InteractFormer: Modeling Agent Interactions for Multi-Agent Action Anticipation

Yiqi Jin, Simon Stepputtis, Katia Sycara, Yaqi Xie
Carnegie Mellon University
{yiqij, sstepput, sycara, yaqix}@andrew.cmu.edu

Abstract

Action anticipation from an ego-centric perspective plays a crucial role in embodied intelligence and vision-based assistants. While previous works have extensively explored single-agent ego-centric action anticipation, multi-agent scenarios—where collaboration among multiple individuals represents a more common real-world situation—have received significantly less attention. In multi-agent settings, since agents collaborate to complete tasks, their actions exhibit both momentary and cross-temporal correlations, making it essential to leverage these relationships effectively. Our work, InteractFormer, focuses on multi-agent scenarios and models the inherent correlations that exist when multiple agents cooperate to complete tasks, jointly predicting future actions across all agents. By capturing the relationships between agents and incorporating visual cross attention, our approach enables more accurate anticipation of collaborative behaviors. Extensive quantitative experiments across various indoor tasks show that our method outperforms state-of-the-art techniques. Moreover, attention visualizations highlight the effectiveness and interpretability of our interaction modeling approach, offering valuable insights into collaborative behavior anticipation.

1. Introduction

Human Action Anticipation, the task of predicting future actions before they are fully executed, is crucial for enhancing the responsiveness, safety, and interactivity of intelligent systems [22]. By enabling proactive decision-making, action anticipation plays a vital role in applications such as autonomous driving, human-robot collaboration, smart surveillance, and human-computer interaction. [8, 18, 27]

Single-agent action anticipation has seen extensive progress enabled by a variety of large-scale egocentric and third-person datasets such as EPIC-KITCHENS[6], EGTEA Gaze+[24], 50-Salads[32], and Ego4D[16], which provide rich annotations of temporally aligned actions. Early methods relied on RNN-based architectures, including LSTMs and GRUs, to model the sequential structure of video observations [1, 2, 9, 11, 23, 25]. More recent approaches leverage Transformer-based architectures to capture long-range spatial-temporal dependencies [12, 13, 15, 20, 31, 34, 35]. With the emergence of foundation models, concurrent works have adopted large language models (LLMs) and video-language models (VLMs) to generate diverse and temporally plausible action sequences [21, 28, 33, 37, 38]. Beyond architectural improvements, there is a growing trend toward exploiting structured or semantic information: some methods construct relational graphs to model interactions between actors and objects, while some explicitly detect interactable objects or model high-level goals to better guide anticipation. These diverse lines of research collectively contribute to the evolving landscape of single-agent action anticipation [3, 4, 17, 26, 29, 30, 36].
Multi-Agent Action Anticipation, however, remains underexplored, with significantly fewer datasets [19] and specialized methods. Unlike the single-agent setting, anticipating the actions of multiple interacting agents introduces additional challenges such as modeling inter-agent dependencies, joint intention understanding, and social plausibility. Only a handful of recent works explicitly address multi-agent anticipation, and they often rely on adaptations of single-agent models without explicitly modeling the relational among agents [34]. This reveals a significant gap in current research and underscores the need for more dedicated benchmarks and modeling approaches tailored to the multi-agent setting.

To address this gap, we propose **InteractFormer**, a model designed to capture both *within-timestep* and *across-timestep* interactions among agents, and to jointly predict future actions of all agents. Specifically, as shown in Fig. 1, we introduce a cross-agent visual attention module that operates across the agent dimension at each time step. Intuitively, interactions between agents are often most directly reflected in the visual domain—for example, through gaze, gesture, or object manipulation—and raw visual in-

