A Self-Supervised Framework for Embodied Active Event Perception

Zhou Chen[1], Sanjoy Kundu[1], Harsimran S. Baweja[2], and Sathyanarayanan N. Aakur[1]

***Abstract*— For robots to truly assist, support, or collaborate with humans, they must perceive and carefully respond to unfolding events in dynamic, human-centric environments. We present EASE, a self-supervised framework for active event perception that unifies spatiotemporal representation learning with embodied control through predictive free energy minimization. Rather than relying on predefined tasks or external rewards, EASE continuously aligns perception and action using intrinsic signals—prediction error and entropy, to adaptively segment and track salient human activities in real time. This tight coupling enables emergent capabilities such as memory-like persistence, re-acquisition of lost targets, and privacy-preserving summarization, making it well suited for assistive and socially responsive robotics. Evaluations in both simulation and real-world settings demonstrate how EASE enables robots to perceive and act in ways that are robust, scalable, and attuned to human-centered care.**

I. Introduction

Understanding and responding to human activity in real-world settings remains a core challenge for autonomous systems designed to support care, companionship, or collaborative tasks. Robots operating in such environments—especially in assistive roles—must not only perceive events as they unfold, but do so in a way that respects human privacy, adapts fluidly to uncertainty, and requires minimal prior knowledge. Traditional methods often rely on predefined action categories, annotated datasets, or post-hoc video processing, which limits adaptability and raises ethical and practical concerns in settings like eldercare or rehabilitation, where privacy and responsiveness are essential.

In this paper, we present EASE, a self-supervised framework for embodied active event perception, inspired by cognitive theories of event segmentation [1] and visuomotor control [2]. EASE integrates perception and action within a predictive free energy minimization loop, allowing robots to dynamically segment, track, and summarize events using only intrinsic signals, without external labels or predefined goals. Unlike prior works that treat event understanding and motor control separately [3], [4], [5], [6], [7], [8], EASE unifies perception with action.

Contributions: We: (i) propose a unified framework for privacy-aware active event perception that integrates segmentation, summarization, and control based on prediction error and entropy, (ii) develop a free energy minimization paradigm that aligns perception and action without relying on task-specific supervision, (iii) introduce an event summarization strategy that captures only the most salient activity while discarding raw video or sensitive data, and (iv) validate EASE across simulation and real-world scenarios, highlighting emergent behaviors such as implicit memory, continuity in tracking, and robustness to distractions—critical for socially responsive robotic care.

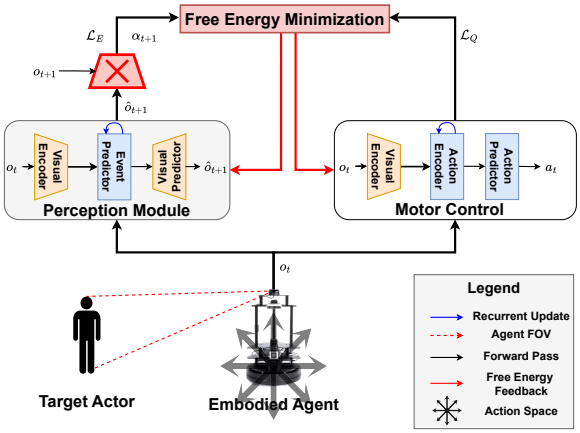

Fig. 1. **Overview.** The perception module processes sensory observations (o_t) to predict future observations ($\hat{o}_{t+1}$) and minimize discrepancies ($\mathcal{L}_E$). The motor control module uses the sensory input (o_t) to generate actions (a_t), minimizing the control loss $\mathcal{L}_Q$. *Free Energy Minimization* provides supervision for both modules for robust active event perception. The prediction loss is used for downstream event perception tasks.

Related Work. Prior work on active tracking includes task-specific supervision, multi-agent systems, and adversarial training with domain-specific rewards [9], [7]. Trehan et al. [8] proposed an energy-based framework combining predictive learning with PID control. For event perception, self-supervised methods have been used for action boundary detection and group dynamics [4], [5], though most focus on representation learning [3], [10], [11] under passive observation. In contrast, EASE integrates prediction errors as intrinsic signals for unified perception and control in dynamic environments. It leverages self-supervised learning strategies such as future prediction and contrastive learning [3], [10], [4], [12], [13], [14], grounded in the Free Energy Principle (FEP) [15], [16], with applications in navigation [17] and robot control [18].

