EMOTIONAL ENHANCEMENT OF CLONED VOICES VIA SPEAKER-EMBEDDING EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Native emotion instructions guide speech planning, but do not directly specify how the acoustic condition derived from a reference voice should change. We introduce a lightweight, reference-conditioned speaker-embedding editor for enhancing emotional expression in frozen text-to-speech synthesis. Trained on paired synthetic embeddings, the editor predicts a residual from the reference embedding and requested emotion, allowing both direction and magnitude to adapt to the reference rather than applying a shared shift. A scalar controls editing strength, and the edited condition enters CosyVoice 3 through its native acoustic interface without updating the synthesizer or requiring target-emotion recordings at inference. By holding speech tokens, acoustic prompts and decoding noise fixed, the method enables acoustic realization to be adjusted independently of changes to the speech plan. Experiments across Chinese and English, in both zero-shot and native emotional-instruction settings, show approximately 6.3–18.2% relative gains in mean target-emotion cosine. Bilingual listening evaluation further supports improved emotion appropriateness and conditional naturalness non-inferiority, alongside reduced speaker similarity at higher editing strength. These findings demonstrate a complementary approach to emotional enhancement that combines reference-dependent editing with adjustable control through a pretrained acoustic interface.

1 INTRODUCTION

Native emotional instructions can guide a cloned voice, but the resulting speech plan is not the only input to its acoustic realization. In CosyVoice 3 (CV3), the language model plans speech tokens while a reference-derived CAM++ vector conditions the Flow-based acoustic model (Du et al., 2025; Wang et al., 2023). We ask whether a learned edit of this second input can strengthen the requested expression while preserving the native plan, with no synthesizer retraining or target-emotion recording at inference.

The distinction has a vocal-production motivation: expressive delivery is organized through prosody and realized through acoustics that also carry speaker traits and phonation cues (Scherer, 2003). A reference-derived speaker condition therefore need not be emotion invariant. CV3 exposes planning and acoustic conditioning as separate computational inputs. Injecting control into token generation may change the plan itself; editing the second input allows us to retain that plan and study its realization separately. This is separation of control pathways, not statistical disentanglement of identity and emotion. We test additional control rather than assume that native instructions fail.

A shared emotion direction and a reference-conditioned edit pose different prediction problems. EmoKnob applies an emotion-level direction to a reference embedding at a user-selected strength (Chen et al., 2024); for a fixed direction and strength, the added increment is the same across references. Our network instead uses the reference to predict the increment itself, with paired target embeddings supervising both magnitude and direction. This allows reference-dependent changes rather than requiring a common shift. The shared and norm-matched arms test the roles of reference-conditioned prediction and residual scaling within the same synthesis pipeline.

The main editor operates on 192-dimensional embeddings. We train it on synthetic neutral/target pairs and apply it to real neutral references without target-domain fine-tuning. New references reuse the fitted editor, and training requires no backpropagation through CV3. The study tests whether

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.

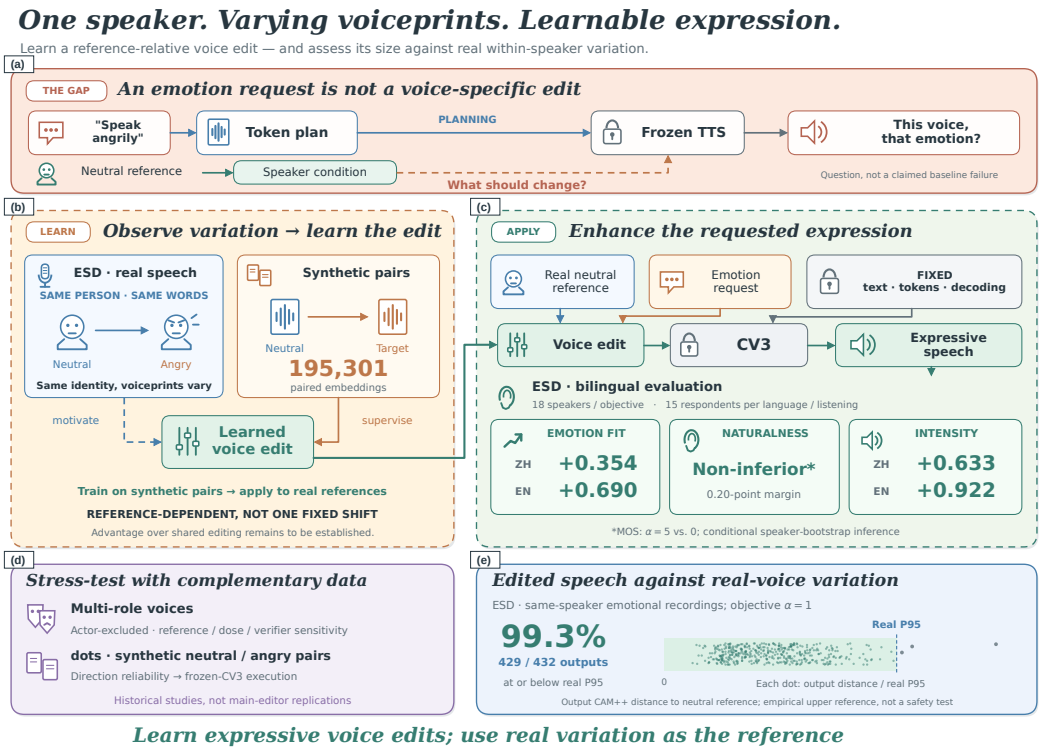


Figure 1: **Why edit the speaker condition, and how is the edit learned?** (a) Emotion requests guide planning; reference-conditioned editing adds control of acoustic realization. (b) Real within-speaker variation motivates editing, while separate synthetic neutral/target pairs supervise the residual predictor. (c) The learned edit enters frozen CV3 under matched text, tokens and decoding; high-dose listening results are detailed in Tables 3 and 7. (d) Historical direction studies provide complementary evidence, not replications of the main editor. (e) The $\alpha = 1$ distance diagnostic is evaluated alongside emotion response on the same outputs in Table 1; its empirical P95 is not a high-dose identity guarantee. Figure 2 details the implementation.

paired endpoint prediction translates into audible enhancement while retaining the pretrained conditioning interface. A separate calibrated ESD extension uses different inputs and is reported independently.

Contributions. **Reference-conditioned endpoint learning:** paired supervision learns a residual instead of imposing a shared edit across references. **Architecture-aligned execution:** the editor uses CV3's existing acoustic condition while retaining its planner, decoder and native instruction interface. **Prediction-to-speech evidence:** held-out embedding prediction and matched synthesis are complemented by bilingual ratings that separate emotion appropriateness, intensity, naturalness and speaker similarity. Speaker-balanced analysis and real emotional references characterize gains and costs; magnitude controls delimit the direction-specific interpretation.

2 Related Work

Emotion and style control. Global Style Tokens learn expressive representations jointly with synthesis (Wang et al., 2018); PromptTTS and InstructTTS use natural-language descriptions (Guo et al., 2023; Yang et al., 2024), and EmoSteer-TTS controls synthesis through activations (Xie et al., 2025). Here the native planner and acoustic model remain frozen, and emotional instructions are retained in the instruction-mode baseline. Our comparison measures the additional response to editing the acoustic condition, not a replacement of instruction following.

Reference dependence and intensity evaluation. EmoKnob estimates shared directions from normalized paired differences and evaluates them with MetaVoice-1B (Chen et al., 2024). It includes an emotion-strength listening test and sample-count/strength sweeps of WER and speaker similarity. Its direction-construction rule does not learn a target-reference-conditioned increment, and its binary-choice listening endpoints do not report separate MOS for naturalness, speaker similarity, emotion appropriateness and intensity. Our learned four-emotion editor uses FiLM conditioning (Perez et al., 2018), while our evaluation separates those four rated dimensions at two edit strengths under native emotional instructions. This distinguishes reference-conditioned prediction and multidimensional dose evaluation from shared-direction extraction and binary emotion judgments. Appendix L.5 details the complementary scope and statistical questions.

Related concurrent work. An unpublished manuscript studies coordinated planning and native-condition transport (Anonymous, 2026). The present study isolates fixed-plan editing with a frozen planner and Flow; Appendix L.6 states the shared principles and evidence boundaries.

3 Reference-Conditioned Speaker-Embedding Editing

3.1 Inputs and frozen synthesis interface

The inference inputs are a neutral reference recording, synthesis text, target emotion and scalar editing strength. Let $e = \mathcal{N}(E(w_{\mathrm{ref}})) \in \mathbb{R}^{192}$ be the normalized CAM++ reference embedding, with $\mathcal{N}(z) = z/\|z\|_2$. Let H denote the speech-token plan produced by the native CV3 interface, and let p, ϵ contain the acoustic prompt and decoding state. The reference-conditioned editor predicts a residual $f_\psi(e, a)$ for emotion a. We average the raw predictions of three independently fitted editors:

$$\hat{r}(e, a) = \frac{1}{3}\sum_{k=1}^{3} f_{\psi_k}(e, a), \qquad U_\alpha = \mathcal{N}(e + \alpha\hat{r}), \qquad \hat{w} = G(H, U_\alpha; p, \epsilon). \tag{1}$$

The original reference bypasses the network and is added to the scaled residual before normalization. Setting $\alpha = 0$ recovers the matched unedited condition. The planner, Flow model and vocoder are frozen; the editor changes only the supplied speaker condition. No target-derived DSP or activation request enters the main model.

Causal organization and native-interface placement. Figure 2 maps a conceptual account of vocal production onto the synthesis interface: expressive planning and speaker-dependent acoustic realization both contribute to observed speech (Scherer, 2003). In CV3, the 192-D CAM++ vector conditions Flow rather than the token LM (Du et al., 2025; Wang et al., 2023). Editing this input preserves a cached plan and reuses the pretrained conditioning path. This computational separation motivates the intervention site; output quality and identity remain empirical questions.

