SUB-PLAY: Adversarial Policies against Partially Observed Multi-Agent Reinforcement Learning Systems

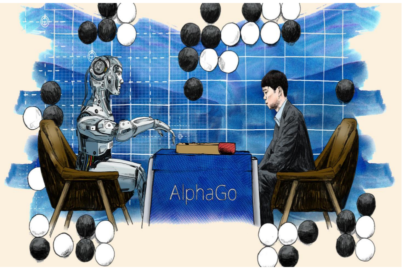
Oubo Ma, Yuwen Pu, Linkang Du, Yang Dai, Ruo Wang,
Xiaolei Liu, Yingcai Wu, and Shouling Ji

浙江大學
ZHEJIANG UNIVERSITY

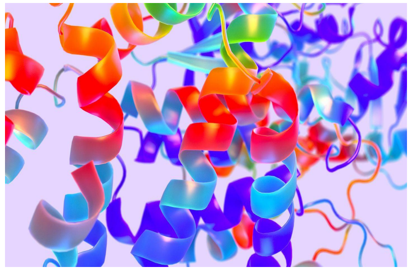
Reinforcement Learning

- **Reinforcement learning** is a machine learning paradigm where an agent learns to make optimal **sequential decisions** in an environment by maximizing cumulative rewards through trial and error.

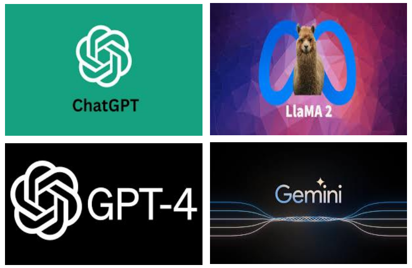
Atari 2600



AlphaGo



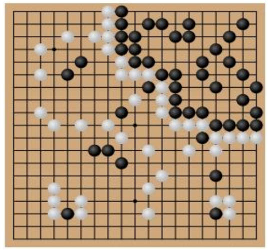
AlphaFold



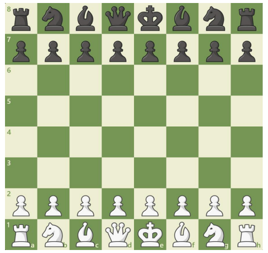
RLHF

Competitive Environment

- A **<u>competitive environment</u>** is a context where multiple agents interact with conflicting objectives, engaging in strategic decision-making to optimize their outcomes.



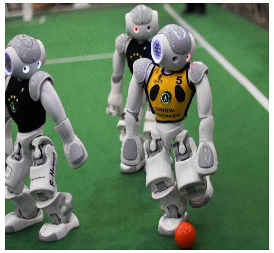
Go



Chess

Poker

Auction



Soccer

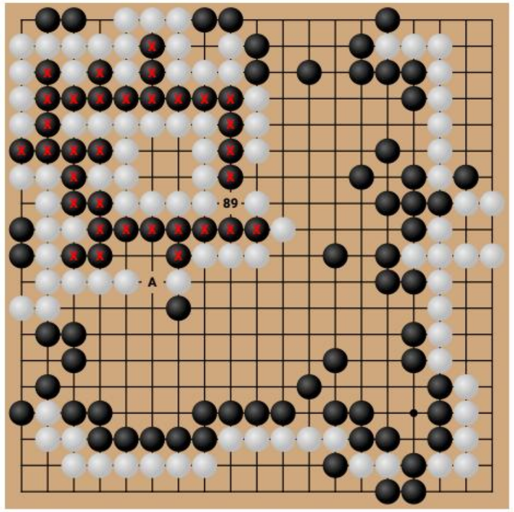
Wargame

Adversarial Policy

- Is it safe to deploy a reinforcement learning system in a competitive environment?

Adversarial Policy

- Is it safe to deploy a reinforcement learning system in a competitive environment?
- The attacker can obtain **adversarial policies** that achieve over a 97% win rate against **KataGo**, an AlphaZero-style superhuman Go AI, with training costs under 14% of KataGo's.



Adversarial Policies Beat Superhuman Go AIs. [Wang et al., ICML 2023]

Adversarial Policy

- **Adversarial policies** are a class of sequential decision-making policies used to minimize the cumulative rewards of a specific reinforcement learning system.

Adversarial Policy

- **Adversarial policies** are a class of sequential decision-making policies used to minimize the cumulative rewards of a specific reinforcement learning system.

- Adversarial policies exist because RL training in competitive environments relies on **Self-play**, which focuses on finding an optimal policy rather than an **equilibrium policy**.

Adversarial Policy

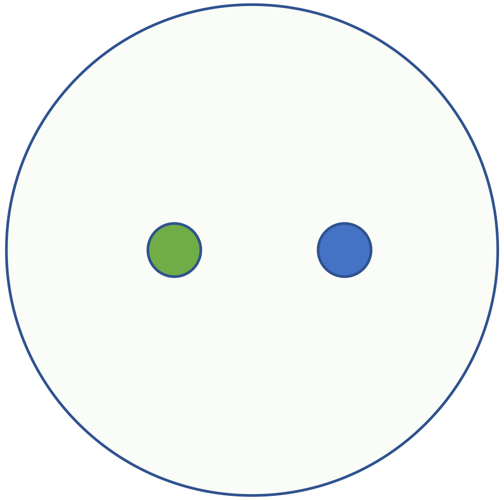
- **Adversarial policies** are a class of sequential decision-making policies used to minimize the cumulative rewards of a specific reinforcement learning system.
- Adversarial policies exist because RL training in competitive environments relies on **Self-play**, which focuses on finding an optimal policy rather than an **equilibrium policy**.
- When an agent employs a non-equilibrium policy, the opponent can increase its rewards by adjusting its own policy. In a competitive environment, one party's gain directly results in the other party's loss, which is the **essence** of adversarial policies.

Research Progress

Research Findings

One-on-one fully observable

competitive environments



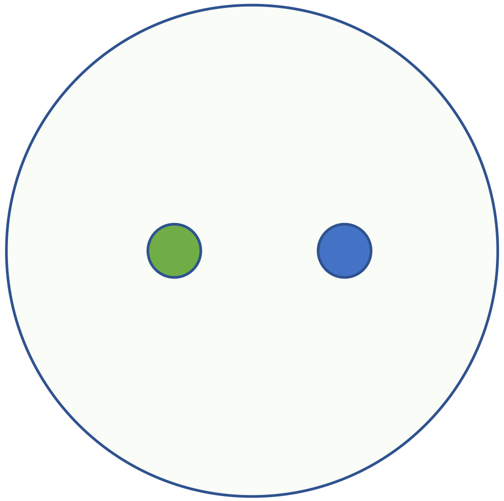
- Adversarial Policies: Attacking Deep Reinforcement Learning. [Gleave et al., ICLR 2020]
- Adversarial Policy Learning in Two-player Competitive Games. [Guo et al., ICML 2021]
- Adversarial Policy Training against Deep Reinforcement Learning. [Wu et al., USENIX 2021]
- Adversarial Policies Beat Superhuman Go AIs. [Wang et al., ICML 2023]
- PATROL: Provable Defense against Adversarial Policy in Two-player Games. [Guo et al., USENIX 2023]
- Rethinking Adversarial Policies: A Generalized Attack Formulation and Provable Defense in RL. [Liu et al., ICLR 2024]