II. EASE: Embodied Active Spatiotemporal Event Perception

Overview. Our framework, EASE, consists of two subsystems that work together for sensory event perception

[1]Z. Chen, S. Kundu, and SN Aakur are with the CSSE Department, Auburn University, Auburn, AL 36849 USA Email: {zzc0053,szk0266,san0028}@auburn.edu

[2] HS Baweja is with the School of Kinesiology, Auburn University, Auburn, AL 36849 USA. Email: hsb0025@auburn.edu

[3]This work was partially supported by the US National Science Foundation Grants IIS 2348689 and IIS 2348690 and the US Department of Agriculture grant 2023-69014-39716-1030191.

and motor control. The overall architecture is illustrated in Figure 1. A *perception module* receives a sequence of sensory observations ($\{o_t\}_{t=1}^{T}$; $o_t \in \mathbb{R}^{H\times W\times C}$) as input and outputs intrinsic signals for event perception and motor control in the form of spatiotemporal uncertainty distributions (α_t) and temporal segmentation (δ_t) cues. These uncertainty and segmentation cues generate an intrinsic signal that serves as the guiding metric for the *motor control module*, which outputs a sequence of actions ($\{a_t\}_{t=1}^{T}$; $a_t \in \mathbb{R}^k$) to minimize the system's energy and stabilize event representations.

A. Learning as Free Energy Minimization

Our framework minimizes system free energy, which quantifies environmental uncertainty, through the joint optimization of perception and action. Inspired by active inference [15], [16] and predictive coding [1], the perception module generates sensory predictions and identifies salient features through uncertainty distributions, while actions actively align observations with predictions. This closed-loop interaction continuously refines event representations while stabilizing sensory input through selective information processing and adaptive behavior.

The *free energy* of the system can be decomposed into two complementary terms: a **prediction-based drive** term, and an **action-driven uncertainty reduction** term, and formalized as an optimization for

$$\arg\min_{a} \left[\|o(a) - \hat{o}\|^2 + \lambda \sum_{i,j} \alpha_{ij}(a) \|o_{ij}(a) - \hat{o}_{ij}\|^2 \right] \quad (1)$$

where $o(a)$ represents the sensory observation at time t influenced by action a, $\hat{o}$ is the predicted observation generated by the perception module, and $\alpha_{ij}(a)$ are uncertainty distribution values dynamically modulated by action a, focusing on salient spatial regions. The first term represents the global sensory prediction error, and the second term computes and integrates the model's spatiotemporal uncertainty to emphasize the reduction in regions of higher surprise. λ is a tradeoff factor. By minimizing these terms jointly, the system learns to optimize its predictions while selecting actions that stabilize sensory input, aligning perception and action in a unified framework.

1) Prediction-based Drive: The perception module learns the spatiotemporal dynamics of the environment through a recurrent, generative model. A visual encoder $\phi(o_t) \longrightarrow \mathbf{f}_t$ encodes the raw visual observation into a spatial feature map at time t. At each timestep, the perception module processes the feature map $\mathbf{f}_t \in \mathbb{R}^{h\times w\times d}$ and predicts the expected feature map $\hat{\mathbf{f}}_{t+1}$ at the next timestep. The anticipated feature map is compared to the actual features to compute the prediction error: $\mathcal{L}_E = \|\mathbf{f}_{t+1} - \hat{\mathbf{f}}_{t+1}\|^2$. This prediction error serves as the primary intrinsic signal driving the system.

Quantifying uncertainty. The prediction errors generated by the perception module also provide a mechanism for capturing uncertainty and guiding focus. Discrepancies between observed and predicted feature maps highlight areas where the system's understanding is incomplete or inaccurate. These spatially distributed errors compute an uncertainty distribution α_{ij} that dynamically allocates focus to salient regions. The uncertainty distribution is computed by

$$\alpha_{ij} = \text{Softmax}\left(\frac{\|\mathbf{f}_{t,ij} - \hat{\mathbf{f}}_{t,ij}\|^2}{\tau} \right) \quad (2)$$

where $\mathbf{f}_{t,ij}$ and $\hat{\mathbf{f}}_{t,ij}$ represent feature vectors at spatial location (i, j), and τ controls sensitivity to prediction errors.

