PROBECRC: RELIABLE PROBE-GUIDED SPEECH ATTRIBUTE STEERING WITH CONFORMAL RISK CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern text-to-speech models generate highly natural speech, yet controlling specific speaker attributes typically requires dedicated model adaptation and training. Existing inference-time alternatives can substantially degrade speech quality and provide no guarantees on the resulting attribute change. We introduce **ProbeCRC**, a probe-guided activation-steering framework with risk-controlled strength selection for pretrained autoregressive text-to-speech models. Layer-wise linear probes identify attribute-related directions, which we use to construct regularized hidden-state interventions. A shared steering parameter, defined in probe-logit space, translates the desired steering strength into layer-specific perturbations. We calibrate this parameter on held-out data using Conformal Risk Control (CRC), enabling attribute manipulation under a user-specified bound on failure risk. Experiments demonstrate effective bidirectional shifts in gender, emotion, and age. ProbeCRC outperforms baseline steering approaches and satisfies the prescribed risk guarantees while minimally affecting speech quality. Demo samples are available online.[1]

1 INTRODUCTION

Modern generative speech models can synthesize remarkably natural and expressive speech (Tan et al., 2024; Wang et al., 2023; Le et al., 2023), yet controlling how that speech is generated remains an open challenge. Recent instruction- and prompt-conditioned speech generation models allow users to specify desired characteristics, such as speaking style, gender, or emotion, through natural-language descriptions alongside reference audio (Guo et al., 2023; Yang et al., 2023; VoxCPM Team, 2026; Li et al., 2026; Zhou et al., 2026). Other controllable synthesis architectures incorporate structured representations, using spherical emotion vectors to control style and intensity (Cho et al., 2024), or rotational embeddings for smooth attribute control (Tian et al., 2025). While these approaches provide flexible control, they rely on control mechanisms learned during model training and do not directly expose or manipulate the internal representations underlying these attributes. Directly intervening in such representations therefore provides a complementary route toward controllable speech generation.

Activation steering modifies a model's intermediate representations at inference time to influence its outputs without updating the underlying model parameters. This paradigm has been increasingly studied in large language models and, more recently, in vision and speech generation (Zou et al., 2023; Li et al., 2023; Pandey et al., 2024; Su et al., 2025; Konovalova et al., 2026). In speech, existing methods have demonstrated that directions associated with properties such as emotion, accent, and prosody can be identified in hidden representation spaces and manipulated during generation (Xie et al., 2025; Lee et al., 2025; Wang et al., 2026a;b; Zhang et al., 2026; Yang & Hansen, 2026). However, determining an effective intervention requires not only identifying *where* and *in which direction* to steer, but also deciding *how strongly* to intervene.

The latter problem is particularly important because steering strength directly controls the magnitude of the intervention. Insufficient steering may fail to induce the desired attribute, whereas excessive

[1] `https://probecrc.netlify.app/`.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

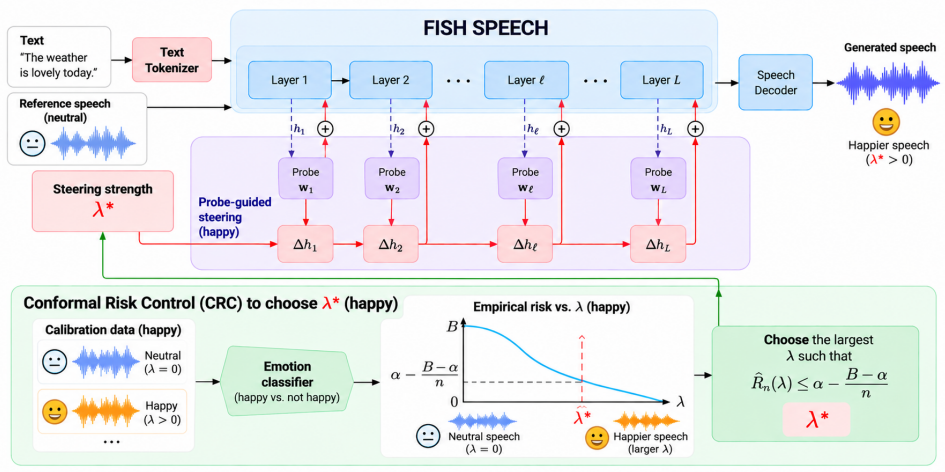


Figure 1: Overview of ProbeCRC, illustrated for steering neutral speech toward a happy emotion.

steering may unnecessarily perturb the generated speech. In standard additive activation steering, intervention strength is typically controlled by a scalar coefficient multiplying a steering direction, whose value may be tuned as a hyperparameter (Li et al., 2023). Alternatively, the coefficient can be determined adaptively from a predefined target probe output (Gao et al., 2026). To the best of our knowledge, existing approaches do not explicitly select steering strength to satisfy a user-specified bound on steering-failure risk.

In this work, we introduce **ProbeCRC**, a probe-guided activation-steering framework with risk-controlled strength selection for controllable autoregressive TTS, which is illustrated in Fig. 1. Layer-wise linear probes identify attribute-related directions, which we use to construct regularized hidden-state interventions. Rather than directly scaling these directions in hidden space, we define a shared steering parameter λ in probe-logit space and translate it into layer-specific perturbations. We then calibrate λ on held-out data using Conformal Risk Control (CRC) (Angelopoulos et al., 2024), enabling attribute manipulation under a user-specified bound on steering-failure risk.

Our main contributions are:

- We introduce a probe-guided activation-steering method that converts a shared target change in probe-logit space into regularized, layer-specific hidden-state interventions.
- We formulate steering-strength selection as a risk-control problem and use CRC to calibrate the steering parameter with respect to a user-specified steering-failure risk.
- We show that ProbeCRC enables bidirectional manipulation of gender, emotion, and age while preserving speech quality and outperforming baseline methods. Moreover, the risk-controlled strength-selection procedure satisfies guarantees while minimally affecting speech quality.

2 RELATED WORK

Activation Steering in Generative Models. Recent work has explored activation steering as an inference-time approach for controlling pretrained models without retraining their underlying parameters. In transformer-based autoregressive models, semantically meaningful attributes can be manipulated by modifying hidden representations along attribute-related directions (Zou et al., 2023). Within this line of work, activation steering has been applied to properties such as sentiment, truthfulness, refusal behavior, language, and multiple simultaneous attributes (Turner et al., 2023; Li et al., 2023; Arditi et al., 2024; Yunfan et al., 2025; Nguyen et al., 2025; Su et al., 2025). The same general principle has also been explored beyond language models, including diffusion-based image generation and autoregressive music generation (Konovalova et al., 2026; Zhao et al., 2026).

A growing line of work uses learned predictors of model activations to construct more targeted interventions. In particular, linear probes have been used not only to decode attributes from hidden representations, but also to derive directions for subsequent activation steering (Li et al., 2023; Gao et al., 2026). These methods demonstrate that learned representations of attributes can guide

inference-time interventions, but differ in how the intervention direction and magnitude are constructed.

Activation Steering for Speech Generation. Controllable speech generation has traditionally been approached through instruction- and prompt-conditioned models, which allow attributes such as speaking style and emotion to be specified through natural-language descriptions or multi-modal prompts (Guo et al., 2023; Yang et al., 2023; VoxCPM Team, 2026; Li et al., 2026; Zhou et al., 2026). These approaches incorporate controllability into the model during training. More recently, activation steering has emerged as an alternative approach for controlling pretrained speech generation models at inference time without retraining their underlying parameters. Existing approaches construct interventions from contrasts between speech representations for attributes such as emotion and accent (Xie et al., 2025; Yang & Hansen, 2026), use learned predictors and their gradients to perform targeted edits of prosody and pronunciation (Lee et al., 2025), or steer interpretable features identified through sparse autoencoders (Koriagin et al., 2026). Other work has investigated where controllable attributes are represented within modern TTS architectures (Wang et al., 2026a;b), or combined hidden-state steering with guidance in the acoustic space for emotion control (Zhang et al., 2026). Our method uses layer-wise probe directions and defines the global steering parameter through a desired change in probe-logit space rather than directly as a hidden-space scaling coefficient.

