DeFi Intent Discovery via Maximum Entropy Inverse Reinforcement Learning

Aizierjiang Aiersilan[2], Jerome Yen[2], Sheng Wang[1,*]

[1]Joint Lab of Finance and Business Intelligence, Guangdong Institute of Intelligence Science and Technology

[2]Computer and Information Science, The University of Macau

***Abstract*—Decentralized Finance (DeFi) transaction sequences can obscure a user's high-level goal behind multi-step smart-contract calls, routing abstractions, and rapidly changing market conditions. Many prior intent-mining pipelines rely on transaction-level semantic labels, which can miss the sequential decision structure of long-horizon strategies. We formulate DeFi intent discovery as sequential reward inference and learn a parametric reward $R_\theta(s, a)$ from expert demonstrations via Maximum Entropy Inverse Reinforcement Learning (MaxEnt IRL). To capture short-horizon market trends under volatility and partial observability, we augment the state with a temporal market-gradient term $\nabla_t \mathbf{m}$, defined as the block-to-block finite difference of market-state variables. We refer to this state design as Chemo-IRL. To enable controlled synthetic evaluation under intent hiding, we introduce Gym-DeFi, a Gymnasium-compatible simulator with a configurable action-obfuscation channel that corrupts observed action identifiers during data generation. We evaluate reward recovery using Reward Recovery Error (RRE) and downstream intent probing using macro-F1 from a fixed post-hoc attribution-based decoder. On this controlled synthetic benchmark under high obfuscation, Chemo-IRL attains the lowest RRE among intent-label-free baselines while remaining competitive on macro-F1 within the same label-free setting. These results should be interpreted as benchmark-level comparisons in a controlled synthetic environment rather than as direct evidence of real-world DeFi deployment performance.**

***Index Terms*—Decentralized Finance, Intent Discovery, Reinforcement Learning, Blockchain Analytics, Sequential Decision Making**

I. Introduction

The Ethereum blockchain [1], [2] and its Layer-2 scaling solutions have created an open, adversarial financial environment where user goals are rarely stated explicitly. A user's intent, the objective underlying a sequence of transactions, is often entangled with low-level contract calls and rapidly evolving market conditions. This makes intent inference challenging, especially when action identifiers (e.g., protocol/function signatures) are uninformative or noisy. More broadly, prior work in cryptocurrency analytics suggests that volatile digital-asset markets are sensitive to external information signals such as news and social-media activity [3].

Many existing approaches emphasize transaction-level semantic annotation or supervised intent labeling of atomic actions. While useful, static classification can obscure temporal dependencies and long-horizon strategies composed of multiple steps. For example, a user may approve a router to spend USDC, swap USDC for WETH, deposit WETH as collateral, and then borrow USDC. A static classifier may label these as separate atomic intents, whereas the composite strategy corresponds to leveraged exposure that is defined by sequential composition and implicit utility maximization.

Rather than directly predicting an intent label from a sequence (e.g., via an LSTM or Transformer classifier), we seek an explicit utility signal that explains expert behavior and supports analysis beyond classification (e.g., preference inference and counterfactual evaluation) [4]. We therefore cast DeFi intent discovery as sequential reward inference in a Markov Decision Process and learn a parametric reward model under the Maximum Entropy principle from expert demonstrations. The learned model assigns higher probability to trajectories that accrue higher discounted reward, while remaining stochastic to account for noise and partial observability.

Our analogy is used as a modeling motivation rather than an IRL objective. In bacterial chemotaxis, agents bias actions using temporal sensing of concentration changes. Analogously, we include a temporal market-gradient term $\nabla_t \mathbf{m}$ in the state to expose short-horizon trends that are informative for multi-step DeFi strategies under volatility and partial observability.

