- *Activation steering.* Contrastive Activation Addition (CAA) (Rimsky et al., 2024) computes steering directions from exemplar differences and scales them at inference.

- *Training-time numeric control.* SteerLM (Dong et al., 2023) finetunes on data labeled with regressor-predicted attribute values, allowing users to set numeric controls such as "positivity=7/10." This achieves excellent calibration but requires labeled data and SFT cycles.

These methods demonstrate smooth, single-attribute control, but rarely extend to *dual-concept* settings. While some (e.g., FUDGE, GeDi, or energy-based methods) can in principle compose multiple guidance signals by assigning separate weights λ_a, λ_b, systematic evaluation of interference between attributes remains limited.

Towards Multi-Concept Fine-Grained Control. Recent frameworks begin to explore smooth, fine-grained evaluation. The Smoothly Controllable Text Generation (SCTG) benchmark (Zhou et al., 2024) defines fine-grained control as the ability to vary an attribute over a 10-point scale, using LLM-as-judge with Elo-style pairwise comparisons to assess calibration and relevance. However, SCTG focuses exclusively on single-attribute scenarios. In contrast, our work explicitly evaluates *dual-concept* fine-grained control, introducing systematic protocols to measure interference when varying one concept while holding another fixed. This perspective highlights the challenges of compositional control and the need for methods robust to attribute entanglement.

Evaluation of Controllability. Evaluation typically relies on automatic classifiers trained to predict style or attribute labels on generated outputs (Moschitti et al., 2014). While efficient, such classifiers often suffer from subjectivity and domain mismatch (Pang, 2019). Human evaluation remains the gold standard but is costly and inconsistent. More recent work explores LLMs themselves as judges (Zheng et al., 2023; Sun et al., 2023), providing scalable and flexible evaluation pipelines. Our evaluation setup builds on this line, using pairwise comparisons with strong judge models to assess fine-grained controllability in both single- and dual-concept scenarios.