Existing approaches consequently devote substantial attention to *where* steering should be applied and *how* an effective intervention should be constructed, using activation differences (Xie et al., 2025; Wang et al., 2026a;b; Yang & Hansen, 2026) or learned attribute predictors (Lee et al., 2025). Steering strength is typically treated as a hyperparameter in speech activation steering. Outside the speech domain, probe-based methods determine it from a target probe probability (Gao et al., 2026), but without explicit risk control. Our work addresses this complementary problem by defining steering strength in probe-logit space and using Conformal Risk Control to calibrate it from held-out data.

3 BACKGROUND

3.1 Additive Activation Steering

Activation steering modifies intermediate model representations during inference while leaving the model parameters unchanged. A common formulation applies an additive intervention at layer ℓ,

$$h'_\ell = h_\ell + \lambda v_\ell, \tag{1}$$

where $h_\ell \in \mathbb{R}^d$ is the ℓth layer hidden representation, $v_\ell \in \mathbb{R}^d$ is a steering direction and $\lambda \in \mathbb{R}$ controls the strength and sign of the intervention.

A widely used approach constructs v_ℓ from the difference between the mean activations of two contrastive groups. Let $\mathcal{H}_\ell^+$ and $\mathcal{H}_\ell^-$ denote the layer-ℓ activations associated with the target and contrasting attributes, respectively. Their mean representations are

$$\mu_\ell^+ = \frac{1}{|\mathcal{H}_\ell^+|} \sum_{h \in \mathcal{H}_\ell^+} h, \qquad \mu_\ell^- = \frac{1}{|\mathcal{H}_\ell^-|} \sum_{h \in \mathcal{H}_\ell^-} h, \tag{2}$$

yielding the mean-difference steering direction $v_\ell^{\mathrm{MD}} = \mu_\ell^+ - \mu_\ell^-$. This contrastive construction can be used directly in the additive intervention of Eq. (1) and has been widely used in activation steering (e.g., Turner et al., 2023), including for speech attributes such as accent and emotion (Yang & Hansen, 2026; Xie et al., 2025). Alternatively, steering directions can be learned directly from labeled activations, for example using linear probes. Regardless of how the direction is obtained, activation steering ultimately modifies the hidden state through an additive intervention whose magnitude is controlled by a scalar coefficient.

3.2 Conformal Prediction and Conformal Risk Control

Conformal prediction (CP) is a distribution-free framework for calibrating the uncertainty of a predictive model using held-out data (Shafer & Vovk, 2008; Angelopoulos & Bates, 2023). Given a pretrained model and an independent calibration set $\{(X_i, Y_i)\}_{i=1}^n$, split conformal prediction constructs a prediction set $C(X)$ for a new input X. Under exchangeability of the calibration and test examples, it satisfies the marginal coverage guarantee $\Pr(Y_{n+1} \in C(X_{n+1})) \geq 1 - \alpha$, where $\alpha \in (0, 1)$ is a user-specified error level, equivalently bounding the marginal miscoverage probability by α. Importantly, this guarantee does not require assumptions on the form of the underlying data distribution.

Conformal Risk Control (CRC) extends conformal calibration from prediction-set miscoverage to general bounded monotone losses (Angelopoulos et al., 2024). Consider a family of prediction rules parametrized by $\lambda \in \Lambda$ and an associated loss function $L(\lambda)$, bounded above by $B < \infty$. The population risk is $R(\lambda) = \mathbb{E}[L(\lambda)]$. Given a set of n calibration examples with associated losses $\{L_i\}_{i=1}^n$, the empirical risk is $\widehat{R}_n(\lambda) = \frac{1}{n}\sum_{i=1}^n L_i(\lambda)$. CRC accounts for the finite calibration sample through the corrected empirical risk

$$h_n(\lambda) = \frac{n}{n+1}\widehat{R}_n(\lambda) + \frac{B}{n+1}. \tag{3}$$

For a prescribed target risk α, the calibrated parameter is selected as

$$\hat{\lambda}_\alpha = \inf\{\lambda \in \Lambda : h_n(\lambda) \leq \alpha\}. \tag{4}$$

For a new test example with loss L_{n+1} this selection provides the following expected-risk guarantee.

Theorem 1 (CRC Risk-Control Guarantee)**.** *Assume that $\{L_i(\lambda)\}_{i=1}^{n+1}$ are exchangeable, non-increasing in λ, and satisfy $\sup_{\lambda\in\Lambda} L_i(\lambda) \leq B < \infty$ almost surely. Assuming additionally that the feasible endpoint satisfies the target risk, the CRC-selected parameter in Eq. (4) satisfies*

$$\mathbb{E}\left[L_{n+1}\left(\hat{\lambda}_\alpha\right)\right] \leq \alpha. \tag{5}$$

4 METHOD

We introduce ProbeCRC for reliable probe-guided speech attribute steering. Our method consists of three stages. First, we train layer-wise linear probes to identify attribute-relevant directions in the model's hidden representation space. Second, we parameterize steering by a target change in probe-logit space and use the learned probes to translate this target into regularized, layer-specific hidden-state interventions. Finally, we use CRC to select the steering strength λ^* according to a binary steering-failure criterion.

4.1 Base Model

We perform steering on Fish-Speech (Liao et al., 2024), an autoregressive text-to-speech model that generates discrete semantic tokens from textual input. Fish-Speech employs a dual autoregressive architecture, consisting of a Slow Transformer that models the sequence-level semantic representation and a Fast Transformer that predicts the individual codebooks. We apply our interventions to the hidden representations of the Slow Transformer during semantic-token generation.

Let $h_\ell \in \mathbb{R}^d$ denote the Slow Transformer hidden representation at layer ℓ, where $\ell = 1, \ldots, L$. The model parameters remain frozen throughout our experiments; only the lightweight attribute probes described below are trained. During inference, the learned interventions are added to these hidden representations without modifying the underlying model parameters.

4.2 Layer-wise Attribute Probes

For each layer ℓ, we independently train a linear probe to predict a binary target attribute from the corresponding hidden representations. The probe is defined as

$$p_\ell(y = 1 \mid h_\ell) = \sigma\left(w_\ell^\top h_\ell + b_\ell\right), \tag{6}$$

where $w_\ell \in \mathbb{R}^d$ and $b_\ell \in \mathbb{R}$ are the learned probe parameters, and $\sigma(\cdot)$ denotes the sigmoid function. The corresponding probe logit is $\phi_\ell(h_\ell) = w_\ell^\top h_\ell + b_\ell$.

Beyond predicting whether the target attribute is encoded at a given layer, the probe provides an attribute-relevant direction in the hidden representation space. Since the probe logit changes linearly along w_ℓ, we use the learned probe weights to construct a layer-specific hidden-state intervention.

4.3 Probe-Guided Regularized Steering

Rather than directly scaling the probe direction in hidden space, we parameterize steering through a desired change in the probe logit. Specifically, we use a shared global parameter $\lambda \in \mathbb{R}$ and define

$$\Delta\phi_\ell = \lambda, \tag{7}$$

for each intervened layer. The sign of λ determines the steering direction, while its magnitude determines the desired change in the attribute logit.