We study intent discovery under action-identifier obfuscation, where logged action identifiers may be corrupted during data generation. Our contributions are:

- **Chemo-IRL.** We cast DeFi intent discovery as sequential reward inference and instantiate MaxEnt IRL with a compact reward model, using a temporal market-gradient state term $\nabla_t \mathbf{m}$ and a deterministic semantic feature interface $\phi(s, a)$.
- **Gym-DeFi**. We introduce a Gymnasium-compatible simulator for controlled and reproducible DeFi-style sequential decision making, including a configurable action-obfuscation channel and standard evaluation interfaces.
- **Reproducible evaluation.** On a lightweight synthetic benchmark, we compare supervised, heuristic, and label-free baselines using reward recovery (RRE) and downstream intent macro-F1 under a fixed decoder, highlighting the trade-off between behavior matching and discrete intent decoding under high obfuscation.

* Corresponding author Email: wangsheng@gdiist.cn.
Code: https://github.com/Ezharjan/ChemoIRL

II. Related Work

A. Intent Mining in Blockchain

Recent blockchain analysis increasingly emphasizes semantic lifting of smart-contract interactions from calldata, logs, and execution traces into higher-level representations. In DeFi, intent is often expressed through compositional, cross-protocol sequences, and prior work has studied how complex transactions can be decomposed into reusable building blocks [5]. The TIM framework proposes an intent taxonomy and assigns intent labels from transaction-level signals [6]. In contrast to semantic annotation that operates at the transaction level, we model user behavior as a sequential decision process and infer a latent reward that explains observed trajectories, with intent treated as a downstream interpretation of the recovered reward.

B. Inverse Reinforcement Learning

Inverse Reinforcement Learning (IRL) seeks to recover a reward function from expert demonstrations [7]. Maximum Entropy IRL resolves reward ambiguity by modeling expert trajectories as a maximum entropy distribution [8], and has inspired modern imitation and distribution-matching baselines such as GAIL, AIRL, SQIL, and ValueDICE [9]–[12]. Several IRL directions are closely related to intent discovery, including learning mixtures of latent rewards from unlabeled demonstrations in multi-intention IRL [13] and adversarial settings where agents may hide strategy from an inverse learner [14]. Recent work also highlights that matching demonstrations may not imply recovering task-relevant rewards, framing a gap between data alignment and task alignment [15], which echoes the trade-off we observe between reward recovery and downstream intent classification.

C. Biological Chemotaxis Models

Bacterial chemotaxis illustrates gradient seeking under uncertainty, where temporal sensing biases motion toward attractants [16], [17]. We use this literature only as a modeling analogy to motivate a temporal gradient term in the state, rather than to model biochemical mechanisms or solve chemotaxis dynamics.

III. Methodology

We use standard Markov Decision Process notation. a_t^{clean} denotes the pre-obfuscation action identifier and a_t denotes the observed identifier used for learning and evaluation.

A. Chemotactic Markov Decision Process

We define the DeFi environment as a Markov Decision Process $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, R^*, \gamma \rangle$, where $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the stochastic transition function induced by smart contract execution and market dynamics (e.g., AMM reserve updates, lending rate changes, and gas conditions), and $R^* : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is an unknown latent reward function. Chemo-IRL learns a parametric reward model $R_\theta(s, a)$ that explains expert demonstrations under the Maximum Entropy principle, with discount factor $\gamma = 0.99$. Figure 1 summarizes the Chemo-IRL workflow described in this section, from the bio-mimetic state and observed action identifiers to semantic feature extraction and MaxEnt IRL reward learning.

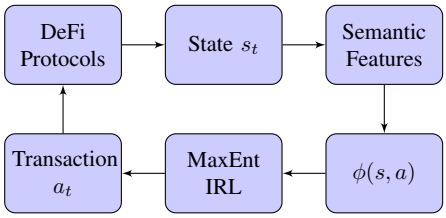

Fig. 1: Chemo-IRL overview. The agent observes a bio-mimetic DeFi state, extracts semantic features $\phi(s, a)$ from observed action identifiers (possibly obfuscated), and learns a reward model under a MaxEnt IRL objective to guide transaction selection.