Research Progress

Research Findings

One-on-one fully observable

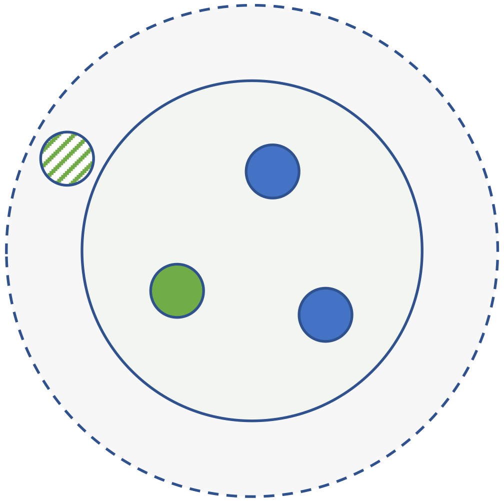
competitive environments



Research Gaps

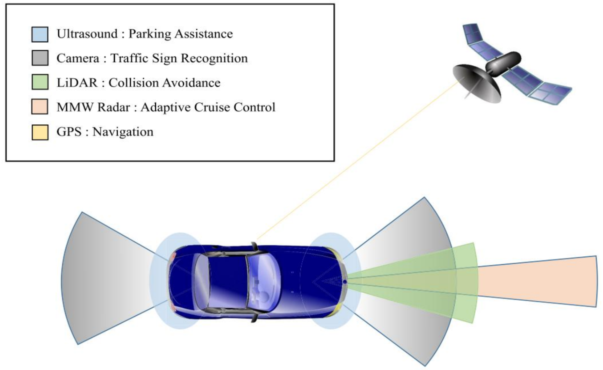
Many-to-many partially observable

competitive environments



- Adversarial Policies: Attacking Deep Reinforcement Learning. [Gleave et al., ICLR 2020]
- Adversarial Policy Learning in Two-player Competitive Games. [Guo et al., ICML 2021]
- Adversarial Policy Training against Deep Reinforcement Learning. [Wu et al., USENIX 2021]
- Adversarial Policies Beat Superhuman Go AIs. [Wang et al., ICML 2023]
- PATROL: Provable Defense against Adversarial Policy in Two-player Games. [Guo et al., USENIX 2023]
- Rethinking Adversarial Policies: A Generalized Attack Formulation and Provable Defense in RL. [Liu et al., ICLR 2024]

Partial Observable Situations

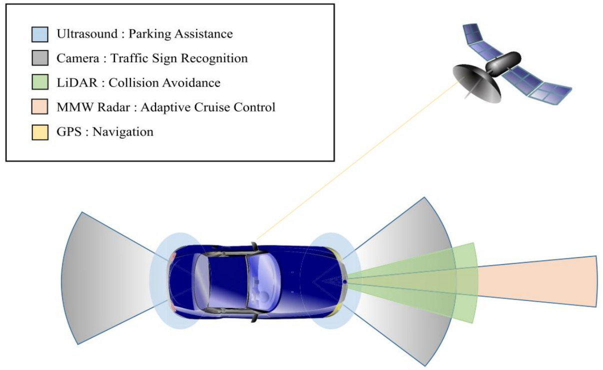


In-vehicle sensors

Fog of war

Imperfect-information games

Research Question



In-vehicle sensors

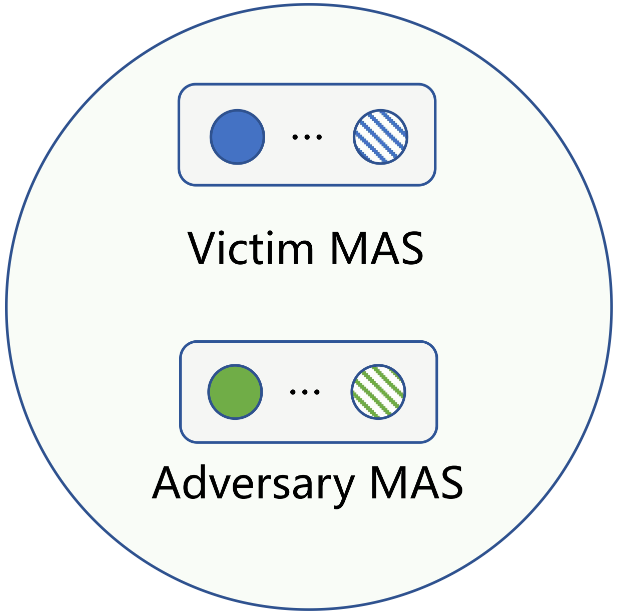
Fog of war

Imperfect-information games

Research Question: Do reinforcement learning systems encounter the risk of adversarial policies in many-to-many competitive environments, especially when the attacker can only obtain partial observations?

Threat Model

- **Environment Description.** A partially observable competitive environment consists of two multi-agent systems (MASs), where one victim MAS implements a multi-agent reinforcement learning (MARL) policy, while the other adversary MAS is controlled by the attacker.



Threat Model

- **Attacker's Goal.**
  - Minimize the performance of the victim MAS on a specific MARL task.

Threat Model

- **Attacker's Goal.**
  - Minimize the performance of the victim MAS on a specific MARL task.

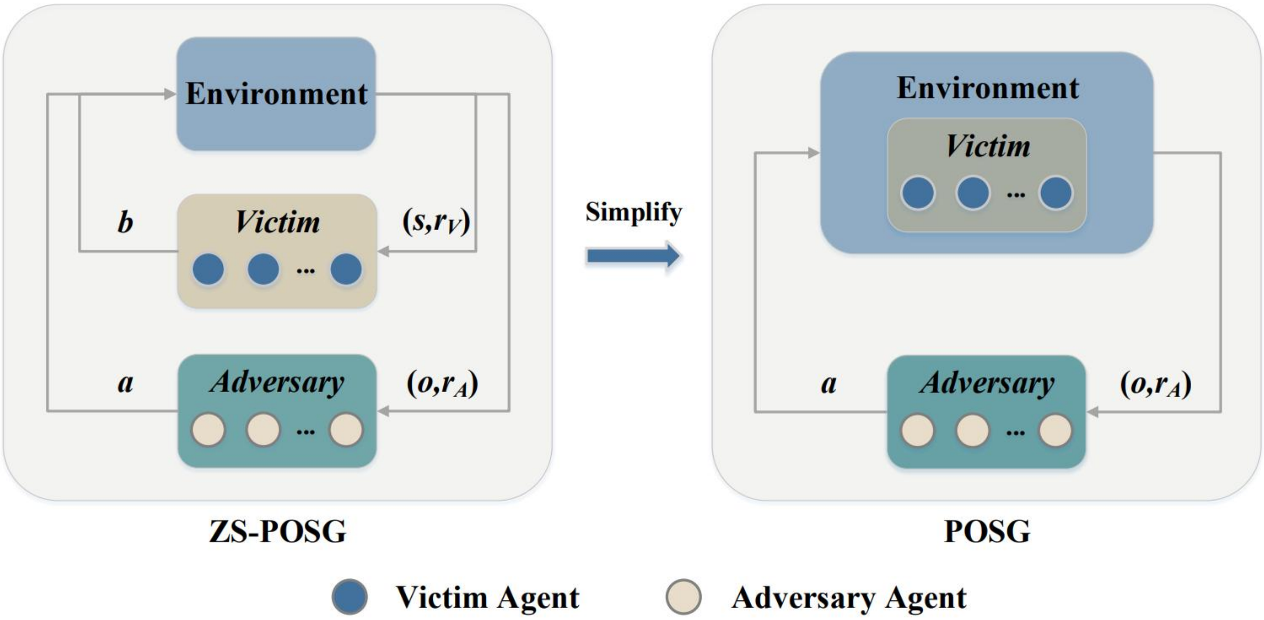
- **Attacker's Capabilities.**
  - The attacker can interact with the victim and obtain partial observations of the environment at each time step.
  - For the attacker, the victim MAS is a black box, except for knowing the number of victim agents.
  - The attacker cannot manipulate the environment or the victim's observations.

