Learning Collective Dynamics of Multi-Agent Systems using Event-based Vision

Minah Lee MINAH.LEE@GATECH.EDU
Uday Kamal UDAY.KAMAL@GATECH.EDU
Saibal Mukhopadhyay SAIBAL.MUKHOPADHYAY@ECE.GATECH.EDU
School of Electrical and Computer Engineering
Georgia Institute of Technology, GA, USA

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

This paper proposes a novel problem: *vision-based perception* to learn and predict the *collective dynamics* of multi-agent systems, specifically focusing on interaction strength and convergence time. Multi-agent systems are defined as collections of more than ten interacting agents that exhibit complex group behaviors. Unlike prior studies that assume knowledge of agent states, we focus on deep learning models to **directly predict collective dynamics from visual data**, captured as frames or events. Due to the lack of relevant datasets, we create a simulated dataset using a state-of-the-art flocking simulator, coupled with a vision-to-event conversion framework. We empirically demonstrate the effectiveness of event-based representation over traditional frame-based methods in predicting these collective behaviors. Based on our analysis, we present event-based vision for Multi-Agent dynamic Prediction (evMAP), a deep learning architecture designed for real-time, accurate understanding of interaction strength and collective behavior emergence in multi-agent systems. Supplementary materials, including detailed descriptions of experiments and video demonstrations, are available here.

Keywords: Multi-Agent System, Event Camera, Swarm Behavior

1. Introduction

The systems of large number (>10) of agents, hereafter referred to as a multi-agent system, are crucial in a wide range of autonomy applications, including swarm robotics (Khamis et al. (2015)) and fleets of autonomous vehicles (Rios-Torres and Malikopoulos (2017)). Inspired by collective behaviors observed in nature, such as fish schools and bird flocks (Figure 1), these systems aim to achieve collective goals through the interaction among individual agents using a set of decentralized rules. While analytical flocking models like Reynolds' (Reynolds (1987)) and Vicsek's (Vicsek et al. (1995)) replicate these behaviors, they depend on precise localization, often impractical in real-world scenarios. Therefore, *real-time prediction of collective behavior*, like how and when agents will achieve a collective goal, is essential for adapting the local rules and controlling multi-agent systems in a real-world environment (Virágh et al. (2014); Shrit and Sebag (2021)) as illustrated in Figure 2. This prediction is valuable in competitive settings like swarm herding (Chipade and Panagou (2020)), where understanding the system dynamics of adversarial agents can enhance strategic control. The prediction is crucial for optimizing resources and minimizing risks in complex operations, such as coordinating astrobots in telescopes (Macktoobian et al. (2021, 2022)), where precise maneuvering and dense formations are important. As swarm operations scale in complexity, the prediction of collective behavior becomes increasingly critical, underscoring the need for advancement in methods for learning and control of multi-agent systems.

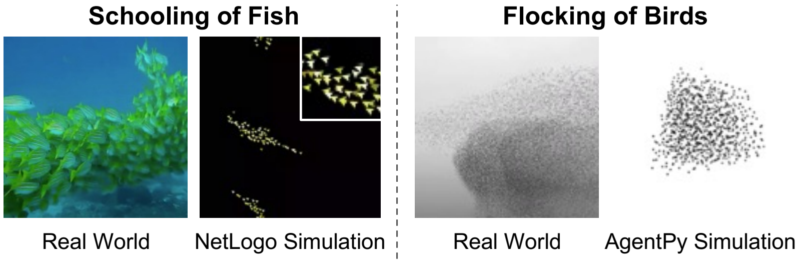


Figure 1: **Sample flocking scenes (BLUE (2021); Geographic (2016)) and simulators (NetLogo (Wilensky (1998)), AgentPy (Foramitti (2021))).**

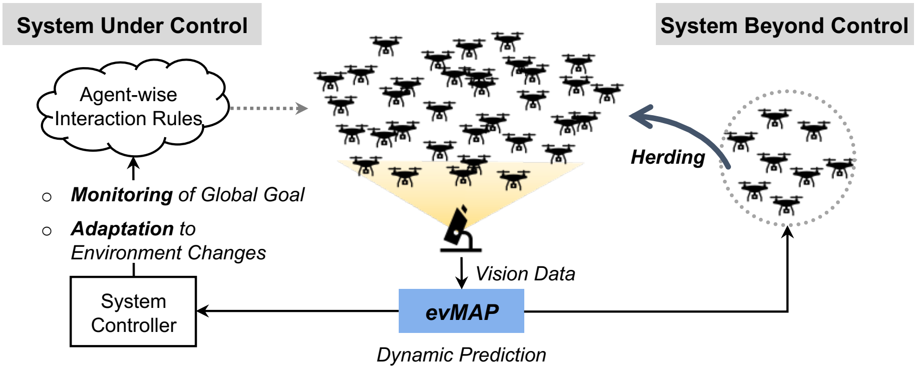


Figure 2: **Application examples of collective dynamic prediction of multi-agent system.** Multi-agent dynamic prediction is helpful for both systems that are under and beyond control.

This paper introduces the novel problem of ***real-time prediction of collective behavior in multi-agent dynamics from visual observations***. While prior studies focus on individual trajectory prediction using prior state knowledge (Vishnu et al. (2023); Luo et al. (2023); Mehr et al. (2023)),integrating object detection and trajectory prediction (Figure 3(a)) poses significant computational and scalability challenges. Dense agent scenes (Figure 1) hinder object detection (Yuan et al. (2023); Su et al. (2023)), and trajectory prediction across $M \times T \times N$ trajectories (M: agents, T: sequence length, N: sequences) becomes infeasible for large systems. Moreover, continuous state-based prediction methods (e.g., GPS-enabled systems) are limited by high energy usage and latency from continuous GPS usage.