2) Uncertainty-based Action Selection: The motor control module uses the uncertainty distribution $\alpha_{ij}$to guide actions by learning a policy that aligns the frame center c_t with high-prediction-error regions u_t, reducing uncertainty and implicitly minimizing free energy over time. It is parametrized by a neural network sharing the perception encoder, enabling access to the feature-level representation $\mathbf{f}_t$. A Deep Q-Network (DQN) [19] estimates Q-values for discrete actions $a_t \in \mathcal{A}$, trained with reward $r_t = -\|c_t - u_t\|$ to encourage focus on high-α_{ij} regions and refine the generative model. The input state $s_t = \{\mathbf{f}_t, \alpha_{ij}\}$ is used to compute the expected cumulative reward. The policy is optimized via temporal difference loss:

$$\mathcal{L}_Q = \mathbb{E}\left[\left(Q(s_t, a_t) - \left(r_t + \gamma \max_{a'} Q(s_{t+1}, a') \right) \right)^2 \right], \quad (3)$$

where r_t is the reward and γ is the discount factor for future rewards.

3) Learning Process: The perception and motor control modules are jointly optimized to minimize the free energy in Equation (1). The prediction loss $\mathcal{L}_E$ reduces discrepancies between predicted and observed features, while the motor policy minimizes the temporal difference loss $\mathcal{L}_Q$ by aligning observations with high-uncertainty regions. These losses share feature representations, coupling perception and action to reduce prediction error.

4) Implementation Details: We use the Stable-Baselines3 [20] DQN implementation with the following key parameters: a batch size of 32, a replay buffer size of 50,000, a learning rate 10^{-5}, and an exploration final epsilon of 0.02. The policy network includes a custom feature extractor based on the first eight layers of EfficientNet-B0 [21], outputting feature maps of size $(320, 7, 7)$. These features are processed by 2 LSTM modules: LSTM-Event, which models temporal dynamics and outputs a (320, 7, 7) feature map, and LSTM-QVal, which aggregates temporal features into a 1024-dimensional vector for the policy. The network architecture includes fully connected layers with 256 and 64 units, followed by the output layer matching the action space size. Training updates the CNN and LSTM modules iteratively based on a two-stage strategy. Early training ($t < 50,000$) focuses on aligning predicted and actual features, while later stages optimize temporal difference loss using computed intrinsic rewards. The framework is trained for $300k$ timesteps on UnrealCV-Gym [22].

B. Event Perception: Segmentation and Summarization

Generic event segmentation and snapshot creation are essential for embodied agents in dynamic settings, enabling

TABLE I

EVENT SEGMENTATION RESULTS IN SIMULATION ENVIRONMENTS.

Model ↓ Env. →	Seg. Mode	City		Urban City		Rand. Room	
		IoU	Acc	IoU	Acc	IoU	Acc
EASE-Hybrid	1	0.52	0.70	0.35	0.49	0.47	0.64
EASE	1	0.41	0.59	0.33	0.49	0.51	0.69
EASE-Supervised	2	0.31	0.46	0.29	0.39	0.33	0.46
EASE-Hybrid	2	0.30	0.42	0.20	0.32	0.23	0.30
EASE	2	0.20	0.33	0.26	0.36	0.24	0.34

*Segmentation modes (Seg. Mode) 1 and 2 denote how events are segmented. 1: using $\mathcal{L}_E$ or using prediction assessment.

them to organize visual input into meaningful segments and generate concise summaries. By focusing on salient events and discarding redundant or sensitive data, it also supports privacy.

To enable generic event segmentation and summarization from streaming videos, the framework leverages prediction errors ($\mathcal{L}_E$) and entropy to detect event boundaries and select representative frames. This process is grounded in the free energy minimization framework, where segments are identified in regions of high uncertainty, and summarization minimizes local prediction errors within those segments. The system detects event boundaries B_t based on the entropy of prediction errors within a sliding window of size N:

$$B_t = \arg\max_t [H_t], \quad \text{with} \quad H_t = -\sum_{i=1}^{N} p_{t,i} \log p_{t,i}, \quad (4)$$

where H_t is the entropy at time t, and $p_{t,i}$ are normalized prediction errors:

$$p_{t,i} = \frac{\mathcal{L}_E(i)}{\sum_{j=1}^{N} \mathcal{L}_E(j)}, \quad \mathcal{L}_E(i) = \|\mathbf{f}_{t+1,i} - \hat{\mathbf{f}}_{t+1,i}\|^2. \quad (5)$$

Peaks in the entropy curve H_t highlight moments of heightened prediction error variability, which are treated as event boundaries. For summarization, the most representative frame S_k for each segment k is selected by minimizing the prediction loss within the segment $[B_k, B_{k+1}]$:

$$S_k = \arg\min_{t \in [B_k, B_{k+1}]} \mathcal{L}_E(t). \quad (6)$$

Here, S_k corresponds to the frame with the lowest prediction error, representing the event's most stable and well-predicted observation.