Problem Formulation

- The attacker's training of adversarial policies in the aforementioned environment can be formalized as a zero-sum partially observable stochastic game (**ZS-POSG**).

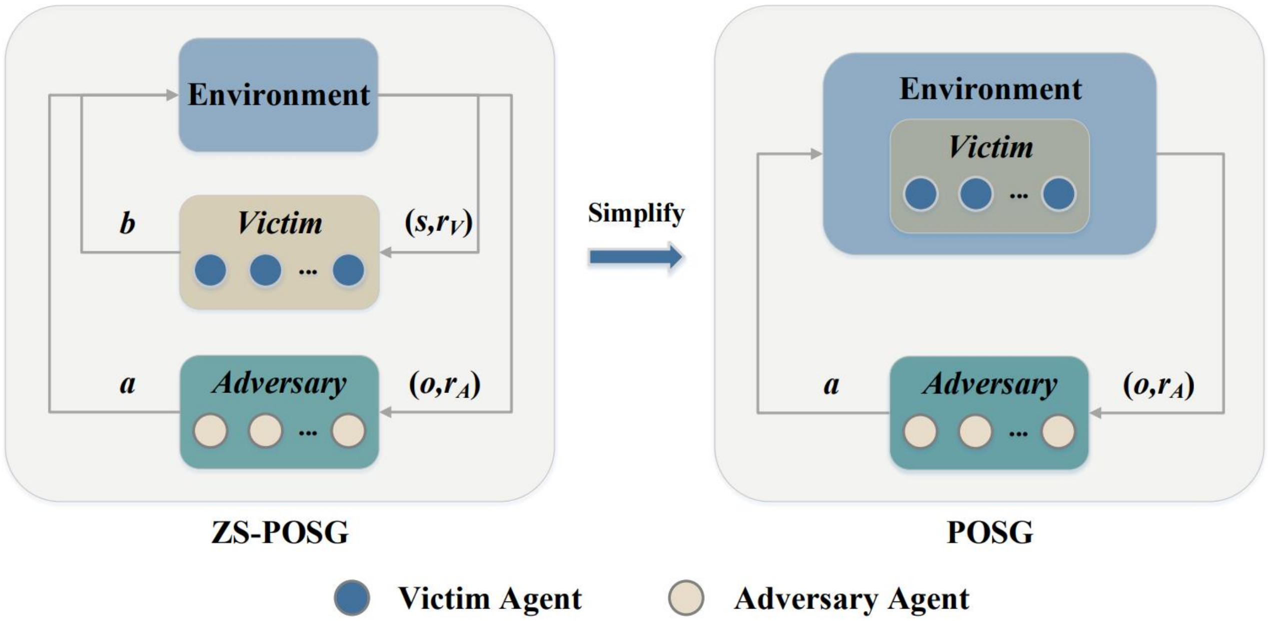
Problem Simplification

- The problem can be simplified from a **ZS-POSG** to a **POSG** if the joint policy of the victim is **fixed**.



Problem Simplification

- The problem can be simplified from a **ZS-POSG** to a **POSG** if the joint policy of the victim is **fixed**.



- Subsequent evaluations demonstrate that even when the fixed assumption is relaxed, the attack remains effective.

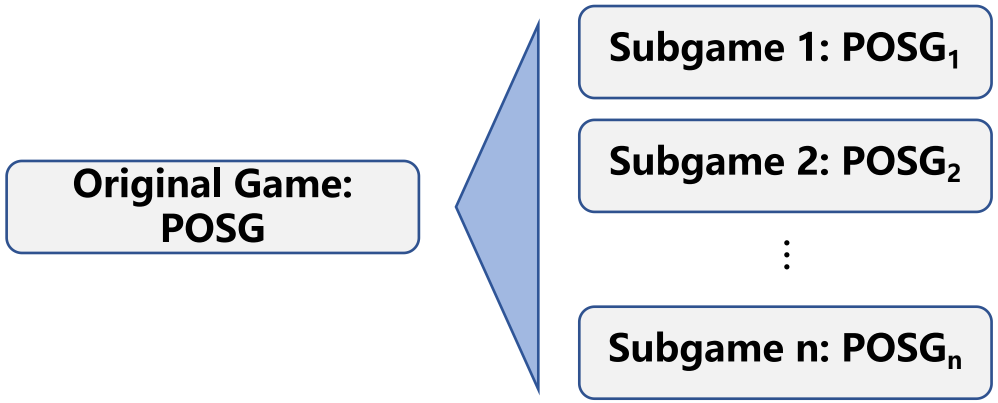
Challenges

- **Challenge I.** How can the attacker address a POSG and generate adversarial policies with limited interactions?

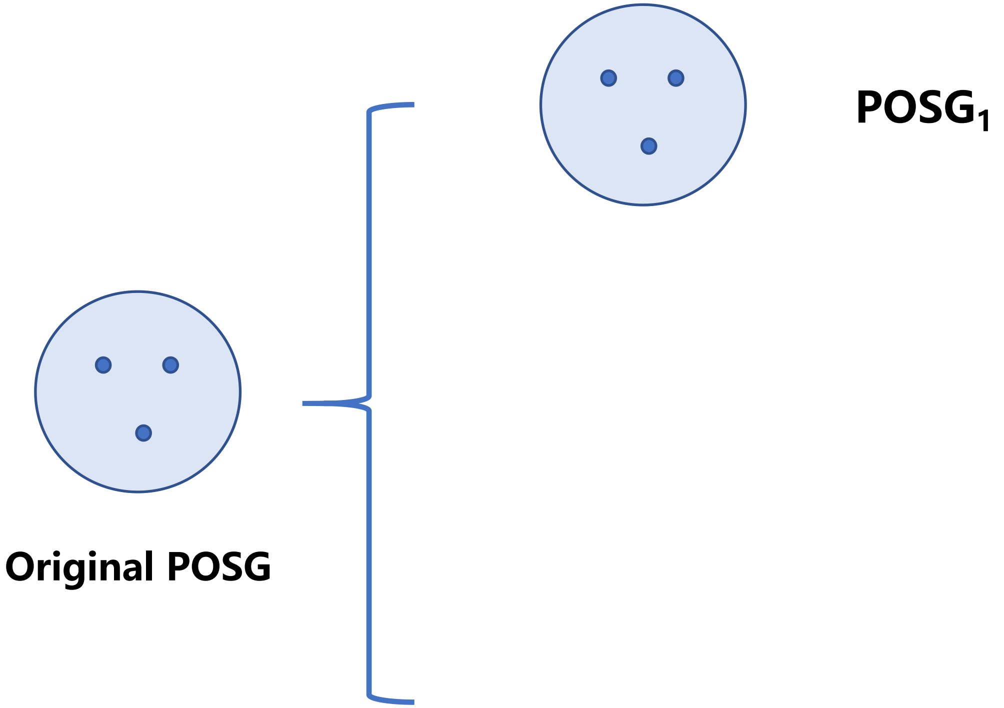
Challenges

- **Challenge I.** How can the attacker address a POSG and generate adversarial policies with limited interactions?

