{c}_A^T)$. The collection of these context vectors, $\mathbf{o}_n$, is then utilized to modify the attention weights in the following steps of our approach.

3.4. Knowledge Guided Attention Mechanism

In this section, we introduce our main contribution, which is a methodology for augmenting the attention mechanism of transformers with symbolic knowledge contained within a KG. Specifically, we augment FUTR with a KG that contains various domain-specific commonsense relations that enable our model to identify the objects present in the scene, establish connections with their respective affordances, and consequently identify the most plausible set of actions that can be performed in the current context. This information is used to modify the weight of each visual feature in the attention.

Having identified the set of concepts in the scene, we can leverage them to adjust our attention weights. This modification allows our model to prioritize the features associated with objects having relevant affordances, giving them higher importance compared to those not present in the scene. We obtain a separate knowledge-guided rectification matrix for our encoder and decoder, namely $\mathbf{R}_e$ and $\mathbf{R}_d$ respectively. This is obtained by processing the context vectors, $\mathbf{o}_n$, through our LSTM-based knowledge-guided rectification functions, $\mathbf{R}_{e/d} = f_{KG}^{e/d}(\mathbf{o}_n)$. This function takes in the zero-padded context vectors and outputs a matrix that signifies the weight we need to assign to each visual feature. Therefore, we modify the standard multi-head attention [37] to obtain our knowledge-guided attention separately for our transformer encoder and decoder.

$$\text{KG-Attn}_{e/d}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\big(\frac{\mathbf{Q}\mathbf{R}_{e/d}\mathbf{K}^T}{\sqrt{d_k}}\big)\mathbf{V} \tag{3}$$

These rectification matrices modify the weights of attention so that features corresponding to the objects with relevant affordances are weighted according to the likelihood of an action being performed with them.

4. Experiments

In this section, we present the assessment of our proposed neuro-symbolic action anticipation method in comparison to the existing benchmark utilized for this task.

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 [5]	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 [39]	- / 6.44	- / 7.20	- / 7.90	- / 9.12	- / 6.12	- / 6.80	- / 7.86	- / 9.02	6.0	7.0
FUTR [17]	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
Ours	**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**
Breakfast										
KG Baseline [5]	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 [39]	- / 11.20	- / 12.24	- / 13.62	- / 13.82	- / 11.08	- / 12.04	- / 12.98	- / 13.22	5.39	9.80
FUTR [17]	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**
Ours	**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

Table 1. Performance of the proposed approach compared to the current SOTA for different values of α and β percent (top row). We also compare to Video-Llama, a multimodal fusion model utilizing large language models, on the action-sequence predictions task.

Datasets. We evaluate the effectiveness of our approach using two publicly available benchmark datasets for action anticipation in kitchen-based videos: the 50Salads dataset [36] with its five splits, densely annotated with 17 fine-grained action labels and 3 high-level activities, and the Breakfast dataset [27] with four splits, categorizing each frame into one of 10 breakfast-related activities using a comprehensive set of 48 fine-grained action labels.

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. To quantify the ability of our model to identify the sequence of 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 and the ground truth order action sequence. Finally, we also employ immediate single next-action prediction as a metric.

Quantitative Evaluation. We compare our approach against a KG-only baseline [5], an LLM-based baseline [39], and the current state-of-the-art in long-term action anticipation [17], where each metric is averaged over all the splits. The details about the implementation of these baselines are available in the supplementary. As can be viewed in Table 1, our approach 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 seven out of the ten metrics we used on the Breakfast dataset. On the MoC metric, we outperform the baseline by up to 9% on 50Salads and 1% on Breakfast.

Qualitative Evaluation. We also showcase an example to compare our approach against [17] by looking at the time-series segmentation of the predicted future actions. Figure 2 depicts an example where the model observes two actions in the 10% observed segment of the video and then tries to predict what actions take place in the next 30% of the video. While our model identifies all four actions along with their durations accurately, the baseline approach fails to identify the last action of *pour the dough into the bowl*.

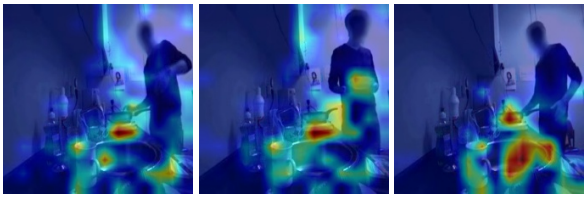

Figure 3. Heatmaps highlighting visual features where the attention is boosted as a result of our knowledge-guided rectification.

This can be attributed to the ability of our approach to focus on the set of objects present in the scene (specifically, the *bowl*), associate them with their respective affordances (of being able to *contain* the dough) and therefore, identify the possible set of actions that could be taken in the future (i.e, *pouring the dough in the bowl*).

