Guided Reward Strategy in Imperfect-Information Games

Jiayang Wang, Yang Xu, Chuangji Zeng
Jiangxi Agricultural University, Nanchang University

1. Research Motivation & Value

- **Extreme Decision Complexity:** Overcoming partial observability & hidden state inference in imperfect-information games.
- **Real-World Mapping:** Applications in intelligent transport, aerospace scheduling, and medical decision-making.
- **Core Challenges:** Tackling information asymmetry, diverse strategies, stochastic dynamics, and short vs. long-term planning.

2. Technical Foundations

- **Spatio-Temporal Modeling:** CNN for card patterns, ResNet for stable gradients, and LSTM/GRU for action sequence dependency.
- **Partially Observable RL:** MDP extension for non-Markovian properties, belief state estimation, and value functions dependent on observation history.
- **Sample Complexity Bottleneck:** Highlighting the necessity of guided rewards to accelerate learning in sparse reward environments.

3. Mahjong Encoding & Replay

- **Multi-View Encoding:** Dual hand representation (4x9 matrix + 34-bit one-hot), integer feature encoding, and n-gram action history hashing.
- **Effective Tile Estimation:** Quantifying win potential by estimating completion rate with unseen tiles.
- **Intelligent Replay Buffer:** Timestamp frequency capping, and dynamic 1:1 win/loss sampling to prevent concept drift and policy bias.

4. Guided Reward Architecture

"Guide → Reward Prediction → RL Policy" Loop

Supervised Guide Network
Soft Label (Win-rate Prediction) + Symmetric Data Augmentation

↓

Dual-Path GRU Reward Prediction
Temporal Difference Encoding (Δ Hidden State)

↓

Hybrid Reward Formula
$R = R_{base} + R_{extended} + R_{predicted}$

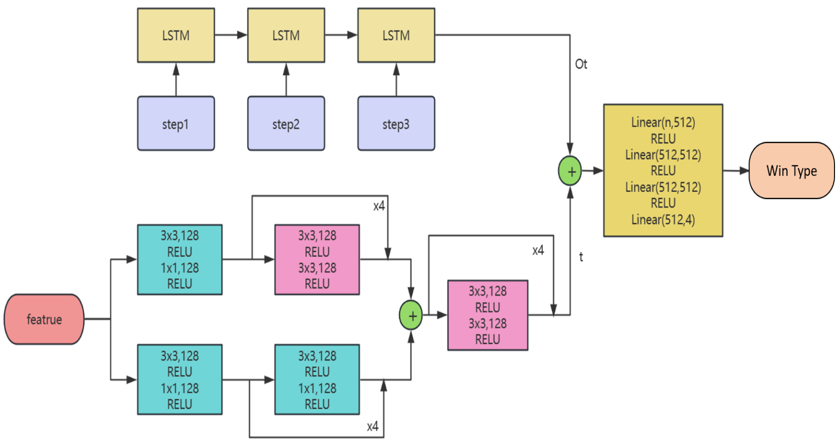
↓

RL Agent (DQN/PPO)
Receives dense feedback for policy updates

5. Experiments & Insights

- **Guide Network Performance:** Achieved 85% test accuracy, a +6% improvement over CNN baseline. Error concentration in rare hand patterns identified.
- **Reward Prediction Convergence:** Dual GRU loss dropped to 1.0 within 75 epochs (vs. >2.0 for CNN). Optimal 128-dim hidden layer found.
- **End-to-End RL Effect:** DQN with hybrid rewards boosted win-rate from 10% to 30% in 1.5M steps, approaching PPO's upper limit.

Metric	Traditional RL	Hybrid Reward RL
Sample Efficiency	Baseline	40% Less Data
Final Win-Rate	18%	30%
Avg. Game Score	+15	+50 (Stable)



6. Conclusion & Next Steps

Key Technical Contributions

- Feature-free guider with prior knowledge from data augmentation.
- Dual-GRU for dense, unbiased credit assignment.
- Universal integration with mainstream RL algorithms (DQN/PPO).
- Dynamic replay buffer to prevent sampling bias and drift.

Future Research Directions

- Refine hand classification to reduce errors on rare patterns.
- Implement active learning for high-uncertainty sample labeling.
- Automate reward function tuning via meta-gradients.
- Validate model transferability to other domains (e.g., Poker, StarCraft).