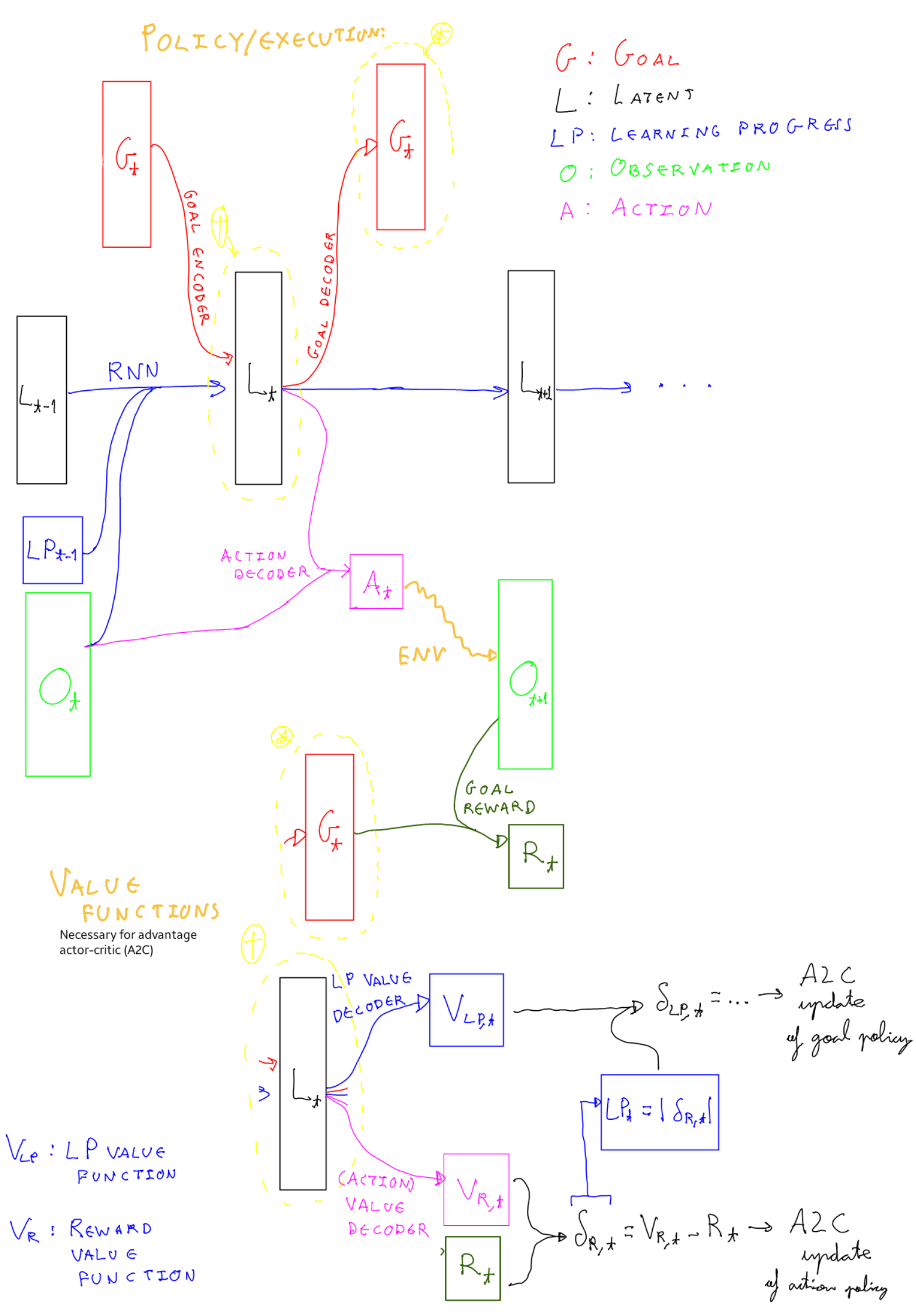
POLICY/EXECUTION:
G : GOAL
L : LATENT
LP : LEARNING PROGRESS
O : OBSERVATION
A : ACTION
GOAL ENCODER
GOAL DECODER
RNN
ACTION DECODER
ENV
GOAL REWARD
VALUE FUNCTIONS
Necessary for advantage actor-critic (A2C)
LP VALUE DECODER
(ACTION) VALUE DECODER
A2C update of goal policy
A2C update of action policy
V_{LP} : LP VALUE FUNCTION
V_R : REWARD VALUE FUNCTION

TRAINING PROCEDURE

- We train GOAL ENCODER and GOAL DECODER to autoencode / represent goals in L
- We train RNN to produce a new L_{t+1} (from L_t, O_t) which encodes a good goal G_{t+1} (good in the sense of having high LP)
- We train ACTION DECODER to produce an action A_t which will lead to an observation O_{t+1} closely matching G_t.