



Audio Demo

Motivation

- LLM-based TTS improves speech expressiveness, making controllable emotional expression increasingly important.
- Existing label- and reference-based methods rely on external conditioning, offering limited interpretability and flexibility once input conditions are fixed.
- Dense activation steering enables training-free emotion control, but captures emotion as a single global representation shift rather than separable feature-level variation.
- We use SAEs to decompose emotion-related variation in the TTS semantic backbone into sparse latent features for interpretable bidirectional control.

Conceptual View

Same text content and speaker identity held fixed

Input speech



SAE latent intervention
selected sparse latents

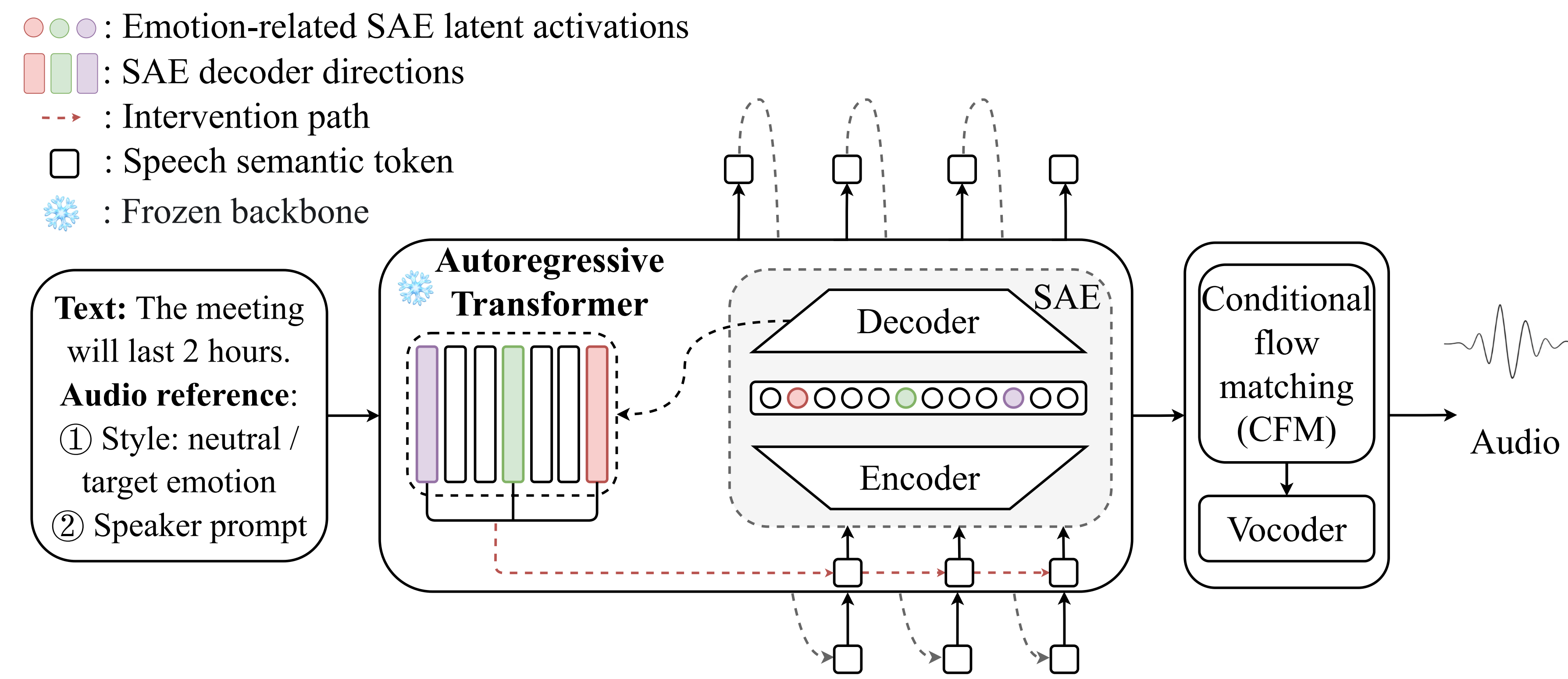
Output speech

Target emotion Neutral

Sparse Autoencoders

- Represents each hidden state with only a small number of active latent features.
- Transforms dense, entangled activations into sparse features that are easier to interpret.
- Provides a feature-level interface for identifying and steering emotion-related variation in the TTS semantic backbone.

