A Risks and Ethics

There are no risks or ethical concerns with this work.

B Licensing

This work is conducted on datasets that are either publicly available or authorized for research use. We ensured that the use of all existing datasets was consistent with their original intended use as specified by their licenses. Similarly, all models used in experiments in this work were used as dictated by their respective licenses. We used Co-pilot and AI Assistants to support human-generated artifacts.

C Extended Related Work

Prompting for Concept Control. Prompt-based methods, including prefix-tuning, soft prompts, and learned prompt vectors, have emerged as lightweight alternatives to full model fine-tuning for controllable text generation. Prefix-tuning has been used to inject attributes without retraining the model (Liu et al., 2024; Gu et al., 2022a), extended to multi-aspect settings through plugin modules and disentanglement objectives (Huang et al., 2022; Zeng et al., 2023). Other approaches learn attribute-specific soft prompts, either with contrastive training (Qian et al., 2022), latent prior manipulation (Gu et al., 2022b), or interference-reducing designs such as Tailor (Yang et al., 2023b). DisCup (Zhang and Song, 2022) further integrates discriminator feedback into prompt learning, while Attribute Alignment (Yu et al., 2021) builds on conditioning mechanisms. These methods show strong controllability but require training effort and often struggle to generalize across multiple attributes.

Representation Engineering and Steering. Representation engineering (RepE) methods manipulate hidden activations to steer model behavior. They have been shown effective in controlling sentiment (e.g., shifting polarity or tone) (Turner et al.; Konen et al., 2024; Cai et al.; Zou et al., 2023), typically using datasets such as GoEmotions (Demszky et al., 2020) or Yelp (Asghar, 2016). Beyond sentiment, RepE has been extended to personality traits, steering along MBTI (Zhang et al., 2024) or OCEAN (Weng et al.) dimensions, influencing reasoning style, honesty, and conversational stance. Other work explores steering for language, style, and genre, including cross-lingual transfer (Guo et al., 2024; Scalena et al.), or stylized generation (Konen et al., 2024; Beaglehole et al., 2025). Recent steering techniques such as Contrastive Activation Addition (CAA) (Rimsky et al., 2024) provide training-free, intensity-scalable control vectors derived from positive/negative exemplars, aligning closely with the idea of numeric sliders. However, most RepE studies focus on one attribute at a time, without probing how multiple steering directions interact. Further to this, Wu et al. (2025) demonstrate that simple prompting often performs much better than many of the more complex RepE methods discussed above.

Style Transfer and Multi-Attribute Control. Supervised text style transfer methods rely on parallel corpora and sequence-to-sequence models (Jhamtani et al., 2017; Mukherjee et al., 2023), but are constrained by scarce paired data. Unsupervised methods for non-parallel data include prototype editing (swapping style markers with target-style phrases) (Mukherjee et al., 2023), or disentanglement strategies that factorize semantics and style, recombining them via back-translation or adversarial training (Shen et al., 2017; Prabhumoye et al., 2018). While effective for coarse style shifts, these approaches are not naturally suited for fine-grained numeric control or multi-attribute specification.

Fine-Grained Control: Single-Attribute Methods. Most work on fine-grained control introduces a continuous "knob" for a *single* attribute, with evaluation focused on calibration along that one dimension. Families include:

- *Decoding-time guidance.* PPLM (Dathathri et al., 2020) backpropagates from an attribute classifier through LM hidden states at generation time; GeDi (Krause et al., 2021) trains small conditional LMs to reweight token probabilities; FUDGE (Yang and Klein, 2021) trains discriminators predicting sequence-level attributes from partial prefixes; and energy/logit methods such as COLD (Qin et al., 2022) and BOLT (Liu et al., 2023) add attribute-specific energies or biases. Each provides a tunable weight parameter, enabling smooth control of attribute intensity.

- *Product-of-experts.* DExperts (Liu et al., 2021) combine base LMs with expert/anti-expert models, where the mixture coefficient α controls strength.