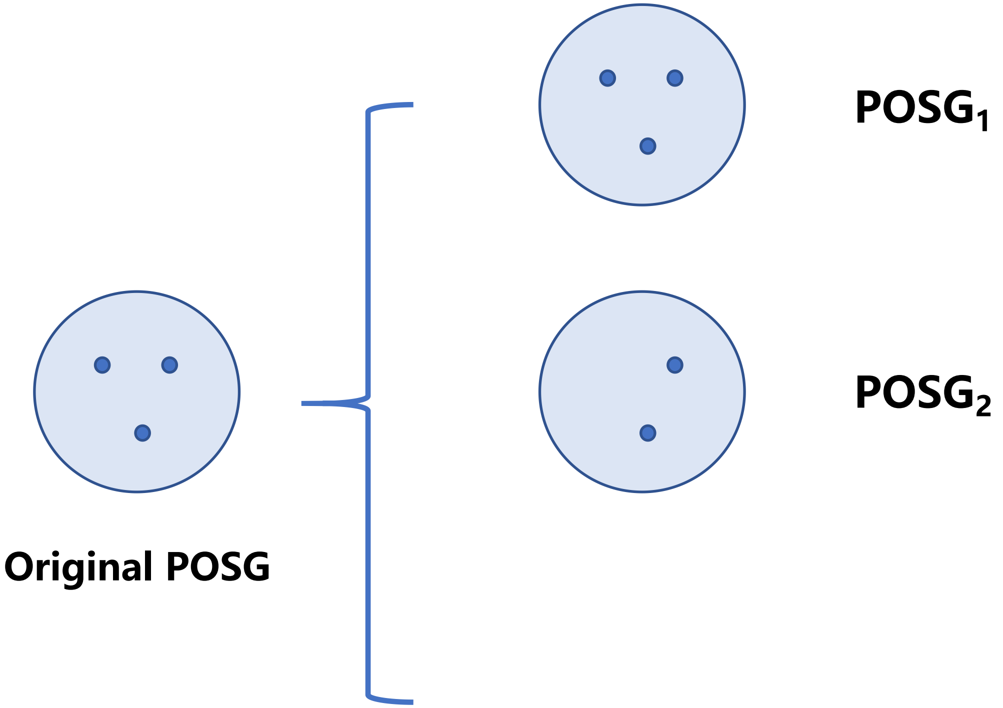
- **Subgame Construction.** We adopt a **divide-and-conquer** strategy by decomposing a complex POSG into multiple simpler POSGs, allowing for a more efficient solution to the overall problem by addressing each subgame individually.



