A Geometric Perspective on Composable Emotion Steering in Text-to-Speech Models

Siyi Wang [1] **James Bailey** [2] **Ting Dang** [1]

Abstract

While prior work has explored emotion control in hybrid text-to-speech systems, the geometric properties of these modules, and their implications for steerability, remain poorly understood. We present the first comparative study of speech language model (SLM) and conditional flow-matching (CFM) modules as activation steering sites for mixed-emotion speech synthesis. We first characterize emotion representations using linear probing and local intrinsic dimensionality (LID), and then evaluate single-site and joint steering for mixed-emotion synthesis. Our results show that SLM offers a clean, low-dimensional emotion-specific subspace with strong speaker–emotion disentanglement, while CFM exhibits poor cross-speaker generalization due to speaker–emotion entanglement. Joint steering increases emotion intensity but degrades proportional control and speech quality on in-distribution data. These findings provide practical guidance for multi-site activation steering in hybrid TTS systems and highlight the importance of representation geometry in controllable speech generation.

1. Introduction

Generating emotionally controllable speech is essential for applications such as conversational agents, audiobook narration, and assistive communication. Human emotional expression is nuanced, often involving mixed affective cues where multiple emotions coexist within a single utterance (Zhou et al., 2022; Cowen & Keltner, 2017), a complexity that current systems generally fail to control effectively. Existing emotion control methods operate through the model's external interface: label-based approaches can enable explicit emotion conditioning but require costly annotated data and retraining (Cho et al., 2025; Gao et al., 2025), while prompt-based methods can describe target emotions but lack precise quantitative control over emotion proportions (Guo et al., 2023; Yang et al., 2025).

Activation steering bypasses these limitations by directly injecting learned direction vectors into intermediate activations at inference time, without retraining (Zou et al., 2023; Turner et al., 2023). This paradigm has shown success in LLMs and text-to-image diffusion (Rodriguez et al., 2025; Rimsky et al., 2024). State-of-the-art TTS systems increasingly adopt hybrid architectures combining a speech language model (SLM) with a conditional flow-matching (CFM) decoder (Du et al., 2024; Anastassiou et al., 2024; Zhou et al., 2026), where the SLM governs high-level prosodic structure and the CFM renders fine-grained acoustics, each a potential site for steering emotional expression. Wang et al. (2026) demonstrate composable mixed-emotion steering via the SLM, while Xie et al. (2025) achieve continuous single-emotion intensity control via CFM. However, no prior work has systematically compared the representation geometry at these two steering sites, examined how geometric properties relate to steering effectiveness, or investigated whether jointly steering both modules yields complementary or interfering effects.

We present the first comparative study of SLM and CFM as activation steering sites for mixed-emotion synthesis. Through linear probing and local intrinsic dimensionality (LID) analysis of both modules' representation geometry, combined with single-site and joint steering experiments on four datasets, our study reveals three key findings: (i) the SLM encodes emotions in geometrically distinct, low-dimensional subspaces with strong cross-speaker generalization, suggesting favorable conditions for emotion-specific intervention; in contrast, the CFM entangles speaker and emotion representations, making clean emotion-only intervention difficult; (ii) SLM steering achieves superior proportional control of mixed emotions, while CFM steering produces stronger overall intensity at the cost of speaker fidelity; (iii) joint steering across both sites amplifies intensity but degrades proportional control in-distribution, due to two independent perturbations interfering rather than complementing each other. These findings offer practical guidelines

[1]The University of Melbourne, Australia [2]Monash University, Australia. Correspondence to: Siyi Wang <siyi.wang.4@student.unimelb.edu.au>.

Published at ICML 2026 Workshop on the Machine Learning for Audio, Seoul, South Korea. PMLR 306, 2026.

for multi-site steering and shed light on the latent geometry of hybrid TTS, informing future work on representation control and interpretability in speech generation.

