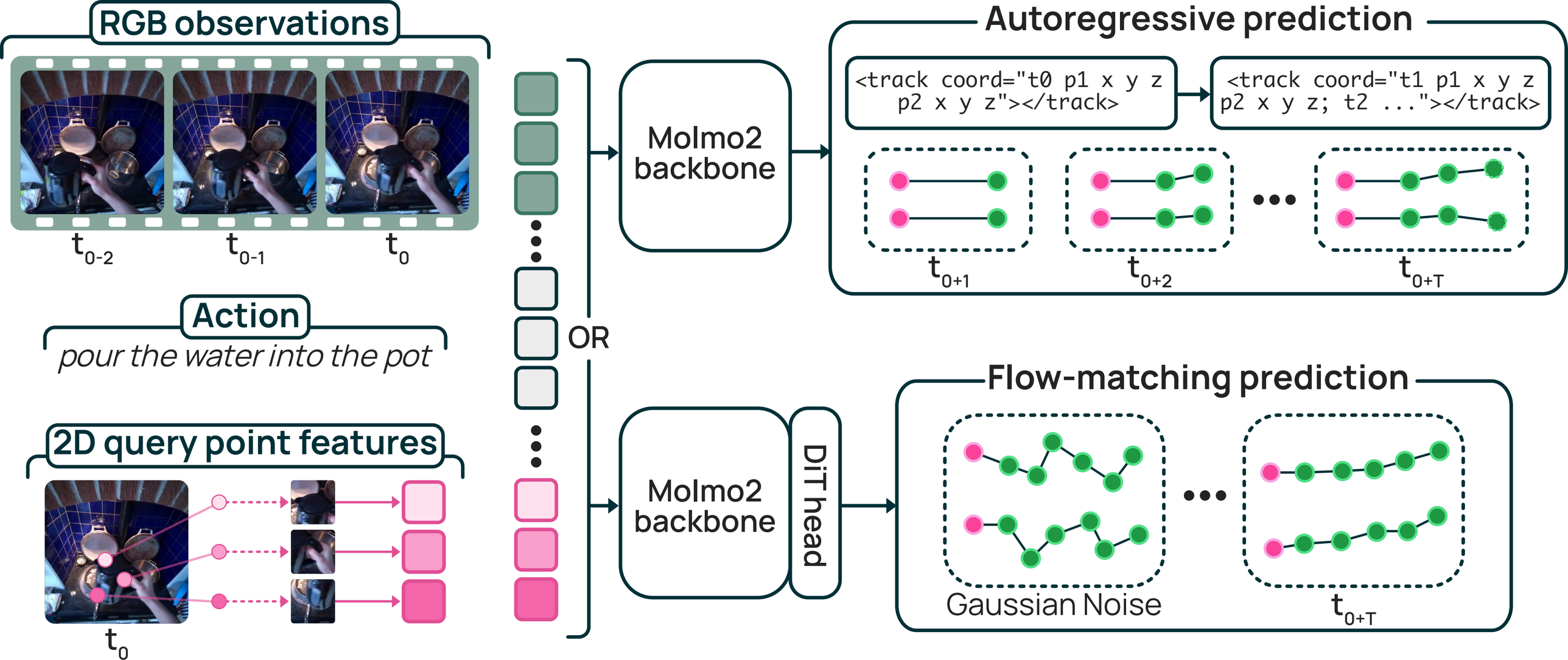
RGB observations
t_{0-2}
t_{0-1}
t_0
Action
pour the water into the pot
2D query point features
t_0
OR
Molmo2 backbone
Autoregressive prediction
<track coord="t0 p1 x y z p2 x y z"></track>
<track coord="t1 p1 x y z p2 x y z; t2 ..."></track>
t_{0+1}
t_{0+2}
t_{0+T}
Molmo2 backbone
DiT head
Flow-matching prediction
Gaussian Noise
t_{0+T}