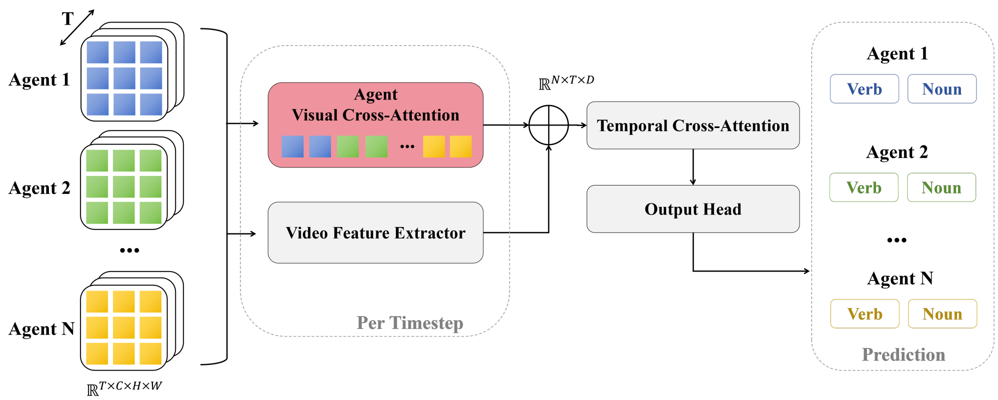

Figure 1. Overview of **InteractFormer**. Given video inputs from multiple agents, the model first applies an agent visual cross-attention module to capture spatial interactions across agents at each time step. The fused features are then processed by a temporal cross-attention module to capture temporal dynamics, enabling joint prediction of future actions.

puts provide richer spatial context than pre-extracted features. Our attention mechanism leverages this, and its attention heatmaps offer interpretable insights into how the model attends to different agents in context.

The output of this visual module is then fused with video features extracted by a pretrained backbone, such as I3D [5]. We further apply a temporal cross-attention module along the time axis. Since each timestep's features have already been enriched by multi-agent visual interaction, this temporal module effectively captures cross-agent, cross-time correlations. Through this two-stage pipeline, InteractFormer models both simultaneous and temporally-evolving interactions among agents, enabling accurate joint future action prediction.In summary, our main contributions are as follows:

- We address the multi-agent action anticipation task by modeling inter-agent interactions, both within and across time, via a simple yet effective architecture, InteractFormer, that incorporates structured attention over raw visual inputs and temporal sequences.
- Extensive experiments demonstrate that our method significantly outperforms prior baselines in multi-agent settings.
- The cross-agent attention module yields interpretable and meaningful patterns in its learned attention maps.

2. Method

2.1. Problem Formulation

We follow most single-agent action anticipation works and define our objective as predicting the future action of agents (consisting of verb and noun) at time t_{interval} after observing video inputs of length t_{obs}.

In our work, we jointly predict the future actions of all agents using the inputs from all agents simultaneously. Formally, given the ego-centric video observations $\mathcal{V} = \{V_1, V_2, \dots, V_{t_{\text{obs}}}\}$ from multiple agents, we aim to predict their future actions $\mathcal{A} = \{a_1, a_2, \dots, a_N\}$ at time $t_{\text{obs}} + t_{\text{interval}}$, where N is the number of agents and each action $a_i = (v_i, n_i)$ consists of a verb v_i and a noun n_i.

2.2. InteractFormer

As illustrated in Fig. 1, our proposed model consists of two core components designed to capture spatial and temporal interaction patterns among agents: an *Agent Visual Cross-Attention* module and a *Temporal Cross-Attention* module.

The first module explicitly models interactions among agents at each time step by applying cross-attention over raw visual inputs. The second module builds on these enriched representations and models the temporal evolution of inter-agent dynamics.

These two stages form a sequential pipeline that first encodes multi-agent spatial interactions, then captures their temporal evolution, ultimately allowing the model to jointly predict the future actions of all agents.

Agent Visual Cross-Attention: Given ego-centric video inputs $\mathcal{V} \in \mathbb{R}^{B\times N\times T\times 3\times H\times W}$, where B is the batch size, N is the number of agents, and T is the number of frames, we first extract spatial visual tokens for each agent independently via a shared patch embedding:

$$\mathbf{X}_i^{(t)} = \text{PatchEmbed}(V_i^{(t)}) \in \mathbb{R}^{P\times d}$$

where P is the number of patches per frame and d is the embedding dimension.