Our comparison assigns $\mathrm{do}(U = U_\alpha)$ with H, p, ϵ fixed. It isolates an implemented realization-side control, not identity and emotion as independent representations. We neither intervene on the token plan nor identify a natural human-emotion effect. Context, communicative function, prosodic grammar and speaker traits are conceptual variables, not extra network inputs. Appendix L formalizes the control distinction, reuse boundary and relationship to the concurrent manuscript.

3.2 Residual supervision and network

For synthetic pair i, let x_i and y_i be normalized neutral and target embeddings. Supervision uses the raw residual $r_i = y_i - x_i$, with training endpoint $\hat{y}_i = \mathcal{N}(x_i + f_\psi(x_i, a_i))$. The loss combines endpoint cosine and residual error:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \omega_i \left[1 - \hat{y}_i^\top y_i + \frac{0.1}{192\rho}\|f_\psi(x_i, a_i) - r_i\|_2^2\right]. \tag{2}$$

Here ρ is the training-set mean squared residual coordinate, floored at 10^{-6}, and ω_i balances source-key/emotion cells and is normalized to unit training mean. Unlike averaging unit difference directions, this supervision retains pairwise displacement length: for unit endpoints, $\|y_i - x_i\|_2^2 =$

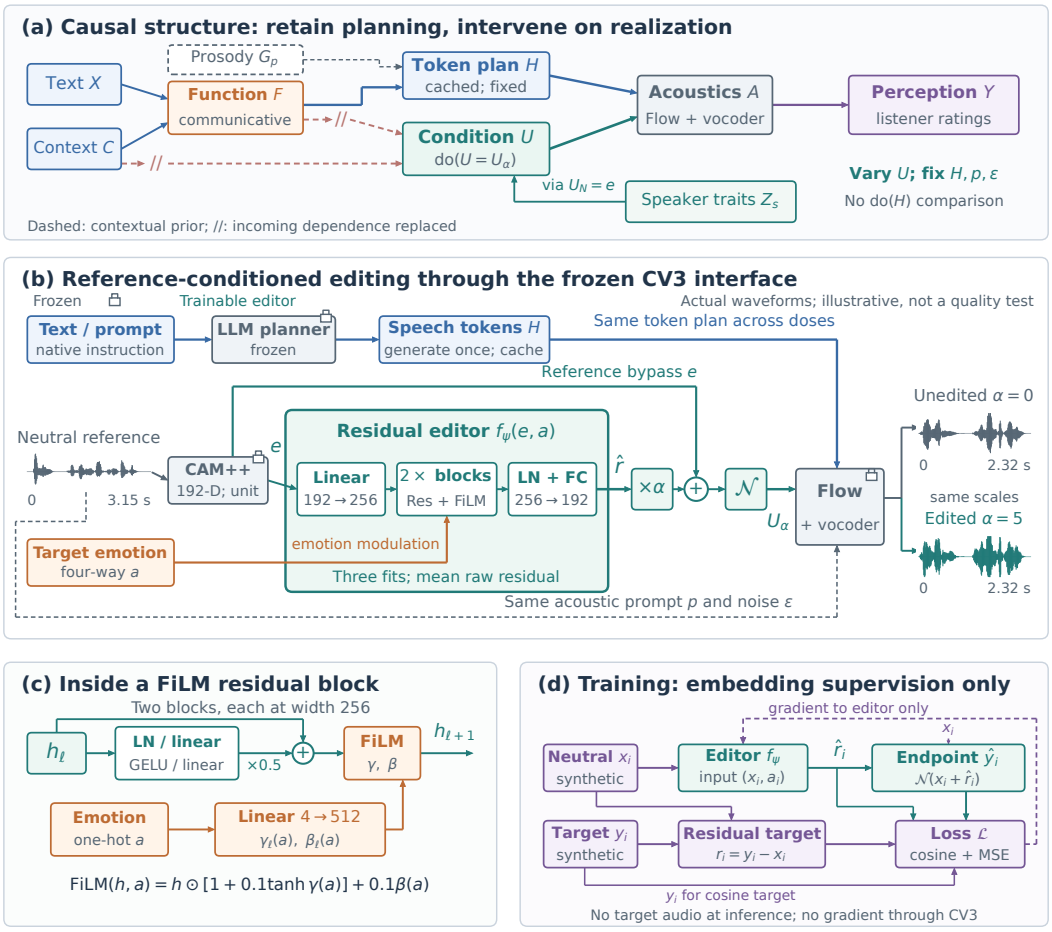


Figure 2: **Causal organization and independent control of acoustic realization.** (a) The conceptual graph separates planning from realization; the implemented comparison fixes H, p, ϵ and assigns only U. Crossed links replace the baseline condition. This is neither an identified human causal model nor identity–emotion disentanglement. (b) Reference embedding, emotion, predicted residual, bypass, strength and normalization produce U_α for frozen CV3. The planner and Flow are unchanged. Unedited/edited traces share time and amplitude scales; the different-content reference is separate. They illustrate execution, not perceptual superiority. (c) Two width-256 FiLM residual blocks predict the edit. (d) Paired embeddings supervise only the editor; no gradient passes through CV3 and no target-emotion recording is supplied at inference.

$2(1 - x_i^\top y_i)$. It learns where a reference-conditioned endpoint should lie, not only which shared direction to apply. Length is not a perceptual-intensity label. The synthetic endpoints did not share fixed tokens and randomness, so they supervise prediction rather than identify a pure causal emotion effect.

A linear projection maps the reference embedding to width 256, followed by two residual blocks and a 192-dimensional output projection. A four-way emotion indicator supplies the FiLM scale and shift. Each block uses a half-scaled residual update, followed by $h \leftarrow h \odot [1 + 0.1 \tanh \gamma(a)] + 0.1\beta(a)$. We train three seeds with AdamW, learning rate 3×10^{-4}, weight decay 0.01, at most 60 epochs and patience 10. Validation source-key macro cosine selects checkpoints. The text-driven evaluation tests the shift from synthetic neutral embeddings during training to real neutral-reference embeddings at inference.

3.3 SHARED AND MAGNITUDE-MATCHED CONTROLS

The shared comparator averages raw target-minus-neutral residuals within each training source-key/emotion cell, then averages source keys equally for an emotion, giving $\bar{r}_a$. We call this the source pooled-residual comparator and label it Shared residual (or Shared) in the tables. It uses the same source training split and synthesis interface as the learned editor; individual residuals are not unit-normalized before averaging.

The norm-matched control rescales the predicted residual to $\|\bar{r}_a\|_2$ before applying Eq. 1. This removes the difference in residual length between pooled and neural arms for a given emotion, without matching the resulting angle or identity cost. Normalization maintains the supplied condition's norm, not speaker identity. Direction, magnitude and perceived intensity must therefore be evaluated separately.

The supplied condition is unit-normalized; we do not claim preservation of the reference vector's pre-normalization norm. Appendix D distinguishes the audited objective caches from the listening study's incomplete render provenance, documented in Appendix C.

4 EXPERIMENTAL SETUP

Source training data. The training collection contains 195,301 deduplicated pairs of synthetic neutral and target embeddings. Connected source-key and exact-text groups define the split: 154,840 training, 19,652 validation and 20,809 test pairs. The 60,365 source keys are not verified identities, and exact-text grouping does not establish near-duplicate or person-level separation. Prior emotion-based selection limits population claims. This collection trains the editor; principal waveform and listening evaluations use real neutral ESD references (Zhou et al., 2021).

Text-driven objective evaluation. We evaluate 18 ESD speakers, nine per language, and four target emotions. One query per speaker–emotion cell is synthesized with three separate neutral references, yielding 216 reference conditions. Two native interfaces, zero-shot and emotional `instruct2`, and four editing arms yield 1,728 waveforms at $\alpha = 1$. Within each interface and condition, all arms share text, tokens, acoustic prompt and decoder noise. Reference recordings differ from the query text; target recordings are evaluation-only. Repeated references and waveforms are not additional speakers. The ESD speakers come from corpus training and development splits; person-level overlap with backbone or scorer training is unknown.

Metrics and aggregation. Emotion2vec target-embedding cosine measures emotional proximity (Ma et al., 2024). CAM++ reference cosine measures proximity in the edited representation, not independent identity verification. Whisper large-v3 supplies Chinese CER and English WER on full outputs, with beam size 5 and NFKC normalization (Radford et al., 2023). We average references within query, the four emotions equally within speaker, then speakers equally within language. Comparisons are within an interface, not between different prompts. CER/WER do not measure naturalness.

Listening protocol. The dose study compares unedited CV3 with FiLM at $\alpha = 1.5$ and 5 under native emotional instructions. Each case presents three versions under position-balanced A–C labels, hiding system and dose. An independent neutral recording anchors naturalness and similarity. Four 1–5 ratings assess emotion appropriateness (E), naturalness (N), speaker similarity (S) and target-emotion intensity (I). Intensity is neither generic arousal nor quality. No target-emotion recording is needed to construct the panel.