Inspired by advancements in deep learning for vision-to-dynamics modeling (Ze et al. (2023); Yarats et al. (2021)), this work focuses on ***directly predicting collective dynamics from visual inputs*** captured via frame and event-based cameras (Figure 3(b)). While existing vision-to-dynamics models have been demonstrated for systems with a few agents, predicting the dynamics of collective multi-agent systems (with more than 10 agents) from vision remains an unexplored area. We propose leveraging advancements in event cameras (Gallego et al. (2020)), which capture per-pixel brightness changes with high temporal resolution, and offer unique advantages in understanding fast-changing dynamics and overcoming limitations of frame-based methods. Event-based vision has recently achieved significant improvements in object recognition, detection, and segmentation, as well as tracking high-speed objects (Gehrig and Scaramuzza (2023); Peng et al. (2023); Hamaguchi et al. (2023); Li et al. (2022a); He et al. (2021)). However, applying it to understand multi-agent dynamics remains largely unexplored (Figure 3(b-2)).

To address these challenges, we propose a novel approach for predicting collective multi-agent dynamics directly from visual inputs, utilizing event-based vision for its effectiveness in dynamic and dense scenarios. We introduce a new dataset and simulation framework for this task, conduct a comprehensive comparison between frame-based and event-based methods, and demonstrate the

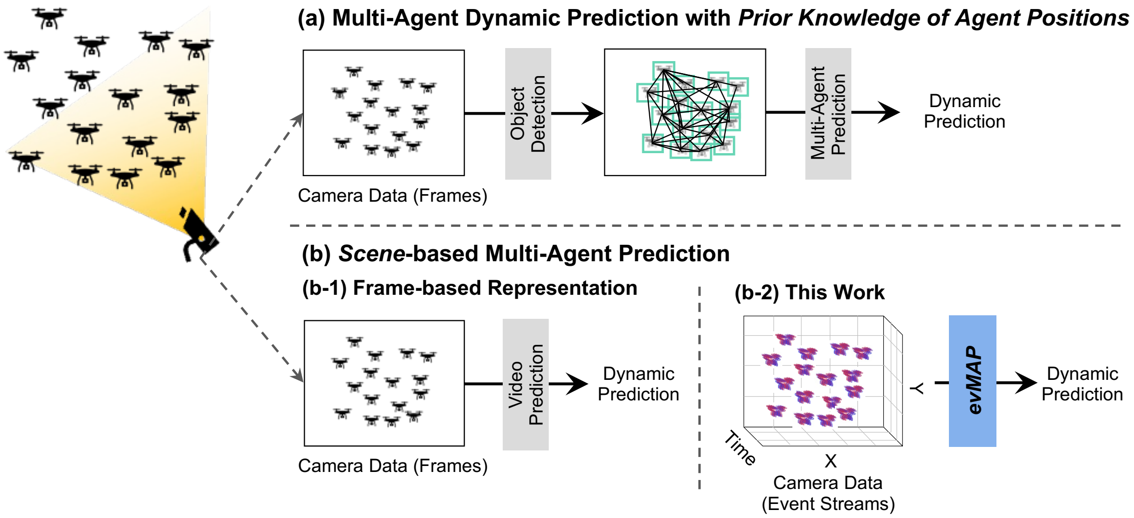

(a) Multi-Agent Dynamic Prediction with *Prior Knowledge of Agent Positions*
Camera Data (Frames)
Object Detection
Multi-Agent Prediction
Dynamic Prediction
(b) *Scene*-based Multi-Agent Prediction
(b-1) Frame-based Representation
Camera Data (Frames)
Video Prediction
Dynamic Prediction
(b-2) This Work
Y
Time
X
Camera Data (Event Streams)
evMAP
Dynamic Prediction

Figure 3: **Several methods for understanding dynamics in a multi-agent system.** (a) Many previous studies in multi-agent prediction require pre-processing for detecting agents. This paper focuses on scene-based perception: compared to (b-1) frame-based methods, (b-2) event-based methods demonstrate their effectiveness in understanding multi-agent dynamics.

advantages of event-based approaches in learning and predicting collective behaviors. Furthermore, we develop evMAP, a transformer-based architecture tailored to efficiently model and predict multi-agent dynamics with high accuracy. This paper makes following unique contributions:

- This paper introduces a novel problem, ***vision to prediction of collective multi-agent dynamics*** for real-time perception and control of multi-agent system. To the best of our knowledge, this is the first study to discuss multi-agent dynamic prediction from visual observation.
- This paper performs a comparative study between frame- and event-based methods, and empirically demonstrate the advantage of event representation in learning and predicting collective behavior of multi-agent systems. Prior works have studied processing event representation for various tasks, but to the best of our knowledge, this is the first work demonstrating event-based methods for predicting multi-agent dynamics.
- We present a new transformer-based deep learning architecture, **evMAP**, designed for efficient learning and prediction of multi-agent collective behaviors.

2. Vision to Multi-Agent Dynamic Prediction

2.1. Problem Formulation

Vision-based multi-agent dynamic prediction aims to forecast the collective dynamics of a multi-agent system with $M(> 10)$ interacting agents. An event-based camera captures a set of events $\mathcal{E}_\tau = \{(x_i, y_i, t_i, p_i) \mid t_i \in \tau, T = \sup t_i\}$ during a time interval τ, where (x_i, y_i) $(0 \leq x_i \leq W, 0 \leq y_i \leq H)$ denotes pixel locations, t_i represents event triggering timestamps, $p_i \in \{-1, 1\}$ indicates the polarity (brightness change). The goal is to predict the collective dynamics $\mathcal{D}_\tau \in \mathbb{R}^n$, such as **agent-wise interaction strength** (Section 4.1) and **collective behavior emergence time** (Section 4.2) , *directly from event camera data*, without mapping events to individual agents.

