智 能 与 计 算 学 部
College of Intelligence and Computing

TJU Deep Reinforcement Learning Lab

天津大学 - 深度强化学习实验室

http://www.icdai.org/

A hierarchical framework for solving the driving tasks

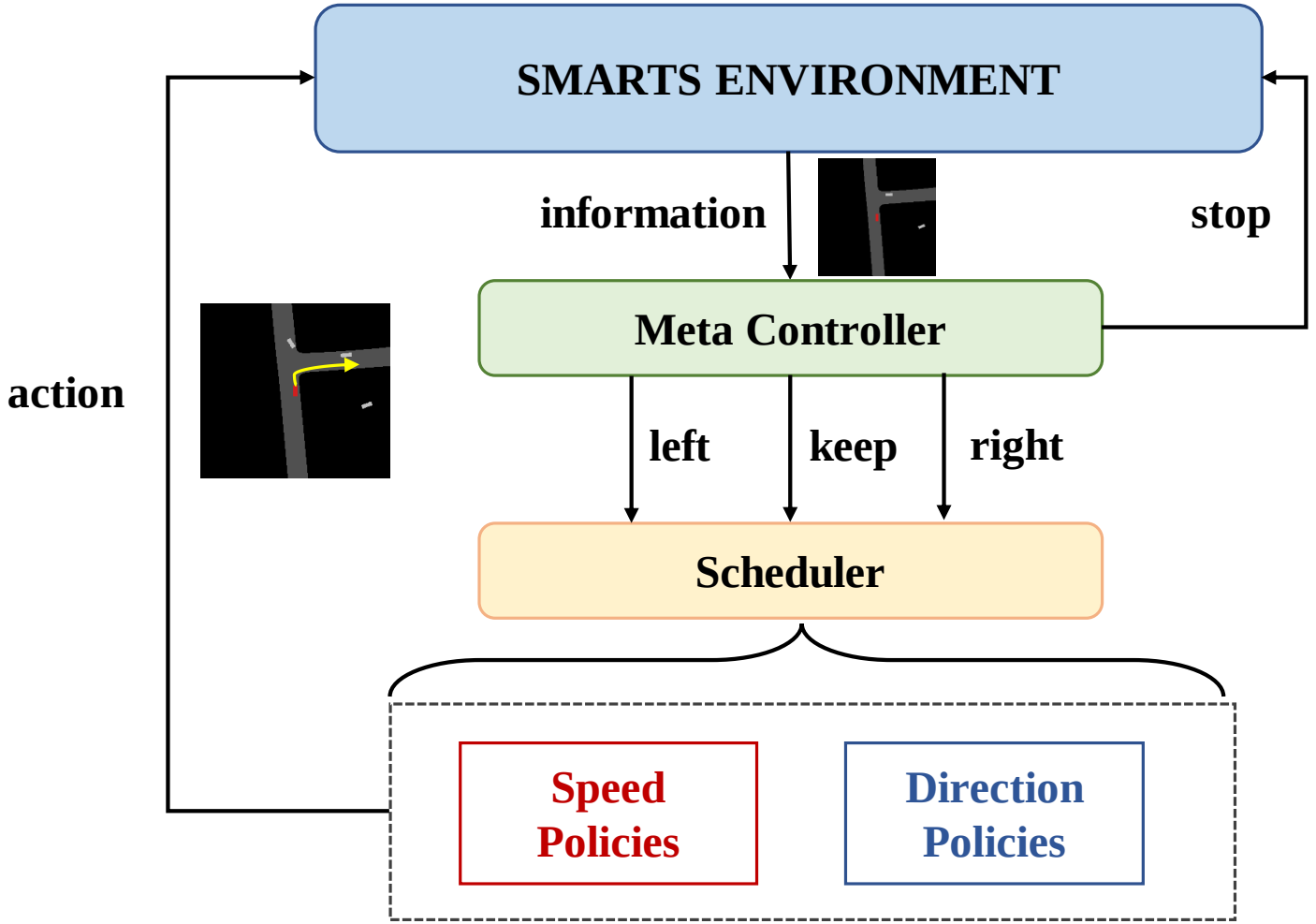
Yi Ma and Hebin Liang

Motivation

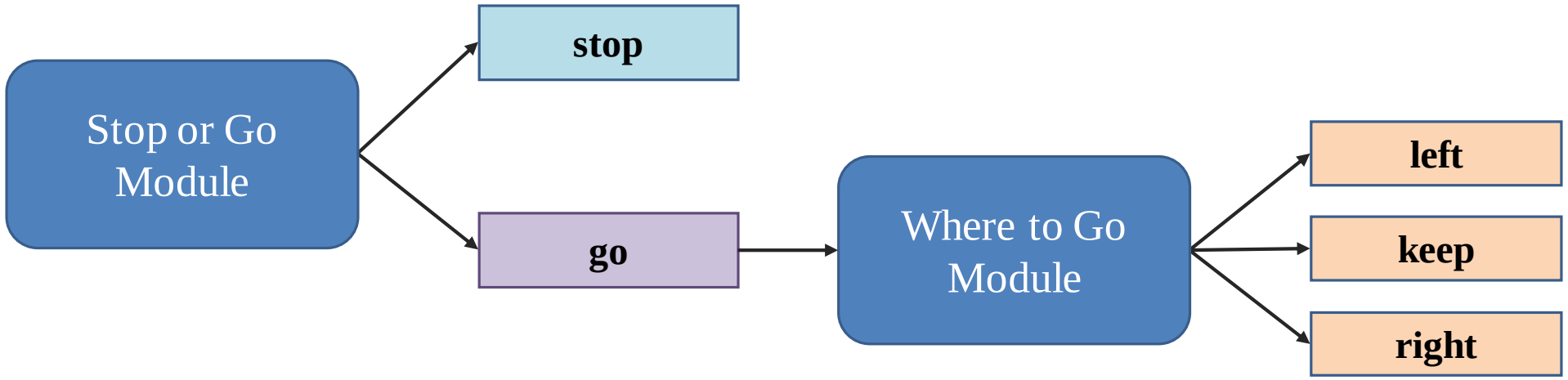
- **End to end offline learning doesn't work well**
  - Hard to tune even in a single scenario
    - Chaotic and redundant (mutated heading information, additional cars) datasets
    - No reward information
  - Hard to generalize

Can we divide the tasks into several sub-tasks and conquer each sub-task?

