Algorithm -

Init - policy (π), critic(V), feature_extractor (G), action predictor (I) , forward model (F), Replay Memory (M)

For episodes in 1 to E:

For t = 1 ….. T:

$g^t = G(s^t)$

Sample action (a^t) = $\pi(g^t)$; U(0,1) >= epsilon
Random ; otherwise

$r^t, s^{t+1} = env.step(a^t)$

Store transition ($s^t, a^t, r^t, G(s^{t+1})$) in M

==========

Sample transition ($s^t, a^t, r^t, G(s^{t+1})$) from M

$g^t = G(s^t)$ # Recompute
$g^{t+1} = G(s^{t+1})$ # From M

ICM Update

$a^{pred} = I(g^t, g^{t+1})$

$L_I = cross\_entropy(a^{pred}, a^t)$

$g^{pred} = F(g^t, a^t)$

$L_F = MSE(g^{pred}, g^{t+1}) = r^{intrinsic}$

Update F and I. Note G(.) is updated at all time steps

Update A2C

$R = r^{intrinsic} + r^t + \gamma V(g^{t+1})$

$a^i = sample(dist = \pi(g^t))$

$\Theta(\pi) = \Theta(\pi) + \nabla(\pi) \log(\pi(a^i \mid g^t))[R - V(g^t)]$
$\Theta(V) = \Theta(V) + \nabla(V)[R - V(g^t)]^2$

If Refresh == True:

$a^{new} = one\_hot(a^i)$

$g^{new} = F(g^t, a^{new})$

Update Transition ($s^t, a^{new}, r^t, g^{new}$)