At each time step t, we compute attention for agent i over the visual tokens of all agents at the same frame. Specifically, for each agent i, its query Q_i is obtained from its own patch tokens $\mathbf{X}_i^{(t)}$, while the key and value matrices K, V are constructed by concatenating the patch tokens from all agents:

Method	mAP ↑ / Top-1 Acc. ↑							
	1×1		1×2		2×1		2×2	
	Verb	Noun	Verb	Noun	Verb	Noun	Verb	Noun
I3D [19]	27.7 / 30.5	23.3 / 38.9	21.1 / 25.6	13.6 / 28.9	18.0 / 24.8	16.6 / 23.3	18.7 / 22.5	14.0 / 25.1
RULSTM [9]	36.9 / 35.4	28.9 / 41.4	24.2 / 29.3	16.2 / 30.4	24.7 / 34.7	20.0 / 26.2	22.8 / 29.7	17.5 / 27.0
LSTR [35]	48.6 / 43.9	37.6 / 50.8	37.1 / 39.2	24.9 / 39.3	32.4 / 41.8	24.9 / 33.8	30.6 / 39.3	22.0 / 34.6
HiMemFormer [34]	48.0 / 44.6	38.1 / 51.5	37.2 / 39.8	25.4 / 40.1	32.4 / 41.7	24.6 / 32.9	30.8 / 39.6	22.6 / 34.2
Ours	49.4 / 45.5	37.9 / **52.0**	**38.3** / 39.6	25.6 / **40.3**	**36.1** / **44.4**	**27.6** / **35.6**	**33.4** / **40.3**	**24.8** / **36.5**
Ours w/o TPV input	48.4 / 45.2	37.9 / 50.2	38.2 / 39.5	25.3 / 39.3	34.9 / 43.6	27.0 / 34.7	31.5 / 39.4	23.8 / 35.1
Ours w/o agent visual attention	47.9 / 44.2	38.1 / 51.5	36.5 / 39.6	24.9 / 40.0	32.3 / 40.1	24.1 / 32.0	30.5 / 37.9	22.3 / 34.1
Ours agent feature attention	**49.8** / **45.7**	**38.3** / 51.9	37.8 / **39.8**	25.5 / 38.8	35.5 / 43.7	27.2 / 33.6	32.8 / 40.0	24.4 / 34.8

Table 1. Performance of different methods on the LEMMA dataset [19]. We report the mAP and Top-1 Acc. across four scenarios described in Section 3.1. All results are averaged over five runs. **We consider mAP to be a more reliable metric**, as the distribution of verbs and nouns in the ground truth is imbalanced.

$$Q_i = W_Q \mathbf{X}_i^{(t)}$$
$$K = W_K \left[\mathbf{X}_1^{(t)}; \dots; \mathbf{X}_N^{(t)}\right] \quad V = W_V \left[\mathbf{X}_1^{(t)}; \dots; \mathbf{X}_N^{(t)}\right]$$

We then compute standard scaled dot-product attention with h heads:

$$\text{Attention}(Q_i, K, V) = \text{softmax}\left(\frac{Q_i K^\top}{\sqrt{d/h}}\right) V$$

The updated tokens are passed through a feed-forward network and residual layers, and the resulting patch tokens are aggregated (e.g., by average pooling) to yield a visual representation $\mathbf{z}_i^{(t)} \in \mathbb{R}^D$ for each agent at each time step.

This cross-agent attention mechanism enables each agent to selectively attend to the relevant visual features of all others, dynamically learning their interactions from raw visual input.

Finally, we fuse the feature sequence $\mathbf{F} \in \mathbb{R}^{B\times N\times T\times D}$ (e.g., extracted by a pretrained I3D model [5]) and the output of the visual cross-attention module $\mathbf{V} \in \mathbb{R}^{B\times N\times T\times D}$ via element-wise addition:

$$\mathbf{F}' = \mathbf{F} + \mathbf{Z}$$

This representation is then used as input to the subsequent temporal modeling stage.

