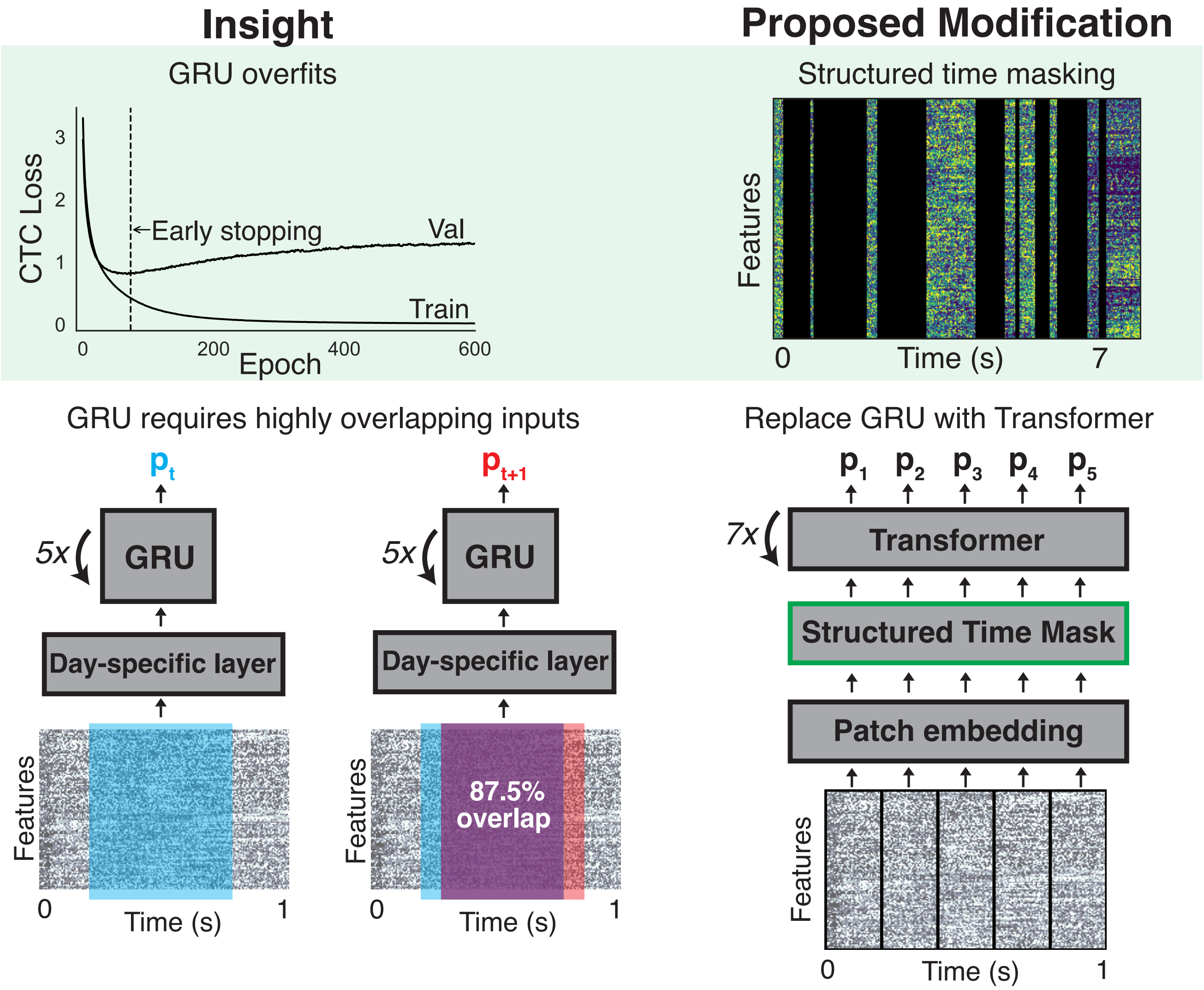

Figure 1: We propose two modifications to the training process and architecture based on two key insights. First, we observe that the baseline GRU overfits early into training (black dashed line indicates where training was stopped by original authors). We propose masking consecutive temporal chunks of neural activity during training to delay overfitting. Second, consecutive inputs to the GRU are highly overlapping in time, and this is necessary to reach optimal performance. We hypothesize that processing redundant inputs increases the computational cost of the GRU, and propose replacing the GRU with a transformer which can effectively process non-overlapping patches of neural data.
We combine these modifications by training the transformer with structured time masking.