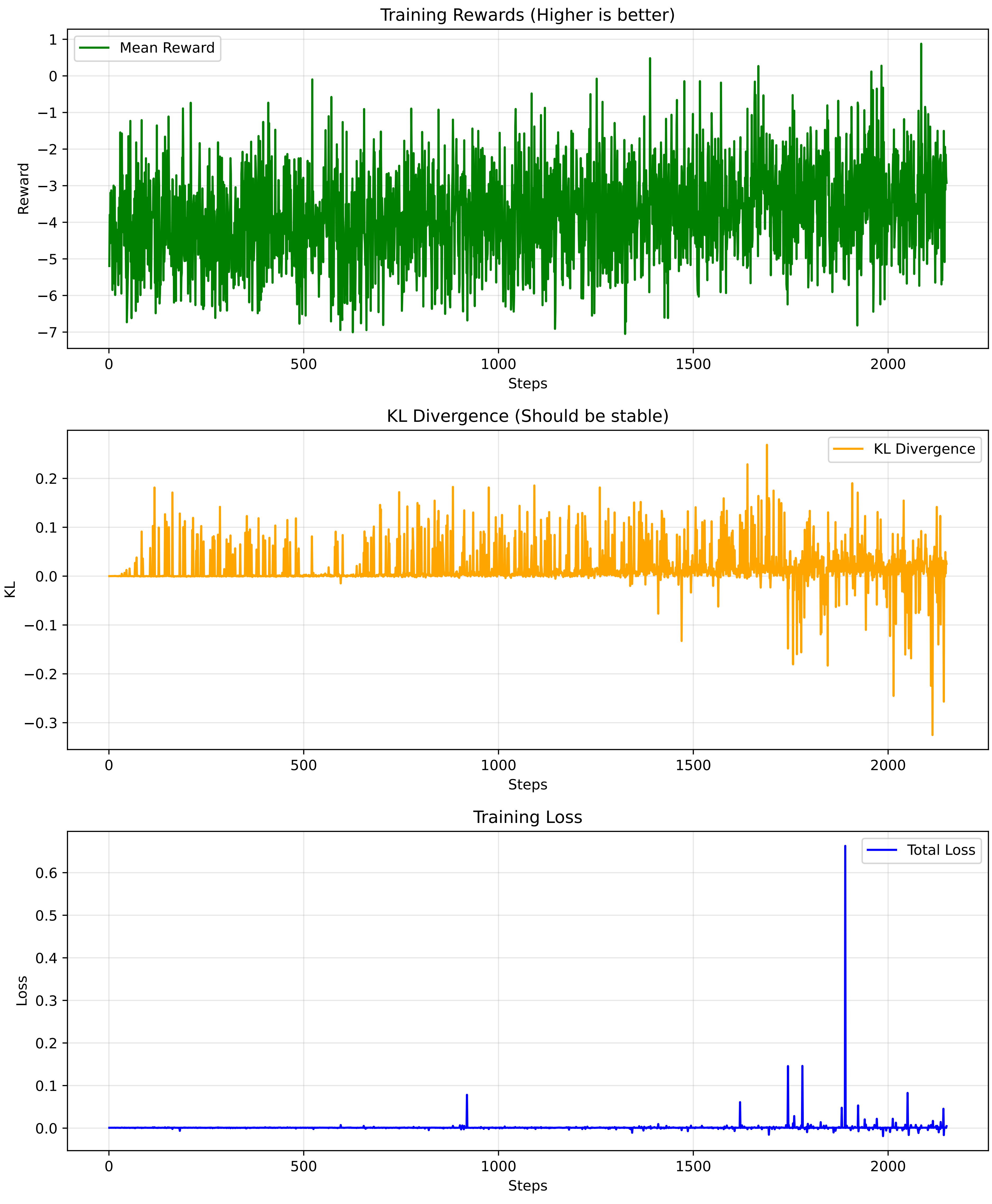

Training Rewards (Higher is better)

KL Divergence (Should be stable)

Training Loss