1) State Space: The state $s_t \in \mathbb{R}^d$ is a composite vector

$$s_t = \mathbf{u}_t \oplus \mathbf{m}_t \oplus \nabla_t \mathbf{m}, \tag{1}$$

where $\mathbf{u}_t \in \mathbb{R}^{d_u}$ represents user state (e.g., wallet token balances, approvals, and position related statistics), $\mathbf{m}_t \in \mathbb{R}^{d_m}$ represents market state (e.g., token prices, AMM pool reserves, lending rate signals, and gas related conditions), and the temporal gradient is

$$\nabla_t \mathbf{m} = \frac{\mathbf{m}_t - \mathbf{m}_{t-1}}{\Delta t}. \tag{2}$$

We set Δt to the simulator block time, which is aligned with Ethereum's typical block interval (approximately 12 seconds). This gradient captures short horizon market changes and implements chemotactic temporal sensing. In our environment, $d_u = 20$ and $d_m = 22$, giving $d = d_u + 2d_m = 64$.

2) Action Space: The action space is hierarchical. Each action corresponds to a transaction type

$$a = \langle \mathcal{P}_{id}, \mathcal{F}_{sig}, b \rangle, \tag{3}$$

where $\mathcal{P}_{id} \in \{1, \ldots, N_p\}$ is the protocol identifier, $\mathcal{F}_{sig} \in \{1, \ldots, N_f\}$ is the function signature (or function family), and $b \in \{1, \ldots, M\}$ is a discretized value bin that coarsely represents transaction size. This yields a discrete action space of size $|\mathcal{A}| = N_p \times N_f \times M$. For experiments, we use $N_p = 5$, $N_f = 20$, and $M = 10$, giving $|\mathcal{A}| = 1000$.

3) Observed actions under obfuscation: We model action-identifier obfuscation by observing a_t generated from a clean identifier a_t^{clean} via the mixture channel

$$a_t \sim (1 - p_{\text{obf}})\, \delta(\cdot = a_t^{\text{clean}}) + p_{\text{obf}}\, \text{Unif}(\mathcal{A}), \tag{4}$$

applied independently at each step. Unless otherwise stated, all learning and evaluation use trajectories $\{(s_t, a_t)\}_{t=0}^{T}$, and semantic features and intent decoding are computed from the observed identifiers.

B. Maximum Entropy IRL Formulation

Following Ziebart et al. [8], we model the trajectory distribution under the Maximum Entropy principle as

$$P(\tau \mid \theta) = \frac{1}{Z(\theta)} \exp\left(\sum_{t=0}^{T} \gamma^t R_\theta(s_t, a_t) \right), \tag{5}$$

where $Z(\theta)$ is the partition function and $\gamma \in [0,1)$ is the discount factor.

We instantiate the reward as a parametric function of semantic features,

$$R_\theta(s,a) = f_\theta\big(\phi(s,a)\big), \tag{6}$$

where $\phi : \mathcal{S} \times \mathcal{A} \to [0,1]^K$ is a K-dimensional semantic feature map. In our implementation, $K = 21$ and the feature taxonomy is aligned with prior intent categories [6]. The reward model f_θ is realized as a lightweight neural network, similar in spirit to deep MaxEnt IRL parameterizations [18].

In practice, computing the partition function and the model expectation is intractable, so we use reward-guided rollouts $\mathcal{D}_\theta$ as negative samples and optimize a contrastive surrogate. Let $G_\theta(\tau) \triangleq \sum_{t=0}^{T} \gamma^t R_\theta(s_t, a_t)$. We minimize

$$\tilde{\mathcal{L}}(\theta) = \mathbb{E}_{\tau\sim\mathcal{D}_\theta}[G_\theta(\tau)] - \mathbb{E}_{\tau\sim\mathcal{D}_E}[G_\theta(\tau)]\,, \tag{7}$$

with gradients computed by backpropagation through R_θ.