2. Method

Modern hybrid TTS systems generate speech in two stages (Du et al., 2024). The speech language model (SLM) autoregressively generates discrete speech tokens $\mathbf{z} = f_{\text{SLM}}(\mathbf{x}, \mathbf{c}_{\text{ref}})$ from text $\mathbf{x}$ and reference audio $\mathbf{c}_{\text{ref}}$, encoding high-level prosodic and semantic structure. The conditional flow-matching (CFM) module then transforms these tokens into a mel-spectrogram $\mathbf{m} = f_{\text{CFM}}(\mathbf{z}, \mathbf{c}_{\text{ref}}, \mathbf{v})$, rendering fine-grained acoustic details.

2.1. Geometry Analysis

The geometry analysis aims to characterize how emotions are organized in the representation spaces of SLM and CFM, and in particular whether they form structures that support compositional control. To achieve reliable mixed-emotion steering, individual emotion directions should ideally be composable, so that their weighted combinations produce meaningful mixed directions (Wang et al., 2026).

Linear Discriminability We use linear probing to analyze the linear separability of emotion representations in the activation space (Alain & Bengio, 2016). Concretely, we train a linear classifier at each layer of the SLM and CFM, and compare their classification performance across layers. Higher classification performance indicates more separable emotion representations, and possibly more reliable steering vectors for compositional steering (Wang et al., 2026).

Local Intrinsic Dimensionality To further reveal the geometric structure of the representation manifold, Local intrinsic dimensionality (LID) is used (Amsaleg et al., 2015). For each sample's activation representation, we compute its K nearest neighbors (in Euclidean distance) in the activation space. Let $r_1, r_2, \cdots, r_K$ denote the distances to these neighbors sorted in ascending order. We then estimate the LID by modeling the growth rate of the neighborhood radius using the Levina–Bickel maximum likelihood estimator (Levina & Bickel, 2004), which captures how quickly the local volume expands around each point. Higher LID indicates a more complex and less constrained local geometry, suggesting that emotion information is distributed across a higher-dimensional space rather than in a compact subspace.

We compute LID for each emotion as well as over all speech samples. The *per-emotion* setting captures the geometry of each emotion-specific subspace, while the all-samples setting captures the overall geometry of the full emotion space (referred to as the *pooled* setting). We define $\Delta\text{LID} = \text{LID}_{\text{pooled}} - \overline{\text{LID}}_{\text{per-emo}}$ as the difference between pooled and average of per-emotion LID. When $\Delta\text{LID} > 0$, pooling emotions increases the estimated manifold dimensionality, indicating that different emotions contribute additional independent directions of variation beyond those captured within individual emotion subspaces. Conversely, when $\Delta\text{LID} < 0$, pooling does not increase dimensionality, suggesting that emotion categories largely lie on a shared manifold. A positive ΔLID is favorable for mixed-emotion steering, as it indicates emotion-specific directions in the representation space that can potentially be composed.

2.2. Activation Steering

Building on these geometric analyses, we next turn to how emotion directions are extracted from activations and used for steering. For each layer l at either SLM or CFM, we first extract the activation difference for each emotion e as:

$$\mathbf{u}_e^{(l)} = \frac{1}{N_e}\sum_{j=1}^{N_e}\mathbf{h}_{e,j}^{(l)} - \frac{1}{N_0}\sum_{i=1}^{N_0}\mathbf{h}_{0,i}^{(l)}, \tag{1}$$

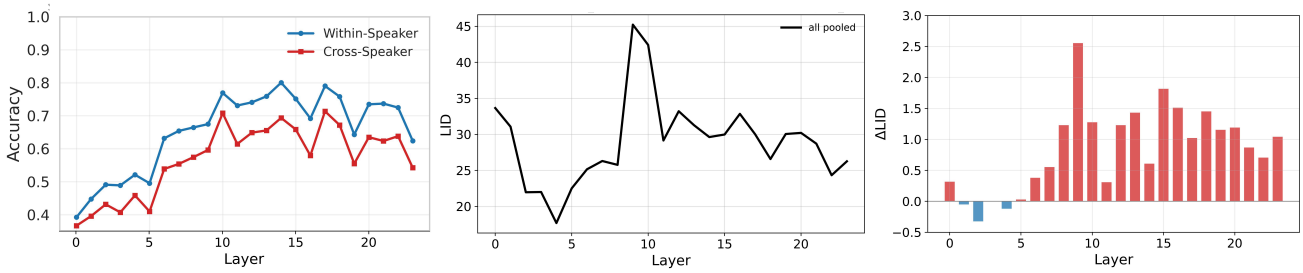
where $\mathbf{h}_{e,j}^{(l)}$ and $\mathbf{h}_{0,i}^{(l)}$ denote activations from emotion-e and neutral samples, respectively. For the SLM, steering vector $\mathbf{v}_e^{(l)} = \mathbf{u}_e^{(l)}$ is extracted from attention output activations at the last-token position of complete utterances (Wang et al., 2026). For the CFM, $\mathbf{u}_e^{(l)}$ is extracted from residual stream activations, L_2-normalized, masked to the top-k emotion-relevant frames identified via an emotion classifier, and aggregated to get $\mathbf{v}_e^{(l)}$ (Xie et al., 2025). For mixed-emotion synthesis, single-emotion vectors are composed via weighted summation: $\mathbf{v}_{\text{mix}}^{(l)} = \sum_e p_e \mathbf{v}_e^{(l)}$, where p_e denotes the proportion summing to 1.