Consider an additive intervention $v_\ell \in \mathbb{R}^d$, $h'_\ell = h_\ell + v_\ell$. Because the probe is linear, the corresponding change in its logit is

$$\phi_\ell(h'_\ell) - \phi_\ell(h_\ell) = w_\ell^\top v_\ell. \tag{8}$$

We seek an intervention that approximates the desired probe-logit change while limiting the magnitude of the perturbation applied to the hidden representation. We therefore define

$$v_\ell^* = \arg\min_{v_\ell} \left(w_\ell^\top v_\ell - \lambda\right)^2 + \eta\|v_\ell\|_2^2, \tag{9}$$

where $\eta > 0$ controls the trade-off between matching the target probe-logit change and limiting the hidden-state perturbation. The resulting intervention is (see Appendix A.1):

$$v_\ell^*(\lambda) = \left(w_\ell w_\ell^\top + \eta I\right)^{-1} w_\ell \lambda. \tag{10}$$

In practice, we solve the corresponding linear system directly, avoiding explicit matrix inversion. At inference time, the intervention is applied as $h'_\ell = h_\ell + v_\ell^*(\lambda)$. The same global λ is shared across the intervened layers, while the layer-specific probe parameters determine how the desired logit change is mapped into each layer's hidden representation space.

4.4 Risk-Controlled Steering Strength Selection

We use CRC to select the global steering strength λ on a held-out calibration set. For each calibration example i, we define the binary steering-failure loss

$$L_i(\lambda) = \mathbf{1}\left\{\hat{Y}_i(\lambda) \neq Y^{\text{target}}\right\}, \qquad L_i(\lambda) \in \{0, 1\}, \tag{11}$$

where $\hat{Y}_i(\lambda)$ is the attribute predicted by the external evaluator after steering with strength λ, and Y^{target} is the desired attribute. Consequently, $R(\lambda) = \Pr\left(\hat{Y}(\lambda) \neq Y^{\text{target}}\right)$, such that the CRC risk directly quantifies the probability of steering failure.

Since $L_i(\lambda) \in \{0, 1\}$, the loss is bounded by $B = 1$. The loss is monotone in the absolute steering strength $|\lambda|$ as increasing $|\lambda|$ strengthens the target attribute and thereby reduces the failure rate. In a few cases, a monotonicity correction is required, as detailed in Appendix A.2. For each target risk α, we therefore select $\hat{\lambda}_\alpha$ according to Eq. (4). Calibration is performed independently for each steering direction, with candidate values ordered from weaker to stronger steering toward the target. The selected $\hat{\lambda}_\alpha$ is then fixed and applied to unseen test speakers.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Model. We use the pretrained OPENAUDIO S1-MINI checkpoint of Fish-Speech with 28 Transformer layers and hidden dimension $d = 1024$.

Datasets. We apply ProbeCRC to three controllable speech attributes: emotion, gender, and age.

For emotion steering, we use three emotional speech datasets: the **Emotional Speech Dataset (ESD)**, **CREMA-D**, and **RAVDESS**. We focus on the opposing emotional states *happy* and *sad* and train the layer-wise emotion probes using speakers from all three datasets, while steering is applied to neutral recordings from these speakers. This allows us to evaluate bidirectional steering toward happy and sad from a neutral starting point.

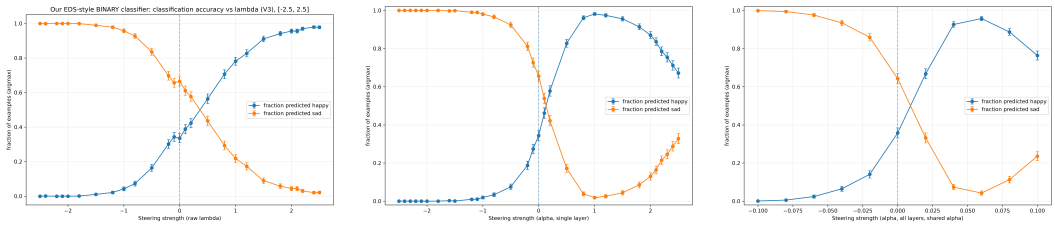
For gender steering, we use the **LibriTTS dataset**. We construct a balanced set of male and female speakers and divide the speakers into two disjoint subsets. The layer-wise gender probes are trained on one half of the speakers, while steering is performed exclusively on the remaining held-out speakers. This speaker-disjoint split allows us to evaluate whether the learned gender directions generalize to speaker identities that were not observed during probe training.

For age steering, we use the **Common Voice dataset**. We train the layer-wise age probes using recordings from two well-separated age groups: speakers labeled as *teens* and speakers aged 70 and above. Steering is then performed on recordings from speakers in their forties, which are not included in either probe-training group. This provides an intermediate-age starting point from which we evaluate steering toward both younger- and older-associated representations.

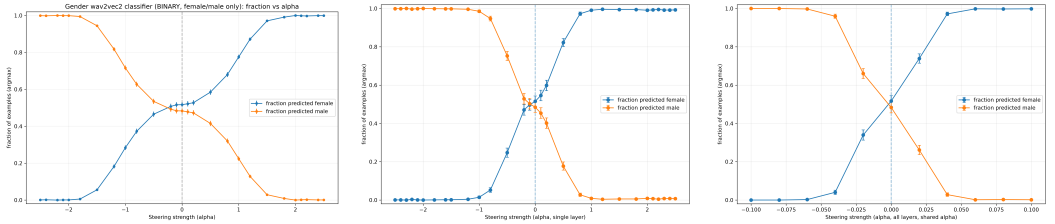
Across all three attributes, the sentences used for probe training are disjoint from those used for steering, ensuring that the learned attribute directions are evaluated on linguistic content not observed during probe training. Further details are provided in Appendix B.1.

Evaluation Metrics. We evaluate attribute control using external speech classifiers for emotion and gender that are independent of the layer-wise probes used for steering. Because reliable automatic age classification from speech is substantially more challenging, we evaluate age steering perceptually through our human listening study rather than using an automatic classifier. Further details and motivation are provided in Appendix B.2. We assess perceptual speech quality using **WV-MOS** and **UTMOSv2**, pretrained speech-quality models that predict perceptual naturalness on the standard 1-5 MOS scale. To assess content preservation, we transcribe each generated utterance using Whisper-medium and compute the word error rate (WER) against the target sentence. For emotion and age, we additionally measure speaker-identity preservation through the cosine similarity between speaker embeddings of the steered and corresponding unsteered ($\lambda = 0$) outputs, extracted using a pretrained WavLM speaker-verification model. Further details are provided in Appendix B.3.

Baselines. We compare ProbeCRC against two complementary classes of baselines. First, we consider conventional difference-of-means activation steering, using both single-layer and multi-layer interventions. Following the formulation and categorization discussed in Sec. 3, the single-layer baseline applies a mean-difference steering direction Yang & Hansen (2026); Xie et al. (2025), while the multi-layer baseline applies layer-specific mean-difference directions across all intervened layers using a shared steering coefficient Wang et al. (2026a;b). Second, we compare against controllable TTS models with dedicated attribute-control mechanisms: IndexTTS2 (Zhou et al., 2026) for emotion and VoxCPM2 (VoxCPM Team, 2026) for gender. Because their control mechanisms differ from our probe-logit steering parameter, we evaluate all methods using the same downstream attribute classifiers rather than directly comparing their numerical control parameters. Full implementation and control-configuration details are provided in Appendix B.4.