C. Semantic Feature Extraction

Raw on-chain traces contain low-level fields such as function selectors and integer arguments that are not directly amenable to intent-level reasoning. We therefore construct a lightweight semantic feature map $\phi(s,a) \in [0,1]^{21}$ from the bio-mimetic state and decoded transaction actions. These features capture interpretable signals related to swapping, lending, liquidity provision, yield seeking, governance, risk exposure, and temporal patterns. All features are computed deterministically and clipped to $[0,1]$, enabling reproducible evaluation of the IRL core without relying on external semantic models.

D. Gym-DeFi Environment

We implement `Gym-DeFi`, a Gymnasium-compatible environment [19] for controlled DeFi-style sequential decision making. The environment maintains the 64-dimensional state $s_t = \mathbf{u}_t \oplus \mathbf{m}_t \oplus \nabla_t \mathbf{m}$, driven by stochastic market dynamics. Each `step` decodes a discrete action into $(protocol\_id, function\_sig, value\_bin)$ with $|\mathcal{A}| = 1000$ and updates the state accordingly; episodes terminate stochastically or at horizon $T_{\max}$.

E. Training Procedure

Chemo-IRL alternates between generating rollouts under the current reward and updating θ so that expert demonstrations achieve higher discounted return than sampled rollouts. Rollouts use ε-greedy selection: with probability ε sample uniformly from $\mathcal{A}$, otherwise choose $a = \arg\max_{a'\in\mathcal{A}} R_\theta(s,a')$, and update θ by minimizing Eq. (7). The full-sweep $\arg\max$ is tractable for $|\mathcal{A}| = 1000$ but scales linearly with $|\mathcal{A}|$.

IV. Experiment

A. Dataset

We generate a synthetic dataset of DeFi trajectories in `Gym-DeFi` under controlled and reproducible conditions. Each trajectory is a variable-length state–action sequence with length sampled between 5 and 15 steps and may terminate early due to stochastic episode termination. States and actions follow the definitions in Section III.

Trajectories are associated with six intent classes aligned with the TIM taxonomy [6], including three simple intents (`swap`, `lend`, `liquidity`) and three complex intents (`yield`, `complex_leverage`, `complex_LP`). Intent labels are used for evaluation and for supervised baselines only. Intent-label-free methods are trained solely from expert trajectories.

To emulate intent hiding at the action-identifier level, we apply the action-obfuscation channel defined in Section III-A3 during data generation, with $p_{\text{obf}} = 0.7$ (Table I). Unless otherwise stated, all methods consume trajectories of the form $\{(s_t, a_t)\}_{t=0}^{T}$, and semantic features and downstream intent decoding are computed from the observed identifiers.

We use an 80/20 train-test split with a fixed seed on 200 trajectories. Conclusions should be interpreted as benchmark-level comparisons under a fixed configuration rather than estimates of real-world on-chain performance.

B. Baselines

We compare Chemo-IRL with supervised intent classifiers, a heuristic baseline, and intent-label-free reward-learning baselines under the same deterministic semantic feature map $\phi(s,a)$, train/test split, and trajectories logged with observed (possibly obfuscated) action identifiers. Supervised baselines are TIM [6] and a supervised MLP trajectory classifier trained with intent labels to predict trajectory intent from aggregated semantic features; we evaluate multiple MLP configurations and report the best-performing configuration in Table II. The heuristic baseline *Action Frequency* predicts intent by majority voting over observed action categories. Intent-label-free baselines include SQIL [11], Linear GAIL [9], and ValueDICE [12], which learn from expert trajectories without intent labels; for these reward-learning methods (and Chemo-IRL), discrete intent labels for macro-F1 are obtained via the fixed attribution-based decoder in Section IV-C.