At inference, steering is applied by modifying the activation at layer l: $\tilde{\mathbf{h}}^{(l)} = f_r(\mathbf{h}^{(l)} + \alpha \cdot \mathbf{v}_{\text{mix}}^{(l)})$, where α controls steering strength and f_r renormalizes the modified activation to preserve the original scale (Turner et al., 2023).

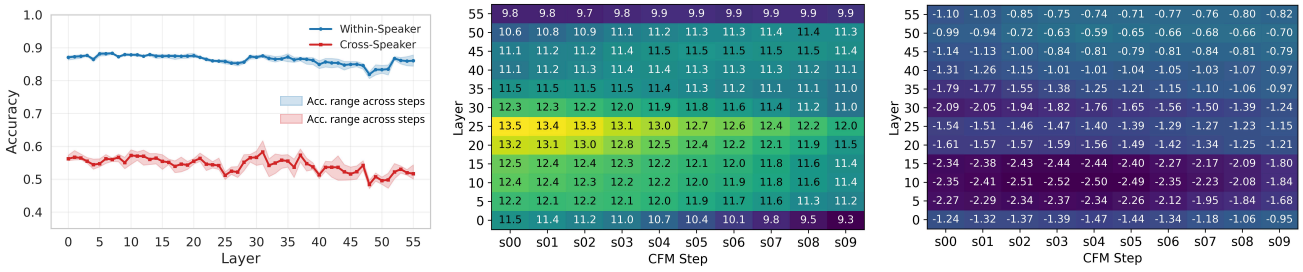
3. Experimental Setup

Model. We use CosyVoice2 (Du et al., 2024) as our backbone. The SLM is a 24-layer Qwen2.5-based (Qwen et al., 2025) transformer and the CFM is a 56-layer DiT with 10 denoising steps. For geometry analysis, we extract activations from all SLM layers and CFM layers across 10 steps. For steering, based on the findings in Section 4.1, we apply steering vectors at SLM layers 14 and 17 (Wang et al., 2026). For CFM steering, since emotion discriminability is uniformly distributed across layers (Section 4.1), we follow Xie et al. (2025) and apply steering vectors at every 5th layer (12 layers in total) across all 10 denoising steps.

Datasets. We use ESD (Zhou et al., 2021), CREMA-D (Cao et al., 2014), and RAVDESS (Livingstone & Russo, 2018) across five emotions (angry, happy, neutral, sad, surprise). For linear probing, we reserve 30% of speakers for cross-speaker evaluation (4,530 utterances). From the remaining speakers, 11,311 utterances are used for probe training and 4,850 utterances for within-speaker evaluation. For LID, we



(a) Emotion discriminability in SLM *(b)* Pooled LID in SLM *(c)* ΔLID in SLM



(d) Emotion discriminability in CFM *(e)* Pooled LID in CFM *(f)* ΔLID in CFM

Figure 1. Geometry analysis of SLM (a–c) and CFM (d–f). (a,d) Per-layer emotion discriminability (linear-probe accuracy); blue = within-speaker, red = cross-speaker; shading in (d) shows the accuracy range across denoising steps. (b,e) Pooled LID across layers (CFM shown per denoising step). (c,f) $\Delta\text{LID} = \text{LID}_{\text{pooled}} - \overline{\text{LID}}_{\text{per-emo}}$ across layers.

sample 4,000 utterances for both per-emotion and pooled estimates, with k=50 neighbors, averaged over 10 resampling trials. Steering vectors are extracted from 50% of speakers and evaluated on CREMA-D (in-distribution) and IEMOCAP (Busso et al., 2008) (out-of-distribution), with mixed-emotion ground truths derived from multi-rater annotation disagreement (Wang et al., 2026).

