InteractFormer: Inter-Agent Spatiotemporal Attention for Multi-Agent Action Anticipation

Extended Abstract

Yiqi Jin
yiqij@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Simon Stepputtis
stepputtis@vt.edu
Virginia Polytechnic Institute and State University
Virginia, USA

Carl Busart
carl.e.busart.civ@army.mil
DEVCOM Army Research Laboratory
USA

Katia Sycara
sycara@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

Yaqi Xie
yaqix@andrew.cmu.edu
Carnegie Mellon University
Pittsburgh, PA, USA

ABSTRACT

Action anticipation in multi-agent scenarios is critical for embodied intelligence but remains under-explored compared to single-agent settings. Effective anticipation requires capturing complex inter-agent correlations across both momentary interactions and temporal evolutions. We propose *InteractFormer*, a model specifically designed to jointly predict future actions of all agents by modeling their inherent cooperation. Our approach captures fine-grained relationships through visual cross-attention and incorporates spatial bounding-box cues to ground inter-agent dynamics. Extensive experiments on two benchmarks—household collaborative tasks (LEMMA) and multi-agent sports (SportsHHI)—demonstrate that InteractFormer consistently outperforms state-of-the-art methods. Visualizations further confirm that our model provides interpretable insights into collaborative behavior.

KEYWORDS

Multi-Agent Modeling; Interaction Modeling; Action Anticipation; Human-Human Interaction Anticipation

ACM Reference Format:
Yiqi Jin, Simon Stepputtis, Carl Busart, Katia Sycara, and Yaqi Xie. 2026. InteractFormer: Inter-Agent Spatiotemporal Attention for Multi-Agent Action Anticipation: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/ZFVD4206

1 INTRODUCTION

Anticipating the intentions of multiple partners is a prerequisite for proactive assistance and safe coordination in human-robot teams [2, 5]. Unlike trajectory forecasting, we focus on *video-based action anticipation with verb-noun semantics*, which provides the critical "what/why" for high-level planning. In shared workspaces, predictions must extend beyond single-actor reasoning to account for inter-agent dependencies—such as cooperation and coordination—that are often visually entangled [3].

Despite its importance, multi-agent anticipation remains limited in both benchmarks and modeling. Most existing methods adapt single-agent architectures (e.g., LSTMs or Transformers) by processing streams independently and employing late-stage fusion [4, 7]. These approaches often overlook the rich spatial-temporal dynamics reflected in raw visual cues (e.g., gaze, shared object manipulation). Even state-of-the-art single-agent models like InAViT [6] or zero-shot VLMs like Qwen-2.5-VL [1] struggle in these settings as they fail to explicitly treat interactions as a central component of the reasoning process.

To address this, we propose **InteractFormer**, which captures inter-agent interactions directly in the visual domain. Our model utilizes a cross-agent visual attention module to allow agents to attend to each other's observations at the token level, augmented with spatial bounding-box information. These representations are then processed via temporal cross-attention to model evolving coordination. We validate InteractFormer on two distinct domains: household collaboration (LEMMA) and team sports (SportsHHI).

Our contributions: (1) We introduce *InteractFormer*, a unified model for inter-agent interaction modeling via structured visual-spatial and temporal attention. (2) We demonstrate SOTA performance across both action anticipation and interaction understanding on two complementary benchmarks. (3) We provide qualitative evidence of the model's interpretability through attention visualizations.

2 PROBLEM AND METHOD

Problem Formulation. We aim to model multi-agent spatio-temporal dynamics for unified behavior prediction. Given a video sequence with N agents over T frames, we denote visual inputs as $\mathcal{V} = \{V_i^{(t)}\}_{i=1,t=1}^{N,T}$, where $V_i^{(t)}$ represents agent-specific views (e.g., egocentric crops). Our model maps these inputs to per-agent representations to support two tasks: (1) **Action anticipation**, predicting future actions $\hat{a}_i = (v_i, n_i)$, and (2) **Interaction Anticipation**, predicting future pairwise interaction class $\hat{y}_{i,j}$ for agent pairs (i, j).

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/ZFVD4206

Method	mAP ↑ / Top-1 Acc. ↑							
	1×1		1×2		2×1		2×2	
	Verb	Noun	Verb	Noun	Verb	Noun	Verb	Noun
HiMemFormer [7]	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
Qwen-2.5-VL (zero-shot) [1]	11.8 / 20.0	11.2 / 22.8	12.1 / 22.0	7.4 / 21.3	11.7 / 19.2	10.0 / 20.7	8.8 / 17.7	7.7 / 18.6
InAViT [6]	48.2 / 45.0	38.0 / 51.2	37.5 / 39.7	25.6 / 40.0	32.6 / 41.9	24.8 / 33.0	31.0 / 39.8	22.8 / 34.1
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**

Table 1: Action Anticipation Results on LEMMA[3].

Method	Backbone	mAP	Recall@20
Anticipation			
Wu et al. (2024)[8]	SlowFast-R50	6.57	26.55
InteractFormer (ours)		7.55	30.91
Wu et al. (2024)[8]	SlowFast-R101	6.77	27.59
InteractFormer (ours)		8.01	30.24
Wu et al. (2024)[8]	ViT-B	9.32	35.26
InteractFormer (ours)		**9.82**	**37.64**

Table 2: Interaction Anticipation Results on SportsHHI[8].