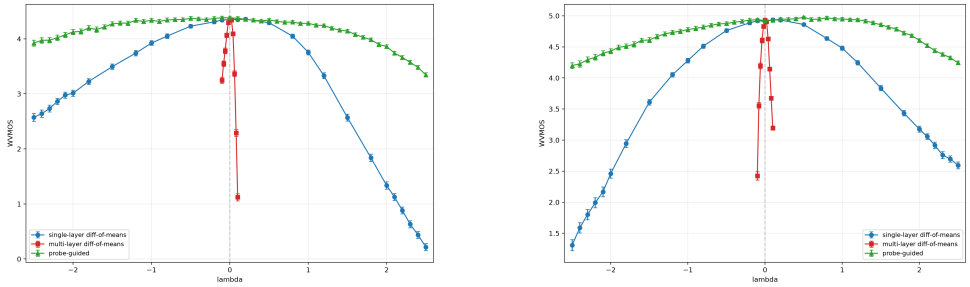


(a) Probe-guided (b) Single-layer (c) Multi-layer



(d) Probe-guided (e) Single-layer (f) Multi-layer

Figure 2: Attribute-classifier response across λ for emotion (top) and gender (bottom).



(a) Emotion (b) Gender

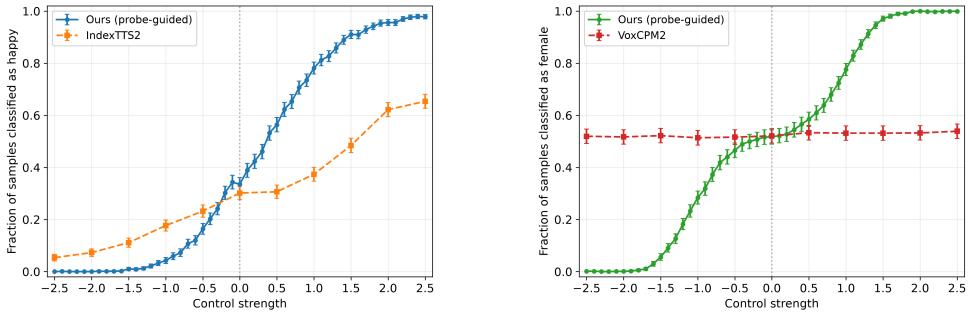
Figure 3: WV-MOS speech quality across steering strengths for probe-guided, single-layer, and multi-layer difference-of-means steering.

6 RESULTS

6.1 STEERING EFFECTIVENESS AND SPEECH QUALITY

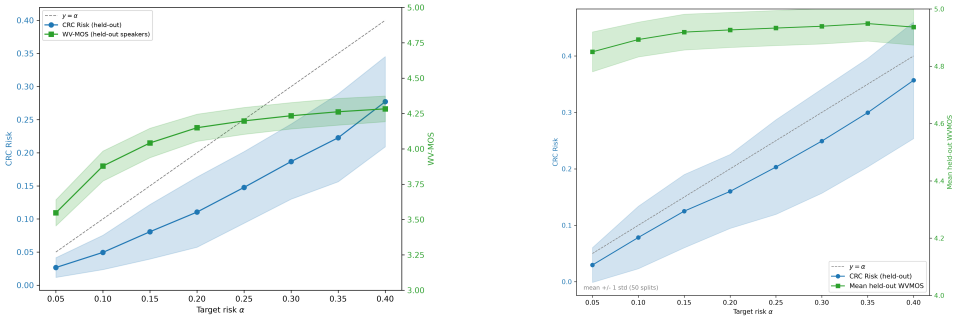
Comparison with Difference-of-Means Steering. We compare probe-guided steering with single- and multi-layer difference-of-means steering in terms of attribute control and speech quality. Figure 2 shows that all three approaches produce strong bidirectional attribute shifts for both emotion and gender. However, Fig. 3 shows that probe-guided steering maintains high predicted speech quality over a substantially wider steering range, whereas quality deteriorates under stronger difference-of-means interventions, particularly in the multi-layer setting. Probe-guided steering therefore achieves strong bidirectional attribute control while preserving substantially higher speech quality across the evaluated steering range, yielding a more favorable control–quality trade-off than the difference-of-means baselines. Additional UTMOSv2 results are provided in Appendix C.1.

Comparison with Controllable TTS Models. We next compare ProbeCRC with the controllable TTS baselines. Figure 4 shows the fraction of outputs classified as the target attribute as control strength is varied. For emotion, both methods respond to increasing control toward happy, but ProbeCRC produces a substantially stronger transition, from near zero under negative steering to nearly one under strong positive steering, whereas IndexTTS2 exhibits a more gradual change. For gender, ProbeCRC similarly produces a clear transition toward female, while VoxCPM2 remains close to balanced across control settings. These results show that, under the same downstream evaluators, probe-guided steering produces a stronger and more complete response to control strength than the evaluated controllable-TTS baselines. Results for the complementary sad and male directions are reported in Appendix C.2.



(a) Emotion: fraction classified as happy (b) Gender: fraction classified as female

Figure 4: Comparison with controllable TTS baselines using the same downstream attribute classifiers. The dashed vertical line indicates the unmodified control setting.



(a) Emotion: steering toward happy (b) Gender: steering toward female

Figure 5: CRC evaluation. Blue curves show empirical steering-failure risk and green curves show WV-MOS. Solid curves show the mean across random splits, shaded regions indicate ± 1 standard deviation, and the dashed line marks the target risk α.

6.2 Risk-Controlled Steering Strength Selection

We next evaluate whether CRC can select steering strengths that achieve a prescribed target risk on unseen speakers while preserving speech quality. We repeat the evaluation over 50 random seeds, each using a different speaker-level split with 80% of speakers for calibration and 20% for testing. The calibration speakers are used to select a global steering strength $\hat{\lambda}$ for each target risk α, and the selected $\hat{\lambda}$ is then evaluated on the test set.

Figure 5 reports the resulting held-out steering-failure risk and WV-MOS for steering toward happy and female. For both attributes, the mean risk remains below the target level across the evaluated range. Speech quality remains high across the CRC-selected operating points: WV-MOS is stable for female steering and improves for happy steering as the target risk is relaxed. Thus, CRC adjusts steering strength to meet different target risk levels without substantially compromising speech quality. Results for the sad and male directions are provided in Appendix C.3.

6.3 Content and Speaker Preservation

For emotion and gender, we evaluate preservation using the steering strength most frequently selected across CRC runs at $\alpha = 0.05$. As shown in Table 1, WER remains low across both attributes, while speaker similarity, where evaluated, remains high. Overall, strong steering largely preserves linguistic content and speaker identity where evaluated.

6.4 Human Evaluation

We conduct a human listening study to assess perceptual attribute changes and the preservation of speech quality and speaker identity. We evaluate steering along the three attributes and compare our probe-guided method against single-layer steering. For age and emotion, we also evaluate unmodified speech exhibiting the target attribute from the original dataset. The study contains four

Table 1: Intelligibility and speaker preservation under strong steering. Speaker similarity is measured relative to the corresponding unsteered ($\lambda = 0$) output. Values are reported as mean $\pm$ SEM.

Attribute	**Target**	**WER (%)**	**Spk. Sim. to $\lambda = 0$**
Emotion	Happy ($\lambda = 2.3$)	2.95 ± 0.29	0.825 ± 0.0026
	Sad ($\lambda = -1.1$)	4.37 ± 0.41	0.932 ± 0.0015
Gender	Female ($\lambda = 1.5$)	2.67 ± 0.23	–
	Male ($\lambda = -1.6$)	3.47 ± 0.39	–

Table 2: Human evaluation of attribute steering. Raw is the mean signed attribute rating (-2 to $+2$), and Dir. is the percentage of judgments in the target direction.