2.2. Multi-Agent Simulation and Visual Data Preparation.

Due to the lack of datasets for collective behavior of large interacting agent groups, we simulate multi-agent dynamics using flocking simulators (NetLogo Wilensky (1998), AgentPy Foramitti

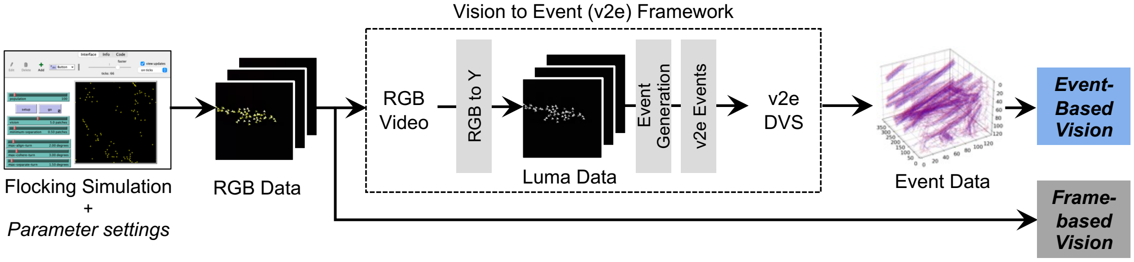


Figure 4: **Simulation framework of frame- and event-based vision for multi-agent dynamic prediction.** Due to the absence of existing dataset, flocking simulations (Wilensky (1998); Foramitti (2021)) and event synthesis toolbox (Hu et al. (2021)) are used to generate multi-agent dynamic sequence and convert from frame to event.

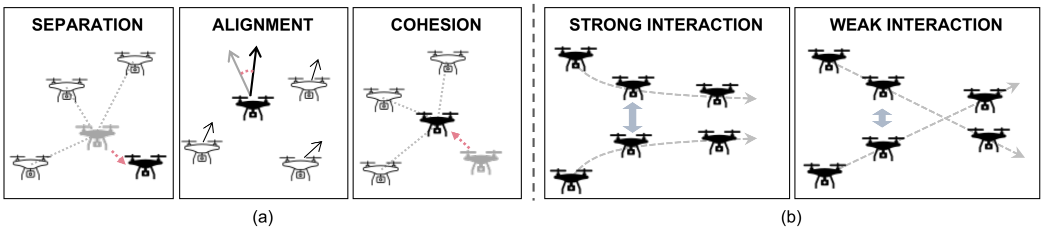


Figure 5: **(a) Three basic rules of flocking behavior.** Each agent avoids repulsion from other agents (*separation*), steers towards the average heading of neighbors (*alignment*), and steers towards the average position of neighbors (*cohesion*) (Reynolds (1987)). (b) (left) Strong interactions cause large changes in agents' velocities and (right) weak interactions relatively maintain individual agents' velocities.

(2021)) and convert frames to event data using the v2e framework Hu et al. (2021) (Figure 4). Simulations are based on Reynolds' rules (Reynolds (1987)) (Figure 5), generating interactions of varying intensities. Frames are used for training frame-based models, while event data trains event-based models. The dataset includes simulations with diverse interaction strengths and behavior convergence times (see Supplementary for details).

2.3. Prediction Target and Evaluation Metric

In this paper, we focus on two tasks: **1. Interaction Strengths between Agents** is critical for understanding collective dynamics as described in Figure 5(b). This includes time-varying changes, which is important to understand changes in the swarm intents. **2. Convergence Time of the Swarms** is the point at which most agents align their movement directions as shown in Figure 6. Initially, agents start with random positions and headings, gradually interacting based on flocking rules until reaching a state where the majority move in a unified direction, marking convergence (see Supplementary for details).

We evaluate predictions using an error ratio y_{pred}/y_{GT}, plotting it over normalized time $t_{observe}/T_c$. The ratio is chosen due to high variation in ground truth values across sequences. The *error over time* (EOT) metric, the area under this curve, quantifies prediction accuracy, with lower EOT indicating better performance (Figure 7).

2.4. Prediction Models

The concept of vision to multi-agent dynamic prediction is new, and there is no existing prediction models designed specifically for this task. We adapt existing video recognition models (SlowFast (Feichtenhofer et al. (2019)), MoViTv1 (Fan et al. (2021)), MoViTv2 (Li et al. (2022b))) and event-based object recognition models (AEGNN (Schaefer et al. (2022)), Eventformer (Kamal et al.

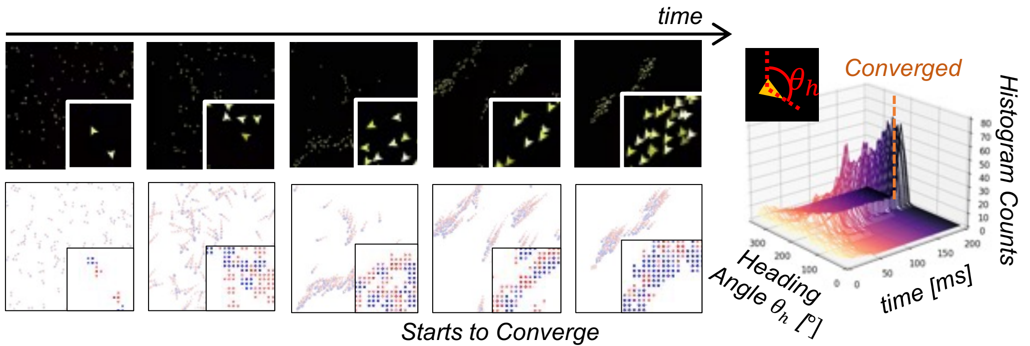


Figure 6: **Multi-agent swarming behavior in frame (top row) and event (bottom row) domains.** Events are aggregated into frames for every 1ms for visualization purposes. The heading directions of all agents converge to a certain direction from a specific time, referred to as *convergence-time*.