As illustrated in heatmaps depicted in Figure 3 (right), our approach demonstrates the capability to focus not only on currently utilized objects but also on those with potential for future use. During the execution of the *stirring the dough* action, the approach enhances features corresponding to both the *pan* and the *bowl*. This specific instance is intriguing as our approach accurately recognizes that the subsequent action after *stirring the dough* would likely involve *placing the dough into the bowl*.

5. Conclusion and Future Work

Our novel knowledge-guided action anticipation approach considers both objects and their affordances in the scene, augmenting it with commonsense relations via a KG. By supplementing the action anticipation transformer with a KG, our methodology outperforms the current state-of-the-art long-term action anticipation in videos with short video contexts, establishing a new benchmark for this task.

Acknowledgements. We would like to acknowledge the support from DARPA under grant HR001120C0036, 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] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, Oct 2019. 1, 2

[2] Yazan Abu Farha, Qiuhong Ke, Bernt Schiele, and Juergen Gall. Long-term anticipation of activities with cycle consistency. *Pattern Recognition*, page 159–173, 2021. 1, 2

[3] Mohammad Sadegh Ali Akbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. Encouraging lstms to anticipate actions very early. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 280–289, 2017. 1

[4] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016. 2

[5] Sarthak Bhagat, Simon Stepputtis, Joseph Campbell, and Katia P. Sycara. Sample-efficient learning of novel visual concepts. *ArXiv*, abs/2306.09482, 2023. 2, 3, 4

[6] Sarthak Bhagat, Shagun Uppal, Zhuyun Yin, and Nengli Lim. Disentangling multiple features in video sequences using gaussian processes in variational autoencoders. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[7] Kaidi Cao, Maria Brbic, and Jure Leskovec. Concept learners for few-shot learning. In *International Conference on Learning Representations*, 2021. 2

[8] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017. 2

[9] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2

[11] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? - anticipating temporal occurrences of activities. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 1, 2

[12] Basura Fernando and Samitha 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)*, Jun 2021. 2

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

[14] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 2

[15] Sayontan Ghosh, Tanvi Aggarwal, Minh Hoai, and Niranjan Balasubramanian. Text-derived knowledge helps vision: A simple cross-modal distillation for video-based action anticipation. In *Findings*, 2022. 1

[16] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2021. 2

[17] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3052–3061, 2022. 1, 2, 4

[18] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 2

[19] Xuejiao Hu, Jingzhao Dai, Ming Li, Chenglei Peng, Yang Li, and Sidan Du. Online human action detection and anticipation in videos: A survey. *Neurocomputing*, 491:395–413, 2022. 2

[20] Matthew Hutchinson and Vijay Gadepally. Video action understanding. *IEEE Access*, 9:134611–134637, 2020. 1

[21] Yangqing Jia, Joshua T Abbott, Joseph L Austerweil, Tom Griffiths, and Trevor Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 2

[22] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9917–9926, 2019. 1, 2

[23] Khimya Khetarpal, Zafarali Ahmed, Gheorghe Comanici, David Abel, and Doina Precup. What can i do here? a theory of affordances in reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020. 1

[24] Taesup Kim, Sungwoong Kim, and Yoshua Bengio. Visual concept reasoning networks. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[25] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. 2020. 2

[26] Hema S. Koppula and Ashutosh Saxena. Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):14–29, 2016. 1

[27] Hilde Kuehne, Ali Arslan, and Thomas 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)*, June 2014. 1, 4

[28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun

Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[29] 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)*, pages 20–28, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society. 2

[30] Lingjie Mei, Jiayuan Mao, Ziqi Wang, Chuang Gan, and Joshua B. Tenenbaum. FALCON: Fast visual concept learning by integrating images, linguistic descriptions, and conceptual relations. In *International Conference on Learning Representations*, 2022. 2

[31] Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, and Du Tran. Leveraging the present to anticipate the future in videos. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2915–2922, 2019. 2

[32] Debaditya Roy and Basura Fernando. Action anticipation using pairwise human-object interactions and transformers. *IEEE Transactions on Image Processing*, 30:8116–8129, 2021. 2

[33] Javier Selva, Anders S. Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas Baltzer Moeslund, and Albert Clap'es. Video transformers: A survey. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2022. 1

[34] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 154–171, Cham, 2020. Springer International Publishing. 1, 2

[35] Fadime Sener and Angela Yao. Zero-shot anticipation for instructional activities. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019. 2

[36] Sebastian Stein and Stephen 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*, UbiComp '13, page 729–738, New York, NY, USA, 2013. Association for Computing Machinery. 1, 4

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc. 1, 2, 3

[38] Chaoxia Wu and Philipp Krähenbühl. Towards long-form video understanding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1884–1894, 2021. 1

[39] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 4

[40] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 408–424, Cham, 2014. Springer International Publishing. 1

[41] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R. Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *ArXiv*, abs/2012.06567, 2020. 1

[42] Atabay Ziyaden, Amir Yelenov, and Alexandr Pak. Long-context transformers: A survey. In *2021 5th Scientific School Dynamics of Complex Networks and their Applications (DCNA)*, pages 215–218, 2021. 1