Thirty complete response files from 16 respondents cover 128 cases and 16 speakers in total; each language has 15 respondents, 64 cases and eight speakers. They yield 1,440 stimulus ratings. Respondents overlap across languages. Fifteen of 16 templates were returned, but overlapping assignments cover every case–dose cell with two to six ratings. After completeness and playback checks, we average listeners within case–dose, cases within speaker, and speakers equally within language, without score-based exclusions. The primary contrast is $5 - 0$; $1.5 - 0$ is secondary. Pointwise 95% intervals use 10,000 speaker-bootstrap draws, conditional on the observed respondents and record-

Table 1: **Expression and reference-distance diagnostics on the same outputs.** FiLM at $\alpha = 1$; nine speakers and 108 outputs per row. ΔEmo is descriptive speaker-balanced mean target-emotion cosine gain over matched unedited synthesis. $d = 1 - \cos$ is distance to the neutral enrollment; real targets match speaker and query text. P95 counts use each output's speaker–emotion/reference real pool. New errors count positive-to-nonpositive attribution transitions; FiLM has 431/432 positive margins, with one pre-existing failure. Columns are separate summaries, not a joint success rate or an identity-safety test.

Lang.	Interface	ΔEmo	d_{gen} / d_{real}	$\leq$ real P95	New errors
ZH	Zero-shot	$+0.0395$	0.1759 / 0.5579	108/108	0
ZH	Instruct2	$+0.1043$	0.3551 / 0.5579	108/108	0
EN	Zero-shot	$+0.0278$	0.2682 / 0.4060	106/108	0
EN	Instruct2	$+0.0488$	0.2628 / 0.4060	107/108	0

ings. N and S use the originally specified 0.20-point non-inferiority margin. Appendix C documents allocation deviations and early inspection; the analysis is exploratory.

5 RESULTS

5.1 EXPRESSION ENHANCEMENT WITH A REAL-VOICE VARIATION REFERENCE

We first evaluate expression gain together with the scale of reference-distance change on the same $\alpha = 1$ outputs. Table 1 aligns emotion scores with real-variation and attribution diagnostics by case, interface and arm. Mean target-emotion cosine increases in all four language–interface groups, including both native emotional-instruction settings. Meanwhile, 429 of the 432 FiLM outputs have reference distances at or below their matched real-emotion P95, and none introduces a closed-set attribution failure relative to unedited synthesis. This evidence supports additional emotional response at a scale interpretable against observed real-voice variation, rather than treating every decrease in neutral-reference similarity as an identity change.

Interpreting the real-voice reference. The post-hoc distance diagnostic compares generated speech and real target-emotion recordings against the same neutral enrollment. Mean FiLM-minus-real distance gaps range from -0.3820 to -0.1378; all four Bonferroni-adjusted interval upper bounds are negative (Appendix D, Table 11). The empirical P95 additionally describes individual-output distance against available same-speaker, same-emotion recordings. Three outputs exceed this upper reference. It is a scalar distance benchmark, not membership in a human expression manifold; emotional intensity is not matched and small reference pools make tail estimates uncertain.

Against eight same-language competing enrollment centroids, FiLM retains positive CAM++ margins on 431/432 outputs. The sole failure is already present at zero dose, and mean margins decrease in all four groups. These are closed-set diagnostics in the edited representation, not independent identity verification. Positive average emotion gains and P95 coverage do not establish how many individual outputs satisfy both criteria. They also do not transfer to the listening study's higher doses.

Matched controls and content fidelity. Table 2 compares unedited, shared and norm-matched synthesis on the same panel. The shared and unedited arms also mostly lie below real P95 (Appendix D, Table 10); coverage alone does not distinguish FiLM. The largest FiLM emotion-cosine gain occurs in Chinese `instruct2`, from 0.5746 to 0.6789. Across the four groups, the gains in Table 1 divided by the corresponding unedited means give approximately 6.3–18.2% relative increases in this cosine score. These are descriptive metric changes, not percentages of perceived emotion or evidence that every speaker improves.

Reference cosine decreases by 0.0075–0.0387 in the FiLM rows. Chinese CER improves in both interfaces, while English WER increases. These outcomes show emotional response with a measurable reference-proximity cost, not an across-metric improvement. The unchanged token plans locate the contrast in acoustic-condition assignment; they do not remove the expression already present in the tokens or identify a natural effect of human emotion.

Table 2: **Text-driven emotional enhancement and its costs.** ESD, 18 speakers, $\alpha = 1$. Emo: target-emotion cosine; Ref.: CAM++ reference cosine; Error: CER (ZH) or WER (EN), in percent. Arms share synthesis conditions within an interface. Bold marks best displayed values within each block, not significance.

Lang.	Interface	Method	Emo↑	Ref.↑	Error↓
ZH	Zero-shot	Unedited CV3	.3060	.8397	1.53
		Shared residual	.3149	**.8419**	1.81
		FiLM	**.3455**	.8241	**1.26**
		FiLM, norm-matched	.3093	.8406	1.53
ZH	Instruct2	Unedited CV3	.5746	**.6836**	6.81
		Shared residual	.6067	.6752	6.87
		FiLM	**.6789**	.6449	**5.30**
		FiLM, norm-matched	.6029	.6761	6.36
EN	Zero-shot	Unedited CV3	.4406	**.7393**	4.83
		Shared residual	.4553	.7351	**4.44**
		FiLM	**.4683**	.7318	5.14
		FiLM, norm-matched	.4556	.7333	5.14
EN	Instruct2	Unedited CV3	.5431	**.7585**	2.56
		Shared residual	.5700	.7512	2.56
		FiLM	**.5919**	.7372	2.96
		FiLM, norm-matched	.5664	.7492	**2.41**

Table 3: **Bilingual listening results.** Native emotional instruction is shared across doses; shaded rows are ours. Each language: 15 response sessions, 64 cases, eight speakers and 240 ratings per dose. E: emotion appropriateness; N: naturalness; S: similarity; I: intensity, not quality (1–5; speaker-balanced means; bold marks the column maximum within a language block and is not significance).

Method	**Control**	α	**E-MOS↑**	**N-MOS↑**	**S-MOS↑**	**I-MOS**
Chinese: 15 response sessions, 64 cases, eight speakers						
CV3 (unedited)	Prompt	0	3.22	3.48	**4.15**	3.36
CV3 + FiLM (ours)	Prompt + U edit	1.5	3.46	3.47	3.99	3.80
CV3 + FiLM (ours)	Prompt + U edit	5	**3.57**	**3.51**	3.70	**3.99**
English: 15 response sessions, 64 cases, eight speakers						
CV3 (unedited)	Prompt	0	3.32	3.89	**4.45**	3.35
CV3 + FiLM (ours)	Prompt + U edit	1.5	3.57	3.92	4.39	3.70
CV3 + FiLM (ours)	Prompt + U edit	5	**4.01**	**4.02**	4.23	**4.27**

5.2 Listening evaluation of emotional expression

The bilingual listening results extend the automatic evidence to perceived expression. Emotion appropriateness and requested-emotion intensity increase at both doses in both languages (Table 3). At $\alpha = 5$, the paired E-MOS gain is $+0.354$ with interval $[+0.262, +0.451]$ in Chinese and $+0.690$ with $[+0.617, +0.753]$ in English. The eight pointwise intervals for E and I across the four language–dose comparisons exclude zero. The per-emotion breakdown and listener-resampling sensitivity are reported in Appendix C, without treating those exploratory comparisons as additional independent confirmations.

Table 7 in Appendix C separates naturalness from speaker similarity. All four N intervals clear the -0.20 margin, supporting conditional non-inferiority rather than equivalence. At the high dose, S changes by -0.448 in Chinese and -0.224 in English. The Chinese interval lies entirely below -0.20, whereas the English interval crosses that margin; neither establishes high-dose S non-inferiority. At the low dose, English S clears the margin and Chinese S does not. The English low-dose interval includes zero, which does not preclude non-inferiority or establish equivalence. Shared editing was not evaluated in this listening study.

5.3 What reference conditioning and magnitude contribute

Paired endpoint learning yields a measurable predictive gain: on held-out synthetic pairs, three FiLM seeds reach cosines 0.90869, 0.90874 and 0.90877 versus 0.88771 for source pooling. The consistent point estimates support learning beyond this fixed comparator on the reported split. They do not isolate the value of direction versus magnitude, establish person-disjoint generalization, or guarantee equivalent gains after synthesis. The real-reference waveform and listening studies therefore test a separate link in the prediction-to-speech chain.

Residual magnitude contributes to the observed expression response. Norm-matched FiLM produces smaller emotion-cosine gains than unrestricted FiLM; its emotion cosine is below pooling in three groups and only 0.0003 higher in the fourth. This comparison tests direction choice at the pooled residual length, but does not separately establish the value of reference-dependent magnitude prediction. Matching length also leaves the final condition angle and output identity cost unmatched (Appendix L.4). The data therefore support joint evaluation of direction and magnitude, while leaving a reference-specific directional advantage or a better trade-off at matched identity cost unresolved.

Independent calibrated ESD extension. A separate calibrated ESD network uses a 576-dimensional voice input, DSP and privileged target-derived activation conditions. At $\alpha = 1$, its target-emotion cosine gain is 0.0921, but joint automatic success is 32.81% versus 56.25% for its shared baseline. Added conditions do not establish corrected superiority over ablations, and activation-distance gain is -0.2509. These results concern the calibrated extension rather than the source-trained main model.

Scope of supplementary evidence. Appendix E retains the calibrated ESD network and its controls. Historical sweeps in Appendices H–J use angular editing with separate verifiers and a separate four-listener study; neither they nor the representation-specific IndexTTS2 results in Appendix K evaluate or transfer the main FiLM model.

6 Discussion and Conclusion

Reference-conditioned speaker editing provides an additional control for emotional realization while preserving CV3's speech plan and pretrained acoustic interface. On the reported synthetic split, the learned editor predicts target endpoints more accurately than source pooling. Applied to real neutral references, it increases mean target-emotion cosine in both languages and native interfaces. The waveform and listening studies evaluate the consequences of those edits beyond condition-space prediction. Real-recording distances and closed-set attribution contextualize reference drift without serving as identity guarantees.