Approach Overview



1. Identify emotion-related latents

Rank SAE features by sentence-level emotion selectivity: $\Delta_i^{(e)} = \mathbb{E}_u[\mathbf{1}_i^{(e)}(u) - \mathbf{1}_i^{(\text{neutral})}(u)]$
 Select the top-m emotion-related latents; main experiments use $m = 6$.

2. Feature-level intervention

For selected emotion latents $j \in \mathcal{F}_e$, shift their activations: $a_j^{\text{new}}(x_{l,t}) = a_j(x_{l,t}) + \alpha_e$
 Positive α_e induces the target emotion; negative α_e suppresses it.

Experiment Setup

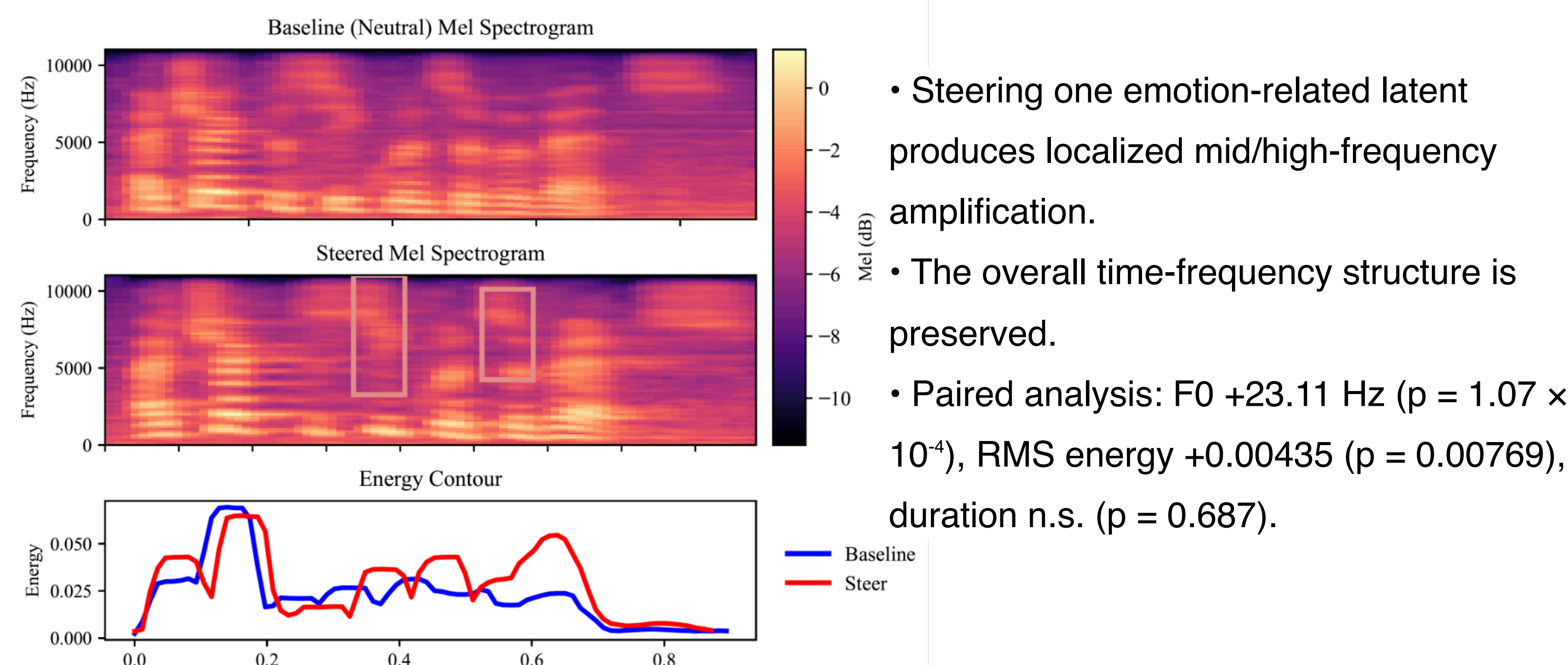
SAE training and controlled data

- Top-k SAE trained on layer-16 semantic-backbone residual activations
- 56,000 emotion-controlled TTS generations
- 4,096 SAE latents; $k = 32$ active features per token
- 400 texts \times 7 emotions \times 20 speaker-timbre references

