1 Additional results

This document contains some additional experiment results, which are supplementary to the discussions in the results paragraph of Sec 4.3 in the main paper.

1.1 Comparison of student network architectures

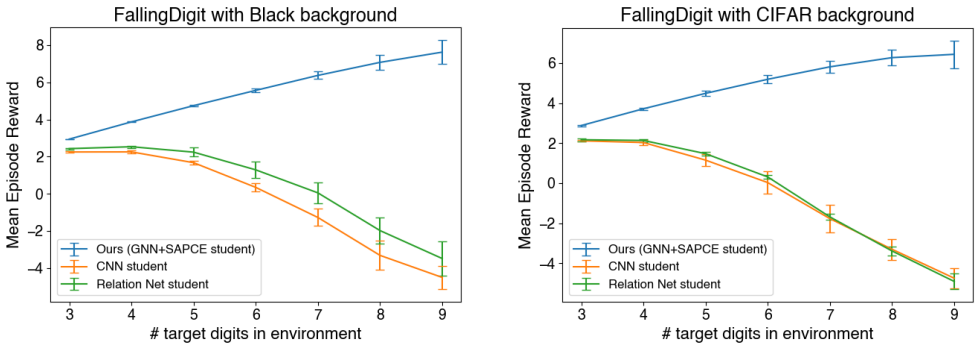

Figure 1: Comparison of different student network architectures on FallingDigit environments. All the students in this figure are refactorized from the same teacher policies (Relation Net) and the same demonstration datasets. For ours, we report the mean episode reward over 9 different runs (3 different RL teacher runs and 3 refactorization runs per teacher). For the other baselines, the result is averaged over 3 runs (3 different RL teacher runs and 1 refactorization run per teacher). The error bar shows standard deviation. We observe that GNN with SPACE generalizes best among all the three student network architectures.

Method	Test on unseen backgrounds
CNN student	2.53(0.66)
Relation Net student	3.92(2.06)
Ours (GNN+SPACE student)	**6.05(2.44)**

Table 1: Comparison of different student network architectures on BigFish. All the students in this table are refactorized from the same teacher policies (CNN). For ours, we report the mean episode reward over 24 different runs (4 different RL teacher runs, 3 different demonstration datasets per teacher and 2 refactorization runs per dataset). For the other baselines, the result is averaged over 4 runs (4 different RL teacher runs, 1 demonstration dataset per teacher and 1 refactorization run per dataset). The standard deviation is in the parentheses. We observe that GNN with SPACE generalizes best among all the three student network architectures.

1.2 Comparison of teacher and student with the same network architecture

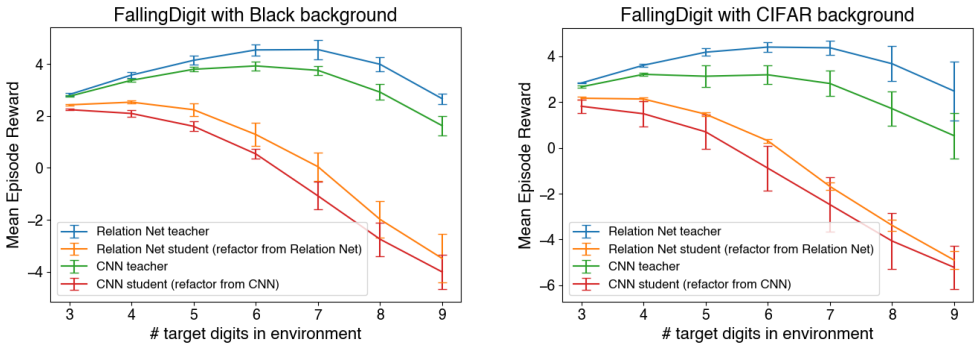

Figure 2: Comparison of teachers and students with the same network architectures on FallingDigit environments. Every student in this figure has the same network architecture with its teacher. All the teachers are trained by DQN and results are are averaged over 3 different runs. All the student results are averaged over 3 runs (3 different RL teacher runs and 1 refactorization run per teacher). The error bar shows standard deviation. We observe that the refactorization process itself does not necessarily improve the performance over the teacher policy.

Method	Test on unseen backgrounds
CNN teacher	4.40(1.90)
CNN student (refactor from CNN)	2.53(0.66)
Relation Net teacher	4.54(1.22)
Relation Net student (refactor from Relation Net)	4.21(1.55)

Table 2: Comparison of teachers and students with the same network architectures on BigFish. Every student in this table has the same network architecture with its teacher. All the teachers are trained by PPO and results are are averaged over 4 different runs. All the student results are averaged over 4 runs (4 different RL teacher runs, 1 demonstration dataset per teacher and 1 refactorization run per dataset). The standard deviation is in the parentheses. We observe that the refactorization process itself does not necessarily improve the performance over the teacher policy.