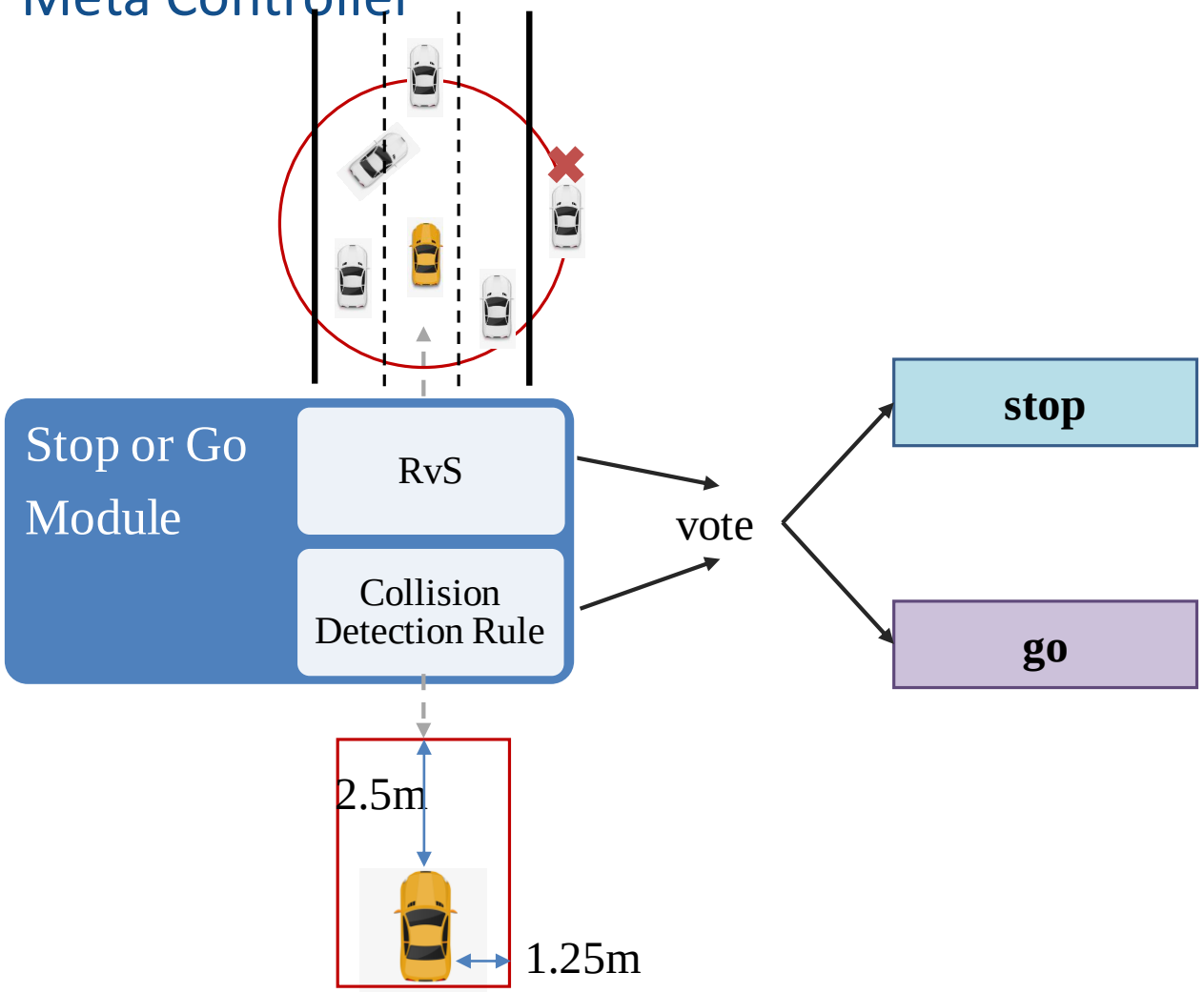
Overall Framework



Meta Controller



Meta Controller



Algorithm 1 RvS-Learning

Input: Dataset of trajectories, $\mathcal{D} = \{\tau\}$
Initialize policy $\pi_\theta(a \mid s, \omega)$.
while not converged **do**
    Randomly sample trajectories: $\tau \sim \mathcal{D}$.
    Sample time index for each trajetory, $t \sim [1, H]$, and sample a corresponding outcome: $\omega \sim f(\omega \mid \tau_{t:H})$.
    Compute loss: $\mathcal{L}(\theta) \leftarrow \sum_{(s_t, a_t, \omega)} \log \pi_\theta(a_t \mid s_t, \omega)$
    Update policy parameters: $\theta \leftarrow \theta + \eta \nabla_\theta \mathcal{L}(\theta)$
end while
return Conditional policy $\pi_\theta(a \mid s, \omega)$

MDP:

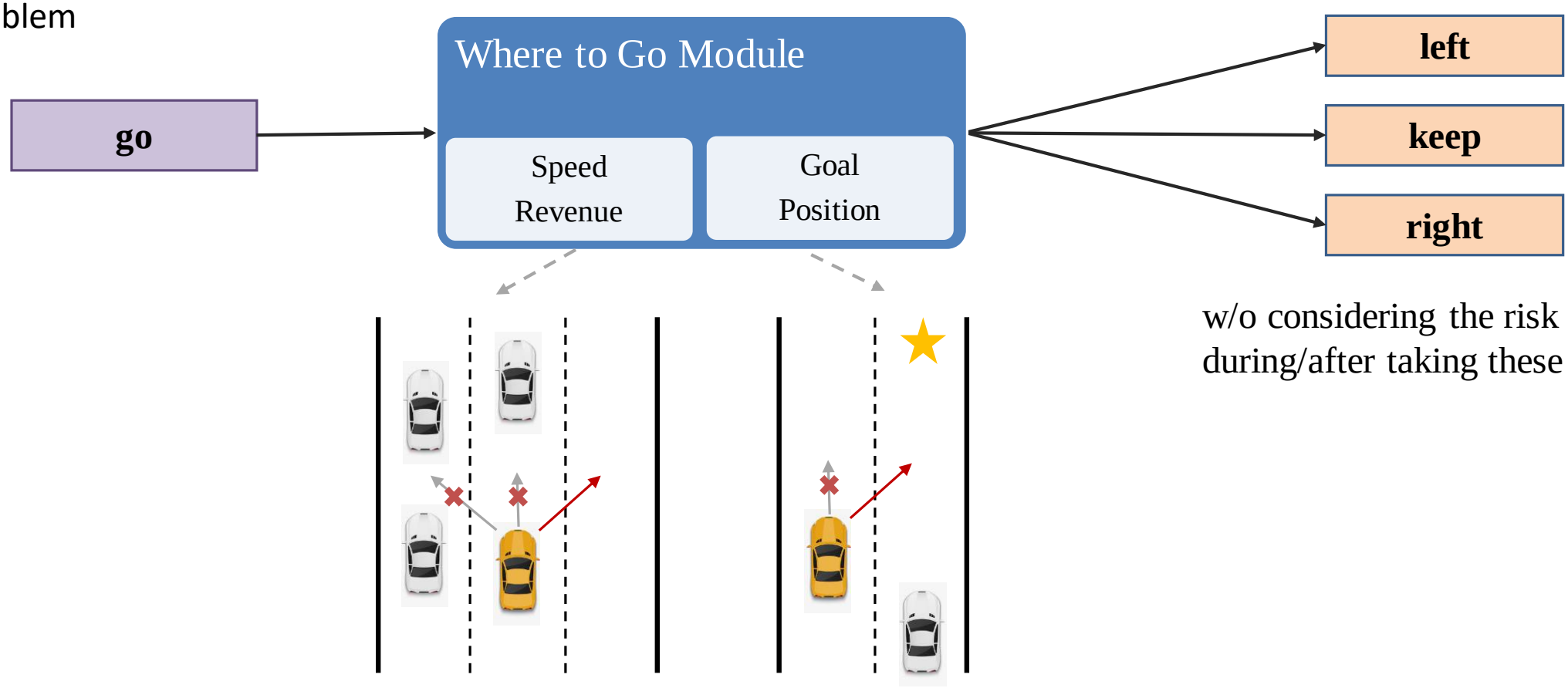
State: the bounding box of ego, the speed limit of the current lane, the relative heading, position and speed of the five closest vehicles to ego, and the bounding box of these vehicles.

Action: The egos whose speed is less than 0.2 times the speed limit are assumed to be in a static state. Otherwise, the car is assumed to be in a moving state.