Attribute	**Direction**	**Method**	**Raw**	**MOS**	**Similarity**	**Dir. (%)**
Age	Older	Original	+1.02	3.55	–	73.3
		Probe	+0.82	3.53	2.93	65.0
		Single	+1.58	2.33	1.30	95.0
	Younger	Original	-0.65	3.28	–	55.0
		Probe	-0.70	3.78	3.00	65.0
		Single	-1.55	1.97	1.43	96.7
Emotion	Happy	Original	+0.85	3.95	3.55	65.0
		Probe	+1.18	3.25	2.72	93.3
		Single	+1.55	2.20	1.53	93.3
	Sad	Original	-0.93	3.78	3.62	58.3
		Probe	-0.92	3.65	3.40	78.3
		Single	-1.33	2.90	1.75	91.7
Gender	Feminine	Probe	+1.59	3.74	–	100.0
		Single	+1.52	2.06	–	100.0
	Masculine	Probe	-1.51	4.04	–	97.5
		Single	-1.68	1.99	–	95.0

speech samples for each steering direction and method. Each candidate sample is presented together with its corresponding unsteered reference generated with $\lambda = 0$. Participants rate the perceived attribute change relative to the reference on a five-point scale from -2 (large negative shift) to $+2$ (large positive shift), where positive denotes *happy*, *older*, or *feminine*, and negative denotes *sad*, *younger*, or *masculine*. In addition, participants rate speech quality (MOS) and, where applicable, speaker-identity similarity. We report the mean signed attribute rating (Raw), MOS, similarity, and directional success, defined as the percentage of judgments in the intended direction. Further evaluation details are provided in Appendix C.4.

Table 2 shows that both methods produce clear perceptual shifts in the intended directions, but with markedly different effects on speech quality and speaker identity. Single-layer steering generally induces stronger attribute changes, particularly for age and emotion, but substantially degrades quality and identity similarity. In contrast, probe-guided steering achieves consistently high directional success while better preserving both properties. Its age and emotion shifts are comparable to those of the original dataset samples and are generally more consistently perceived in the intended direction, although this trend is less pronounced for the older direction. The gender results are particularly strong, with near-perfect directional success and the highest quality ratings. Overall, probe-guided steering provides effective attribute control with a substantially better trade-off between steering strength and perceptual preservation than single-layer steering.

7 Conclusion

We introduced **ProbeCRC**, a framework that combines probe-guided activation steering with risk-controlled strength selection for controllable speech generation. ProbeCRC uses learned attribute probes to guide hidden-state interventions toward a desired change in probe-logit space, while Conformal Risk Control selects the steering strength according to a prescribed steering-failure risk. Across emotion, gender, and age, ProbeCRC enables effective bidirectional attribute control while largely preserving speech quality, linguistic content, and speaker identity. Compared with conventional activation steering and controllable TTS baselines, our results demonstrate a favorable trade-off between attribute control and speech preservation, with human evaluation further confirming that the induced changes are perceptually meaningful.

AI USE STATEMENT

In this work, we used generative AI tools to assist with the implementation and debugging of research code, the drafting and editing of portions of the manuscript, and the organization and presentation of experimental results. All AI-assisted code was reviewed, modified where necessary, and tested by the authors. The scientific content and interpretation of the results were determined by the authors, with generative AI used to assist with English translation, linguistic refinement, and improving the clarity and readability of the presentation.

Generative AI tools were also used to assist with literature search and summarization, as well as with aspects of the implementation and presentation of the listening study. References identified with AI assistance were checked against their sources.

The authors take responsibility for the final content of this work, including its text, code, analyses, and conclusions.

ETHICS STATEMENT

Our evaluation includes a human listening study conducted with participants recruited through Prolific. The study protocol was reviewed and approved by the institutional ethics committee. Participants freely chose to take part in the study and were compensated for their participation. Responses were analyzed without identifying individual participants. The study involved evaluation of generated speech samples and did not require participants to provide sensitive personal information.

Our work focuses on controlled manipulation of perceptual speech attributes, specifically emotion, gender, and age. The method was developed and evaluated for controllable speech generation, with the goal of modifying predefined speech attributes while preserving the linguistic content and speaker identity as much as possible. All experiments in this work are restricted to benign attribute-control settings and do not involve steering toward harmful or offensive content.

REFERENCES

Anastasios N Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Deok-Hyeon Cho, Hyung-Seok Oh, Seung-Bin Kim, Sang-Hoon Lee, and Seong-Whan Lee. Emosphere-tts: Emotional style and intensity modeling via spherical emotion vector for controllable emotional text-to-speech. In *Interspeech*, pp. 1810–1814, 2024.

Ziqi Gao, Yaotian Zhu, Qingcheng Zeng, Xu Zhao, Ziqing Wang, Feng Ruan, and Kaize Ding. Raptor: Ridge-adaptive logistic probes. *arXiv preprint arXiv:2602.00158*, 2026.

Zhifang Guo, Yichong Leng, Yuxiang programmes Wu, Xu Tan, Jiawei Chen, Yanqing Liu, Sheng Zhao, and Xiang-Yang Li. Prompttts: Controllable text-to-speech with natural language descriptions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Rohan Narayan Koli. *Classification of speaker's age, gender and nationality using transfer learning*. PhD thesis, Dublin, National College of Ireland, 2021.

Nina Konovalova, Andrey Kuznetsov, and Aibek Alanov. Shift: Steering hidden intermediates in flow transformers. *arXiv preprint arXiv:2604.09213*, 2026.

Nikita Koriagin, Georgii Aparin, Nikita Balagansky, and Daniil Gavrilov. Interpreting and steering a text-to-speech language model with sparse autoencoders. *arXiv preprint arXiv:2606.10029*, 2026.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Kouzelis, Kushal Lakhotia, Michelle Annapureddy, Channing Sarkar, Abdelrahman Hassan, et al. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 37932–37954, 2023.

Kyowoon Lee, Artyom Stitsyuk, Gunu Jho, Inchul Hwang, and Jaesik Choi. Counterfactual activation editing for post-hoc prosody and mispronunciation correction in tts models. In *Proc. Interspeech*, 2025.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Yunpei Li, Xun Zhou, Jinchao Wang, Lu Wang, Yong Wu, Siyi Zhou, Yiquan Zhou, Yining Wang, Yaogen Yang, Zhetao Hu, et al. Indextts 2.5 technical report. *arXiv preprint arXiv:2601.03888*, 2026.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*, 2024.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Multi-attribute steering of language models via targeted intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20619–20634. Association for Computational Linguistics, 2025.

Karran Pandey, Paul Guerrero, Matheus Gadelha, Yannick Hold-Geoffroy, Karan Singh, and Niloy J. Mitra. Diffusion handles: Enabling 3d edits for diffusion models by lifting activations to 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7695–7704, 2024.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of machine learning research*, 9(3), 2008.

Jingran Su, Jingfan Chen, Hongxin Li, Yuntao Chen, Qing Li, and Zhaoxiang Zhang. Activation steering decoding: Mitigating hallucination in large vision-language models through bidirectional hidden state intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12964–12974, 2025.

Xu Tan, Jiawei Chen, Hao Liu, Jian Cong, Chen Zhang, Yanqing Liu, Xi Wang, Yichong Leng, Yuanhao Yi, Lei He, et al. Naturalspeech: End-to-end text-to-speech synthesis with human-level quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(6):3863–3875, 2024.

Fengping Tian et al. Marco-voice technical report. *arXiv preprint arXiv:2508.02038*, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

VoxCPM Team. Voxcpm2 technical report. *arXiv preprint arXiv:2606.06928*, 2026.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhengyang Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text-to-speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Siyi Wang, James Bailey, and Ting Dang. A geometric perspective on composable emotion steering in text-to-speech models. *arXiv preprint arXiv:2607.00946*, 2026a.