InteractFormer Architecture. InteractFormer consists of two core components designed to capture inter-agent dependencies without relying on coarse pre-extracted features.

- **Agent Visual Cross-Attention:** Instead of processing agents independently, this module allows each agent to attend to the patch-level visual tokens of all other agents at each timestep. To ground these cues, we incorporate a **BBox Cross-Attention** module that conditions visual features on agent bounding box trajectories, enabling reasoning over spatial proximity and relative configurations.
- **Temporal Cross-Attention:** A temporal transformer processes these spatially-enriched features to capture evolving inter-agent dynamics, such as coordination and role-switching, ensuring socially and temporally coherent predictions.

Task-Agnostic Adaptation. InteractFormer is modular and adapts to diverse benchmarks with minimal changes. For **LEMMA** (egocentric), we treat egocentric streams as primary inputs and use the third-person view as an auxiliary stream for spatial enrichment via bounding boxes. For **SportsHHI** (third-person), we apply a ViT-based encoder on full frames and use bounding box centers for 2D positional embeddings to preserve spatial awareness in crowded team-sports scenes. The same core architecture then predicts either agent-wise actions or pairwise interaction labels.

3 EXPERIMENTAL RESULTS

Experimental Setup. We evaluate InteractFormer on two distinct multi-agent benchmarks: (1) **LEMMA** [3], focusing on household collaborative tasks with egocentric and third-person views. We follow the official setup to jointly predict the next verb-noun action for all agents. (2) **SportsHHI** [8], featuring dense human-human interactions in team sports. To align with the anticipation setting, we define a *pseudo-anticipation* task using only the first 25% of each video clip to predict the interaction label.

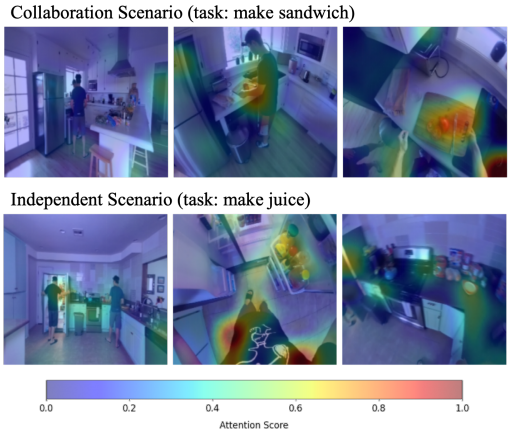

Figure 1: Attention visualizations on LEMMA[3].

Experimental Results. Results on LEMMA (Table 1) show InteractFormer outperforms Transformer [6, 7] and zero-shot VLM [1] baselines, with the most significant gains in multi-agent scenarios (2×1 and 2×2). Similarly, on SportsHHI (Table 2), our model surpasses the SOTA [8] across all backbones even with only 25% visual observation, effectively inferring interactions via early behavioral and spatial cues. Attention visualizations (Figure 1) further demonstrate interpretability; the model correctly shifts focus toward the partner's perspective and shared objects during collaboration, while prioritizing its own FPV in independent scenarios. This confirms InteractFormer's ability to selectively condition on meaningful inter-agent dynamics based on social context.

4 CONCLUSION AND FUTURE WORK

Overall, *InteractFormer* provides a unified framework for multi-agent anticipation by treating interaction modeling as a primary component of visual reasoning. Our results across diverse benchmarks highlight the importance of fine-grained spatial-temporal dependencies for accurate social perception. Moving forward, we plan to extend this architecture to handle more complex scenarios, such as joint activity modeling and long-term temporal forecasting, to further enhance its applicability in real-world perception tasks.

ACKNOWLEDGMENTS

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

REFERENCES

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923* (2025).

[2] Guy Hoffman, Tapomayukh Bhattacharjee, and Stefanos Nikolaidis. 2024. Inferring Human Intent and Predicting Human Action in Human–Robot Collaboration. *Annual Review of Control, Robotics, and Autonomous Systems* 7 (2024). https://doi.org/10.1146/annurev-control-071223-105834

[3] Baoxiong Jia, Yixin Chen, Siyuan Huang, Yixin Zhu, and Song-Chun Zhu. 2020. LEMMA: A Multi-view Dataset for LEarning Multi-agent Multi-task Activities. In *European Conference on Computer Vision (ECCV)*.

[4] Yitian Li, Qi Zhu, and Shuai Yan. 2022. MAT: Multimodal Anticipation Transformer for Action Anticipation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[5] Maithili Patel and Sonia Chernova. 2022. Proactive Robot Assistance via Spatio-Temporal Object Modeling. In *Proceedings of the 6th Conference on Robot Learning (CoRL)*. 881–891.

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

[7] Zirui Wang, Xinran Zhao, Simon Stepputtis, Woojun Kim, Tongshuang Wu, Katia Sycara, and Yaqi Xie. 2024. HiMemFormer: Hierarchical Memory-Aware Transformer for Multi-Agent Action Anticipation. In *NeurIPS Workshop on Video-Language Models*.

[8] Tao Wu, Runyu He, Gangshan Wu, and Limin Wang. 2024. Sportshhi: A dataset for human-human interaction detection in sports videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18537–18546.