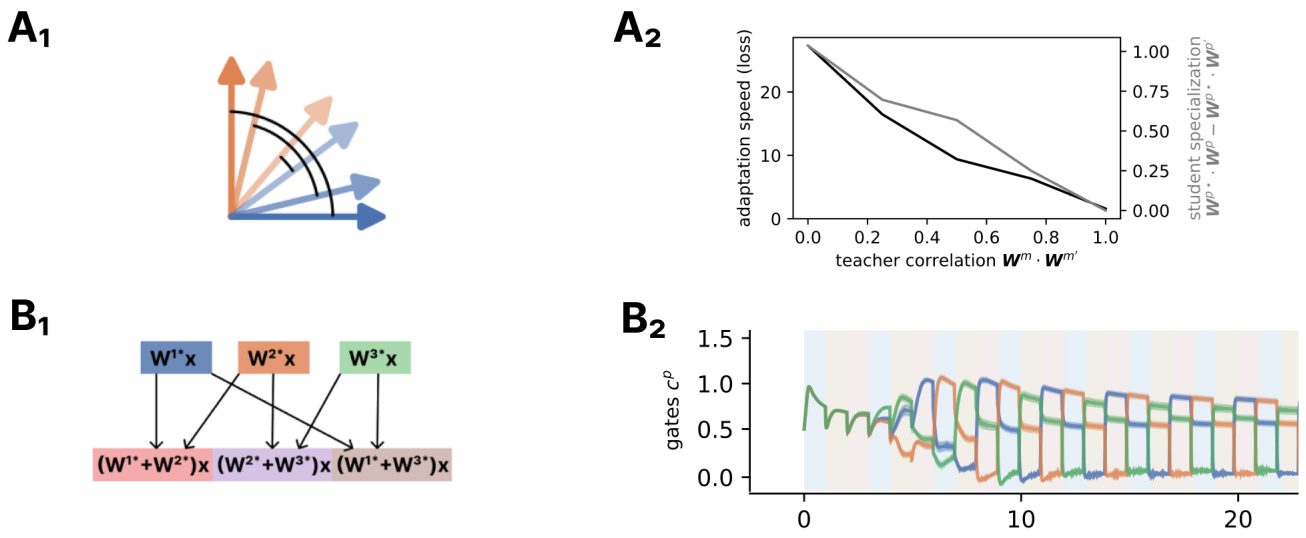

Figure R1: Robustness to relaxing orthogonality between teachers. A *(Left)* Changing teacher correlation. (*Right*) Adaptation speed as measured by the loss after a block switch (*black*) and student specialization (*gray*), both as a function of the teacher correlation. Teacher correlation is measured in terms of cosine similarity. 0 represents the orthogonal case studied in the paper. **B** (*Left*) Example of partially correlated task, generated from shared structure of compositional teachers. (*Right*) The gating variables converge to select the respective constituting teachers for the compositional task at hand.

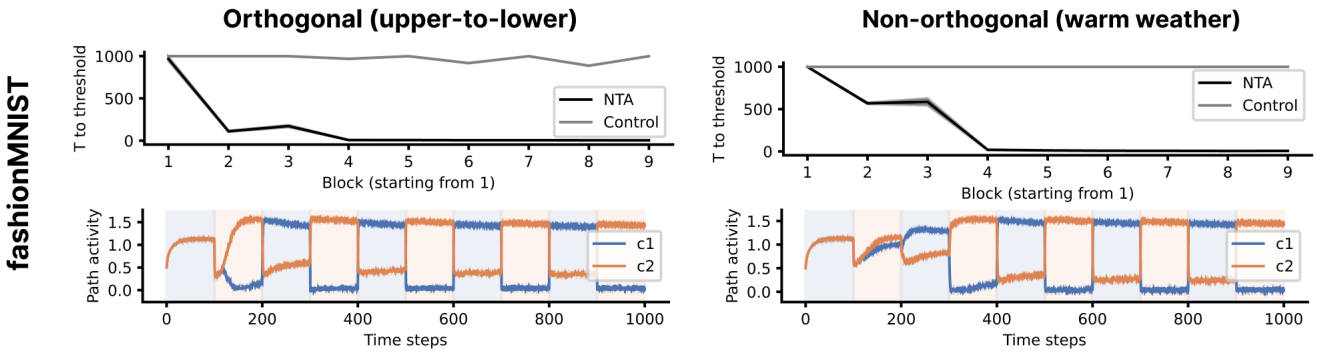

Figure R2: Adaptation to a natural image-based, correlated task. NTA quickly adapts across fashionMNIST for (*left*) an orthogonal sorting based on upper-to-lower items of clothing and (*right*) a correlated sorting for warm-to-cold weather clothing. The panels show (*top*) number of batches until accuracy threshold of 85% is reached (set to 1000 if it is never reached within a block of 1000 batches) and (*bottom*) activity of the context nodes.

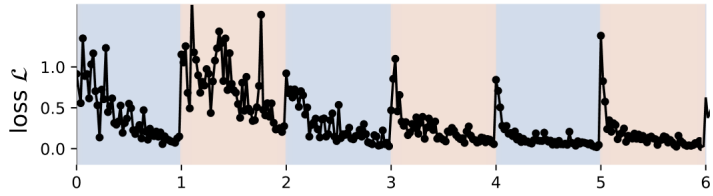

Figure R3: Few-shot adaptation after block switches. Like Fig. 1A in the manuscript, but with coarsely discretized time to examine the adaptation after a single sample. As this drastically reduces signal-to-noise ratio , we average over 100 samples. Markers indicate a single step of gradient descent on one sample.

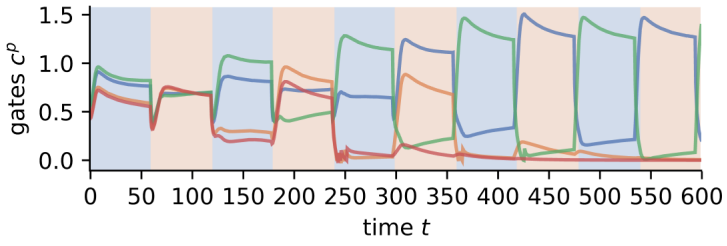

Figure R4: Redundant paths become inactive when representation is costly. Gating variables like in Fig. 1B, but with more paths than teacher tasks (P=4 > M=2). Students that are preferably aligned due to the random initialization scpecialize to the M=2 teachers, whereas other gates decay to 0.