III. EXPERIMENTAL SETUP

We evaluate the EASE framework in both simulated and real-world environments to assess its performance across active tracking, event segmentation, and summarization tasks.

Simulation Environment. Training and evaluation are conducted in UnrealCV-Gym. We train on the Flexible-Room environment for real-world experiments, which offers adjustable clutter and difficulty settings. We train on the Random Room environment for simulation experiments and evaluate on the City1 and UrbanCity environments.

Real-world Evaluation. Real-world experiments are conducted on the Interbotix LoCoBot platform. The setup simulates an office-like environment with distractors such as windows, furniture, and stairs. Three actors perform unscripted daily activities, such as walking, adjusting thermostats, or working on laptops, in 4-minute episodes that contain at least 10 actions, each lasting about 15 seconds. Three annotators review recordings to mark event boundaries, assess tracking success, and assess summarization quality, ensuring realistic and diverse conditions for robust evaluation.

TABLE II

PERFORMANCE EVALUATION ON THE ACTIVE TRACKING TASK.

Model ↓ Env. →	City		Urban City		Rand. Room	
	AR	AL	AR	AL	AR	AL
TLD+PID [8]	12	**90**	19	**115**	25	**147**
MIL+PID [8]	**32**	59	24	50	21	43
MOSSE+PID [8]	16	56	**49**	68	**28**	62
Smart-Target [6]	232	473	233	466	403	458
Random-Target [6]	214	455	204	464	409	455
AD-VAT+ [7]	**326**	483	**322**	488	427	493
EASE-Supervised	248	**500**	290	**490**	**459**	**500**
PredLearn-PID [8]	114	343	71	349	115	319
EASE-Hybrid	233	**500**	**253**	**496**	**438**	**500**
EASE	**273**	**500**	155	443	214	491

Quantitative Evaluation Metrics. Following prior work [23], event segmentation is evaluated using precision, recall, and F1 score, comparing the detected boundaries with the ground-truth within tolerance windows. Strict evaluation uses narrow tolerances (e.g., 2 to 15 frames, corresponding to 0.5 seconds at 30 FPS), while relaxed evaluation allows broader tolerances (15 to 45 frames), reflecting the complexities of active event perception. Tracking performance in the simulation environment is measured using total environment reward and average episode length, following prior work [6], [7]. For real-world tracking evaluation, we use the average qualitative judgment from the annotators, who grade each frame as 1 (tracking) or 0 (not tracking).

Baselines. We evaluate three versions of our framework: (i) **EASE**, the fully self-supervised model, (ii) **EASE-Hybrid**, trained with both self-supervised losses and simulation rewards for enhanced tracking, and (iii) **EASE-Supervised**, trained solely on environmental rewards, representing state-of-the-art active tracking methods. Segmentation and summarization for EASE and EASE-Hybrid use $\mathcal{L}_E$. For EASE-Supervised, we use state transition differences in the controller LSTM's hidden state as the perception signal, as done in Predictability Assessment [23].

IV. RESULTS AND DISCUSSION

A. Evaluation in Simulated Environments

We train and evaluate our model and baselines in increasingly challenging environments within UnrealCV-Gym. Training is performed in the Random Room environment. Following Luo *et al.* [24], random augmentation is applied using version v4 for training and v0 for testing. To assess robustness and adaptability, we also evaluate performance in the more complex City1 and Urban environments, which introduce greater visual clutter and distractions.