Temporal Cross-Attention: Given the fused representation $\mathbf{F}' \in \mathbb{R}^{B\times N\times T\times D}$ from the previous stage, we reshape the input per agent and apply multi-head self-attention across the temporal dimension. Each agent's feature sequence $\{\mathbf{f}'^{(1)}, \dots, \mathbf{f}'^{(T)}\}$ is treated as a set of tokens, and standard attention with residual connections and a feed-forward network is used to capture temporal dependencies.

This module enables each agent to reason over the temporal evolution of not only its own behavior but also its interactions with others. Since the input features already encode inter-agent visual context, the temporal attention models how these cross-agent relationships develop over time, thereby capturing motion patterns, coordination cues, and short-term intention dynamics in a multi-agent setting.

3. Experiments

3.1. Experimental Setup

We validate our method using the LEMMA dataset [19]. LEMMA contains 324 multi-agent activity videos captured from both third-person view (TPV) and first-person view (FPV) of each agent, with well-annotated compositional atomic actions. The dataset is organized into four scenarios: single-agent single-task (1×1), single-agent multi-tasks (1×2), multi-agent single-task (2×1), and multi-agent multi-tasks (2×2) videos. Following previous works [6, 7, 10, 14, 34, 35], we decompose action prediction into verb and noun prediction, and evaluate performance using mean Average Precision (mAP) and Top-1 Acc. metrics.

3.2. Implementation Details

We set the observation duration t_{obs} to 16 seconds and the anticipation interval t_{interval} to 1 second. Following prior work, we use a pretrained I3D model [5] as the video feature extractor. As for the input, we use the TPV and all agents' FPVs as the input to the agent channel. All attention modules in the model use 8 heads and a hidden size of 1024. For the visual cross-attention module, the patch size is set to 32. We train the model using the AdamW optimizer with a constant learning rate of 5×10^{-5} and a weight decay of 5×10^{-3}. All experiments are conducted on a single NVIDIA RTX 6000 GPU.

3.3. Results and Discussion

Quantitative Results: We compare our method with I3D baseline provided in LEMMA [5, 19], single-agent RNN-based and Transformer-based action anticipation methods [34, 35]. Note that LSTR[35] and HiMemFormer[34] are online models that see ground-truth labels after each prediction. We apply their methods to our offline task, and as expected, their performance is noticeably lower than reported in the original paper[34]. Nevertheless, the comparison remains fair under our setting. As shown in Tab. 1, our method outperforms the baselines across all four scenarios. Notably, the improvements are more significant in the multi-agent scenarios of 2×1 and 2×2. This suggests that by explicitly modeling multi-agent interactions, our approach excels in predicting multi-agent behaviors, demonstrating the effectiveness and superiority of our method.

Attention Visualization: We visualize the attention of the agent cross-attention module to explore the model's understanding of interactions among multiple agents. Figure 2 presents two examples under the 2×1 scenario, showing the Attention heatmaps for Agent1 (center perspective) with respect to the inputs from TPV, Agent1, and Agent2 perspectives (from left to right). In the two observed videos, the agents are performing either cooperative actions or individual actions.

In the upper example, Agent1 is about to receive the cutting board from Agent2. The model assigns high attention weights to both agents' perspectives, visually focusing on Agent2 and the object in hand. In the bottom example, the two agents are currently doing separate actions, and Agent1 focuses mainly on its own perspective. This demonstrates the model's ability to comprehensively consider the relationships between multiple agents' inputs and interaction scenarios, validating the effectiveness of our approach. See more examples in the supplementary material.

Ablation Study: We compare the full model with three variants: (1) removing TPV input, (2) removing the agent visual cross-attention module, and (3) replacing agent visual cross-attention with agent-wise attention applied on extracted video features. The results are shown in Tab. 1.

We observe that adding TPV input brings only marginal improvements. This is expected, as TPV cameras in the LEMMA dataset [19] are often distant and suffer from occlusions (see Fig. 2), providing limited additional information beyond the FPV inputs.

Removing the agent visual cross-attention module leads to performance drops across all four scenarios, with particularly notable degradation in multi-agent settings.