Goal: We set the output of $f(w|\tau_{t:H})$ to be the waypoints and headings in the subsequent next 5 steps

Meta Controller

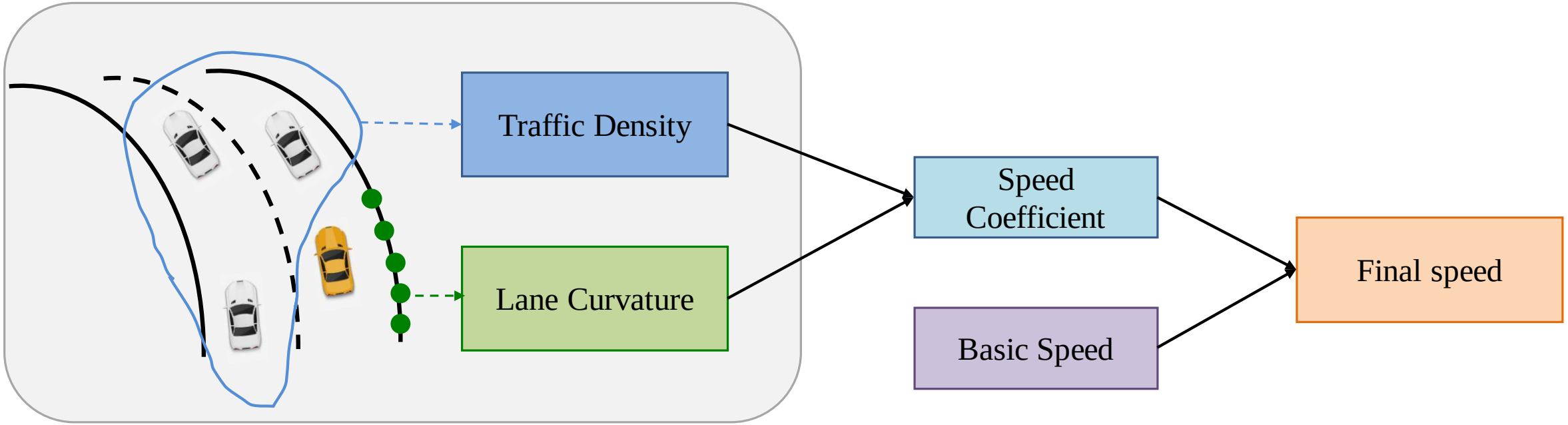
Bandit Problem



w/o considering the risk during/after taking these actions

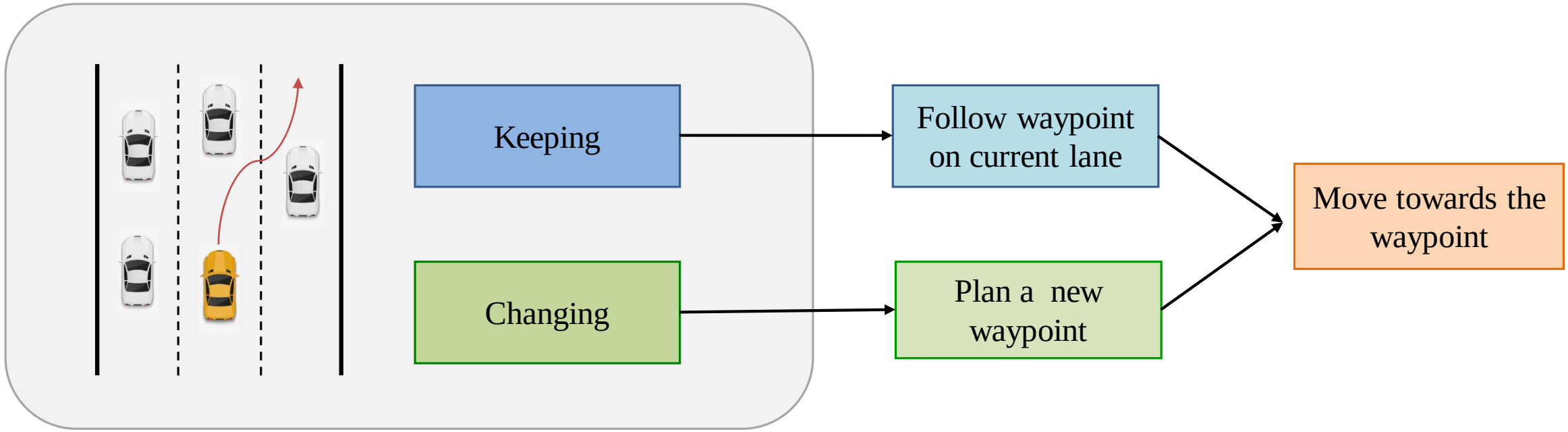
Scheduler: Speed Policies

Classification Problem



For simplicity, the range of speed coefficient is discretized. That is the reason why our method performs poorly above humanness metric.