Evaluation metrics. We evaluate along two axes. For emotion control we use: **E-SIM**, cosine similarity between Emotion2Vec embeddings (Ma et al., 2024) of synthesized and ground-truth speech; **TEP**, mean probability assigned to target emotions by the Emotion2Vec classifier; ρ, Spearman correlation between the ranking of emotion probability increases and the ground-truth emotion ranking; and **H-Rt**, fraction of samples where the ground-truth dominant emotion shows the largest probability increase. For speech quality we report: **S-SIM**, cosine similarity between WavLM speaker embeddings (Chen et al., 2022) of synthesized and reference speech; and **WER**, word error rate via Whisper-Large-v3 (Radford et al., 2023).

4. Results

4.1. Geometry Comparison

Linear discriminability. In the SLM, within-speaker accuracy reaches 0.80 and cross-speaker accuracy 0.71, yielding a small mean gap of 0.08 (Figure 1a). Discriminability peaks in the mid-to-late layers (10–17), indicating that emotion information is concentrated in a localized, speaker-invariant subspace. In contrast, the CFM achieves similarly high within-speaker accuracy (0.89) but much lower cross-speaker accuracy (0.62), resulting in a larger mean gap of 0.32 (Figure 1d). Moreover, discriminability is broadly uniform across layers and denoising steps, with no clear peak. These results suggest that SLM representations are more separable and generalizable across speakers, while CFM representations are more entangled with speaker identity and diffusely distributed, making the SLM a more suitable site for extracting robust emotion steering directions.

LID trend. In the SLM (Figure 1b black line), pooled LID exhibits a compression–expansion pattern, first decreasing, then increasing, and finally stabilizing in later layers, consistent with geometric dynamics observed in transformer representations (Valeriani et al., 2023). In contrast, the CFM (Figure 1e) consistently exhibits an increase followed by a decrease in LID across layers at every denoising step, suggesting that intermediate layers construct richer representations with more complex local geometry before compressing them toward the prediction target. Furthermore, LID progressively decreases across denoising steps, indicating that the representation manifold becomes increasingly structured and lower-dimensional as CFM iteratively refines towards final output (Lipman et al., 2022).

Emotion subspace structure (ΔLID). ΔLID reveals a fundamental contrast between the two modules (Figure 1c,1f). In the SLM, ΔLID is near zero in early layers (0–5), then becomes consistently positive from layer 6 onward (mean: +0.84), indicating that combining emotion categories increases manifold dimensionality, and that emotions occupy distinct directions and contribute additional geometric struc-

Table 1. Geometric comparison of SLM and CFM as steering sites on CosyVoice2. Probe accuracies are best-layer values; within–cross gap is averaged across all layers and steps.

Property	SLM	CFM
Hidden Dim.	896	256
Probe acc. (within / cross)	0.80 / 0.71	0.89 / 0.62
Mean within–cross gap	0.08	0.32
Manifold dim. (LID)	$\sim$28	$\sim$13
ΔLID	Positive (+0.84)	Negative (-1.48)
Discriminability peak	Mid-to-late	Uniform

Table 2. Steering results for mixed-emotion speech synthesis on CosyVoice2. Best in **bold**, second underlined.