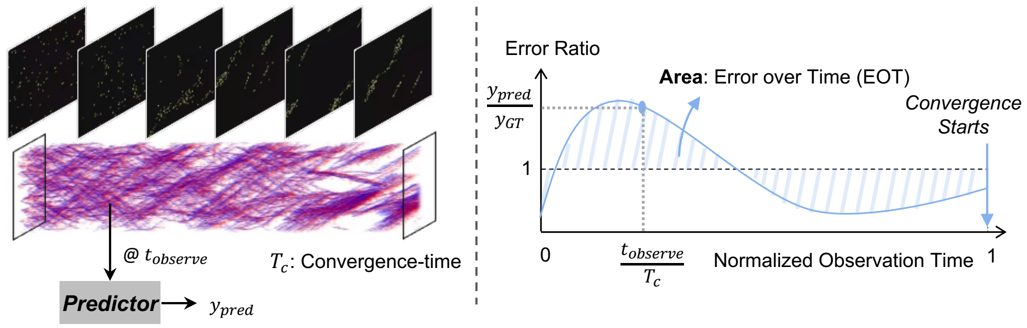
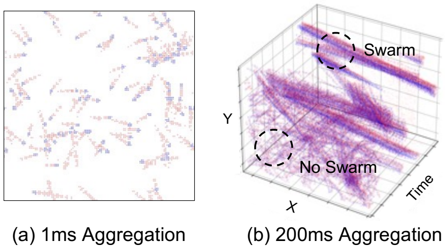

@ $t_{observe}$
T_c: Convergence-time
Predictor
y_{pred}
Error Ratio
$\frac{y_{pred}}{y_{GT}}$
1
0
Area: Error over Time (EOT)
Convergence Starts
$\frac{t_{observe}}{T_c}$
Normalized Observation Time
1

Figure 7: (Left) Multi-agent dynamic prediction over time in frame sequence or event stream, (Right) **Illustration of error ratio and EOT** as quantitative measures for predicting interaction strength and convergence-time.

(2023))) by replacing their classification heads with regression heads trained to predict interaction intensity and convergence time using L1 loss (see Supplementary for detail).



(a) 1ms Aggregation (b) 200ms Aggregation

Figure 8: **Event data from Multi-Agent Dynamics**

Unique Challenges for Multi-Agent Dynamic Prediction. Predicting collective multi-agent dynamics from event data presents unique challenges compared to tasks like object detection (Gehrig and Scaramuzza (2023)) or segmentation (Alonso and Murillo (2019); Sun* et al. (2022)). While these tasks often rely heavily on spatial information from short observation windows Yang et al. (2023), understanding agent interactions and collective behavior demands long-term spatiotemporal patterns (Figure 8). For instance, predicting interaction strength or convergence time requires observing how agents' interactions evolve over time, as collective behaviors often emerge gradually. Effective prediction also requires balancing the use of short-term observations for immediate dynamics and long-term observations for tracking system-level changes. To address these challenges, we propose evMAP, a specialized event-based model designed to capture both short-term and long-term spatiotemporal information, enabling accurate predictions of interaction dynamics and collective behaviors in multi-agent systems.

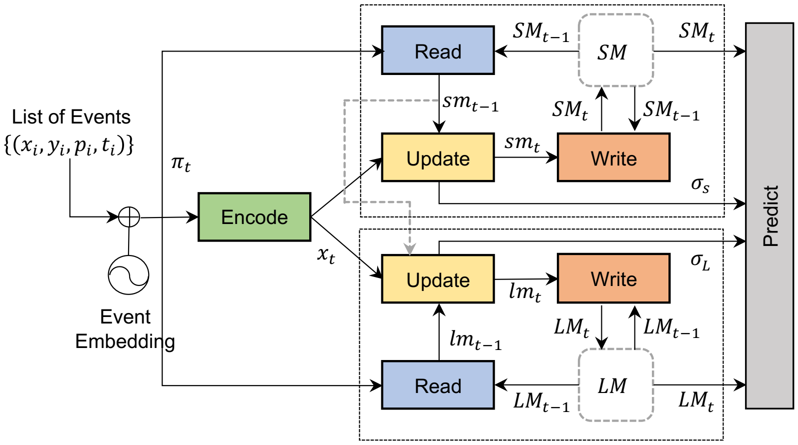


Figure 9: **evMAP Architecture.** evMAP contains two memories to effectively store recent spatial and long-term spatiotemporal observations, and adaptively use them for accurate prediction under dynamic changes.

3. *e*vent-based *v*ision for Multi-Agent dynamic Prediction

We introduce evMAP, a deep learning model designed for understanding the collective behavior of multi-agent systems. As shown in Figure 9, evMAP processes event embeddings π_t using an encoder to compute pair-wise spatiotemporal interactions x_t within event streams. These interactions, combined with prior observations stored in short-term (SM) and long-term (LM) memories, update the spatiotemporal information (sm_t, lm_t). The updated memories are then adaptively merged to predict collective dynamics using classification or regression heads.

Event Embedding: An event stream is chunked by time intervals, with each chunk encoded into a low-dimensional feature space (position, polarity, normalized time, and chunk ID) and mapped to a higher-dimensional space via a learnable Fourier feature-based positional encoder (Figure 10(a)).

Encode: Self-attention among the positional embedding π_t is employed to extract pair-wise interactions within an event chunk. Refine algorithm from Eventformer (Kamal et al. (2023)) is adopted for an efficient computations.

Read & Write: The `Read` operation retrieves relevant past memories M_{t-1} (from SM_{t-1} or LM_{t-1}) using cross-attention with current observations π_t (Figure 10(b-1)). The `Write` operation updates M_{t-1} with current spatiotemporal information m_t (sm_t or lm_t), without the need for `Erase`, as `Update` manage memory adjustments (Figure 10(b-2)).

Update: New spatiotemporal information m'_t (sm'_t or lm'_t) and update gate σ_m (σ_S or σ_L) are each computed based on the current encoded observation x_t and sm_{t-1}, and updated based on the past information (sm_{t-1} or lm_{t-1}) as follows:

$$m'_t = \tanh(W_{xm}x_t, W_{mm}sm_{t-1}) \quad \sigma_m = \sigma(W_x x_t + W_m sm_{t-1}) \quad m_t = (1-\sigma_m) \odot m_{t-1} + \sigma_m \odot m'_t$$

where $\odot$ denotes element-wise multiplication (Figure 10(c)). Small σ_m value corresponds to less new information to be updated, whereas a high σ_m value suggests more new information. `Update` process is executed separately on both short-term sm_t and long-term memories lm_t.

