Beyond steering in the model's native activation space, sparse autoencoders (SAEs) have been used to learn sparse feature spaces in which concepts may be more disentangled and interpretable (He et al., 2025). However, identifying features with SAE does not necessarily yield reliable controllability (Wehner et al., 2025), and SAE-based decompositions can impose systematic constraints such as restricting coefficients to be non-negative (Mayne et al., 2024). In addition, SAE-based approaches often require pretrained SAEs for the target LLM, which can limit practicality.

Finally, a related inference-time control paradigm is representation fine-tuning, a PEFT family that optimizes lightweight modules to reshape representations under a task loss (Wu et al., 2024b; Yin et al., 2024; Wu et al., 2024a). Such methods can be competitive with steering baselines on recent benchmarks (Wu et al., 2025). We therefore include a representative approach, Representation Editing (RED) (Wu et al., 2024a), in our experimental comparisons.