Siyi Wang, Shihong Tan, Siyi Liu, Hong Jia, Gongping Huang, James Bailey, and Ting Dang. CoCoEmo: Composable and controllable human-like emotional tts via activation steering. *arXiv preprint arXiv:2602.03420*, 2026b.

Tianxin Xie, Shan Yang, Chenxing Li, Dong Yu, and Li Liu. Emosteer-tts: Fine-grained and training-free emotion-controllable text-to-speech via activation steering. *arXiv preprint arXiv:2508.03543*, 2025.

Dongchao Yang, Songxiang0 Liu, Rongjie Huang, Guangzhi Lei, Chao Weng, Dan Su, and Dong Yu. Instructtts: Modelling expressive tts with text instructions. *arXiv preprint arXiv:2301.13662*, 2023.

Mu Yang and John H. L. Hansen. Activation steering for accent-neutralized zero-shot text-to-speech. *arXiv preprint arXiv:2603.05977*, 2026.

Xie Yunfan, Lixin Zou, Dan Luo, Min Tang, Chenliang Li, Xiangyang Luo, and Liming Dong. Mitigating language confusion through inference-time intervention. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 8418–8431. Association for Computational Linguistics, 2025.

Xu Zhang, Longbing Cao, and Zhangkai Wu. DUET: Unified dual-space emotion control for diffusion and flow-matching driven text-to-speech. *arXiv preprint arXiv:2606.00066*, 2026.

Daniel Zhao, Daniel Beaglehole, Julian McAuley, Taylor Berg-Kirkpatrick, and Zachary Novack. Steering autoregressive music generation with recursive feature machines. In *International Conference on Learning Representations (ICLR)*, 2026.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. 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.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Mathematical Details

A.1 Derivation of the Regularized Steering Intervention

Recall that

$$\phi_\ell(h_\ell) = w_\ell^\top h_\ell + b_\ell, \tag{12}$$

and that the additive intervention is

$$h'_\ell = h_\ell + v_\ell. \tag{13}$$

Therefore,

$$\phi_\ell(h'_\ell) - \phi_\ell(h_\ell) = w_\ell^\top (h_\ell + v_\ell) + b_\ell - \left(w_\ell^\top h_\ell + b_\ell\right) \tag{14}$$

$$= w_\ell^\top v_\ell. \tag{15}$$

Starting from the regularized objective in Eq. (9),

$$v_\ell^* = \arg\min_{v_\ell} \left(w_\ell^\top v_\ell - \lambda\right)^2 + \eta \|v_\ell\|_2^2. \tag{16}$$

Taking the gradient with respect to v_ℓ and setting it to zero gives

$$2w_\ell \left(w_\ell^\top v_\ell - \lambda\right) + 2\eta v_\ell = 0. \tag{17}$$

Rearranging terms yields

$$\left(w_\ell w_\ell^\top + \eta I\right) v_\ell = w_\ell \lambda. \tag{18}$$

Since $\eta > 0$, the system matrix is invertible, and therefore

$$v_\ell^*(\lambda) = \left(w_\ell w_\ell^\top + \eta I\right)^{-1} w_\ell \lambda, \tag{19}$$

which gives Eq. (10).

A.2 Monotonicity of the Steering Loss

CRC requires the loss to vary monotonically with the calibration parameter. In our setting, λ denotes the steering strength, with its sign determining the target direction. Monotonicity is therefore considered separately for each target direction, with candidate values of λ ordered from weaker to stronger steering toward that target. For positive steering directions this corresponds to increasing λ, whereas for negative steering directions it corresponds to increasing the magnitude of negative λ.

Under this direction-specific ordering, stronger steering toward the target is expected to reduce the steering-failure loss. Thus, for two candidate values λ_1 and λ_2, where λ_2 corresponds to stronger steering toward the same target than λ_1, the desired monotonicity condition is

$$L_i(\lambda_1) \geq L_i(\lambda_2). \tag{20}$$

In practice, individual steering-loss trajectories may exhibit small deviations from exact monotonicity. We therefore replace each observed trajectory by a conservative monotone upper envelope. Let $\lambda' \succeq \lambda$ denote candidate steering strengths that are at least as strong as λ toward the same target direction. We define

$$\widetilde{L}_i(\lambda) = \sup_{\lambda' \succeq \lambda} L_i(\lambda'). \tag{21}$$

Conservativeness. Since λ itself belongs to the set over which the supremum is taken,

$$\widetilde{L}_i(\lambda) = \sup_{\lambda' \succeq \lambda} L_i(\lambda') \geq L_i(\lambda). \tag{22}$$

Thus, the transformed loss never underestimates the original loss.

Monotonicity. Suppose that λ_2 corresponds to stronger steering toward the target than λ_1. Then

$$\{\lambda' : \lambda' \succeq \lambda_2\} \subseteq \{\lambda' : \lambda' \succeq \lambda_1\}. \tag{23}$$

The supremum over the larger set cannot be smaller, and therefore

$$\widetilde{L}_i(\lambda_1) \geq \widetilde{L}_i(\lambda_2). \tag{24}$$

Hence, the corrected loss is non-increasing with steering strength toward the target while remaining conservative with respect to the original loss.

Empirical Monotonicity of Steering Loss. We additionally examine how closely the raw steering-loss trajectories satisfy this condition before applying the monotone correction. Across five random speaker-level splits, the raw loss is already highly monotone (Table 3). The average monotonicity rates are approximately 93.5% and 92.1% when steering toward happy and sad, respectively, and exceed 96% in both gender directions. The small variation across splits indicates that this behavior is stable with respect to the speakers used for calibration.

Table 3: Fraction of steering-loss trajectories that are already monotone before correction. The mean and range are computed across five random speaker-level splits.

Target	Mean (%)	Range (%)
Happy	93.55	93.42–93.74
Sad	92.05	91.43–92.35
Female	96.32	96.10–96.70
Male	96.44	96.00–96.80

B Additional Experimental Details

B.1 Dataset and Probe Training Details

For emotion, the probe data comprise happy and sad recordings from ESD[2], CREMA-D[3], and RAVDESS[4]. ESD contributes recordings from 10 speakers with 350 utterances per emotion. CREMA-D contributes recordings from 91 speakers with one utterance per speaker and emotion, and RAVDESS contributes recordings from 24 speakers with two utterances per speaker and emotion. The external binary emotion classifier, described in Appendix B.2, is trained separately using 100 speakers for training and 25 for validation.

For gender, the LibriTTS[5] speakers are first divided equally between probe training and steering evaluation using a gender-stratified speaker-level split that preserves the original female-to-male ratio within each half.

For age, the probe data are drawn from Common Voice[6] and contain 3,742 utterances, balanced between teenagers and speakers aged 70+ (1,871 per class).

For all three attributes, the data used for probe training are further divided into 80% training and 20% validation subsets.

Probe Training. The layer-wise linear probes are trained for 200 epochs using full-batch optimization and binary cross-entropy with logits. We use Adam with a learning rate of 10^{-3} and weight decay of 10^{-4}.

B.2 Attribute Classifiers

Emotion Classifier. For emotion evaluation, we train a binary happy/sad classifier starting from the pretrained wav2vec 2.0 encoder of SUPERB/WAV2VEC2-BASE-SUPERB-ER[7]. We attach a freshly initialized classification head and train the model in two stages: one warmup epoch with the encoder frozen, followed by full fine-tuning of the entire network for 10 epochs. Training uses ground-truth happy and sad recordings pooled from CREMA-D, ESD, and RAVDESS.

The classifier is trained independently of the layer-wise probes.