Predict: Short-term memory (SM_t) captures recent dynamics, while long-term memory (LM_t) stores broader behavioral patterns. Depending on system dynamics, the model adaptively emphasizes SM_t or LM_t using update gates σ_S and σ_L. The final prediction combines both memories:

$$\sigma_{SL} = \sigma(\sigma_S - \sigma_L) y' = FC(\sigma_{SL} \odot SM_t + (1-\sigma_{SL}) \odot LM_t)$$

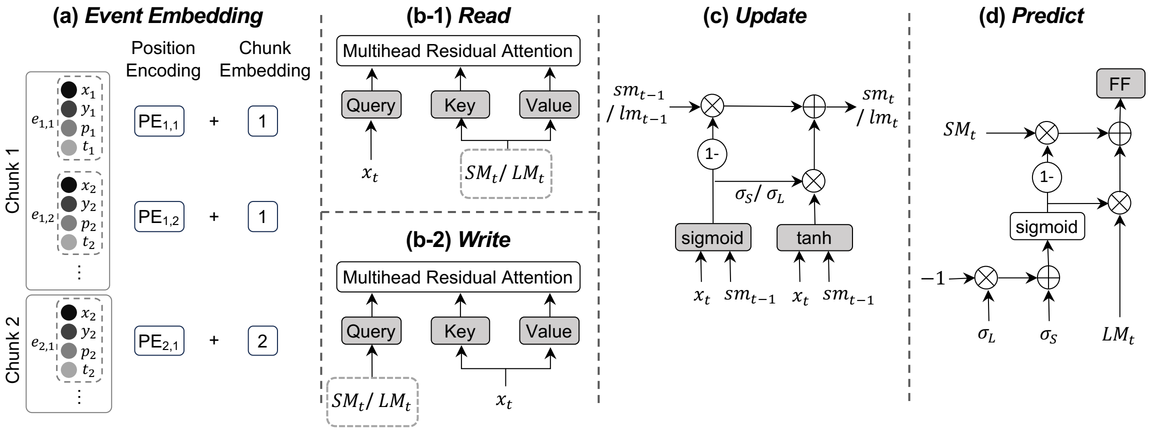


Figure 10: **Computation Blocks.** `Read`, `Write`, `Update` are processed separately for short-term and long-term memories. Grey colored blocks include a fully connect layer.

Comparison to Frame-based Models. Traditional video recognition models such as like SlowFast Feichtenhofer et al. (2019), MoViTv1 Fan et al. (2021), and MoViTv2 Li et al. (2022b) process sequences of images, capturing spatial information such as shape, color, and texture. However, these features are less relevant for multi-agent systems, where agents are small (fewer than 50 pixels) and densely grouped (Figure 1). Moreover, the background often occupies most of the image, making it harder for these models to focus on agent interactions. By contrast, evMAP processes event streams, which inherently encode agent dynamics with high temporal resolution, filtering out irrelevant background information and enabling more efficient modeling of multi-agent behaviors.

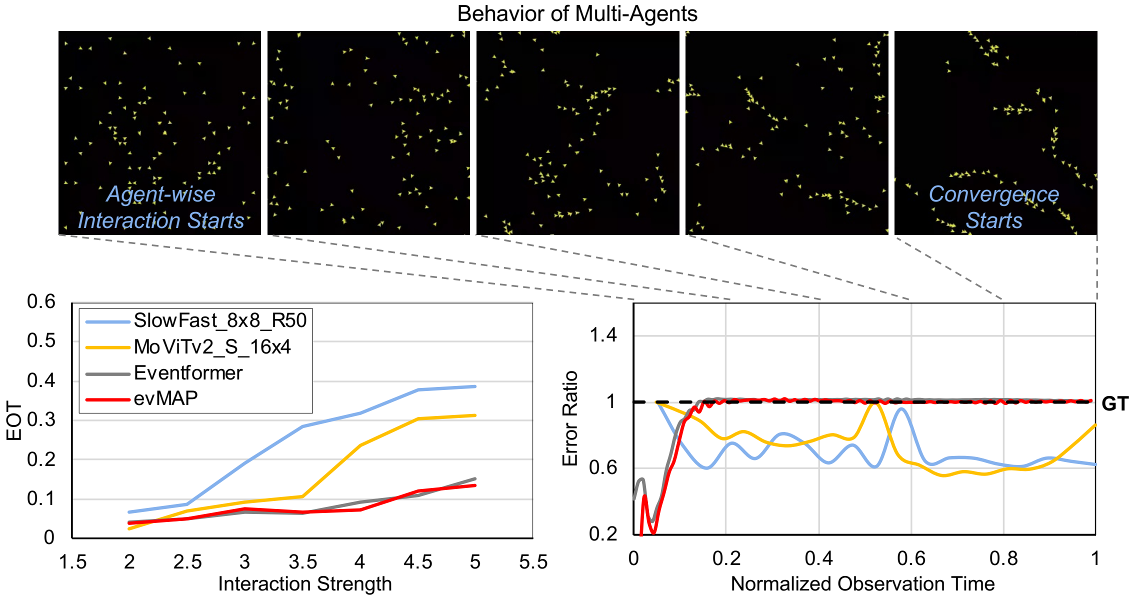
Comparison to Eventformer. Eventformer (Kamal et al. (2023)) incorporates GRU-based computation with associative memory that reads and writes information linked to GRU's hidden states, emphasizing spatiotemporal features. However, its update mechanism applies gating operations twice—once in the GRU and again in the erase computation—potentially leading to an overload of past information. This is helpful for object recognition tasks in static scenes as well as multi-agent dynamic predictions with fixed dynamic rules, but not for multi-agent system where its dynamics change. In contrast, evMAP uses two separate associative memories to store near-past ***spatio***temporal and long-range spatio***temporal*** information, ensuring the preservation of information from two different perspectives. Moreover, `Predict` module in evMAP facilitates the adjustments to dynamic system changes, in which relative experiments demonstrated in Section 4.3.

