Policy Validation
Score Learning & Calibration
Policy Deployment
Generative Policy $\pi(\boldsymbol{A}|\boldsymbol{O})$
Successful Rollouts
Learning + Calibration
Observation-Based Score
Calibration
Action-Based Score
$F_A(\boldsymbol{\tau}_{:t}) \wedge F_O(\boldsymbol{\tau}_{:t})$
FIPER
Failure Prediction at Runtime
Score
Threshold
Failure Warning
Success