Data	Config	E-SIM↑	TEP↑	ρ ↑	H-Rt↑	S-SIM↑	WER↓
CREMA-D	No-steer	.743	.065	–	–	.871	1.07
	CFM $\alpha=1.0$	.767	.097	.098	.691	.858	**0.76**
	CFM $\alpha=2.0$	.786	.160	.193	.717	.807	0.79
	SLM $\alpha=3.0$	.762	.100	.166	.709	**.872**	1.01
	SLM $\alpha=5.0$	.779	.149	**.209**	**.724**	.870	0.78
	Joint $\alpha=1.0$	.767	.131	.112	.695	.859	1.02
	Joint $\alpha=2.0$	**.787**	**.163**	.176	.711	.808	1.06
IEMOCAP	No-steer	.903	.197	–	–	.888	6.70
	CFM $\alpha=1.0$	.910	.218	.138	.729	.885	6.08
	CFM $\alpha=2.0$	.909	.272	.117	.721	.844	6.15
	SLM $\alpha=3.0$	.911	.228	.186	.744	**.891**	**5.86**
	SLM $\alpha=5.0$	**.915**	.253	**.215**	**.755**	.890	6.27
	Joint $\alpha=1.0$	.912	.237	.193	.746	.884	6.05
	Joint $\alpha=2.0$	.911	**.274**	.170	.737	.845	6.29

ture beyond that of individual emotion manifolds. In the CFM, ΔLID is negative across all 56 layers and 10 steps (mean: -1.48), indicating that pooling emotions does not increase dimensionality and that emotion categories largely reside on a shared acoustic manifold. This contrast suggests that SLM contains more distinct emotion subspaces and is more favorable for compositional steering. The overall comparison is summarized in Table 1.

4.2. Steering Comparisons

Table 2 compares steering applied at the SLM only, CFM only, and both modules jointly. Our SLM-only and CFM-only conditions instantiate the steering approaches of (Wang et al., 2026) and (Xie et al., 2025) respectively. For each steering site, we vary the steering strength α and report the best-performing configuration under comparable preserved speech quality (S-SIM within 10% of baseline, WER increase <0.5).

Emotion control. Emotion embedding similarity (E-SIM) and target emotion intensity (TEP) improve over the baseline for both SLM and CFM steering, with comparable performance, indicating that both sites effectively align generated speech with target emotional embeddings. Joint steering yields the highest TEP across datasets, as combined perturbations reinforce overall emotion intensity.

For fine-grained proportional control of each emotion (ρ, H-Rt), SLM steering consistently outperforms CFM on both datasets (Table 2), consistent with geometric analysis. The positive ΔLID and low-dimensional emotion subspace in SLM support cleaner compositional steering, enabling the precise control over emotion mixing. However, joint steering degrades proportional control on in-distribution data, suggesting that steering both modules simultaneously complicates the control over individual emotion ratios.

Speech quality cost. S-SIM degrades noticeably under CFM steering, while SLM steering preserves speaker identity. This is consistent with CosyVoice2 architecture: the SLM is not conditioned on speaker embeddings, whereas the flow-matching module is explicitly conditioned on speaker embeddings and reference speech (Du et al., 2024), so perturbing CFM activations directly interferes with speaker-dependent representations. This aligns with the speaker–emotion entanglement revealed by geometry analysis (§4.1). WER remains stable for single-site methods but increases slightly under joint steering. Overall, *SLM steering provides a better balance between controllability and preservation, making it the more suitable site for emotion steering*.

Steering analysis. The results in §4.2 show that joint steering does not provide additive gains over single-site steering, which we attribute to three factors. (i) Distribution shift: SLM steering moves activations away from the neutral manifold before they reach the CFM module, causing a mismatch for CFM steering vectors, especially in mixed-emotion settings. (ii) Speaker entanglement: CFM steering additionally perturbs speaker-dependent acoustics due to speaker–emotion entanglement in the flow-matching space. (iii) Uncoordinated perturbation: independent interventions at both sites accumulate noise rather than compose, leading to interference that reduces proportional control despite increasing overall emotion intensity.

Future directions. (i) CFM vectors could be extracted conditioned on SLM-steered output so that the extraction distribution matches actual inference-time conditions. (ii) Steering vectors in the CFM could be orthogonalized against speaker directions (Ravfogel et al., 2020; Bartoszcze et al., 2025) to mitigate speaker–emotion entanglement. (iii) Independent per-site α tuning or frame-level adaptive steering may allow the two modules to complement each other more effectively. (iv) Extending this analysis to architecturally distinct systems (e.g., IndexTTS2) would test generality. (v) The geometry–steering relationship, currently characterized at the module level, could be analyzed per-layer and per-step to pinpoint the most steerable directions within each module.