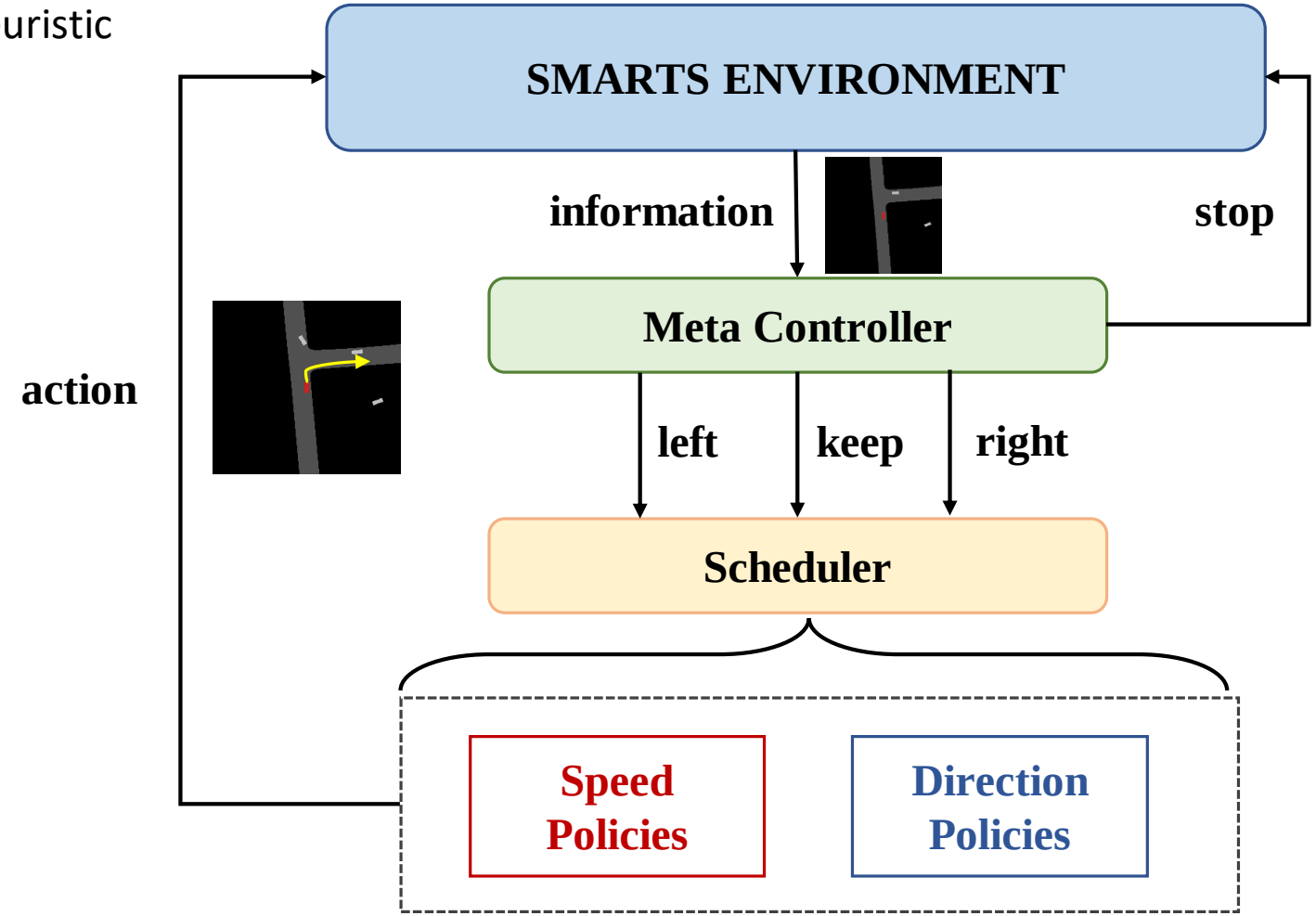
Scheduler: Direction Policies

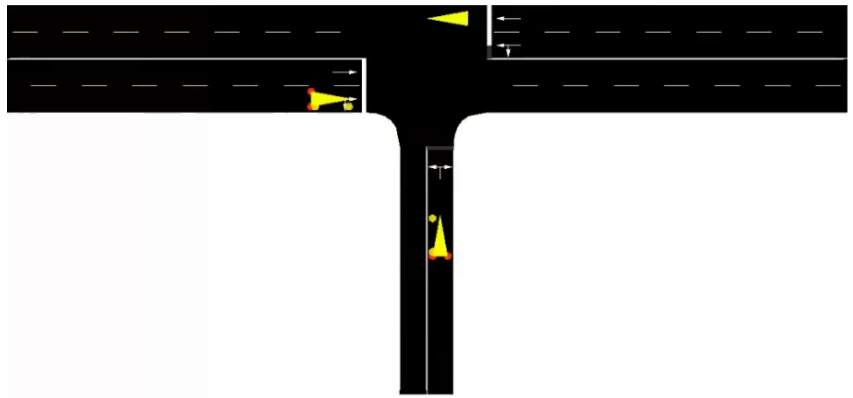


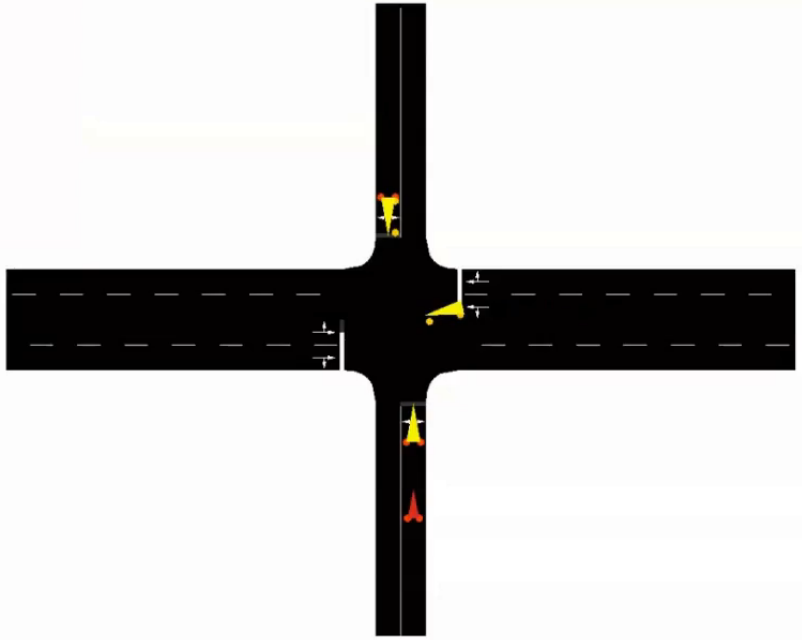
Note that the moving of the lower-level scheduler cannot be interrupted by Where to Go Module until the lane changing is completely finished.

Overall Framework

Hybrid learning and heuristic







智 能 与 计 算 学 部
College of Intelligence and Computing

TJU Deep Reinforcement Learning Lab

天津大学 - 深度强化学习实验室

http://www.icdai.org/