Interestingly, replacing the visual-level cross-attention with agent-wise attention over the feature representations yields mixed results: performance improves in single-agent scenarios but decreases in multi-agent ones. This is intuitive—when only a single agent is present, the benefit of using raw visual input is limited. However, in multi-agent settings, modeling interaction at the visual level is more effective, as it captures spatial relations and inter-agent cues more directly and expressively.

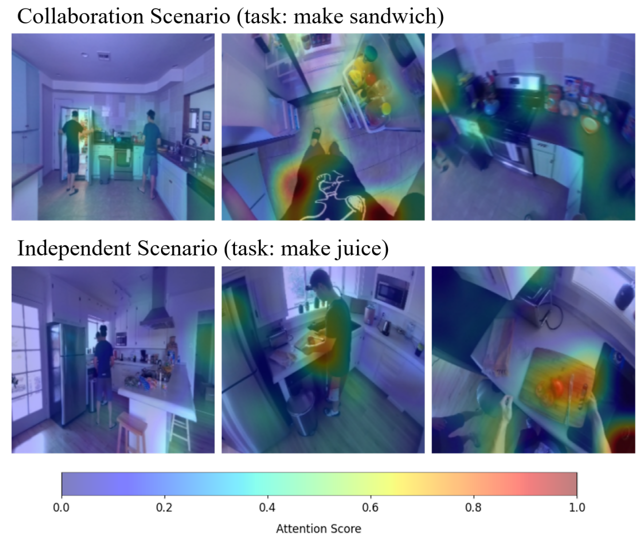

Figure 2. Attention visualizations of two examples: one where the two agents are collaborating, and the other where the two agents are working separately.

4. Conclusion and Discussion

We propose **InteractFormer**, an innovative model that captures interactions among multiple agents, leveraging the correlation of their actions during collaborative tasks to jointly predict future actions. Experiments demonstrate that our approach outperforms strong baselines on the LEMMA dataset. Moreover, our attention analysis and ablation studies validate the effectiveness and interpretability of the multi-agent interaction module.

While our results are promising, this work also highlights key limitations in current benchmarks for multi-agent action anticipation. First, existing well-annotated datasets like LEMMA contain only a small number of agents (at most two), which restricts the modeling of richer social dynamics. Second, there is a lack of explicit supervision regarding inter-agent interaction labels—our model captures interactions through architecture design, but cannot benefit from direct supervision signals.

We believe these challenges reflect broader issues in the field: the lack of large-scale, high-quality datasets that capture complex, multi-agent collaborative behaviors remains a fundamental bottleneck. In the future, we envision applying InteractFormer to more diverse scenarios with richer interaction structures, enabled by the emergence of better benchmarks. We also see opportunities to incorporate explicit reasoning modules or symbolic representations once datasets with annotated inter-agent relations become available.

Acknowledgements

This work has been funded in part by the Army Research Laboratory (ARL) award W911NF-23-2-0007 and W911QX-24-F-0049, DARPA award FA8750-23-2-1015, and ONR award N00014-23-1-2840.

References

[1] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1

[2] Yazan Abu Farha, Qiuhong Ke, Bernt Schiele, and Juergen Gall. Long-term anticipation of activities with cycle consistency. In *Pattern Recognition: 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, September 28–October 1, 2020, Proceedings 42*, pages 159–173. Springer, 2021. 1

[3] Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. Hiervl: Learning hierarchical video-language embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23066–23078, 2023. 1

[4] Sarthak Bhagat, Samuel Li, Joseph Campbell, Yaqi Xie, Katia Sycara, and Simon Stepputtis. Let me help you! neuro-symbolic short-context action anticipation. *IEEE Robotics and Automation Letters*, 2024. 1

[5] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset, 2018. 2, 3

[6] 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, 2018. 1, 3

[7] Dima Damen, Hazel Doughty, and Giovanni Maria et al. Farinella. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 3

[8] Iram Fatima, Muhammad Fahim, Young-Koo Lee, and Sungyoung Lee. A unified framework for activity recognition-based behavior analysis and action prediction in smart homes. *Sensors*, 13(2):2682–2699, 2013. 1