5. Conclusion

We presented the first comparative study of SLM and CFM modules as activation steering sites in hybrid TTS, providing a geometry-to-application analysis of how each module encodes and controls emotion. Our findings reveal distinct roles: the SLM provides clean, speaker-invariant emotion subspaces suited for proportional mixed-emotion control, while the CFM module contributes rich acoustic detail but entangles emotion with speaker identity. Joint steering amplifies intensity but introduces interference, highlighting the need for coordinated multi-site strategies. By connecting representation geometry to steering outcomes, this work offers both an analytical framework and practical guidance for emotion control in hybrid TTS architectures.

References

Alain, G. and Bengio, Y. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Amsaleg, L., Chelly, O., Furon, T., Girard, S., Houle, M. E., Kawarabayashi, K.-i., and Nett, M. Estimating local intrinsic dimensionality. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 29–38, 2015.

Anastassiou, P., Chen, J., Chen, J., Chen, Y., Chen, Z., Chen, Z., Cong, J., Deng, L., Ding, C., Gao, L., et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

Bartoszcze, L., Munshi, S., Sukidi, B., Yen, J., Yang, Z., Williams-King, D., Le, L., Asuzu, K., and Maple, C. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint arXiv:2502.17601*, 2025.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008.

Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A., and Verma, R. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022.

Cho, D.-H., Oh, H.-S., Kim, S.-B., and Lee, S.-W. Emosphere++: Emotion-controllable zero-shot text-to-speech via emotion-adaptive spherical vector. *IEEE Transactions on Affective Computing*, 2025.

Cowen, A. S. and Keltner, D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114 (38):E7900–E7909, 2017.

Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H., et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Gao, X., Zhang, C., Chen, Y., Zhang, H., and Chen, N. F. Emo-dpo: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Guo, Z., Leng, Y., Wu, Y., Zhao, S., and Tan, X. Prompttts: Controllable text-to-speech with text descriptions. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Levina, E. and Bickel, P. Maximum likelihood estimation of intrinsic dimension. *Advances in neural information processing systems*, 17, 2004.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Livingstone, S. R. and Russo, F. A. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391, 2018.

Ma, Z., Zheng, Z., Ye, J., Li, J., Gao, Z., Zhang, S., and Chen, X. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Ravfogel, S., Elazar, Y., Gonen, H., Twiton, M., and Goldberg, Y. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7237–7256, 2020.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Rodriguez, P., Blaas, A., Klein, M., Zappella, L., Apostoloff, N., Suau, X., et al. Controlling language and diffusion models by transporting activations. In *International Conference on Learning Representations*, volume 2025, pp. 89812–89855, 2025.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Valeriani, L., Doimo, D., Cuturello, F., Laio, A., Ansuini, A., and Cazzaniga, A. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.

Wang, S., Tan, S., Liu, S., Jia, H., Huang, G., Bailey, J., and Dang, T. Cocoemo: Composable and controllable human-like emotional tts via activation steering. *arXiv preprint arXiv:2602.03420*, 2026.

Xie, T., Yang, S., Li, C., Yu, D., and Liu, L. Emosteer-tts: Fine-grained and training-free emotion-controllable text-to-speech via activation steering. *arXiv preprint arXiv:2508.03543*, 2025.

Yang, G., Yang, C., Chen, Q., Ma, Z., Chen, W., Wang, W., Wang, T., Yang, Y., Niu, Z., Liu, W., et al. Emovoice: Llm-based emotional text-to-speech model with freestyle text prompting. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10748–10757, 2025.

Zhou, K., Sisman, B., Liu, R., and Li, H. Seen and unseen emotional style transfer for voice conversion with a new emotional speech dataset. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 920–924. IEEE, 2021.

Zhou, K., Sisman, B., Rana, R., Schuller, B. W., and Li, H. Speech synthesis with mixed emotions. *IEEE Transactions on Affective Computing*, 14(4):3120–3134, 2022.

Zhou, S., Zhou, Y., He, Y., Zhou, X., Wang, J., Deng, W., and Shu, J. Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 35139–35148, 2026.

Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.