The listening study supports expression gains at both tested doses. At the higher dose, emotion appropriateness rises from 3.22 to 3.57 in Chinese and 3.32 to 4.01 in English, while speaker similarity falls. Naturalness meets the stated non-inferiority margin under conditional speaker resampling. The two doses characterize a measured trade-off rather than an optimized operating point. Reusing the fitted editor avoids per-reference model updates within the tested CV3 setting. Shared and norm-matched controls leave the direction-specific advantage unresolved; the supported contribution is learned, adjustable enhancement through the native realization interface.

Limitations. Listening allocation is unbalanced and early results were inspected; speaker-bootstrap intervals do not model respondent and speaker populations jointly. CAM++ is the editing representation, ESD is acted speech, and real/generated emotional intensity is not matched. Adaptive dosing, arbitrary emotional references, arousal control and cross-backbone transfer of the trained editor remain unvalidated. Appendix A maps the separate calibrated and historical evidence and its study-specific scope.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix F specify the main editor and source training; Section 4 and Appendix B describe matched synthesis and objective evaluation. Appendix C documents the listening protocol, realized allocation, analysis deviations and panel-key reconstruction. Appendix D details the objective feature-cache audit. These records support reproduction of the reported score analyses. The reported checks cover response validation and stimulus-identifier decoding, rather than independent verification of the original audio-to-render mapping. Training-source identity overlap with backbone or evaluator data remains unverified.

ETHICS STATEMENT

The listening study includes 16 respondents (Appendix C). The response files alone do not verify consent, compensation, eligibility screening or ethics approval; no unverified approval is asserted. Expressive voice cloning carries impersonation risks: deployment requires speaker authorization, synthetic-audio disclosure and compliance with corpus and model licenses. The evaluation uses acted emotional speech.

AI USE STATEMENT

Generative AI assistants supported method and analysis implementation, response-data formatting, experimental-design feedback, result interpretation, literature summaries, figure creation, reference formatting, and manuscript drafting and revision. Generative speech models also produced the synthetic training material and evaluation stimuli described in the experimental protocols; these are distinct from human listener ratings. Verification included rerunning analyses from archived responses, cross-checking reported numbers against result files, and matching collected stimulus identifiers to reconstructed keys. These checks do not constitute independent waveform-level replication. The authors are responsible for the final text, claims and artifacts.

REFERENCES

Anonymous. PATHOS-TTS: Causal emotion editing across instruction-guided speech-token planning and native acoustic conditioning. Unpublished manuscript under review; author identities withheld for double-blind review, 2026.

Haozhe Chen, Run Chen, and Julia Hirschberg. EmoKnob: Enhance voice cloning with fine-grained emotion control. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8170–8180. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.466. URL https://aclanthology.org/2024.emnlp-main.466/.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, et al. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025.

Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. 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.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15747–15760, 2024.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11671. URL https://ojs.aaai.org/index.php/AAAI/article/view/11671.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023.

Klaus R Scherer. Vocal communication of emotion: A review of research paradigms. *Speech communication*, 40(1-2):227–256, 2003.

Hui Wang, Siqi Zheng, Yafeng Chen, Luyao Cheng, and Qian Chen. Cam++: A fast and efficient network for speaker verification using context-aware masking. In *Proc. Interspeech 2023*, pp. 5301–5305, 2023.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A. Saurous. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5180–5189, 2018. URL https://proceedings.mlr.press/v80/wang18h.html.

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, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2913–2925, 2024.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. 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.

A Evidence Map and Model Boundaries

Table 4 maps each experiment to its model, data and supported interpretation. The source-trained 192-dimensional FiLM editor is the main method. The 576-dimensional calibrated ESD network, historical angular directions and IndexTTS2 representation-specific directions are separate constructions. All comparisons must retain their own data support, dose definition and statistical unit.

Table 4: Which evidence answers which question. Counts of conditions or waveforms are not independent speaker counts.

Evidence	Inputs / scale	Supported role
Main objective panel	192-D FiLM; 18 speakers, 216 conditions, $\alpha = 1$	Emotional response with text-generated fixed plans; shared and norm-matched controls
Bilingual listening	Same editor family; 15 respondents per language (16 in total), 128 cases; $\alpha = 1.5, 5$	Perceived expression, naturalness and similarity under observed allocation
Real-variation diagnostic	Main objective panel and matched real targets; $\alpha = 1$	Reference-distance context and closed-set CAM++ attribution, not high-dose MOS preservation
Calibrated ESD extension	576-D voice input, DSP and target-derived activation; 16 outer speaker folds	Separate privileged-support test with full ablations and negative joint result
Historical B/C studies	Statistical angular edits; ESD / character voices	Dose dependence and verifier sensitivity, not the current FiLM editor
Source prefix	4,096 training queries, 90,112 waveforms	Descriptive training-domain behavior, not held-out transfer
IndexTTS2 extension	Wav2Vec2-BERT directions; 18-speaker panel	Representation-specific response, not transfer of the trained CAM++ model

The main model's fitted architecture and learning setup are described in Section 3 and Appendix F. Cross-study alpha values are not equivalent perceptual doses. The real-variation diagnostic does not use the historical dots result as evidence for current FiLM performance. Appendix C distinguishes the reconstructed response-key mapping from a fully verified waveform provenance chain.

A.1 Study-Specific Scope

Historical listening results cannot substitute for listening tests on the current editor. Repeated enrollment references in the text-driven study are conditions rather than additional speakers, and its 18 speakers come from corpus train and dev splits with person-level membership in editor, scorer or backbone training unestablished, so those numbers are development results without a population-level significance claim. The DSP-conditioned study derives its activation requests from target recordings, whereas the source-trained editor receives only an embedding and an emotion label, and the text-driven application shifts that input again to single real neutral enrollment clips; none of the three establishes automatic activation prediction from neutral speech. The main setting is a neutral reference with a requested target emotion, not an arbitrary emotional reference. Extending it would require representing the source–target emotion relationship and testing identity and expression across source states; the present neutral-reference results do not validate that extension. Nor do retrospective dose ceilings establish an input-based adaptive dose policy. The IndexTTS2 experiment changes both the representation and the direction estimator, so it cannot isolate architecture from training origin; its leave-one-speaker-out directions share contributor data across folds, and one norm-matched random direction per condition is a limited perturbation control. The descriptor-block decomposition of the sad failure in Section H is post hoc and was not a preregistered gate.

B Text-Driven Transfer with Neutral References

The transferred editor. The editor here is pretrained on a separate synthetic collection: 195,301 deduplicated synthetic-neutral/synthetic-target embedding pairs, split by connected source-key and exact-text groups into 154,840/19,652/20,809 train/validation/test pairs. It predicts a residual from a

192-dimensional synthetic-neutral embedding and an emotion label only, without the ESD editor's 576-dimensional input, calibration gap, DSP profile or activation request. On held-out synthetic pairs FiLM reaches endpoint cosines 0.90869/0.90874/0.90877 over three seeds against pooled cosine 0.88771, a descriptive embedding advantage of about 0.021 that is not a waveform or MOS gain. The 60,365 source keys are not verified identities, tokens and synthesis randomness were not held fixed across endpoints, and prior emotion-based selection prevents reading the pretraining data as an unbiased emotion intervention. A fixed 4,096-query source-training diagnostic subset (90,112 waveforms) gives FiLM a mean target-label gain of 11.45 points against 6.06 for source pooling at $\alpha = 1$, with reference-cosine decreases of 0.0094 against 0.0029; these are training-domain results. Appendix F gives the pretraining configuration, learning curve and source-prefix protocol.

Protocol and available inputs. This study tests the realization-side assignment when H is produced by a native CV3 interface rather than taken from a target recording. We assess the frozen editor on real neutral ESD enrollment references without target-domain fine-tuning, varying language, requested emotion, speaker, enrollment recording and native control mode. Reference diversity here does not include varying the source emotion. We evaluate 18 speakers, nine per language, four emotions, one query per speaker–emotion cell, each repeated under three separate neutral enrollment recordings whose texts differ from the query, giving 216 conditions; a repeated reference is a condition, not an additional speaker. Each condition supplies one normalized CAM++ vector e_s and the same recording as acoustic prompt, so the arms share the same original enrollment vector and acoustic prompt; only the speaker condition supplied after editing differs. The editor receives e_s and the requested emotion and returns $\mathrm{normalize}(e_s + \alpha\hat{r})$ from the mean residual of three frozen fits, avoiding outcome-based seed selection. Target recordings are evaluation-only; speech tokens are generated from text. Two CV3 interfaces are evaluated separately, zero-shot synthesis and emotional `instruct2`, whose prompt semantics differ, so valid contrasts are edited versus unedited *within* an interface; tokens, acoustic prompt, worker and decoder noise are shared across arms within a cell. Unedited CV3 is the system baseline, including its emotional instruction in `instruct2`. The source pooled-residual arm averages the training residuals to test an emotion-level shared increment; FiLM predicts a reference-conditioned increment, and norm-matched FiLM controls its length. These arms isolate design choices within the same synthesis system. The four arms produce 1,728 waveforms at $\alpha = 1$; the earlier two-speaker dose grid is in Appendix G.

Scoring and aggregation. Full-utterance emotion2vec supplies target-embedding cosine, and CAM++ cosine to the enrollment clip is a diagnostic rather than an independent verifier because it is also the editing space. Local Whisper large-v3 transcribes complete outputs, scored as Chinese CER and English WER with beam size 5 and NFKC normalization. We average references within a query, queries within speaker–emotion cells, the four emotions equally, then speakers within a language, so speakers rather than renderings carry the weight. Speakers come from corpus train and dev splits and person-level membership in editor, scorer or backbone training is not established; these remain development results without a significance claim.