1) Event Perception with Growing Amounts of Clutter: We evaluate EASE's performance in tracking and segmenting human actions across three simulation environments—City,

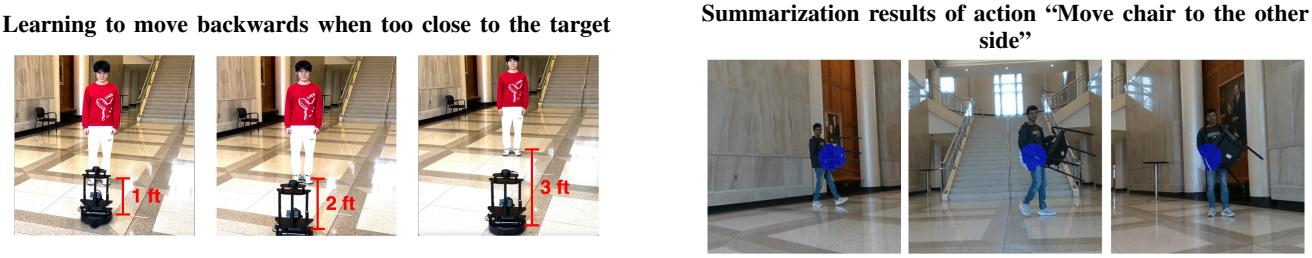

Fig. 2. **Qualitative visualization** of emergent properties from free energy minimization. Left: EASE learns to move back when too close to the target. Right: Summarization output (from robot POV) of the target performing action "Move chair from one side of the room to another", where the over-segmentation of constituent actions (e.g., pick up the chair, walk, stop, and put down the chair) is clearly visible. Note: the area of high uncertainty is marked with a blue circle.

Urban City, and Random Room—using IoU and action-level accuracy (Acc). Table I shows that EASE-Hybrid achieves the highest scores in simpler environments like City, likely due to its combined training. However, in more complex scenes, EASE maintains competitive performance, highlighting the strength of its self-supervised learning objective ($\mathcal{L}_E$) for event segmentation and active perception.

2) Active Object Tracking: We also evaluate EASE on active object tracking—dynamically following salient actors in complex environments. As shown in Table II, EASE outperforms traditional PID-based trackers and matches or surpasses reinforcement learning methods, especially in challenging scenarios. While EASE-Hybrid performs well in simpler settings due to reward-driven learning, the fully self-supervised EASE model maintains strong, consistent performance—despite not being explicitly trained for tracking—highlighting balance between perception and control.

B. Real-world Event Perception

We extend the evaluation of EASE to real-world scenarios to assess its performance on active event segmentation and summarization tasks. Table III summarizes the results across segmentation (strict and relaxed) and summarization metrics, alongside tracking success rates.

1) Temporal Event Segmentation: Table III shows segmentation results for EASE, EASE-Hybrid, and EASE-Supervised under both strict and relaxed evaluation settings. The strict setting demands precise boundary predictions, while the relaxed setting allows more tolerance, better reflecting real-world complexity. The hybrid model achieves higher precision by predicting fewer boundaries and avoiding over-segmentation, but at the cost of lower recall in rapidly changing segments. In contrast, the fully self-supervised EASE model is more sensitive to movement changes, leading to higher recall but lower precision. The supervised model offers a middle ground with balanced precision and recall.

2) Summarization: Event summarization complements segmentation by distilling continuous streams into concise keyframes, aiding efficient review under storage, computation, and privacy constraints. We evaluate summarization using three human-rated metrics—Temporal Coverage, Redundancy, and Quality—each scored 1–5. As shown in Table III, EASE achieves the highest Temporal Coverage, though its motion sensitivity sometimes increases Redundancy. The supervised model scores best in Quality due to structured training, while the hybrid model offers a balanced trade-off. These results highlight how EASE's fine-grained segmentation supports rich summaries.

C. Qualitative Analysis

EASE demonstrates emergent, memory-like behaviors by maintaining focus on salient targets through prediction-driven action, as shown in supplementary video. While effective, it can momentarily lose focus in low-motion scenes, highlighting a trade-off of its unsupervised approach. A supplementary video demonstrating qualitative results in real-world scenarios is provided to help readers better understand EASE's effectiveness in event perception and active tracking.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we introduced EASE, a novel framework for active event perception that unifies spatiotemporal representation learning and embodied control through a free energy minimization paradigm. By leveraging self-supervised learning, EASE adaptively segments, summarizes, and tracks dynamic events in simulation and real-world environments without relying on annotations or extrinsic rewards. The quantitative and qualitative results highlight EASE's ability to balance fine-grained event sensitivity with robust motor control. Moving forward, we aim to enhance EASE by capturing the hierarchical nature of event segmentation, improving its adaptability to subtle or novel actions, and extending it to multi-agent systems for collaborative tasks.

TABLE III

REAL-WORLD PERFORMANCE EVALUATION OF EASE FOR ACTIVE EVENT PERCEPTION TASKS.