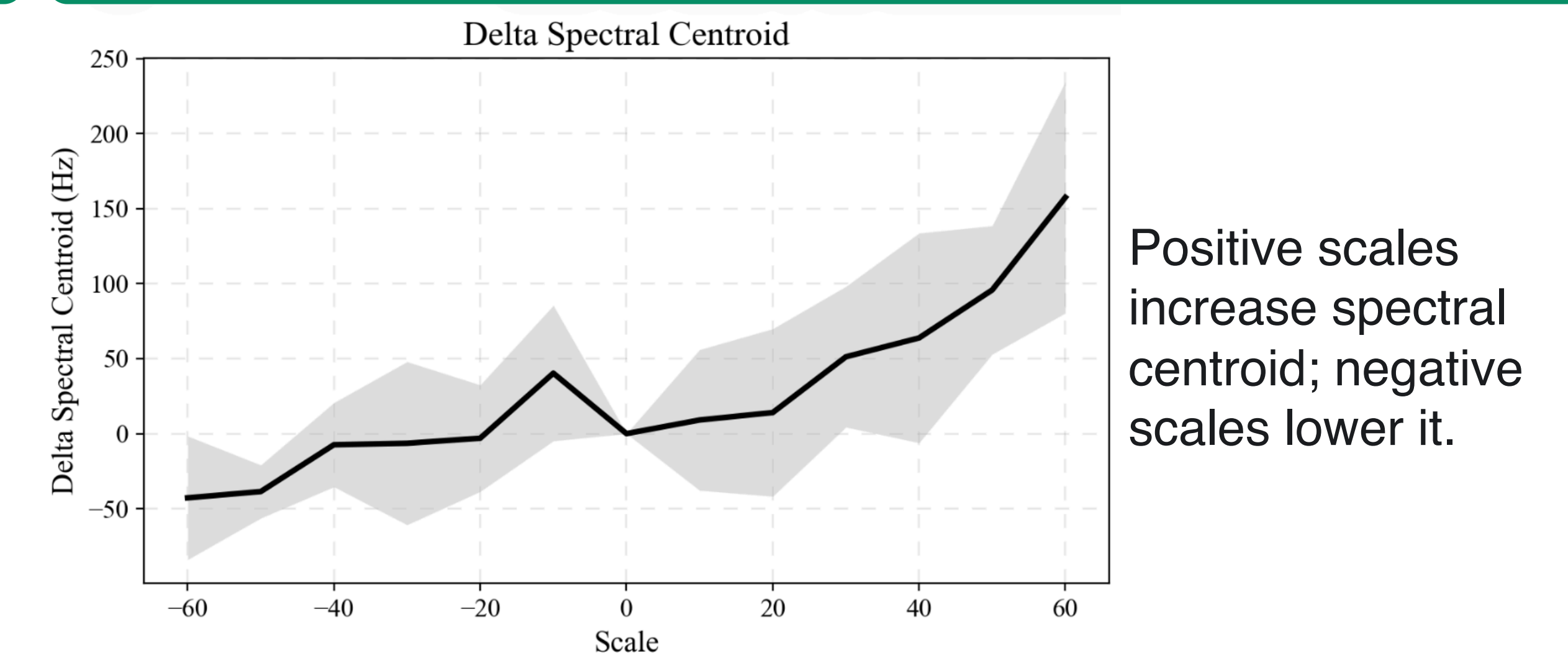
Controlled steering evaluation

- Matched text-speaker pairs; only emotional style varies
- Selectivity analysis: 43,408 matched generations
- Steering: 100 matched cases per target emotion
- Targets: anger, happiness, sadness; induction and suppression
- Metrics: Emo-SIM, WER, Spk-SIM, human EMOS/NMOS