Source-to-real-reference response across diverse conditions. At $\alpha = 1$, FiLM raises mean target-emotion cosine over unedited synthesis in all four language–interface groups, by 0.0278–0.1043, though per-speaker counts in Table 5 show that positive group means need not imply improvement for every speaker. The largest effect is Chinese `instruct2` (0.5746 to 0.6789), an additional response even when an emotional instruction is already present. Reference cosine decreases in every group, by 0.0075–0.0387, showing a systematic expression–reference-proximity trade-off. The real-variation and attribution diagnostics in Appendix D contextualize these decreases without treating them as direct measurements of perceptual identity loss. Mean CER/WER changes are small and mixed: Chinese error decreases in both interfaces, including 6.81 to 5.30 under `instruct2`, while English error increases by at most 0.40 points. These means establish neither perceptual quality nor statistical non-inferiority.

Magnitude and direction remain distinct. Rescaling the FiLM residual to the pooled norm reduces the mean emotion-cosine gains to 0.0033–0.0283, and in Chinese zero-shot synthesis only 6 of 9 speakers improve. Residual magnitude therefore matters to the observed response. The source pooled-residual comparator also increases target-emotion cosine for most speakers, with smaller gains and generally smaller reductions in CAM++ reference cosine than FiLM. This ablation does

Table 5: Text-driven CV3 development results on 18 ESD speakers (9 per language), three separate neutral enrollment references per query, $\alpha = 1$. References are averaged within a query, queries within speaker–emotion cells, four emotions equally, then speakers within a language. Shared residual is the source pooled-residual comparator; FiLM and norm-matched FiLM are source-trained conditional implementations; Unedited CV3 is the matched native-interface baseline. CAM is reference cosine, not an independent identity verifier. Error is mean utterance CER (ZH) or WER (EN), in percent. Improved counts speakers with positive mean emotion-cosine change. Bold marks column-wise best displayed values within each language–interface block, including ties and the unedited baseline. No significance or non-inferiority claim is made.

Lang.	Interface	Method	e2v↑	CAM↑	Error↓	Improved
ZH	Zero-shot	Unedited CV3	0.3060	0.8397	1.53	—
ZH	Zero-shot	Shared residual	0.3149	**0.8419**	1.81	8/9
ZH	Zero-shot	FiLM	**0.3455**	0.8241	**1.26**	**9/9**
ZH	Zero-shot	FiLM (norm-matched)	0.3093	0.8406	1.53	6/9
ZH	Instruct2	Unedited CV3	0.5746	**0.6836**	6.81	—
ZH	Instruct2	Shared residual	0.6067	0.6752	6.87	**8/9**
ZH	Instruct2	FiLM	**0.6789**	0.6449	**5.30**	**8/9**
ZH	Instruct2	FiLM (norm-matched)	0.6029	0.6761	6.36	7/9
EN	Zero-shot	Unedited CV3	0.4406	**0.7393**	4.83	—
EN	Zero-shot	Shared residual	0.4553	0.7351	**4.44**	**9/9**
EN	Zero-shot	FiLM	**0.4683**	0.7318	5.14	8/9
EN	Zero-shot	FiLM (norm-matched)	0.4556	0.7333	5.14	8/9
EN	Instruct2	Unedited CV3	0.5431	**0.7585**	2.56	—
EN	Instruct2	Shared residual	0.5700	0.7512	2.56	**9/9**
EN	Instruct2	FiLM	**0.5919**	0.7372	2.96	**9/9**
EN	Instruct2	FiLM (norm-matched)	0.5664	0.7492	**2.41**	**9/9**

not establish that the neural editor selects an appropriate per-reference dose, that only magnitude matters, or that its direction is better at a matched identity cost. It motivates evaluating direction and magnitude jointly under the operating constraints in Section H.

IndexTTS2: representation-specific editing. A separate experiment edits the 1024-dimensional standardized Wav2Vec2-BERT layer-17 features returned by `get_emb`, using real-recording emotional-minus-neutral means over other speakers (16–17 contributors) with all panel reference and target recordings excluded. This is a new representation-specific statistical editor, not transfer of the trained CAM++ network. Editing the emotion-reference branch at strength 1 raises default-mode emotion cosine by 0.1655 (ZH) and 0.2074 (EN), against 0.0088 and 0.0036 for a norm-matched random control, with all nine speakers per language positive and CAM reference cosine falling by 0.0470 and 0.0240; editing only the speaker-reference branch gives much smaller changes. Seven arms, two native-control settings and 3,024 waveforms are reported in Appendix K. These results support branch-specific response, not universal transfer or quality improvement, and no listener results are available.

C Bilingual Dose Listening: Protocol, Coverage and Uncertainty

Materials and scales. The protocol compares unedited CV3 with source-trained FiLM residual editing at $\alpha = 1.5$ and 5. The retained generation protocol specifies a shared native emotional instruction, cached speech-token plan, neutral conditioning clip, decoder noise and case seed across doses; the waveform-level provenance boundary is stated below. Texts are drawn from neutral transcripts; target-emotion recordings are not required. Three versions of a case are presented together under position-balanced A–C labels, hiding system and dose. An independent neutral recording, distinct from the conditioning clip, anchors naturalness and speaker similarity. E measures appropriateness for the specified emotion, N naturalness, S similarity to the reference speaker, and I strength of the specified emotion. All use integer scales from 1 to 5. Loudness or exaggeration is

Table 6: Speaker-balanced MOS means with pointwise 95% speaker-bootstrap intervals. This is the same cohort as Table 3, not an additional experiment.

Lang.	α	E	N	S	I
ZH	0	3.22 [3.11, 3.32]	3.48 [3.36, 3.61]	4.15 [4.02, 4.26]	3.36 [3.23, 3.48]
ZH	1.5	3.46 [3.35, 3.56]	3.47 [3.39, 3.58]	3.99 [3.85, 4.15]	3.80 [3.75, 3.86]
ZH	5	3.57 [3.45, 3.69]	3.51 [3.42, 3.62]	3.70 [3.55, 3.86]	3.99 [3.80, 4.16]
EN	0	3.32 [3.20, 3.42]	3.89 [3.76, 4.02]	4.45 [4.40, 4.50]	3.35 [3.24, 3.43]
EN	1.5	3.57 [3.46, 3.69]	3.92 [3.85, 3.98]	4.39 [4.30, 4.47]	3.70 [3.64, 3.75]
EN	5	4.01 [3.92, 4.10]	4.02 [3.90, 4.11]	4.23 [4.08, 4.36]	4.27 [4.21, 4.32]

not automatically better, and strong sadness need not be loud. Instructions request headphones and a quiet environment; the exported files cannot verify compliance. Dose choices followed unblinded pilot listening; omitting $\alpha = 3$ does not establish equivalence with $\alpha = 5$.

Actual response coverage. The analyzed batch, supplied on 24 September 2026, contains 30 complete response files: 15 per language and 48 stimulus rows per file. Each session contains 16 cases, each at three doses, giving 1,440 stimulus records and 5,760 individual dimension scores. The language-specific counts use distinct normalized respondent identifiers, not independent ratings. Fourteen respondents completed both languages and two completed one language each, giving 16 respondents in total and 15 per language. The two language cohorts are therefore not independent. The upload date is not an independently recorded recruitment closing time.

The planned layout had eight templates per language and one respondent per template. Actual returns cover seven Chinese and all eight English templates, with multiple respondents on several templates. The missing Chinese template does not remove cases because templates overlap: all 128 cases, 16 speakers and 384 case–dose cells are observed. Each language has eight speakers and 64 cases, with two query texts per speaker–emotion cell. There are 48 case–dose cells with two ratings, 192 with three, 48 with four, and 96 with six. These are counts of cells, not people. Each language contributes 240 ratings per dose. No missing score is imputed and no response is excluded on the basis of its score.

Validation and stimulus-key recovery. Every submitted file contains its 48 expected stimuli exactly once, consistent panel and respondent identifiers, valid scores and candidate/reference playback flags. Flags indicate recorded playback events, not verified listening duration or eligibility. Only the supplied 30 files are included; earlier draft snapshots are not appended as additional respondents. For this analysis, we reconstructed stimulus assignments deterministically from the retained dose plan and panel-builder rule, whose identifiers hash the seed, panel, case and arm. All submitted identifiers match the reconstructed keys. This check validates identifier decoding; it does not independently verify the mapping from the original audio files to stimulus keys and rendering checkpoints.

Primary estimand and uncertainty. For each language, listener ratings are averaged within case and dose, case means within speaker, and speakers equally. Paired contrasts first subtract $\alpha = 0$ for the same case, then apply the same speaker weighting. The primary comparison is $5 - 0$; $1.5 - 0$ is secondary. The scorer resamples eight speaker means with replacement 10,000 times, with a fixed seed rule based on 20260922, and reports percentile 95% intervals. Table 6 adds mean-score intervals using the same cluster unit and number of draws, with fixed presentation seed rules based on 20260924; original paired intervals are unchanged. All intervals are pointwise, not simultaneous across dimensions, languages, doses or emotions. They condition on the collected respondents and cases, and do not incorporate complete crossed listener–speaker population uncertainty. No ordinal mixed-effects model was fitted.

Table 7 reports all within-case paired changes using this estimand; the following sections detail the non-inferiority interpretation and alternative respondent weighting.

Naturalness and similarity non-inferiority. The analysis rule requires the paired 95% lower bound to exceed -0.20 for each of N and S. For $\alpha = 5$, N lower bounds are -0.115 (ZH) and

Table 7: **Paired listening changes and uncertainty.** Each cell gives the change from matched unedited CV3 and its pointwise 95% speaker-bootstrap interval. E/N/S/I follow Table 3. Speakers are equally weighted within language; inference is conditional on observed respondents and cases, not a crossed listener–speaker population analysis.

