π_θ	Policy
a^t	Action
v_i^t	Velocity
s^t	State
M_i	Modal Data
X^t	Cell Positions
D^t	Discrepancy
R^t	Reward
$\hat{A}^t$	Advantage