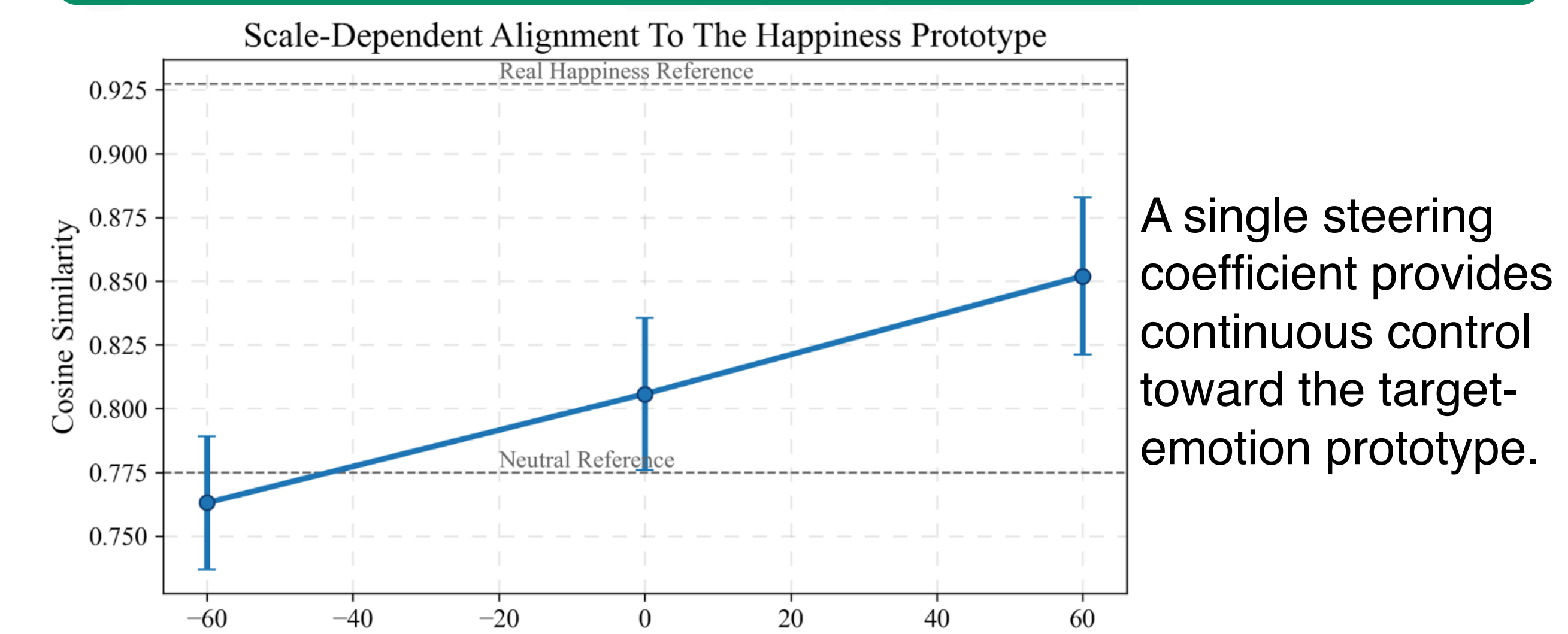
Pitch/Energy Validation



Brightness Validation



Continuous Intensity Control



Bidirectional Steering

Emotion induction: Neutral \rightarrow Target

Method	Anger			Happiness			Sadness		
	Emo \uparrow	WER \downarrow	Spk \uparrow	Emo \uparrow	WER \downarrow	Spk \uparrow	Emo \uparrow	WER \downarrow	Spk \uparrow
VALL-E-X	0.831	3.1	0.302	0.697	5.3	0.320	0.869	7.8	0.352
Spark-TTS	0.857	2.7	0.488	0.770	8.6	0.463	0.907	2.3	0.523
EmoVoice	0.806	4.1	0.358	0.728	3.4	0.342	0.850	4.0	0.386
CosyVoice	0.813	3.9	<u>0.569</u>	0.712	<u>2.9</u>	0.597	0.799	2.4	<u>0.605</u>
Random SAE	0.892	1.4	0.628	0.813	6.0	0.461	0.858	<u>1.7</u>	0.637
Global	<u>0.910</u>	0.1	0.552	0.879	4.0	0.495	0.876	1.9	0.516
SAE-Emotion	0.912	0.3	<u>0.569</u>	0.885	2.2	<u>0.515</u>	<u>0.880</u>	1.5	0.481

Emotion Suppression: Target \rightarrow Neutral

Method	Anger			Happiness			Sadness		
	Emo \uparrow	WER \downarrow	Spk \uparrow	Emo \uparrow	WER \downarrow	Spk \uparrow	Emo \uparrow	WER \downarrow	Spk \uparrow
Random SAE	0.841	0.8	0.342	0.886	2.14	0.343	<u>0.939</u>	0.77	0.427
Global	<u>0.915</u>	<u>2.6</u>	0.392	<u>0.920</u>	1.48	0.379	0.933	1.63	<u>0.436</u>
SAE-Emotion	0.939	2.8	<u>0.374</u>	0.924	2.31	0.301	0.941	<u>0.80</u>	0.441

Perceptual Quality

Method	EMOS	NMOS
SAE-Emotion	3.22	3.49
Global Steering	3.10	3.38
Random SAE	1.82	3.22

Randomized blind listening study with 20 raters; 0–5 scale,

Takeaway

- Emotional variation in the TTS semantic backbone is distributed across multiple sparse SAE latents.
- A small subset of emotion-related latents shows distinct activation and acoustic patterns under controlled conditions.
- Selectivity-ranked latent intervention enables bidirectional emotion control without backbone updates, while largely preserving content, naturalness, and speaker identity.