Lang.	Contrast	ΔE	ΔN	ΔS	ΔI
ZH	$1.5-0$	$+0.238$ $[+0.158, +0.333]$	-0.014 $[-0.124, +0.085]$	-0.154 $[-0.289, -0.038]$	$+0.448$ $[+0.290, +0.591]$
ZH	$5-0$	$+0.354$ $[+0.262, +0.451]$	$+0.022$ $[-0.115, +0.143]$	-0.448 $[-0.624, -0.286]$	$+0.633$ $[+0.331, +0.888]$
EN	$1.5-0$	$+0.253$ $[+0.185, +0.307]$	$+0.031$ $[-0.052, +0.115]$	-0.060 $[-0.133, +0.008]$	$+0.349$ $[+0.276, +0.432]$
EN	$5-0$	$+0.690$ $[+0.617, +0.753]$	$+0.133$ $[+0.034, +0.247]$	-0.224 $[-0.357, -0.107]$	$+0.922$ $[+0.839, +1.008]$

Table 8: Respondent-weighted paired differences with pointwise 95% respondent-bootstrap intervals, separately within language. The observed speaker/case assignments are held fixed.

Lang.	Contrast	ΔE	ΔN	ΔS	ΔI
ZH	$1.5-0$	$+0.242$ [+0.146, +0.346]	$+0.037$ [-0.067, +0.146]	-0.167 [-0.367, +0.004]	$+0.471$ [+0.358, +0.583]
ZH	$5-0$	$+0.375$ [+0.212, +0.550]	$+0.083$ [-0.096, +0.258]	-0.442 [-0.742, -0.179]	$+0.654$ [+0.475, +0.838]
EN	$1.5-0$	$+0.242$ [+0.133, +0.358]	$+0.037$ [-0.054, +0.121]	-0.067 [-0.146, +0.004]	$+0.346$ [+0.200, +0.492]
EN	$5-0$	$+0.671$ [+0.379, +1.008]	$+0.129$ [-0.071, +0.375]	-0.225 [-0.412, -0.062]	$+0.904$ [+0.633, +1.200]

$+0.034$ (EN), whereas S lower bounds are -0.624 and -0.357. Both languages therefore support N non-inferiority under this conditional analysis, but neither supports S non-inferiority at the high dose. The secondary low-dose N intervals also clear the margin; for S, the low-dose English lower bound (-0.133) clears the margin while the Chinese one (-0.289) does not. Failing to establish non-inferiority does not by itself show that a loss exceeds the margin: the English high-dose and Chinese low-dose intervals cross -0.20. The Chinese high-dose interval $[-0.624, -0.286]$, however, lies entirely below the margin under speaker resampling. Low-dose English S has mean -0.060 and interval $[-0.133, +0.008]$, supporting non-inferiority despite including zero; this is not an equivalence test. Higher I scores reflect stronger expression, not improved quality. No margin is applied to E or I.

Listener-resampling sensitivity. For each respondent and language, paired case differences are averaged, followed by an equal-respondent mean and a 10,000-draw respondent bootstrap. This changes the weighting because panel allocation is unbalanced; it is not the same estimand or a substitute for crossed inference. Table 8 retains positive E/I intervals at both doses. High-dose S remains negative, but English high-dose N now includes zero, unlike the speaker-bootstrap interval. At the low dose, the respondent-weighted S intervals now include zero in both languages (ZH $[-0.367, +0.004]$, EN $[-0.146, +0.004]$). Low-dose Chinese N also changes mean sign. The sensitivity analysis supports the broad expression result, not an assertion that every sign or inferential conclusion is invariant.

Analysis deviations and evidence scope. Condition-linked outcomes had been viewed for an earlier two-listener draft before this batch was assembled. The observed allocation also differs from the one-response-per-template and separate-language-cohort plan. We retain the original doses and non-inferiority margin, but label this as an exploratory observed-allocation analysis rather than a preregistered first look or the completed original design. The study compares FiLM with unedited CV3, not shared editing, and does not validate automatic dose selection. Its higher doses differ from the $\alpha = 1$ objective and real-variation diagnostic. Historical four-listener audibility results are a separate study. Recruitment eligibility, consent, compensation and applicable ethics review require author-held records and cannot be inferred from these TSV files.

Table 9: Exploratory high-dose emotion-appropriateness changes by requested emotion. Each language–emotion cell has eight speakers and 16 cases. Pointwise intervals are unadjusted for these eight comparisons; this breakdown is not independent confirmation.

Emotion	ZH ΔE, $5-0$	EN ΔE, $5-0$
Angry	$+0.479$ [+0.094, +0.896]	$+0.854$ [+0.583, +1.156]
Happy	$+0.531$ [+0.318, +0.755]	$+0.677$ [+0.406, +0.938]
Sad	$+0.245$ [+0.094, +0.422]	$+0.354$ [+0.167, +0.542]
Surprised	$+0.161$ [+0.005, +0.286]	$+0.875$ [+0.656, +1.073]

Table 10: Expression and attribution diagnostics for all objective arms at $\alpha = 1$. Each row uses 108 output conditions. ΔEmo is the speaker-balanced mean emotion-cosine change. P95 and positive-margin columns are descriptive counts; new errors are paired unedited-positive to edited-nonpositive transitions. These columns are not a joint success test.

Lang.	Interface	Method	ΔEmo	$\leq$ P95	$m > 0$	New
ZH	Zero-shot	Unedited CV3	$+0.0000$	108/108	108/108	0
ZH	Zero-shot	Shared residual	$+0.0089$	108/108	108/108	0
ZH	Zero-shot	FiLM	$+0.0395$	108/108	108/108	0
ZH	Zero-shot	FiLM, norm-matched	$+0.0033$	108/108	108/108	0
ZH	Instruct2	Unedited CV3	$+0.0000$	108/108	108/108	0
ZH	Instruct2	Shared residual	$+0.0320$	108/108	108/108	0
ZH	Instruct2	FiLM	$+0.1043$	108/108	108/108	0
ZH	Instruct2	FiLM, norm-matched	$+0.0283$	108/108	108/108	0
EN	Zero-shot	Unedited CV3	$+0.0000$	105/108	107/108	0
EN	Zero-shot	Shared residual	$+0.0148$	105/108	107/108	0
EN	Zero-shot	FiLM	$+0.0278$	106/108	107/108	0
EN	Zero-shot	FiLM, norm-matched	$+0.0151$	106/108	107/108	0
EN	Instruct2	Unedited CV3	$+0.0000$	107/108	108/108	0
EN	Instruct2	Shared residual	$+0.0269$	107/108	108/108	0
EN	Instruct2	FiLM	$+0.0488$	107/108	108/108	0
EN	Instruct2	FiLM, norm-matched	$+0.0233$	108/108	108/108	0

D Matched Real-Recording Variation and Closed-Set Attribution

Why contextualize reference distance? CAM++ conditions vary with expression as well as speaker traits. Lower similarity to a neutral reference can therefore reflect expression change, identity-related change, or both. This post-hoc analysis contextualizes the objective experiment's reference-distance scale; it does not replace perceptual speaker-similarity ratings. The analysis uses the existing 18-speaker text-driven experiment at $\alpha = 1$: 216 conditions, two interfaces and four arms, or 1,728 output records. It does not re-render speech or analyze the listening study's $\alpha = 1.5/5$ waveforms.

Alignment with emotional-response measurements. The joint presentation in Table 1 uses retained objective records keyed by case, interface and editing arm. All 1,728 keys match the real-variation analysis, with identical query, reference, speaker and emotion fields; generated CAM++ distances agree exactly. Emotion gains are within-case changes relative to unedited synthesis, averaged with the same speaker weighting. The P95 and attribution columns are separate output counts. No combined per-output success criterion is defined. Table 10 gives the same summaries for all four arms; their high P95 coverage is not evidence of a FiLM-specific advantage.

Matched endpoints. For each condition, let e_{ref} be its exact neutral enrollment embedding, e_{gen} the output embedding and e_{real} the same speaker's real target-emotion recording of the query text. All are unit-normalized. We compute

$$d_{\mathrm{gen}} = 1 - e_{\mathrm{gen}}^{\top} e_{\mathrm{ref}}, \quad d_{\mathrm{real}} = 1 - e_{\mathrm{real}}^{\top} e_{\mathrm{ref}}, \quad \Delta_d = d_{\mathrm{gen}} - d_{\mathrm{real}}. \tag{3}$$

Table 11: **Reference distance relative to matched real speech.** FiLM at $\alpha = 1$, nine speakers per language. Distances are $1 - \cos$ to the same neutral enrollment; real targets match speaker and query text. Gap is $d_{\text{gen}} - d_{\text{real}}$. The 98.75% marginal speaker-bootstrap intervals use Bonferroni adjustment for four comparisons. Negative gaps describe mean distance, not identity preservation.

Lang.	Interface	d_{gen}	d_{real}	Gap [98.75% interval]
ZH	Zero-shot	0.1759	0.5579	-0.3820 [-0.4271, -0.3333]
ZH	Instruct2	0.3551	0.5579	-0.2028 [-0.2570, -0.1477]
EN	Zero-shot	0.2682	0.4060	-0.1378 [-0.1742, -0.0784]
EN	Instruct2	0.2628	0.4060	-0.1432 [-0.1861, -0.0976]

The comparison matches speaker, query text and reference, not frame timing or emotional intensity. The enrollment text differs from the query in both comparisons; this is not the historical same-text neutral/angry pair construction in Figure 1. The real endpoint is evaluator-only and is not fed into the main editor. Across conditions there are 72 distinct real target recordings and 54 distinct enrollment recordings. Reuse across references and interfaces does not create additional people or recordings.

