Learning to Perceive and Act: Active Event Understanding via Predictive Free Energy Minimization

Zhou Chen, Sanjoy Kundu, Harsimran Baweja, Sathyanarayanan Aakur
Auburn University
Auburn, AL, 36849
{zzc0053,szk0266,hsb0025,san0028}@auburn.edu

Abstract

Learning to perceive and act in dynamic environments is a fundamental challenge for embodied agents. Traditional models for event understanding often rely on offline training, annotated datasets, or fixed action policies, limiting their adaptability and real-world robustness. We present EASE, a self-supervised framework that unifies event perception and active control via predictive free energy minimization. EASE predicts feature-level sensory dynamics and quantifies uncertainty to guide adaptive motor policies, closing the perception-action loop without external supervision. By minimizing prediction errors and dynamically attending to high-uncertainty regions, EASE enables agents to segment, track, and summarize salient events in streaming visual input. Our experiments demonstrate that EASE achieves emergent behaviors such as implicit memory and target continuity, outperforming conventional trackers on both simulation and real-world benchmarks while preserving privacy through in-device, streaming inference. These results highlight EASE's promise for scalable, privacy-conscious event understanding in dynamic environments.

1. Introduction

Understanding dynamic events is challenging for autonomous agents, especially when conventional models rely on predefined actions or annotations. Storing raw video for post-processing also raises privacy concerns, highlighting the need for real-time event summarization without retaining identifiable data. Inspired by cognitive theories of event perception [20] and visuomotor control [8], we propose EASE, a self-supervised framework that unifies spatiotemporal representation learning and embodied control through free energy minimization. While prior works have addressed event perception [1, 2, 19] and visuomotor control [18, 22, 23] as separate problems—often relying on predefined action categories, annotated datasets, or task-specific supervision. EASE leverages predictive coding-inspired mechanisms to dynamically align perception and action by minimizing perceptual prediction errors.

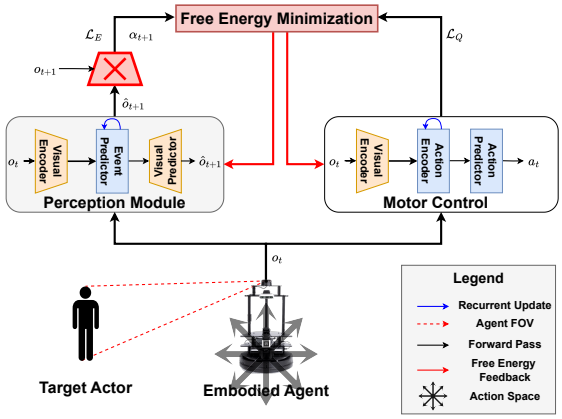

Figure 1. **Overview.** Perception predicts future observations ($\hat{o}_{t+1}$) from sensory observations (o_t), minimizing discrepancy ($\mathcal{L}_E$). Motor control generates actions (a_t) from sensory input (o_t), minimizing control loss ($\mathcal{L}_Q$). Free Energy Minimization supervises both modules.

Prior work on active tracking includes task-specific supervision, multi-agent systems, and adversarial training with domain-specific rewards [21, 23]. Trehan et al. [18] 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 [2, 19], though most focus on representation learning [1, 3, 10] 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 using future prediction [1, 2, 5, 9, 10, 15], grounded in the Free Energy Principle (FEP) [6, 7], with embodied applications [11, 12].

Our contributions in this work are as follows: (i) we propose a unified framework for embodied active event perception that integrates generic event segmentation, sum-

marization, and active tracking using prediction error and entropy as intrinsic signals, (ii) we develop a free energy minimization paradigm that couples perception and action to enable dynamic adaptation to high-uncertainty regions and salient actors, (iii) we introduce an inherently privacy-preserving snapshot summarization strategy that retains only salient events while discarding redundant or sensitive data, and (iv) we validate EASE through extensive experiments in both simulation and real-world settings, demonstrating its effectiveness across tasks such as generic boundary detection, active tracking, and summarization without requiring annotations or external supervision.

2. Proposed Framework: EASE

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

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 [6, 7] and predictive coding [20], 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 is as an optimization for

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

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.

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 error is the primary intrinsic signal for 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 given by $\alpha_{ij} = \text{Softmax}\left(\frac{\|\mathbf{f}_{t,ij} - \hat{\mathbf{f}}_{t,ij}\|^2}{\tau}\right)$, where $\mathbf{f}_{t,ij}$ and $\hat{\mathbf{f}}_{t,ij}$ are feature vectors at spatial location (i, j), and τ controls sensitivity to prediction errors. 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) [4] 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 using:

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

where r_t is the reward and γ is the discount factor.

Learning. 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 and refine both modules.