C. Evaluation Metrics

Macro-F1 (Overall / Simple / Complex): We report macro-averaged F1 over the six intent classes on the held-out test set. We additionally report macro-F1 on the subset of trajectories whose ground-truth labels belong to the three *simple* intents (swap, lend, liquidity) and the three *complex* intents (yield, complex_leverage, complex_LP), respectively. For reward-learning methods, macro-F1 is treated as a downstream probe obtained via a fixed label-free decoder, rather than a quantity directly optimized during training.

Trajectory-level intent prediction for reward models (attribution-based intent probe): For methods that learn a reward $R_\theta(s,a) = f_\theta(\phi(s,\tilde{a}))$, we derive a trajectory intent label using feature attribution. For a semantic feature dimension k, define the per-step attribution

$$A_k(s_t,a_t) = \phi_k(s_t,\tilde{a}_t) \cdot \frac{\partial R_\theta(s_t,\tilde{a}_t)}{\partial \phi_k(s_t,\tilde{a}_t)}.$$

We then accumulate discounted attributions along a trajectory τ,

$$S_k(\tau) = \sum_{t=0}^{T} \gamma^t A_k(s_t, a_t),$$

and predict the intent as the taxonomy group with the largest total attribution score. This yields trajectory-specific intent assignments without training an additional classifier on top of the learned reward.

Reward Recovery Error (RRE): We quantify reward recovery via feature expectation mismatch. Let

$$\mu(\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \gamma^t \phi(s_t, a_t). \tag{8}$$

Given expert test trajectories $\mathcal{D}_{\text{test}}$ and agent rollouts $\mathcal{D}_\theta$ generated under the learned reward, we compute

$$\text{RRE} = \frac{\|\mu(\mathcal{D}_{\text{test}}) - \mu(\mathcal{D}_\theta)\|_2}{\|\mu(\mathcal{D}_{\text{test}})\|_2 + 10^{-6}}. \tag{9}$$

Lower RRE indicates closer matching between expert and induced behavior. For purely supervised baselines that do not learn a reward, RRE is not applicable.

D. Hyperparameters

Table I summarizes the hyperparameters used in all experiments.

TABLE I: Hyperparameter Configuration

Component	Parameter	Value
Data / Split	# trajectories	200
	Train/Test split	0.8 / 0.2
	Obfuscation prob. p_{obf}	0.7
	Random seed	42
Environment	State dim d	64
	Action dim $\lvert\mathcal{A}\rvert$	1000
	Max horizon $T_{\max}$	15
	Discount γ	0.99
Chemo-IRL	Reward net	2-layer MLP (h=64)
	Optimizer	Adam (wd 10^{-4})
	Learning rate	5×10^{-3}
	Training iterations I	60
Rollouts	ε-greedy exploration	0.05

Unless otherwise stated, the environment and rollout-related settings are shared across Chemo-IRL and the intent-label-free reward-learning baselines to ensure a consistent comparison under the same benchmark interface.

V. Results

Experimental context. Results are reported on synthetic `Gym-DeFi` trajectories with deterministic semantic features and action-identifier obfuscation $p_{\text{obf}} = 0.7$. The benchmark uses a fixed split and seed, so the numbers are best interpreted as comparisons under a controlled configuration rather than estimates on real blockchain data.

TABLE II: Overall macro-F1 and reward recovery on obfuscated trajectories. TIM and the supervised MLP baseline use intent labels and provide an upper bound. Among intent-label-free methods, SQIL attains the highest macro-F1, while Chemo-IRL achieves the lowest RRE. † denotes our method.

Method	Overall Macro-F1	RRE↓
TIM (Supervised)	**0.927**	–
Supervised MLP (best)	0.901	–
Action Frequency (Heuristic)	0.092	–
SQIL (Intent-label-free)	**0.474**	0.809
ValueDICE (Intent-label-free)	0.359	1.049
Linear GAIL (Intent-label-free)	0.240	0.938
Chemo-IRL† (Intent-label-free)	0.394	**0.478**

A. Reward recovery and downstream intent macro-F1

Table II reports overall macro-F1 for all methods. Reward Recovery Error (RRE; lower is better) is reported only for reward-learning methods, and is not applicable to purely supervised or heuristic baselines. For intent-label-free reward-learning methods, macro-F1 is computed via the fixed attribution-based decoder.