Model	Segmentation (Strict)			Segmentation (Relaxed)			Summarization			Tracking
	Precision	Recall	F1	Precision	Recall	F1	Coverage	Redundancy	Quality	Success (%)
EASE	14.71	**47.62**	22.47	24.75	**60.12**	35.07	**4.58**	3.58	4.17	94.42
EASE-Hybrid	**38.31**	27.82	**32.06**	**51.67**	36.98	**42.88**	4.08	**4.17**	**4.28**	90.99
EASE-Supervised	21.87	21.54	21.64	37.04	35.39	36.05	4.27	4.05	4.12	**98.01**

REFERENCES

[1] J. M. Zacks and B. Tversky, "Event structure in perception and conception." *Psychological bulletin*, vol. 127, no. 1, p. 3, 2001.

[2] H. Idei, W. Ohata, Y. Yamashita, T. Ogata, and J. Tani, "Emergence of sensory attenuation based upon the free-energy principle," *Scientific reports*, vol. 12, no. 1, p. 14542, 2022.

[3] S. N. Aakur and S. Sarkar, "A perceptual prediction framework for self supervised event segmentation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1197–1206, 2018.

[4] ——, "Action localization through continual predictive learning," in *European Conference on Computer Vision*, 2020.

[5] S. Trehan and S. N. Aakur, "Self-supervised multi-actor social activity understanding in streaming videos," *ArXiv*, vol. abs/2406.14472, 2024.

[6] F. Zhong, P. Sun, W. Luo, T. Yan, and Y. Wang, "AD-VAT: An asymmetric dueling mechanism for learning visual active tracking," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=HkgYmhR9KX

[7] ——, "Ad-vat+: An asymmetric dueling mechanism for learning and understanding visual active tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1467–1482, 2021.

[8] S. Trehan and S. N. Aakur, "Towards active vision for action localization with reactive control and predictive learning," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 783–792.

[9] F. Zhong, P. Sun, W. Luo, T. Yan, and Y. Wang, "Ad-vat: An asymmetric dueling mechanism for learning visual active tracking," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:53533716

[10] R. Mounir, R. Gula, J. Theuerkauf, and S. Sarkar, "Spatio-temporal event segmentation for wildlife extended videos," in *International Conference on Computer Vision and Image Processing*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:251286809

[11] O. Cetintas, G. Brasó, and L. Leal-Taixé, "Unifying short and long-term tracking with graph hierarchies," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 877–22 887.

[12] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, S. Levine, and G. Brain, "Time-contrastive networks: Self-supervised learning from video," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 1134–1141.

[13] I. Dave, R. Gupta, M. N. Rizve, and M. Shah, "Tclr: Temporal contrastive learning for video representation," *Computer Vision and Image Understanding*, vol. 219, p. 103406, 2022.

[14] H. Kuang, Y. Zhu, Z. Zhang, X. Li, J. Tighe, S. Schwertfeger, C. Stachniss, and M. Li, "Video contrastive learning with global context," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3195–3204.

[15] K. Friston, "The free-energy principle: a unified brain theory?" *Nature reviews neuroscience*, vol. 11, no. 2, pp. 127–138, 2010.

[16] K. Friston, S. Samothrakis, and R. Montague, "Active inference and agency: optimal control without cost functions," *Biological cybernetics*, vol. 106, pp. 523–541, 2012.

[17] T. Parr, G. Pezzulo, and K. J. Friston, *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022.

[18] G. Pezzulo, F. Rigoli, and K. J. Friston, "Hierarchical active inference: a theory of motivated control," *Trends in cognitive sciences*, vol. 22, no. 4, pp. 294–306, 2018.

[19] J. Chung, "Playing atari with deep reinforcement learning," *Comput. Ence*, vol. 21, pp. 351–362, 2013.

[20] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

[21] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[22] W. Qiu, F. Zhong, Y. Zhang, S. Qiao, Z. Xiao, T. S. Kim, and Y. Wang, "Unrealcv: Virtual worlds for computer vision," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 1221–1224.

[23] M. Z. Shou, S. W. Lei, W. Wang, D. Ghadiyaram, and M. Feiszli, "Generic event boundary detection: A benchmark for event segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 8075–8084.

[24] W. Luo, P. Sun, F. Zhong, W. Liu, T. Zhang, and Y. Wang, "End-to-end active object tracking and its real-world deployment via reinforcement learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 6, pp. 1317–1332, 2019.