Implementation Details We use the Stable-Baselines3 [14] DQN implementation with: a batch size of 32, a replay buffer size of 50,000, a learning rate 10^{-5}, and an epsilon of 0.02. The policy network uses

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

*Seg. Modes 1 and 2 denote how events are segmented.

Table 1. Event Segmentation Results in Simulation Environments.

EfficientNet-B0 [17] for feature extraction, processed by 2 LSTM modules: LSTM-Event for modeling temporal dynamics and LSTM-QVal, to aggregate temporal features for the policy. Training updates the CNN and LSTM modules iteratively, with early training focusing on event perception and later refinement of the control policy. The framework is trained for $300k$ timesteps on UnrealCV-Gym [13].

2.1. Event Segmentation and Summarization

Generic event segmentation and snapshot creation are essential for embodied agents in dynamic settings, enabling 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: $B_t = \arg\max_t [H_t]$, with $H_t = -\sum_{i=1}^{N} p_{t,i} \log p_{t,i}$,, where H_t is the entropy at time t, and $p_{t,i}$ are normalized prediction errors: $p_{t,i} = \frac{\mathcal{L}_E(i)}{\sum_{j=1}^{N} \mathcal{L}_E(j)}$, $\mathcal{L}_E(i) = \|\mathbf{f}_{t+1,i} - \hat{\mathbf{f}}_{t+1,i}\|^2$. 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 $[B_k, B_{k+1}]$: $S_k = \arg\min_{t \in [B_k, B_{k+1}]} \mathcal{L}_E(t)$. S_k corresponds to the frame with the event's most stable observation.

3. Experimental Evaluation

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 FlexibleRoom 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 and adjusting thermostats, in 4-minute episodes that contain at least 10 actions. Three annotators review and mark event boundaries, assess tracking success, and summarization quality.

Model ↓ Env. →	City		Urban City		Rand. Room	
	AR	AL	AR	AL	AR	AL
TLD+PID [18]	12	**90**	19	**115**	25	**147**
MIL+PID [18]	**32**	59	24	50	21	43
MOSSE+PID [18]	16	56	**49**	68	**28**	62
Smart-Target [22]	232	473	233	466	403	458
Random-Target [22]	214	455	204	464	409	455
AD-VAT+ [23]	**326**	483	**322**	488	427	493
EASE-Supervised	248	**500**	290	**490**	**459**	**500**
PredLearn-PID [18]	114	343	71	349	115	319
EASE-Hybrid	233	**500**	**253**	**496**	**438**	**500**
EASE	**273**	**500**	155	443	214	491

Table 2. Performance Evaluation on the Active Tracking Task.

Evaluation Metrics and Baselines. Following prior work [16], 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 (2 to 15 frames), 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 [22, 23]. For real-world tracking evaluation, we use the average qualitative judgment from the annotators, who grade each frame as 1 (tracking) or 0 (not). 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 as the perception signal, as done in PA [16].

Simulation Results. We train and evaluate our model and baselines in increasingly challenging environments within UnrealCV-Gym. We evaluate EASE's performance in tracking and segmenting human actions across three simulation environments—City, Urban City, and Random Room—using IoU and action-level accuracy (Acc). Table 1 shows that EASE-Hybrid achieves the highest scores

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**

Table 3. Real-World Performance Evaluation of EASE for Active Event Perception Tasks.

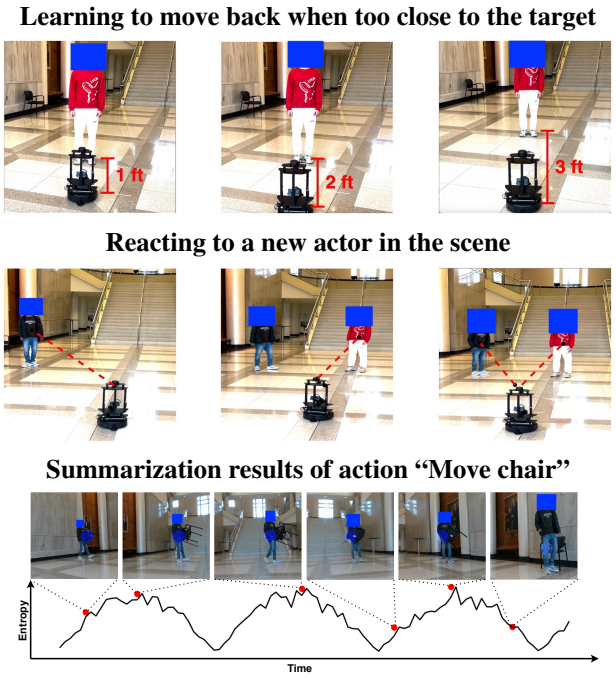

Figure 2. **Qualitative visualization** of emergent behaviors. Top: The robot learns to move backward when too close to the target. Middle: When a new actor is introduced, EASE tracks both old and new targets. Bottom: Summarization output of user actions.