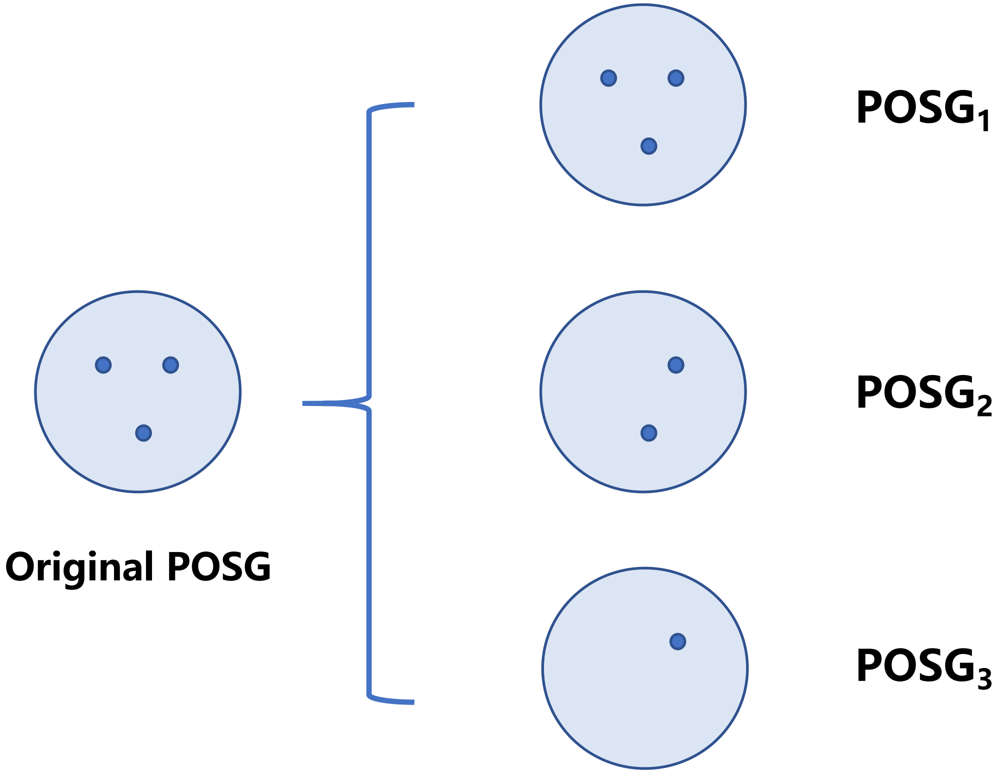
An Example of Subgame Construction



An Example of Subgame Construction

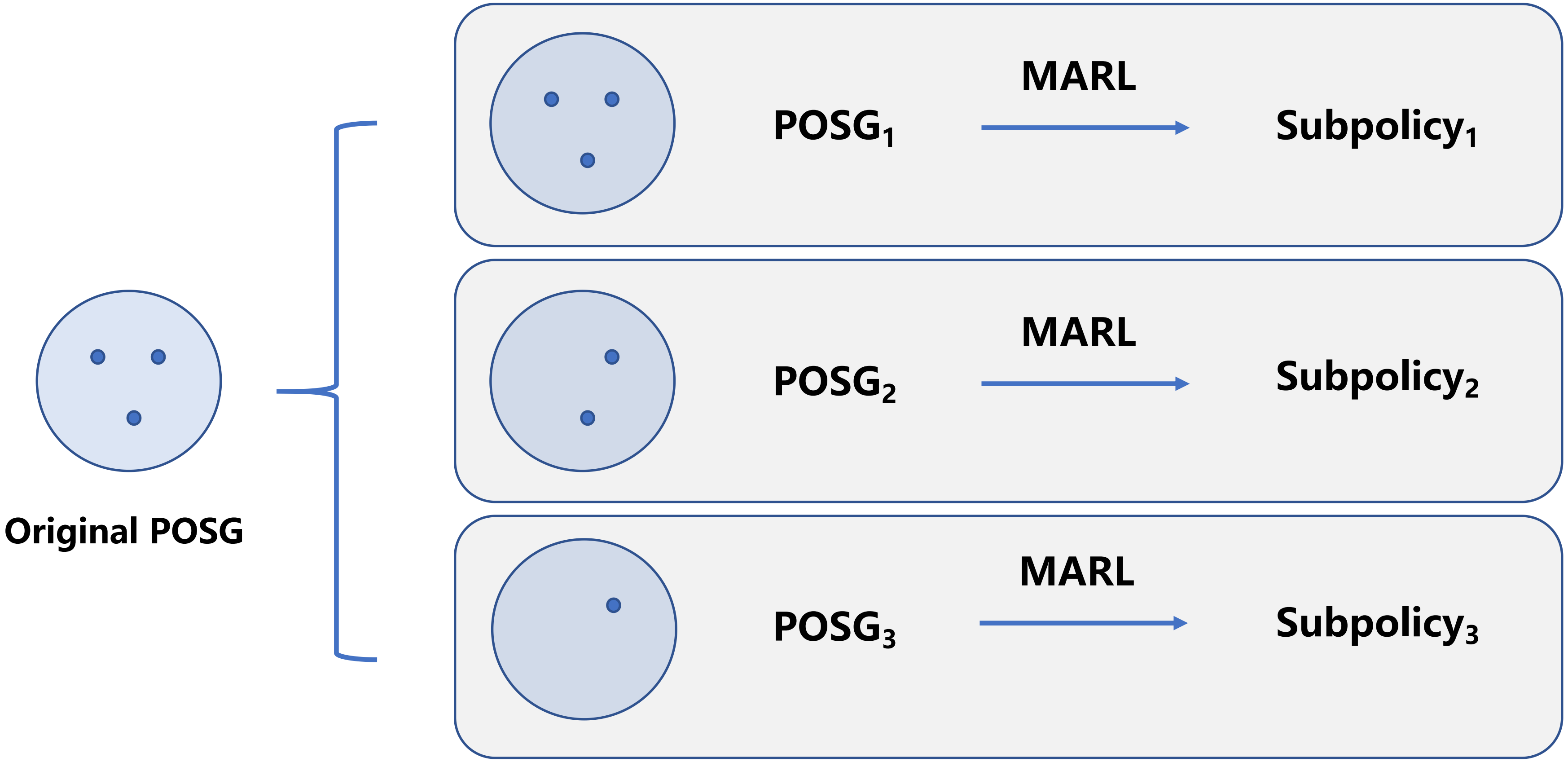


An Example of Subgame Construction



- From the perspective of the observation space, each subgame is **disjoint.**

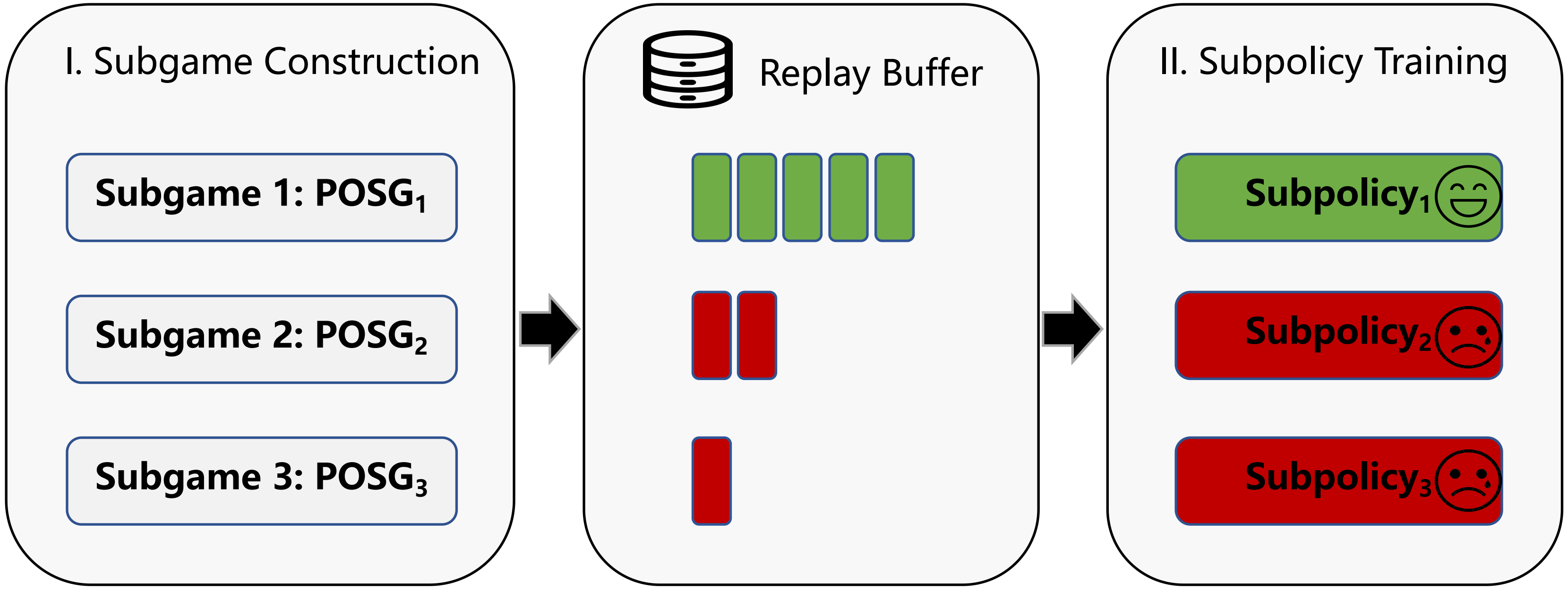
Subpolicy Training



- **Training Strategy.** The attacker needs to initialize a replay buffer for each subgame to store interaction data (**transition**) and train each subpolicy separately.

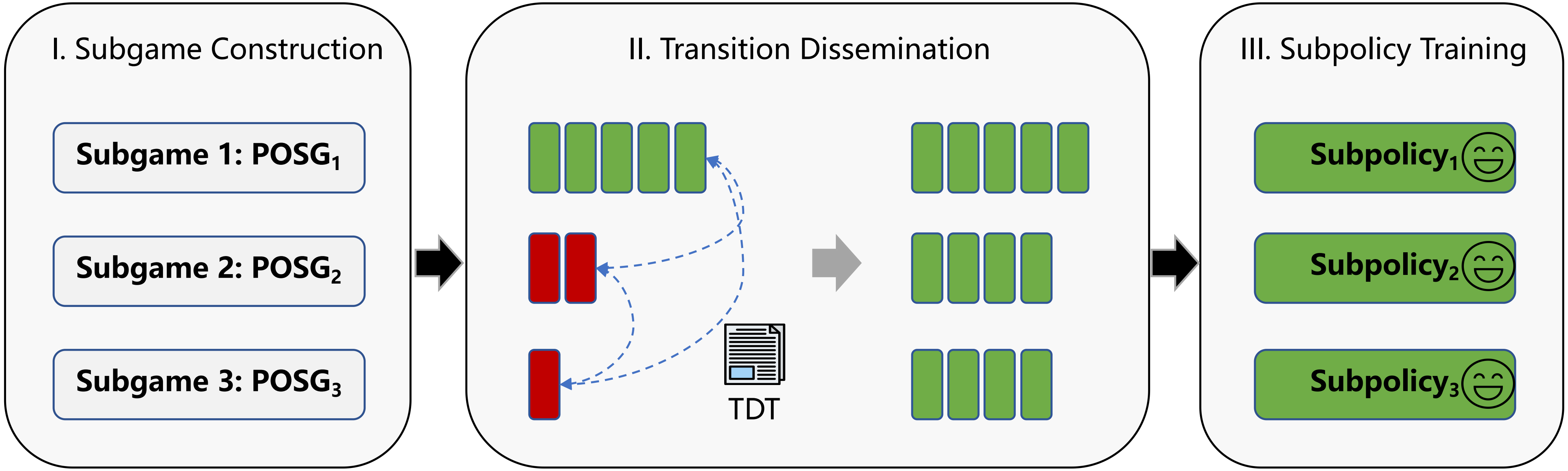
Challenges

- **Challenge II.** In most scenarios, subgames occur at different frequencies, which may result in some subgames lacking sufficient transitions for training.