[9] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1, 3

[10] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *Proceedings of the European conference on computer vision (ECCV) workshops*, pages 0–0, 2018. 3

[11] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Predicting the future: A jointly learnt model for action anticipation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5562–5571, 2019. 1

[12] Harshayu Girase, Nakul Agarwal, Chiho Choi, and Karttikeya Mangalam. Latency matters: Real-time action forecasting transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18759–18769, 2023. 1

[13] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021. 1

[14] Rohit Girdhar and Kristen Grauman. Anticipative video transformer, 2021. 3

[15] 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

[16] Kristen Grauman, Andrew Westbury, and Erin et al. Byrne. Ego4d: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[17] Hongji Guo, Nakul Agarwal, Shao-Yuan Lo, Kwonjoon Lee, and Qiang Ji. Uncertainty-aware action decoupling transformer for action anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18644–18654, 2024. 1

[18] Kelsey P Hawkins, Nam Vo, Shray Bansal, and Aaron F Bobick. Probabilistic human action prediction and wait-sensitive planning for responsive human-robot collaboration. In *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, pages 499–506. IEEE, 2013. 1

[19] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. Lemma: A multi-view dataset for learning multi-agent multi-task activities. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 4

[20] Qiuhong Ke, Mario Fritz, and Bernt Schiele. Time-conditioned action anticipation in one shot. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9925–9934, 2019. 1

[21] Sanghwan Kim, Daoji Huang, Yongqin Xian, Otmar Hilliges, Luc Van Gool, and Xi Wang. Palm: Predicting actions through language models. In *European Conference on Computer Vision*, pages 140–158. Springer, 2024. 1

[22] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12*, pages 201–214. Springer, 2012. 1

[23] Yu Kong, Shangqian Gao, Bin Sun, and Yun Fu. Action prediction from videos via memorizing hard-to-predict samples. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 1

[24] Yin Li, Miao Liu, and James 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)*, pages 619–635, 2018. 1

[25] Siyuan Brandon Loh, Debaditya Roy, and Basura Fernando. Long-term action forecasting using multi-headed attention-based variational recurrent neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2419–2427, 2022. 1

[26] Esteve Valls Mascaró, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6048–6057, 2023. 1

[27] Angelos Mavrogiannis, Rohan Chandra, and Dinesh Manocha. B-gap: Behavior-guided action prediction for autonomous navigation. *arXiv preprint arXiv:2011.03748*, 1 (2), 2020. 1

[28] Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, and Kwonjoon Lee. Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18580–18590, 2024. 1

[29] Himangi Mittal, Nakul Agarwal, Shao-Yuan Lo, and Kwonjoon Lee. Plausible action anticipation using large video-language models. *arXiv preprint arXiv:2405.20305*, 2024. 1

[30] Razvan-George Pasca, Alexey Gavryushin, Muhammad Hamza, Yen-Ling Kuo, Kaichun Mo, Luc Van Gool, Otmar Hilliges, and Xi Wang. Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18286–18296, 2024. 1

[31] Debaditya Roy, Ramanathan Rajendiran, and Basura Fernando. Interaction region visual transformer for egocentric action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6740–6750, 2024. 1

[32] 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*, pages 729–738, 2013. 1

[33] Shijie Wang, Qi Zhao, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Vamos: Versatile action models for video understanding. In *European Conference on Computer Vision*, pages 142–160. Springer, 2024. 1

[34] Zirui Wang, Xinran Zhao, Simon Stepputtis, Woojun Kim, Tongshuang Wu, Katia Sycara, and Yaqi Xie. Himemformer: Hierarchical memory-aware transformer for multi-agent action anticipation. In *NeurIPS Workshop on Video-Language Models*, 2024. 1, 3, 4

[35] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection, 2021. 1, 3

[36] Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, and Chen Sun. Object-centric video representation for long-term action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6751–6761, 2024. 1

[37] 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. 1

[38] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. Antgpt: Can large language models help long-term action anticipation from videos? *arXiv preprint arXiv:2307.16368*, 2023. 1