4. Simulation Results

4.1. Agent-wise Interaction Strength Detection

Figure 11(a) shows EOT (discussed in Section 2.3) for interaction strength detection. Both frame-based (SlowFast, MoViTv2) and event-based (Eventformer, evMAP) models perform similarly for weak interactions, but event-based methods excel in detecting stronger interactions due to their high temporal resolution. ***Event representation*** effectively capture the direction and speed of agents, which frames cannot (Figure 6).

Ability of Early Prediction. As shown in Figure 11(b), frame-based models have delays in making initial predictions due to fixed observation windows. This delay is more pronounced under stronger interactions where convergence times are shorter. Frame-based models struggle to dif-



(a) EOT

(b) Error Ratio

Figure 11: (a) **EOT of interaction strength prediction** across weak to strong interactions, (b) **Interaction strength prediction errors** from various frame- and event-based models (interaction strength=3). Definitions of error ratio, EOT, normalized observation time are discussed in Section 2.3.

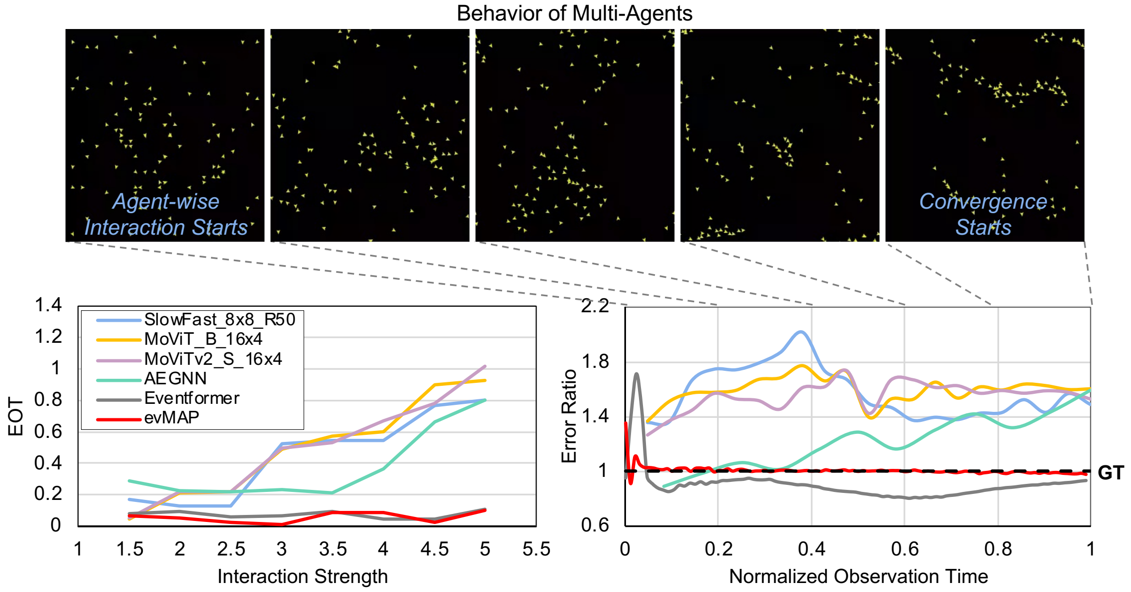
ferentiate interaction strengths, often overestimating weak and underestimating strong interactions. Event-based methods, particularly evMAP, provide accurate predictions early in the observation period (before 25% of the interaction time) across all interaction strengths (see Appendix for more experiment results).

4.2. Emergent Time Prediction of Collective Behavior

Figure 12(a) evaluates convergence-time prediction across different interaction strengths. Frame-based methods (SlowFast, MoViTv1, MoViTv2) perform well under weak interactions (EOT ¡ 0.5), where agent behaviors are simple and predictable. However, as interaction strength increases, their accuracy drops significantly, showing difficulties in capturing complex agent dynamics. Event graph-based methods (AEGNN) maintain high accuracy for weak and intermediate interactions but struggle under strong interactions due to computational constraints that require event reduction. This reduction limits their ability to model non-linear agent behaviors, reducing prediction accuracy.

Event-based methods using transformer architectures (Eventformer and evMAP) consistently outperform other approaches, particularly under strong interactions. Their ability to process the ***full set of event data*** without reduction enables them to capture subtle temporal patterns and abrupt behavioral changes. This advantage is crucial in scenarios with strong interactions, where agents exhibit rapid and complex dynamics.

Ability of Early Prediction. Figure 12(b) shows the prediction performance over time. Under weak interactions, frame-based methods (SlowFast, MoViTv1, MoViTv2) can make accurate early predictions (error ratio 0.95–1.05) within the first 20% of pre-convergence interaction time, reflecting the simplicity of agent dynamics. However, as interaction strength increases, their early prediction quality degrades, failing to capture emerging collective behaviors. AEGNN also struggles to provide accurate early predictions under strong interactions due to event reduction.



(a) EOT (b) Error Ratio

Figure 12: (a) **EOT for convergence-time prediction** across weak to strong interactions. (b) **Convergence-time prediction errors** from various frame- and event-based models (interaction strength=3). Definitions of error ratio, EOT, normalized observation time are discussed in Section 2.3.