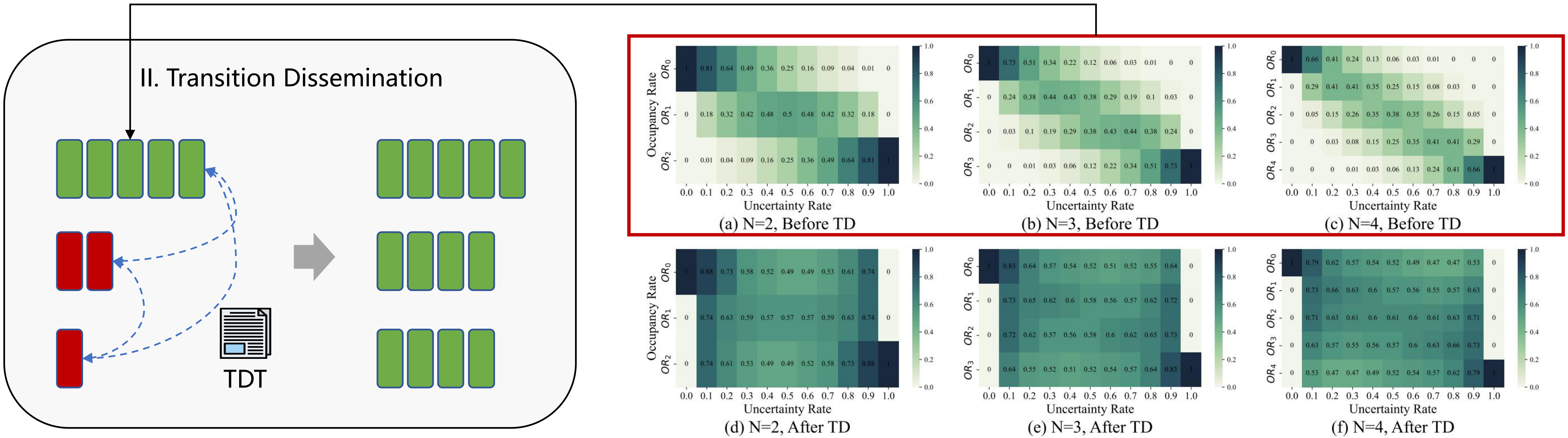


Challenges

- **Transition Dissemination.** Adversary agents generate a **transition dissemination table (TDT)** based on predefined rules and share transitions with one another according to the probabilities outlined in this table.

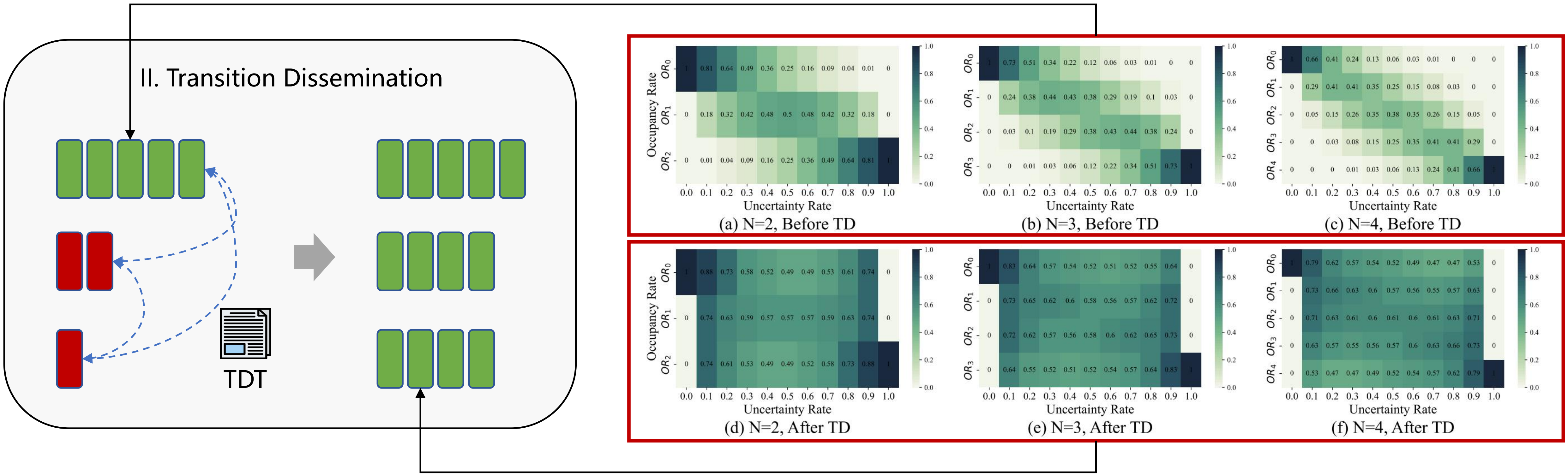


Transition Dissemination



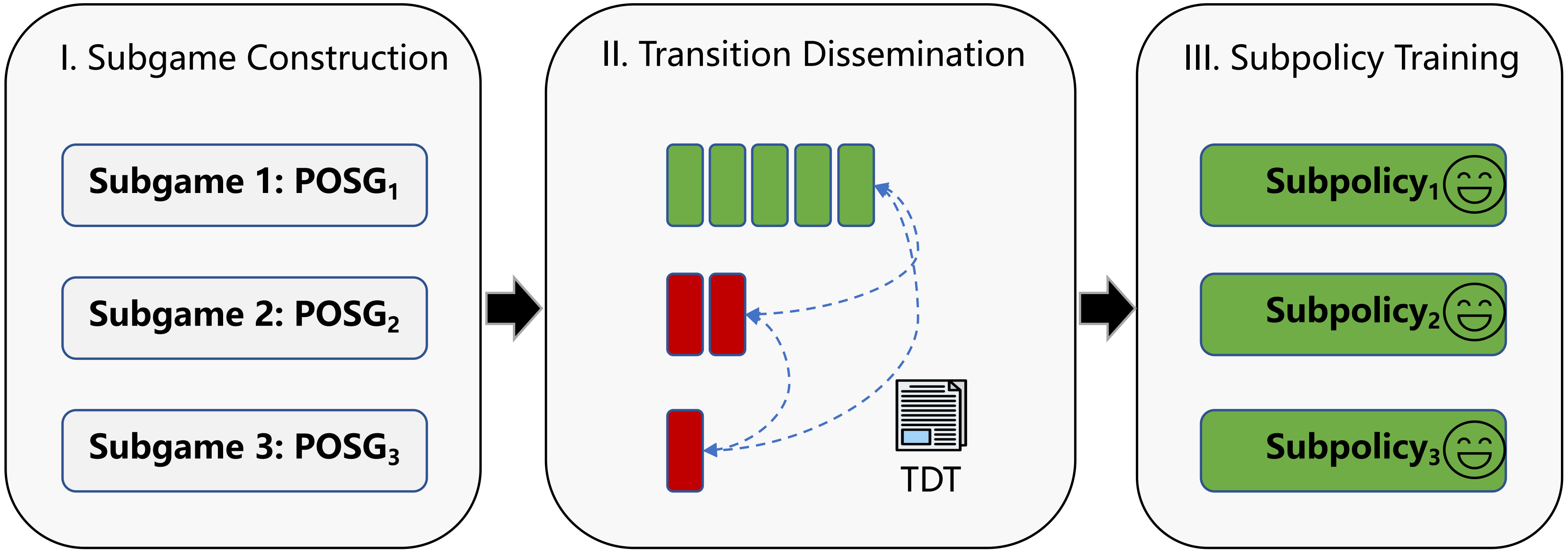
- The number of transitions for each subgame is **uneven**.