Aggregation and inference. References are averaged within query, queries within speaker–emotion, four emotions equally within speaker, and speakers equally within language. Paired bootstrap resamples the nine speakers within each language 50,000 times, using seed 20260924 (ZH) and 20260925 (EN) and preserving matched comparisons. The four main FiLM language–interface comparisons use Bonferroni-adjusted 98.75% marginal percentile intervals, corresponding to an approximate 95% family coverage target. These address mean distance differences, not tolerance intervals for individual utterances. Analysis choices were recorded before calculating this comparison, but after the earlier objective and listening outcomes had been examined; the analysis is post hoc, not preregistered. A gap interval crossing zero would be inconclusive rather than evidence of equivalence.

Table 11 reports the four FiLM comparisons underlying the mean-distance columns in Table 1. The full control-arm results and empirical coverage diagnostics follow below.

Empirical within-speaker distance distributions. For each speaker, requested emotion and enrollment clip, we also compute distances to all available real utterance embeddings in the same speaker–emotion pool. Their empirical 5th, 50th and 95th percentiles describe observed variation. No outcome-based sample selection is applied. Twelve of the 216 conditions use pools containing fewer than 20 emotional recordings, so their tail quantiles are particularly uncertain. Other-neutral distance excludes the exact enrollment clip; its speaker-balanced mean is 0.2531 in Chinese and 0.2851 in English. This supplies context for recording/content variation within neutral speech, not a pure causal decomposition of emotion.

Table 12 reports all four methods, their total reference distance and their extra distance relative to unedited synthesis. A small extra distance must not be substituted for the total generated-to-reference distance. FiLM lies at or below its corresponding empirical real P95 for 429/432 conditions: 100% in both Chinese interfaces, 98.1% in English zero-shot and 99.1% in English `instruct2`. These are descriptive coverage proportions. They neither put every sample within a natural range nor constitute calibrated identity-safety thresholds. Unedited and shared controls also mostly lie below these upper quantiles; this observation is not a FiLM-specific advantage.

A distinct closed-set attribution diagnostic. For speaker s, let c_s be the unit-normalized average of that speaker's distinct enrollment embeddings. We compute $m = e_{\text{gen}}^{\top} c_s - \max_{t \neq s} e_{\text{gen}}^{\top} c_t$ over the other eight enrolled speakers of the same language. Positive m means that this output is closer to its own centroid than any competitor in this fixed pool. This is not an independent verification model, an open-set identification rate or a perceptual judgment. The enrollment clips also condition synthesis, so the diagnostic does not evaluate generalization to unseen enrollment sessions.

FiLM retains positive margins on 431/432 conditions, with no additional failures relative to unedited synthesis. The sole failure is an English zero-shot angry condition whose margin changes from -0.1156 to -0.0812; it fails at both endpoints. Mean margins decrease in all four groups despite unchanged positive-margin counts. Thus unchanged closed-set attribution does not mean the

Table 12: All-arm reference-distance diagnostics on the objective panel. Extra d is edited minus matched unedited distance. P95 coverage uses each condition's own speaker–emotion/reference real distribution, not a pooled cross-language threshold.

Lang.	Interface	Method	Total d	Extra d	$\leq$ real P95
ZH	Zero-shot	Unedited	0.1603	+0.0000	100.0%
ZH	Zero-shot	Shared	0.1581	-0.0021	100.0%
ZH	Zero-shot	FiLM	0.1759	+0.0156	100.0%
ZH	Zero-shot	Norm-matched FiLM	0.1594	-0.0008	100.0%
ZH	Instruct2	Unedited	0.3164	+0.0000	100.0%
ZH	Instruct2	Shared	0.3248	+0.0084	100.0%
ZH	Instruct2	FiLM	0.3551	+0.0387	100.0%
ZH	Instruct2	Norm-matched FiLM	0.3239	+0.0075	100.0%
EN	Zero-shot	Unedited	0.2607	+0.0000	97.2%
EN	Zero-shot	Shared	0.2649	+0.0042	97.2%
EN	Zero-shot	FiLM	0.2682	+0.0075	98.1%
EN	Zero-shot	Norm-matched FiLM	0.2667	+0.0060	98.1%
EN	Instruct2	Unedited	0.2415	+0.0000	99.1%
EN	Instruct2	Shared	0.2488	+0.0074	99.1%
EN	Instruct2	FiLM	0.2628	+0.0213	99.1%
EN	Instruct2	Norm-matched FiLM	0.2508	+0.0094	100.0%

Table 13: Closed-set CAM++ attribution at $\alpha = 1$. Each block has 108 output conditions from nine speakers. New errors are matched transitions from positive unedited margin to nonpositive FiLM margin, not differences between rounded group rates.

		Mean margin		Positive margin		New
Lang.	Interface	Unedited	FiLM	Unedited	FiLM	errors
ZH	Zero-shot	0.4874	0.4793	108/108	108/108	0
ZH	Instruct2	0.3714	0.3447	108/108	108/108	0
EN	Zero-shot	0.2614	0.2575	107/108	107/108	0
EN	Instruct2	0.2455	0.2315	108/108	108/108	0

speaker-related representation is unchanged. Nor can distances or margins at $\alpha = 1$ establish S-MOS non-inferiority at $\alpha = 5$.

Reproducibility audit and remaining scope. The audit checks retained case-manifest and real-embedding fingerprints, complete 1,728-record coverage, baseline deltas, contract signatures and cached content digests. It rechecks 54 enrollment embeddings and every generated cosine across 3,510 cache files. Maximum normalized reference-coordinate discrepancy is 2.03×10^{-6}, and recomputed generated cosine differs from the result table by at most 1.23×10^{-15}. The run checks feature caches rather than re-hashing every waveform or independently rerunning checkpoint inference. It does not independently verify the separate MOS audio-to-render mapping. ESD is acted speech, and matched real/generated emotional intensity is not established. The supported conclusion is smaller mean CAM++ reference distance and no additional errors in this fixed attribution pool, not universal identity preservation, equality with human expression, or an advantage over shared editing at matched constraints.

E Separate Calibrated ESD Extension and Ablations

This appendix concerns the separate 576-dimensional calibrated ESD extension, not the main 192-dimensional source-trained editor. Table 14 preserves the headline controlled-waveform comparison; the figures retain dose and conditioning analyses.

Inputs and experimental support. The controlled ESD editor concatenates the 192-dimensional neutral-render embedding x_i, native neutral condition $U_{N,s(i)} = e_{s(i)}$, and neutral calibration gap $g_{s(i)}$. The gap is the mean real-minus-rendered displacement from neutral support pairs. The target

Table 14: Separate calibrated ESD extension at $\alpha = 1$. CAM is target-speaker cosine gain, Emo target emotion-embedding cosine gain, Act activation-distance reduction, and Joint the seed-averaged automatic joint criterion. This is not the main FiLM model or a MOS comparison.

Editor	CAM gain	Emo gain	Label gain	Act gain	Joint (%)
Shared direction	.0178	.1225	.2003	-.1359	56.25
Calibrated full network	.0006	.0921	.1288	-.2509	32.81
No DSP	-.0001	.0948	.1312	-.2501	33.33
No activation	-.0001	.0941	.1273	-.2537	32.29
Emotion only	.0001	.0918	.1268	-.2486	33.33

embedding y_i supplies training supervision, not a waveform feedback gradient. At evaluation this branch retains target-derived H and uses target-derived acoustic activation requests, so it is not a single-reference, target-free system. The source-pretrained model has different inputs and supervision (Appendix F). Training endpoints $x_i + g_{s(i)} + \hat{r}_i$ must not be confused with normalized synthesis assignments $U_\alpha = \mathcal{N}(U_{N,s(i)} + \alpha\hat{r}_i)$.

Acoustic conditions. The 13-dimensional condition contains four emotion indicators, two continuous speaker-relative pitch/energy values, two activation-band indicators, and five neutral-enrollment DSP descriptors. For descriptor f_k, the requested acoustic component is $a_k = \mathrm{clip}((f_k - \mu_{s,k})/\sigma_{s,k}, -5, 5)$, with neutral-reference robust location $\mu_{s,k}$ and scale $\sigma_{s,k}$ including scale floors. Activation is the mean of the two components; an outer-source-speaker threshold defines the band. The five descriptors are median log fundamental frequency, active energy, pitch interquartile range, voiced fraction and periodicity. They describe neutral enrollment recordings, not a time-varying trajectory or psychological arousal ground truth.

Architecture and optimization. The 576-dimensional standardized voice input is mapped to width 256 by a linear layer and GELU. Two residual blocks each precede condition-dependent FiLM, with multiplicative factor $1 + 0.1\tanh(\gamma(c_i))$ and additive term $0.1\beta(c_i)$. Here γ and β are learned condition-to-feature mappings producing the per-feature scale and shift coefficients, respectively. Layer normalization and a linear projection output 192 residual dimensions. Input and condition standardization uses training statistics. AdamW uses learning rate 10^{-3}, weight decay 0.01, dropout 0.1, at most 200 epochs and patience 25. An inner source-speaker validation set selects the epoch count before refitting on all outer-source speakers. Output and modulation layers are zero-initialized, so checkpoint selection begins at epoch two. Test-speaker target outcomes do not select checkpoints. The activation threshold includes inner-validation speakers among outer-source data, so it is not fit-only preprocessing. Three seeds are evaluated.