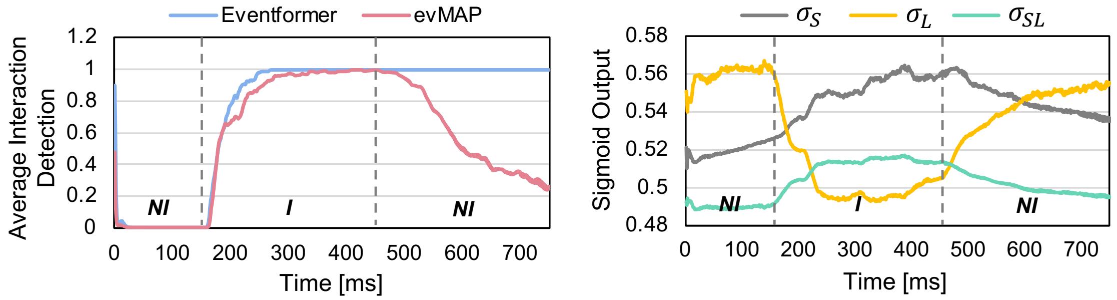


Figure 13: **(top) Agent-wise interaction detection of Eventformer (Kamal et al. (2023)) and evMAP.** Detection 0 indicates absence of interaction and 1 indicates the presence of interaction. **(bottom) Updates and adaptation of short-term and long-term memory in evMAP.** *NI* denotes system without agent-wise interaction and *I* denote system with interaction.

Event-based methods, particularly evMAP, demonstrate robust early prediction capabilities across all interaction strengths. evMAP achieves accurate predictions within the first 10% of pre-convergence time, leveraging its ***associative memory*** to effectively integrate observations over time. While its error ratio for strong interactions may increase closer to convergence, the absolute prediction error remains low due to shorter convergence times under strong interactions. evMAP also adapts well to weak and intermediate interactions by balancing its focus on spatial dynamics and accumulated long-term observations (see Appendix for more experiment results).

4.3. System Dynamic Change Detection

Previous experiments show that processing event representations through associative memory enables early and accurate understanding of multi-agent dynamics. However, a key limitation is the difficulty of adapting quickly to sudden changes in system dynamics, such as abrupt environmental shifts or external interventions, due to reliance on past observations.

Table 1: **Performance comparison among several frame- and event-based methods.** For frame-based methods, an input image size of 224×224 is considered. Each sequence contains, on average, 200 frames or 550K events.

Model	Represent.	Architecture	Avg. EOT ↓ (Interaction Strengths)	Avg. EOT ↓ (Convergence Time)	Params [MB] ↓	TFLOPS /sequence ↓
SlowFast	Image Frame	CNN	0.2453	0.491	53.0	27.5
MoViT	Image Frame	Transformer	-	0.559	36.4	14.1
MoViTv2	Image Frame	Transformer	0.1641	0.562	34.3	12.9
AEGNN	Event Graph	Graph	-	0.390	0.0298	0.190
Eventformer	Event Set	Transformer	$\underline{0.0825}$	$\underline{0.074}$	**0.011**	**0.0192**
evMAP	Event Set	Transformer	**0.0796**	**0.055**	$\underline{0.013}$	$\underline{0.0216}$

To evaluate adaptability, we modified Eventformer (Kamal et al. (2023)) and evMAP by replacing their regression heads with classification heads, training them with cross-entropy loss to distinguish between interactions present or absent. As shown in Figure 13, both models quickly (∼15ms) detect the onset of interactions at 150ms. For evMAP, σ_L (long-term memory updates) decreases, reflecting the stability of collective behavior, while σ_S (short-term memory updates) increases, emphasizing spatial dynamics.

When interactions cease at 450ms, Eventformer fails to detect this change, continuing to predict interactions. In contrast, evMAP successfully recognizes the shift by shifting focus from spatial to temporal dynamics, reflecting reduced spatiotemporal correlations. While evMAP's confidence slightly decreases due to past observations, it effectively adjusts to new dynamics, ensuring accurate adaptation to system changes.

Computation Complexity. Table 1 compares computational costs. Frame-based models (SlowFast, MoViTv1, MoViTv2) have significantly higher FLOPS due to dense image processing. Event-based methods (AEGNN, Eventformer, evMAP) achieve efficient real-time performance by directly processing event data. Eventformer and evMAP offer the lowest computational costs, as they do not require event-graph formulations.

Simulated Data and Real-World Applicability. Due to the difficulty of capturing real-world swarms of more than ten agents with both visible and event cameras, experiments use simulated data based on established interaction rules Reynolds (1987). Event-based data accounts for noise and non-idealities, reflecting real-world conditions Hu et al. (2021). We have also evaluated evMAP on simulated data that imitates real-world scenarios, as detailed in the Supplementary.

5. Conclusions

This paper introduces the novel problem of ***vision to prediction of collective multi-agent dynamics***, focusing on real-time perception and control. Multi-agent systems, defined as collections of over 10 interacting agents, are studied through their collective dynamics rather than individual states. We perform a comparative analysis of frame- and event-based methods, highlighting the advantages of event representations for learning and predicting multi-agent behaviors. To the best of our knowledge, this is the first study to explore multi-agent dynamic prediction using visual data and to demonstrate the effectiveness of event-based methods. Additionally, we present evMAP, a transformer-based architecture that leverages dual memory latents to adaptively capture and predict collective behaviors under dynamic changes.

Acknowledgments

This work is supported by the Defense Advanced Research Projects Agency (DARPA) under Grant Numbers GR00019153. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defence, DARPA, or the U.S. Government.

References

Inigo Alonso and Ana C. Murillo. Ev-segnet: Semantic segmentation for event-based cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2019.

ROYAL BLUE. Amazing school of fish. https://www.youtube.com/watch?v=vcPlPAAsfP0, 2021. [Online; accessed March 2024].

Vishnu S. Chipade and Dimitra Panagou. Multi-swarm herding: Protecting against adversarial swarms. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 5374–5379, 2020. doi: 10.1109/CDC42340.2020.9303837.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

Joël Foramitti. Agentpy: A package for agent-based modeling in python. *Journal of Open Source Software*, 6(62):3065, 2021.