Gender Classifier. For gender evaluation, we use AUDEERING/WAV2VEC2-LARGE-ROBUST-24-FT-AGE-GENDER[8], an off-the-shelf wav2vec 2.0-large-robust model fine-tuned for age and gender prediction. Its training data do not include LibriTTS, which is the dataset used for our gender-steering experiments. The classifier therefore serves as an external evaluator of gender expression. We report the fraction of generated utterances classified as female as our primary gender-steering metric.

[2] https://www.kaggle.com/datasets/nguyenthanhlim/emotional-speech-dataset-esd
[3] https://www.kaggle.com/datasets/ejlok1/cremad
[4] https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio
[5] https://www.openslr.org/60/
[6] https://github.com/common-voice/common-voice
[7] https://huggingface.co/superb/wav2vec2-base-superb-er
[8] https://huggingface.co/audeering/wav2vec2-large-robust-24-ft-age-gender

Age Evaluation. Age classification from speech is particularly challenging due to high inter-speaker variability, recording conditions, and the weak correspondence between chronological age and acoustic characteristics. Prior work reports substantially lower performance for age estimation than for attributes such as gender, with age-group classification accuracies around 52% Koli (2021). We therefore do not rely on an automatic age classifier to evaluate steering effectiveness. Instead, we assess the perceptual effect of age steering directly through the human listening study described in Section 6.4, where participants judge whether the steered speech sounds older or younger relative to the unsteered reference.

B.3 Speech Quality, Content Preservation, and Speaker Similarity Metrics

Speech quality. We evaluate perceptual speech quality using **WV-MOS**[9] and **UTMOSv2**[10], pre-trained speech-quality assessment models that predict perceptual naturalness on the standard 1-5 MOS scale.

Word error rate (WER). Each generated WAV is transcribed with Whisper-medium[11] (greedy decoding, English) and compared against the target sentence used for generation. Reference and hypothesis text are lowercased and stripped of punctuation before alignment; WER is the standard word-level edit-distance ratio: (substitutions + deletions + insertions) / (reference length).

Speaker similarity. For emotion steering, we additionally test whether steering shifts speaker identity. For each example, we embed (i) the steered output at the evaluated λ and (ii) the corresponding unsteered ($\lambda = 0$) output of the same speaker and sentence with a WavLM[12] x-vector speaker-verification model and report the cosine similarity between the two embeddings. A value near 1 indicates that the steered output remains recognizable as the same speaker as the unsteered baseline; lower values indicate steering has altered perceived speaker identity.

B.4 Baseline Configurations

Difference-of-Means Steering. We evaluate difference-of-means activation steering in both single- and multi-layer settings. For the single-layer baseline, we apply the intervention at the layer with the highest probe validation accuracy for the corresponding attribute. For the multi-layer baseline, the mean-difference intervention is applied across all layers using a shared steering coefficient.

IndexTTS2. IndexTTS2 provides an explicit emotion-conditioning mechanism based on an emotion vector together with a scalar intensity weight $\alpha \in [0, 1]$. To evaluate it over the same signed control range used for our method, we use the sign of λ to specify the target emotion and its magnitude to specify the conditioning intensity. Positive values target happy, whereas negative values target sad. The intensity is defined as

$$\alpha = \min\left(1, \frac{|\lambda|}{\lambda_{\max}}\right), \tag{25}$$

where $\lambda_{\max}$ is the largest steering magnitude in the evaluated range. Thus, $\lambda = 0$ corresponds to no explicit emotion conditioning, while the largest positive and negative magnitudes correspond to maximum happy and sad conditioning, respectively.

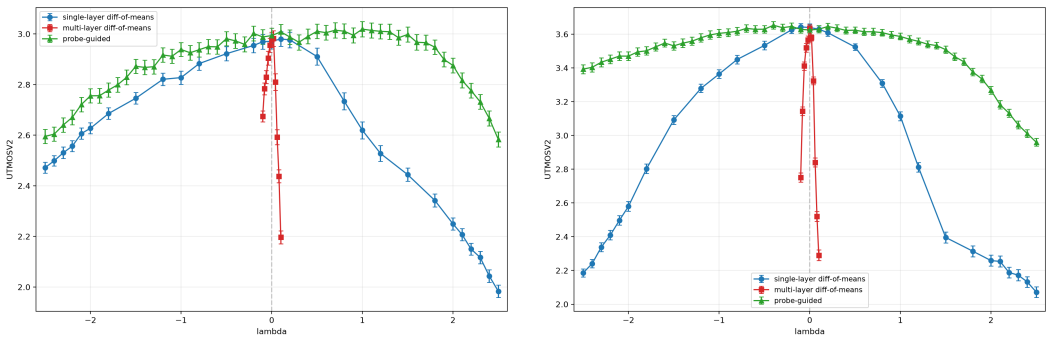
VoxCPM2. VoxCPM2 does not expose an analogous continuous numerical parameter for gender control. We therefore condition gender through natural-language instructions prepended to the input text while retaining the reference audio used for speaker conditioning. We construct a fixed sequence of instructions representing increasing masculine or feminine control strength. Positive values of λ correspond to increasingly feminine instructions, negative values correspond to increasingly masculine instructions, and $\lambda = 0$ uses no gender instruction.

[9] https://github.com/AndreevP/WV-MOS

[10] https://github.com/sarulab-speech/UTMOSv2

[11] https://github.com/openai/whisper

[12] https://huggingface.co/microsoft/wavlm-base-plus-sv



(a) Emotion (b) Gender

Figure 6: UTMOSv2 speech-quality scores under probe-guided and difference-of-means steering for emotion and gender as a function of intervention strength.

Because the IndexTTS2 intensity parameter, the VoxCPM2 text instructions, and our probe-logit steering parameter have different semantics and scales, their numerical control values should not be interpreted as directly comparable. The comparison instead evaluates how the generated speech responds to increasing control strength under the same downstream attribute classifiers.

B.5 CRC Evaluation Protocol

Evaluation Protocol. We evaluate CRC across 50 random seeds. For each seed, speakers are randomly partitioned at the speaker level into an 80% calibration set and a 20% held-out test set, with no speaker appearing in both sets within the same split.

For each prescribed target risk α, we use only the calibration speakers to select the global steering strength $\hat{\lambda}$. Once selected, $\hat{\lambda}$ is fixed and applied to the corresponding held-out speakers. The steering-failure risk and WV-MOS are then computed exclusively on these held-out speakers. The procedure is repeated independently for each of the 50 random speaker-level splits and for each steering direction.

For emotion, steering is applied to neutral speech, and calibration is performed independently for the happy and sad directions. Gender calibration is likewise performed independently for the female and male directions. Candidate steering strengths are restricted to $|\lambda| \leq 2.5$.

The reported curves aggregate the resulting held-out evaluations across all 50 random splits. Solid curves indicate the mean across splits, and the shaded regions indicate ± 1 standard deviation. Thus, the reported variation reflects sensitivity to the particular speakers assigned to the calibration and held-out sets.