Figure 2 visualizes the overall macro-F1 results in Table II. It highlights the performance gap between supervised methods (upper bound) and intent-label-free reward-learning methods under high action-identifier obfuscation, and shows that the action-frequency heuristic degrades sharply in this setting.

Supervised vs. intent-label-free. Supervised methods achieve high macro-F1 (TIM: 0.927; best supervised MLP: 0.901) since they directly optimize intent classification with access to labels. Under high action-identifier obfuscation, the action-frequency heuristic collapses (0.092), indicating that simple parsing of action identifiers is insufficient in this setting.

Reward recovery vs. macro-F1. Among intent-label-free methods, SQIL yields the highest macro-F1 (0.474), while Chemo-IRL achieves the lowest RRE (0.478), improving over SQIL (0.809), Linear GAIL (0.938), and ValueDICE (1.049). This indicates that Chemo-IRL matches expert feature expectations more closely under action-identifier noise. At the same time, macro-F1 is derived from a fixed post-hoc decoder rather than optimized directly, so reward recovery and macro-F1 can diverge under taxonomy-aligned discrete decoding.

This trade-off is summarized in Figure 3, which plots intent-label-free methods in the (macro-F1, RRE) plane. Chemo-IRL attains the lowest RRE (0.478), indicating the closest feature-expectation matching to expert trajectories, while SQIL achieves higher macro-F1 (0.474) but with substantially larger RRE (0.809). These results suggest that, under the fixed post-hoc decoder, better reward recovery does not necessarily translate into the highest taxonomy-aligned macro-F1.

Robustness under obfuscation. Across intent-label-free methods, Chemo-IRL maintains competitive macro-F1 while achieving the lowest RRE, indicating closer feature-expectation matching under action-identifier noise than simple heuristics and distribution-matching baselines.

Fig. 2: Overall macro-F1 on the test split. Supervised methods (TIM and the best supervised MLP baseline) provide an upper bound. Among intent-label-free methods, SQIL attains the highest macro-F1, while Chemo-IRL achieves the lowest reward recovery error (Table II).

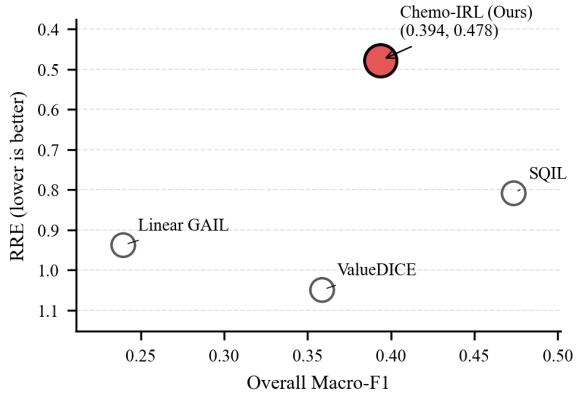

Fig. 3: Trade-off between intent prediction (overall macro-F1) and reward recovery (RRE; lower is better) among intent-label-free methods. Chemo-IRL achieves the lowest RRE (0.478) with competitive macro-F1 (0.394). (Note: the y-axis is inverted so lower RRE appears higher.)

B. Ablation Studies

We ablate two components of Chemo-IRL: (i) the biomimetic temporal gradients and (ii) the semantic feature map. To keep ablations comparable in compute, we run all variants under the same reduced configuration (15 reward-learning iterations with learning rate $\alpha = 0.01$). Since this setting differs from the main configuration, we focus on relative changes across variants rather than comparing absolute values to Table II.

Temporal gradients. Removing temporal gradients reduces overall macro-F1 by $\Delta = -0.180$ (-32.3% relative drop) and collapses complex-intent macro-F1 to zero, suggesting that chemotactic temporal sensing is critical for capturing multi-step strategies.