Transition Dissemination

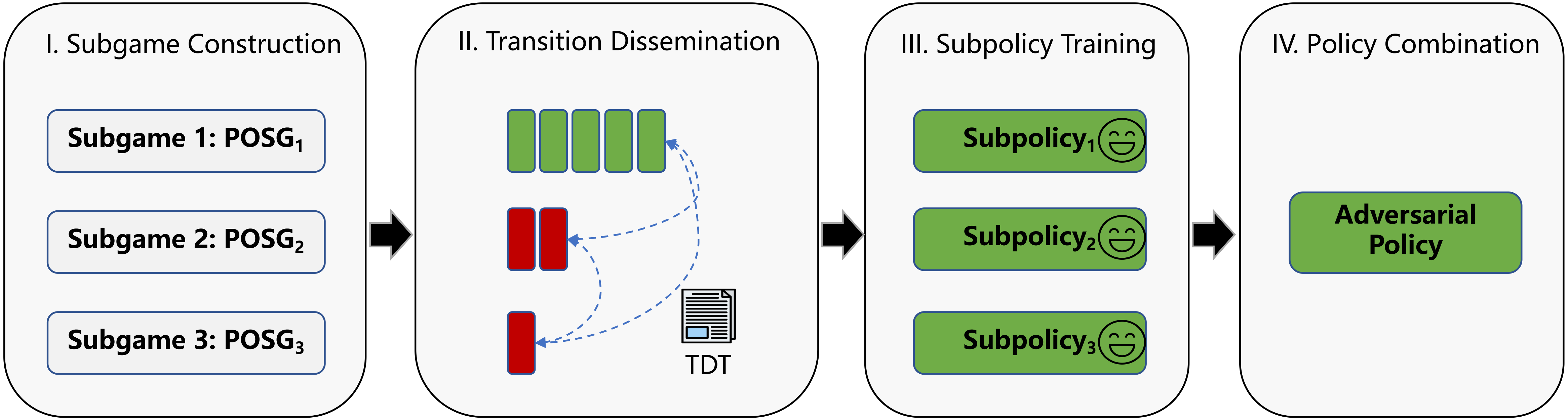


- Transition Dissemination **balances** the number of transitions in each replay buffer across different scenarios.

SUB-PLAY


I. Subgame Construction
Subgame 1: $POSG_1$
Subgame 2: $POSG_2$
Subgame 3: $POSG_3$
II. Transition Dissemination
TDT
III. Subpolicy Training
Subpolicy_1
Subpolicy_2
Subpolicy_3

SUB-PLAY



- **Policy Combination.** Since there is no requirement for **stealthiness**, the attacker implements the policy combination in a **hard-coded manner**.

Evaluation Settings

- **Environment.** (Multi Particle Environments (MPE) framework developed by OpenAI)
- **Tasks.** (Predator-Prey, World Communication)
- **Partially Observable Limitations.** (Uncertainty, Distance, Region)
- **Multi-Agent Settings.** (1v3, 2v3, 3v3, 2v2, 4v2)
- **MARL Algorithms.** (DDPG, MADDPG)
- **Comparison Methods.** (Self-play, Victim-play)
- **Metrics.** (Catch Rate, Collision Frequency)

Attack Performance

- **Uncertainty Limitation.** SUB-PLAY reduces the victim's performance to 51.98% of the baseline and outperforms other methods in **96.0%** (48/50) of scenarios.

- **Distance Limitation.** SUB-PLAY reduces the victim's performance to 55.71% of the baseline and outperforms other methods in **97.5%** (39/40) of scenarios.

- **Region Limitation.** SUB-PLAY reduces the victim's performance to 59.07% of the baseline and outperforms other methods in **100.0%** (10/10) of scenarios.

Ablation Study

Table 2: The ablation results of components in *SUB-PLAY* measured by two metrics (CR↓/CF↓). Acronyms: Subgame Construction (SC), Transition Dissemination (TD), Policy Meritocracy (PM).

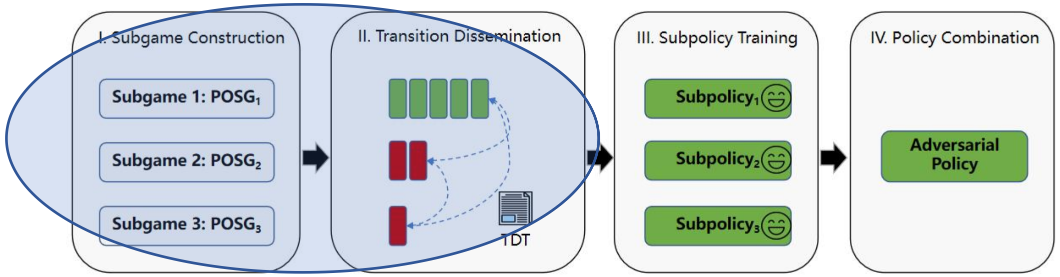
Methods	Limitations				
	Uncertainty (0.25)	Uncertainty (0.50)	Distance (0.5)	Distance (2.0)	Region (1)
Self-play	0.920 / 14.280	0.916 / 13.998	0.936 / 14.349	0.935 / 14.187	0.704 / 4.486
Victim-play	0.782 / 7.823	0.727 / 7.215	0.728 / 6.163	0.670 / 4.891	0.718 / 3.763
SUB-PLAY (SC)	0.830 / 8.402	0.759 / 7.604	0.765 / 6.296	0.708 / 5.982	0.835 / 6.563
SUB-PLAY (SC+TD)	0.617 / 3.740	0.627 / 4.438	0.700 / 6.552	0.672 / 4.675	0.688 / 3.309
SUB-PLAY (SC+PM)	0.731 / 6.059	0.708 / 6.318	0.735 / 6.113	0.677 / 4.576	0.561 / 1.634
SUB-PLAY (SC+TD+PM)	**0.579 / 3.053**	**0.583 / 3.228**	**0.563 / 3.075**	**0.589 / 3.264**	**0.489 / 1.397**

Ablation Study

Table 2: The ablation results of components in *SUB-PLAY* measured by two metrics (CR↓/CF↓). Acronyms: Subgame Construction (SC), Transition Dissemination (TD), Policy Meritocracy (PM).

Methods	Limitations				
	Uncertainty (0.25)	Uncertainty (0.50)	Distance (0.5)	Distance (2.0)	Region (1)
Self-play	0.920 / 14.280	0.916 / 13.998	0.936 / 14.349	0.935 / 14.187	0.704 / 4.486
Victim-play	0.782 / 7.823	0.727 / 7.215	0.728 / 6.163	0.670 / 4.891	0.718 / 3.763
SUB-PLAY (SC)	0.830 / 8.402	0.759 / 7.604	0.765 / 6.296	0.708 / 5.982	0.835 / 6.563
SUB-PLAY (SC+TD)	0.617 / 3.740	0.627 / 4.438	0.700 / 6.552	0.672 / 4.675	0.688 / 3.309
SUB-PLAY (SC+PM)	0.731 / 6.059	0.708 / 6.318	0.735 / 6.113	0.677 / 4.576	0.561 / 1.634
SUB-PLAY (SC+TD+PM)	**0.579 / 3.053**	**0.583 / 3.228**	**0.563 / 3.075**	**0.589 / 3.264**	**0.489 / 1.397**

- The results show that subgame construction alone leads to inferior attack performance, but **combining** it with transition dissemination significantly improves performance.



Scalability Evaluation

- The attack performance of SUB-PLAY is **positively** correlated with the number of subgames, while the improvement gradually diminishes.
- The training cost of SUB-PLAY scales **linearly** with the number of subgames.