Guillermo Gallego, Tobi Delbrück, Garrick Orchard, Chiara Bartolozzi, Brian Taba, Andrea Censi, Stefan Leutenegger, Andrew J Davison, Jörg Conradt, Kostas Daniilidis, et al. Event-based vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

Mathias Gehrig and Davide Scaramuzza. Recurrent vision transformers for object detection with event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

National Geographic. Flight of the starlings: Watch this eerie but beautiful phenomenon — short film showcase. https://www.youtube.com/watch?v=V4f_1_r80RY&t=29s, 2016. [Online; accessed March 2024].

Ryuhei Hamaguchi, Yasutaka Furukawa, Masaki Onishi, and Ken Sakurada. Hierarchical neural memory network for low latency event processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Botao He, Haojia Li, Siyuan Wu, Dong Wang, Zhiwei Zhang, Qianli Dong, Chao Xu, and Fei Gao. Fast-dynamic-vision: Detection and tracking dynamic objects with event and depth sensing. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.

Yuhuang Hu, Shih-Chii Liu, and Tobi Delbruck. v2e: From video frames to realistic dvs events. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021.

Uday Kamal, Saurabh Dash, and Saibal Mukhopadhyay. Associative memory augmented asynchronous spatiotemporal representation learning for event-based perception. In *The Eleventh International Conference on Learning Representations*, 2023.

Alaa Khamis, Ahmed Hussein, and Alaa Elmogy. Multi-robot task allocation: A review of the state-of-the-art. *Cooperative Robots and Sensor Networks*, 2015.

Jianing Li, Jia Li, Lin Zhu, Xijie Xiang, Tiejun Huang, and Yonghong Tian. Asynchronous spatio-temporal memory network for continuous event-based object detection. *IEEE Transactions on Image Processing*, 2022a.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022b.

Wenjie Luo, Cheol Park, Andre Cornman, Benjamin Sapp, and Dragomir Anguelov. Jfp: Joint future prediction with interactive multi-agent modeling for autonomous driving. In Karen Liu, Dana Kulic, and Jeff Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 1457–1467. PMLR, 14–18 Dec 2023. URL https://proceedings.mlr.press/v205/luo23a.html.

Matin Macktoobian, Francesco Basciani, Denis Gillet, and Jean-Paul Kneib. Learning convergence prediction of astrobots in multi-object spectrographs. *Journal of Astronomical Telescopes, Instruments, and Systems*, 2021.

Matin Macktoobian, Francesco Basciani, Denis Gillet, and Jean-Paul Kneib. Data-driven convergence prediction of astrobots swarms. *IEEE Transactions on Automation Science and Engineering*, 2022.

Negar Mehr, Mingyu Wang, Maulik Bhatt, and Mac Schwager. Maximum-entropy multi-agent dynamic games: Forward and inverse solutions. *IEEE Transactions on Robotics*, 39(3):1801–1815, 2023. doi: 10.1109/TRO.2022.3232300.

Yansong Peng, Yueyi Zhang, Zhiwei Xiong, Xiaoyan Sun, and Feng Wu. Get: Group event transformer for event-based vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Craig W. Reynolds. Flocks, herds and schools: A distributed behavioral model. In *Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, 1987.

Jackeline Rios-Torres and Andreas A. Malikopoulos. A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps. *IEEE Transactions on Intelligent Transportation Systems*, 2017.

Simon Schaefer, Daniel Gehrig, and Davide Scaramuzza. Aegnn: Asynchronous event-based graph neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Omar Shrit and Michèle Sebag. I2sl: Learn how to swarm autonomous quadrotors using iterative imitation supervised learning. In *Progress in Artificial Intelligence: 20th EPIA Conference on Artificial Intelligence, EPIA 2021, Virtual Event, September 7–9, 2021, Proceedings 20*, pages 418–432. Springer, 2021.

Yongzhi Su, Yan Di, Guangyao Zhai, Fabian Manhardt, Jason Rambach, Benjamin Busam, Didier Stricker, and Federico Tombari. Opa-3d: Occlusion-aware pixel-wise aggregation for monocular 3d object detection. *IEEE Robotics and Automation Letters*, 8(3):1327–1334, 2023. doi: 10.1109/LRA.2023.3238137.

Zhaoning Sun*, Nico Messikommer*, Daniel Gehrig, and Davide Scaramuzza. Ess: Learning event-based semantic segmentation from still images. *European Conference on Computer Vision. (ECCV)*, 2022.

Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet. Novel type of phase transition in a system of self-driven particles. *Physical review letters*, 75(6):1226, 1995.

Csaba Virágh, Gábor Vásárhelyi, Norbert Tarcai, Tamás Szörényi, Gergő Somorjai, Tamás Nepusz, and Tamás Vicsek. Flocking algorithm for autonomous flying robots. *Bioinspiration & biomimetics*, 9(2):025012, 2014.

Chalavadi Vishnu, Vineel Abhinav, Debaditya Roy, C. Krishna Mohan, and Ch. Sobhan Babu. Improving multi-agent trajectory prediction using traffic states on interactive driving scenarios. *IEEE Robotics and Automation Letters*, 8(5):2708–2715, 2023.

Uri Wilensky. Netlogo flocking model. http://ccl.northwestern.edu/netlogo/models/Flocking, 1998.

Yan Yang, Liyuan Pan, and Liu Liu. Event camera data pre-training. *CoRR*, abs/2301.01928, 2023. doi: 10.48550/arXiv.2301.01928. URL https://doi.org/10.48550/arXiv.2301.01928.

Denis Yarats, Amy Zhang, Ilya Kostrikov, Brandon Amos, Joelle Pineau, and Rob Fergus. Improving sample efficiency in model-free reinforcement learning from images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10674–10681, 2021.

Xiang Yuan, Gong Cheng, Kebing Yan, Qinghua Zeng, and Junwei Han. Small object detection via coarse-to-fine proposal generation and imitation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6317–6327, October 2023.

Yanjie Ze, Nicklas Hansen, Yinbo Chen, Mohit Jain, and Xiaolong Wang. Visual reinforcement learning with self-supervised 3d representations. *IEEE Robotics and Automation Letters*, 8(5):2890–2897, 2023. doi: 10.1109/LRA.2023.3259681.