Semantic features. Replacing semantic features with random vectors reduces overall macro-F1 by $\Delta = -0.150$ (-26.9% relative drop) and also eliminates complex-intent detection, suggesting that the designed semantic feature map carries non-trivial signal beyond chance.

VI. Discussion and Limitations

A. Discussion

A key empirical observation is the divergence between reward recovery and downstream intent macro-F1 among intent-label-free methods. Figure 3 summarizes this trade-off: Chemo-IRL attains the lowest RRE, whereas SQIL attains higher taxonomy-aligned macro-F1 under the fixed decoder but with substantially higher RRE. Because macro-F1 depends not only on reward quality but also on the alignment between the fixed decoder and the intent taxonomy, improving reward recovery does not necessarily maximize macro-F1 under action-identifier noise. From an application perspective, this suggests a hybrid workflow: supervised semantic labeling when intent labels are available, and reward-based behavioral analysis (e.g., preference inference or counterfactual evaluation) when they are not.

The diagnostic ablation also supports the inclusion of temporal market gradients in the state representation: removing them sharply degrades complex-intent probing performance. While this does not establish a definitive causal claim beyond the benchmark, it is consistent with the interpretation that temporal-gradient information helps capture multi-step strategies in a volatile, stateful environment.

B. Limitations and Future Work

Limitations. First, our evaluation is conducted in a controlled synthetic environment with deterministic semantic features. While this setting enables reproducible analysis of reward recovery trends under a clearly specified obfuscation channel, it does not fully capture real on-chain intent hiding mechanisms (e.g., proxy routing, multi-hop execution, and cross-protocol compositionality), and therefore does not support direct claims about real-world deployment.

Second, the benchmark is intentionally lightweight (200 trajectories) and evaluated under a fixed split and seed. We do not report confidence intervals or multi-seed variance in this version, so results should be interpreted as comparisons

under a fixed controlled configuration rather than statistically comprehensive estimates.

Third, for reward-learning methods, discrete intent labels are derived from a post-hoc attribution-based probe rather than being optimized directly for classification. This can limit macro-F1 even when reward recovery (RRE) is strong, and it partially explains the observed macro-F1–RRE trade-off.

Fourth, the current action abstraction discretizes transaction types into a fixed action set and our rollout policy uses a full action sweep to select $\arg\max_{a\in\mathcal{A}} R_\theta(s,a)$, which is tractable for $|\mathcal{A}| = 1000$ but may not scale to richer DeFi action spaces without candidate pruning or proposal mechanisms.

Future work. First, we will harden the benchmark by generating synthetic datasets where atomic action identities are decorrelated from intent, forcing models to rely on state transitions and temporal gradients rather than action templates. Second, we plan to validate Chemo-IRL on real transaction traces with a controlled annotation pipeline. A key direction is to replace hand-crafted semantic features with LLM-assisted semantic extraction from raw calldata/logs into structured intent-relevant representations while maintaining schema validation and caching for reproducibility. Third, bridging the gap between reward recovery (RRE) and downstream intent probing (macro-F1) may benefit from learning a lightweight intent decoder on top of recovered reward/attributions or incorporating limited supervision for calibration. Finally, extending to multi-agent settings (e.g., MEV-aware dynamics) is a promising direction for modeling strategic interactions beyond single-agent trajectories.

VII. Conclusion

We presented Chemo-IRL, a MaxEnt IRL framework for DeFi intent discovery from transaction trajectories. By augmenting the state with temporal market gradients, the model can condition recovered reward on short-horizon trends rather than static snapshots. On the `Gym-DeFi` synthetic benchmark with high action-identifier obfuscation, Chemo-IRL achieves the lowest RRE among intent-label-free baselines while remaining competitive on macro-F1, revealing a benchmark-level trade-off between reward recovery and taxonomy-aligned intent probing under a fixed post-hoc decoder. The paper contributes a sequential reward-inference formulation, a reproducible DeFi benchmark with configurable obfuscation, and an explicit comparison of the reward-recovery versus intent-decoding trade-off.