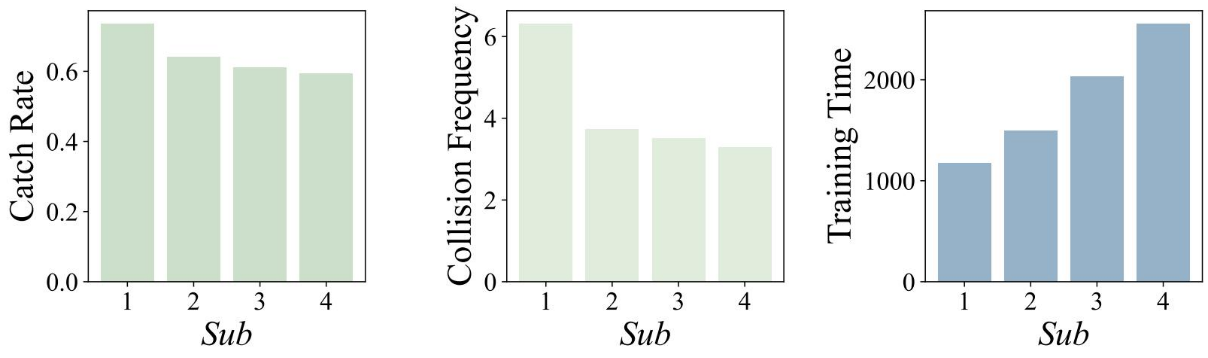
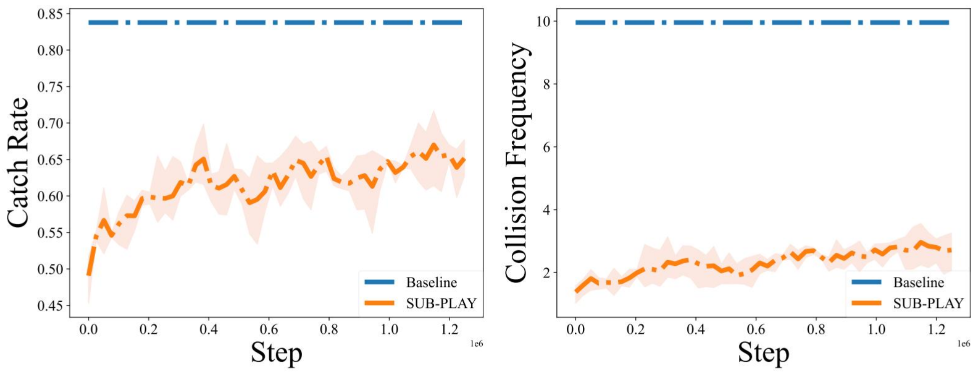


Figure 13: Scalability evaluation.

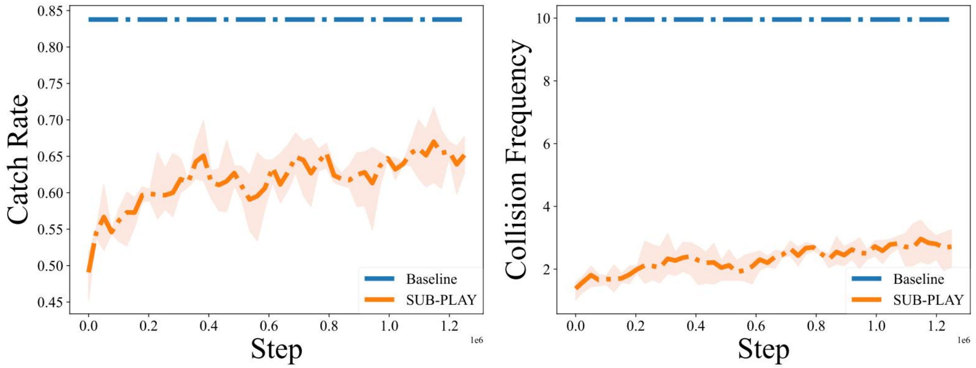
Potential Defenses - Fine-Tuning

- The continuous **fine-tuning** of the victim cannot resist SUB-PLAY.



Potential Defenses - Fine-Tuning

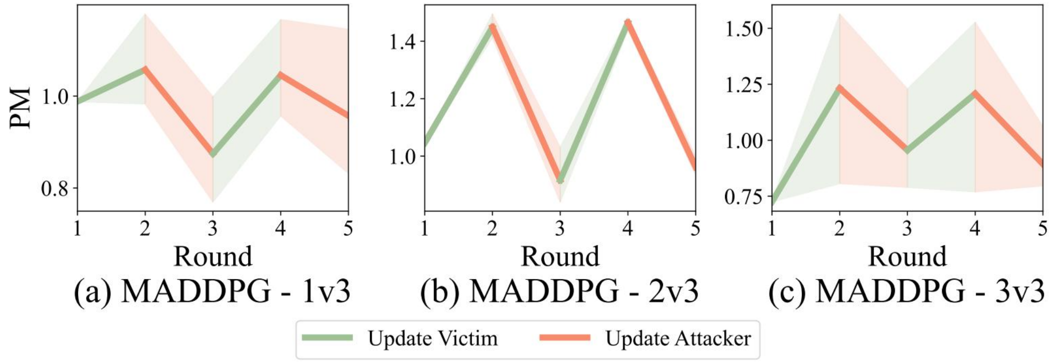
- The continuous **fine-tuning** of the victim cannot resist SUB-PLAY.



- This is due to the RL policies before and after fine-tuning remain close in the **policy space**, which has minimal impact on the generation of adversarial policies.

Potential Defenses - Adversarial Retraining

- Naive **adversarial retraining** cannot resist SUB-PLAY, as it theoretically fails to guarantee that a RL policy will gradually converge to an equilibrium policy.



(a) MADDPG - 1v3 (b) MADDPG - 2v3 (c) MADDPG - 3v3

Potential Defenses - Policy Ensemble

- Deploying RL policies as a **<u>policy ensemble</u>** and dynamically updating the policy pool can partially mitigate the threat of SUB-PLAY, as it effectively **<u>confuses</u>** the attacker's target.

Access	100%			33%		
Scenarios	1v3	2v3	3v3	1v3	2v3	3v3
Uncertainty						
0.00	-0.07	+0.02	-0.04	-2.74	+4.09	-0.89
0.25	+0.02	-0.25	+0.10	-9.86	-13.58	-12.45
0.50	+0.00	-0.02	+0.08	-9.68	-9.14	-17.01
0.75	-0.01	-0.07	+0.04	-15.55	-2.56	+2.85
1.00	+0.00	+0.04	+0.08	-25.78	-0.55	+9.68
Distance						
0.5	-0.09	-0.15	-0.03	-16.17	-7.99	-11.98
1.0	-0.12	-0.12	-0.01	-30.15	-5.65	+0.25
1.5	-0.29	-0.12	-0.02	-20.24	-9.36	-32.69
2.0	-0.13	-0.28	+0.14	-16.01	-20.51	-43.39
Region						
1	-0.08	-0.24	+0.00	-7.99	-37.44	-17.94

Conclusion

- We propose a novel **black-box** attack, SUB-PLAY, which reveals the security threats posed by adversarial policies in **partially observable** competitive environments.

- SUB-PLAY is **algorithm-agnostic**, making it suitable for both centralized and decentralized MARL paradigms.

- We discuss three potential defenses, highlighting that practitioners in RL should not only focus on improving algorithm performance but also pay attention to **deployment details**, which is crucial in mitigating security threats posed by adversarial policies.

Q&A

Oubo Ma, Yuwen Pu, Linkang Du, Yang Dai,
Ruo Wang, Xiaolei Liu, Yingcai Wu, and
Shouling Ji

mob@zju.edu.cn

浙江大学网络系统安全与隐私实验室
NETWORK SYSTEM SECURITY & PRIVACY LAB

Paper