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 segmentation and active perception. We also evaluate EASE on active object tracking—dynamically following salient actors in complex environments. As shown in Table 2, 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 it.

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 3 summarizes the results across segmentation (strict and relaxed) and summarization metrics, alongside tracking success rates. Table 3 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. 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 3, 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, while other models prioritize minimal redundancy

Qualitative Analysis EASE demonstrates emergent, memory-like behaviors by maintaining focus on salient targets through prediction-driven action. While effective, it can momentarily lose focus in low-motion scenes, highlighting a trade-off of its unsupervised approach. Figure 2 highlights some of these behaviors.

4. 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.

Acknowledgement. This research was supported in part by the US NSF grants IIS 2348689 and IIS 2348690 and USDA Grant 2023-69014-39716-1030191.

References

[1] Sathyanarayanan N. Aakur and Sudeep Sarkar. A perceptual prediction framework for self supervised event segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1197–1206, 2018. 1

[2] Sathyanarayanan N. Aakur and Sudeep Sarkar. Action localization through continual predictive learning. In *European Conference on Computer Vision*, 2020. 1

[3] Orcun Cetintas, Guillem Brasó, and Laura Leal-Taixé. Unifying short and long-term tracking with graph hierarchies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22877–22887, 2023. 1

[4] Jonathan Chung. Playing atari with deep reinforcement learning. *Comput. Ence*, 21:351–362, 2013. 2

[5] Ishan Dave, Rohit Gupta, Mamshad Nayeem Rizve, and Mubarak Shah. Tclr: Temporal contrastive learning for video representation. *Computer Vision and Image Understanding*, 219:103406, 2022. 1

[6] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010. 1, 2

[7] Karl Friston, Spyridon Samothrakis, and Read Montague. Active inference and agency: optimal control without cost functions. *Biological cybernetics*, 106:523–541, 2012. 1, 2

[8] Hayato Idei, Wataru Ohata, Yuichi Yamashita, Tetsuya Ogata, and Jun Tani. Emergence of sensory attenuation based upon the free-energy principle. *Scientific reports*, 12 (1):14542, 2022. 1

[9] Haofei Kuang, Yi Zhu, Zhi Zhang, Xinyu Li, Joseph Tighe, Sören Schwertfeger, Cyrill Stachniss, and Mu Li. Video contrastive learning with global context. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3195–3204, 2021. 1

[10] Ramy Mounir, Roman Gula, Jörn Theuerkauf, and Sudeep Sarkar. Spatio-temporal event segmentation for wildlife extended videos. In *International Conference on Computer Vision and Image Processing*, 2021. 1

[11] Thomas Parr, Giovanni Pezzulo, and Karl J Friston. *Active inference: the free energy principle in mind, brain, and behavior*. MIT Press, 2022. 1

[12] Giovanni Pezzulo, Francesco Rigoli, and Karl J Friston. Hierarchical active inference: a theory of motivated control. *Trends in cognitive sciences*, 22(4):294–306, 2018. 1

[13] Weichao Qiu, Fangwei Zhong, Yi Zhang, Siyuan Qiao, Zihao Xiao, Tae Soo Kim, and Yizhou Wang. Unrealcv: Virtual worlds for computer vision. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1221–1224, 2017. 3

[14] Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8, 2021. 2

[15] Pierre Sermanet, Corey Lynch, Yevgen Chebotar, Jasmine Hsu, Eric Jang, Stefan Schaal, Sergey Levine, and Google Brain. Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE, 2018. 1

[16] Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8075–8084, 2021. 3

[17] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[18] Shubham Trehan and Sathyanarayanan 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*, pages 783–792, 2022. 1, 3

[19] Shubham Trehan and Sathyanarayanan N. Aakur. Self-supervised multi-actor social activity understanding in streaming videos. *ArXiv*, abs/2406.14472, 2024. 1

[20] Jeffrey M Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological bulletin*, 127(1):3, 2001. 1, 2

[21] Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. Ad-vat: An asymmetric dueling mechanism for learning visual active tracking. In *International Conference on Learning Representations*, 2018. 1

[22] Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. AD-VAT: An asymmetric dueling mechanism for learning visual active tracking. In *International Conference on Learning Representations*, 2019. 1, 3

[23] Fangwei Zhong, Peng Sun, Wenhan Luo, Tingyun Yan, and Yizhou Wang. Ad-vat+: An asymmetric dueling mechanism for learning and understanding visual active tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1467–1482, 2021. 1, 3