Acknowledgment

This work was supported by the Guizhou Provincial Science and Technology Plan Project (Grant No. QianKeHeZhongDa [2025]031) and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2023B1515130002).

References

[1] G. Wood *et al.*, "Ethereum: A secure decentralised generalised transaction ledger," *Ethereum project yellow paper*, vol. 151, no. 2014, pp. 1–32, 2014.

[2] V. Buterin *et al.*, "A next-generation smart contract and decentralized application platform," *white paper*, vol. 3, no. 37, pp. 2–1, 2014.

[3] X. Chen, S. Wang, X. Yang, S. Wang, and J. Yen, "How to use social media for bitcoin price prediction: A multi-source data fusion method based on cstnet," in *International Conference on Intelligent Computing*. Springer, 2025, pp. 82–94.

[4] B. Li, J. Yen, and S. Wang, "Uncovering financial statement fraud: A machine learning approach with key financial indicators and real-world applications," *IEEE Access*, vol. 12, pp. 194 859–194 870, 2024.

[5] S. Kitzler, F. Victor, P. Saggese, and B. Haslhofer, "Disentangling decentralized finance (defi) compositions," *ACM Transactions on the Web*, vol. 17, no. 2, pp. 1–26, 2023.

[6] Q. Mao, Y. Zhang, J. Chen, W. Zhou, and J. Yan, "Know your intent: An autonomous multi-perspective llm agent framework for defi user transaction intent mining," *arXiv preprint arXiv:2511.15456*, 2025.

[7] A. Y. Ng and S. J. Russell, "Algorithms for inverse reinforcement learning," in *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, p. 663–670.

[8] B. D. Ziebart, A. L. Maas, J. A. Bagnell, A. K. Dey *et al.*, "Maximum entropy inverse reinforcement learning." in *Aaai*, vol. 8. Chicago, IL, USA, 2008, pp. 1433–1438.

[9] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.

[10] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *arXiv preprint arXiv:1710.11248*, 2017.

[11] S. Reddy, A. D. Dragan, and S. Levine, "Sqil: Imitation learning via reinforcement learning with sparse rewards," *arXiv preprint arXiv:1905.11108*, 2019.

[12] I. Kostrikov, O. Nachum, and J. Tompson, "Imitation learning via off-policy distribution matching," *arXiv preprint arXiv:1912.05032*, 2019.

[13] A. Bighashdel, P. Meletis, P. Jancura, and G. Dubbelman, "Deep adaptive multi-intention inverse reinforcement learning," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021, pp. 206–221.

[14] K. Pattanayak, V. Krishnamurthy, and C. Berry, "Inverse-inverse reinforcement learning. how to hide strategy from an adversarial inverse reinforcement learner," in *2022 IEEE 61st Conference on Decision and Control (CDC)*. IEEE, 2022, pp. 3631–3636.

[15] W. Zhou and W. Li, "Rethinking inverse reinforcement learning: from data alignment to task alignment," *Advances in Neural Information Processing Systems*, vol. 37, pp. 27 647–27 688, 2024.

[16] H. C. Berg and D. A. Brown, "Chemotaxis in escherichia coli analysed by three-dimensional tracking," *nature*, vol. 239, no. 5374, pp. 500–504, 1972.

[17] V. Sourjik and N. S. Wingreen, "Responding to chemical gradients: bacterial chemotaxis," *Current opinion in cell biology*, vol. 24, no. 2, pp. 262–268, 2012.

[18] M. Wulfmeier, P. Ondruska, and I. Posner, "Maximum entropy deep inverse reinforcement learning," *arXiv preprint arXiv:1507.04888*, 2015.

[19] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. De Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG *et al.*, "Gymnasium: A standard interface for reinforcement learning environments," *arXiv preprint arXiv:2407.17032*, 2024.