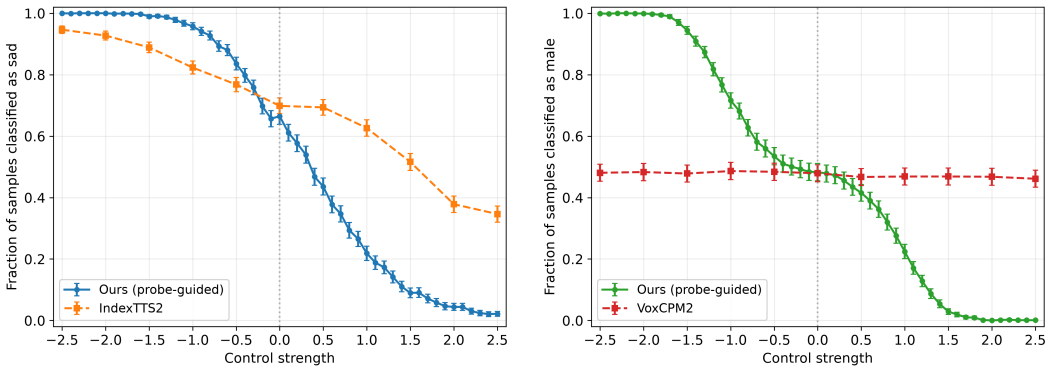
C Additional Results

C.1 Additional Speech Quality Evaluation

We additionally evaluate speech quality using UTMOSv2. Figure 6 reports the results for emotion and gender across the evaluated steering ranges. Consistent with the WV-MOS results, probe-guided steering maintains higher predicted speech quality under stronger interventions, whereas the difference-of-means baselines, particularly the multi-layer variant, result in lower speech quality.

C.2 Additional Controllable-TTS Results

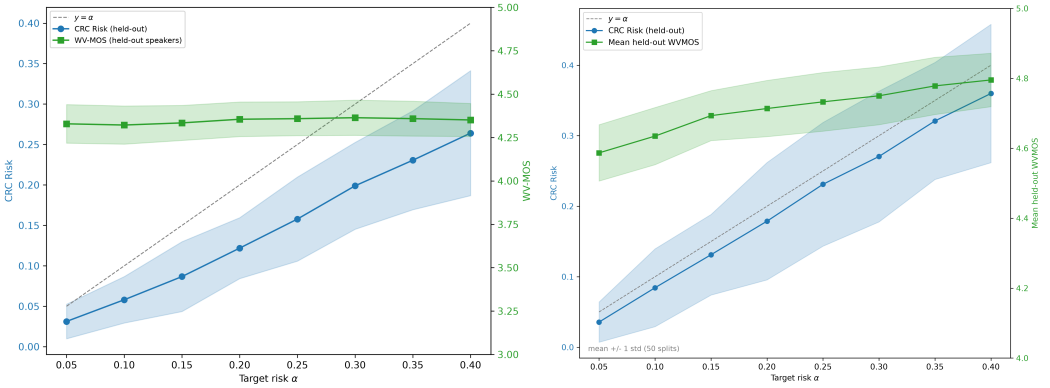
Opposite Target Directions. Figure 7 reports the complementary target directions omitted from the main paper. For emotion, both methods respond to increasing control toward sad, but our method exhibits a substantially stronger transition. The fraction of our outputs classified as sad changes from nearly one under strong negative steering to near zero under strong positive steering, whereas IndexTTS2 shows a more gradual change across the evaluated control range. For gender, our method



(a) Emotion: fraction classified as sad

(b) Gender: fraction classified as male

Figure 7: Comparison with controllable TTS baselines for the complementary steering directions using the same downstream attribute classifiers. The dashed vertical line indicates the unmodified control setting.



(a) Emotion: steering toward sad

(b) Gender: steering toward male

Figure 8: Additional CRC evaluation on held-out speakers. Blue curves show empirical steering-failure risk and green curves show WV-MOS. Solid curves indicate the mean across random splits, with shaded regions denoting ± 1 standard deviation; the dashed line denotes the prescribed target risk $y = \alpha$.

produces a clear transition in the fraction classified as male, while VoxCPM2 remains close to balanced across its control settings.

C.3 Additional CRC Results

Opposite Steering Directions. Figure 8 reports the complementary results for steering toward sad and male. As in the happy and female directions reported in Fig. 5, the mean held-out steering-failure risk remains below the prescribed target risk line across the evaluated range. The empirical risk increases as the target risk is relaxed, indicating that the CRC-selected operating point adapts to the requested risk level.

Predicted speech quality also remains high across these operating points. WV-MOS varies only slightly across target risks for sad steering and remains high for male steering while the held-out risk changes with the prescribed target risk.

Together with Fig. 5, these results show consistent behavior across both directions of emotion and gender steering. Across 50 independently sampled speaker-level calibration/test splits, the CRC-selected intervention adapts to the prescribed target risk while maintaining high predicted speech quality on held-out speakers.

C.4 HUMAN EVALUATION PROTOCOL

Listening-Study Protocol For each attribute direction and evaluated method, we use four speech samples. Each candidate is presented together with its corresponding unsteered ($\lambda = 0$) reference, and participants are asked to compare the perceived target attribute between the two samples.

Attribute judgments are collected on a five-point comparative scale,

$$\{-2, -1, 0, +1, +2\},$$

where 0 indicates no perceived change, and increasing magnitude indicates a stronger perceived difference. The sign convention is fixed across attributes:

Attribute	Negative	Positive
Emotion	Sad	Happy
Age	Younger	Older
Gender	Masculine	Feminine

Thus, for example, a rating of $+2$ for emotion indicates that the candidate sounds substantially happier than the reference, whereas a rating of -2 indicates that it sounds substantially sadder.

Participants additionally rate speech quality using a mean opinion score (MOS). For emotion and age steering, they also rate speaker-identity similarity between the candidate and its reference. Speaker similarity is not collected for gender steering.

Steering Strength Selection For emotion and gender, the evaluated steering strength is determined from the CRC calibration runs. We use the majority-selected λ, i.e., the value selected most frequently across CRC runs at $\alpha = 0.05$. For steering toward sadness, we use a stricter value of α due to the stronger baseline tendency of the samples toward the sad class, as illustrated in Fig. 2. For age, we instead evaluate the strongest steering strengths considered, using $\lambda = -2.5$ toward older speech and $\lambda = 2.5$ toward younger speech.

Natural-Speech Baselines We include a natural-speech (*Original*) condition whose construction differs slightly between emotion and age.

For emotion, the dataset contains recordings from the same speakers in both neutral and target emotional states. The neutral recording therefore serves as the reference, while a natural recording of the same speaker expressing the target emotion (happy or sad) serves as the candidate. This provides a natural reference for the magnitude of an actual emotional change while preserving speaker identity.

For age, recordings of the same speaker at different ages are naturally unavailable. We therefore use speakers in their forties as references and compare them with naturally occurring speakers from the corresponding target age group, while matching the gender of the reference speaker: teenagers for the younger direction and speakers aged 70+ for the older direction. Since the reference and candidate are different speakers in this condition, speaker similarity is not reported for the Original age baseline.

No Original baseline is included for gender.

Participants The human evaluation included 15 participants. For each stimulus, participants provided an attribute-comparison response, together with MOS and, where applicable, speaker-similarity ratings.

An attribute response is considered *opposite-direction* if its sign is inconsistent with the predefined target direction of the corresponding stimulus. For example, a positive response to a stimulus targeting sadness is considered opposite-direction, whereas a negative response is in the intended direction. A response of zero ("about the same") is considered neither intended-direction nor opposite-direction.

Reported Metrics For each attribute direction and method, results are aggregated across the four evaluated speech samples and the 15 participant judgments per sample.

We report the following metrics:

- **Raw**: the mean signed attribute judgment on the $[-2, +2]$ scale. Positive values indicate movement toward happy, older, or feminine speech, whereas negative values indicate movement toward sad, younger, or masculine speech.
- **MOS**: the mean perceptual speech-quality rating.
- **Similarity**: the mean speaker-identity similarity rating, reported for emotion and for the steered age conditions.
- **Directional success**: the percentage of all attribute judgments whose sign agrees with the intended target direction. Zero-valued responses are counted as unsuccessful.

Each reported result therefore contains up to $4 \times 15 = 60$ individual judgments. Directional success captures how consistently listeners perceive the intended direction of change, while the Raw score additionally captures the perceived magnitude of that change.