Controls and aggregation. Five variants comprise the full network, no DSP, no activation, DSP-only conditioning (removing emotion and activation labels), and emotion-only conditioning. All keep the same main voice inputs and architecture; removing explicit DSP does not remove acoustic information carried by embeddings. A wrong-request diagnostic negates pitch/energy components and recomputes the band, which need not swap if the threshold is nonzero. Additional controls include legacy additive and pooled directions, a norm-matched random direction, and a separately identified endpoint-interpolation variant. The latter is not direct residual injection. The 16 outer speaker folds, five variants and three seeds give 240 fits; 64 selected queries generate 3,200 records, not 3,200 independent queries. Seeds are averaged within query before speaker aggregation and bootstrap inference. Primary one-sided superiority tests share the original 200-comparison Holm correction family; intervals are pointwise and conditional on fitted folds.

Interpretation of the full-network gains. At $\alpha = 1$, target emotion-cosine gain is 0.0921 (95% interval [0.0637, 0.1201]; Holm $p = 0.0122$), and target-distribution JS distance reduces by 0.1139 (Holm $p = 0.0334$). Label-probability gain is 0.1288 but corrected $p = 0.2209$. Target-speaker cosine gain is 0.0006, interval [-0.0143, 0.0154]. At $\alpha = 0.5$, speaker gain is 0.0108, interval [0.0038, 0.0182], emotion gain 0.0486, and joint success 36.46%, against 32.81% at $\alpha = 1$. The lower-strength window is descriptive, not a validated deployment policy.

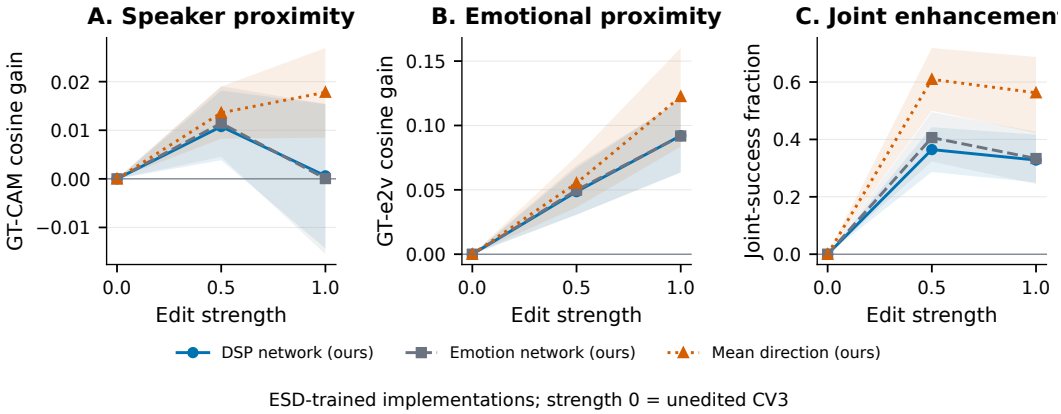


Figure 3: **Emotion and joint gains need not increase together.** Separate calibrated ESD extension, with pointwise 95% speaker-bootstrap intervals after within-query seed averaging. DSP-network and emotion-network denote calibrated variants, not the main source-trained FiLM; their comparator is the ESD pooled residual fitted under this extension's protocol. Joint success requires positive target-speaker cosine, emotion cosine, label-probability and JS gains plus positive margin on the same waveform; activation is excluded. Only the displayed strengths were tested.

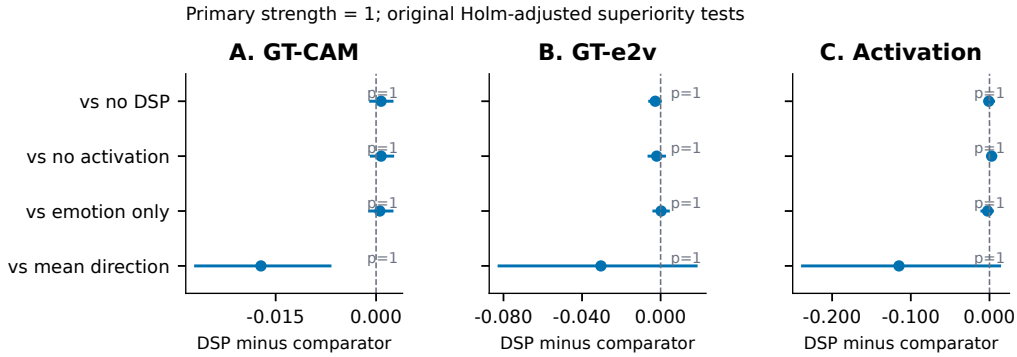


Figure 4: **Explicit conditions do not establish an incremental advantage.** Paired full-network-minus-comparator differences in the separate calibrated ESD extension at $\alpha = 1$, averaged across emotions, with pointwise speaker-bootstrap intervals. Displayed adjusted tests retain the original 200-test correction. Absence of superiority is not equivalence. The pooled direction is a method comparator; unedited CV3 is the system baseline.

Corrected comparisons against no-DSP, no-activation and emotion-only variants do not establish additional benefit. The cached sensitivity analysis in Appendix E.1 shows small but nonzero request responses; it does not establish correct activation control. Activation-distance gain at $\alpha = 1$ is -0.2509, with interval [-0.4099, -0.1056]. Therefore emotional scores alone cannot be read as complete acoustic-request compliance.

E.1 Cached Request-Sensitivity Audit

The audit covers 48 full-model prediction caches, 1,518 unique held-out queries and 4,554 query–seed predictions. Negating the requested pitch/energy components and recomputing their activation band changes the predicted residual by a speaker-equal mean relative L_2 distance of 0.0110, with mean correct-versus-wrong residual cosine 0.999889. The perturbation therefore produces small but nonzero changes in the predicted edit, which rules out an identically disconnected condition path

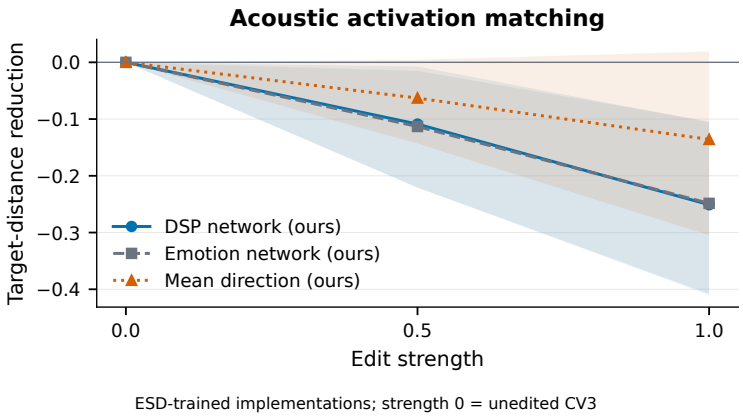


Figure 5: Requested acoustic-activation matching in the separate calibrated ESD extension. Positive target-distance reduction means closer pitch/energy proxies. Shading denotes pointwise speaker-bootstrap intervals after seed averaging; activation does not enter the joint-success metric.

but does not establish that the path implements correct control. Two confounds remain: reflected requests may fall outside the training distribution, and these cached queries are a broader set than the 64 waveform queries. In the original waveform experiment the correct-versus-wrong activation-distance gain differs by -0.00071 (pointwise 95% interval $[-0.00804, 0.00584]$; original Holm $p = 1$). No additional superiority tests or training runs were introduced by this audit.

F Source Training and Training-Domain Diagnostics

Pretraining configuration and learning curve. We train width-256, two-block concat and FiLM editors, each with three seeds, using cosine endpoint and residual-scale-normalized MSE supervision. AdamW uses learning rate 3×10^{-4} and weight decay 0.01, for up to 60 epochs with patience 10; validation source-key macro cosine selects checkpoints. Pooled directions use the corresponding training subset. Evaluation averages within source-key/emotion cells and then equally across source keys. The seed-71 FiLM model reaches endpoint cosine 0.90291 on 15,550 training pairs and 0.90644 on 46,643 pairs using nested subsets; these learning-curve points have one seed and are not independent confirmations. These embedding endpoints must be separated from the waveform experiments: the training-domain diagnostic below evaluates the pretrained models through frozen CV3 without claiming that the synthetic embedding-test gain transfers unchanged to audio, and the earlier ESD transfer run uses a different domain and is not the basis for selecting the source-domain alpha grid.

Source-trained editing transfers within the neutral-reference setting. The objective is to learn an editor from synthetic-neutral/target pairs and apply it to real neutral references without target-domain emotional-pair training. This tests source-to-real neutral-reference transfer, not conversion from arbitrary source emotions. Our source mean-direction implementation, fitted on the same source training split, is an internal comparator for the learned editor. The system baseline is matched unedited CV3, not the mean-direction implementation. A direction fitted on other speakers from a target corpus has access to target-domain supervision and is a separate adaptation reference, not an equal-information zero-shot competitor. Explicit DSP features remain a candidate input, not an already demonstrated prerequisite for generalization.

Training-domain observation set. For each query in the reported prefix, six pretrained checkpoints (concat and FiLM, three seeds each) and the source pooled direction are applied at $\alpha \in \{1, 1.25, 1.5\}$, with one unedited baseline. Each query retains the same neutral-reference plan H, synthesis seed and remaining conditions across U assignments. Checkpoints are rendered separately

Table 15: Source-training waveform response on a fixed diagnostic subset: 4,096 queries and 90,112 renderings from the first 16 synthesis batches. Values are descriptive gains over a matched unedited baseline. Label gains are percentage points (pp); reference gains are cosine differences, not identity-retention rates. All listed methods are our source-trained editing implementations; measurements from separately rendered seeds are averaged within query. Bold marks the largest gain among the listed methods at the same α, not significance or improvement on every metric. These queries were used in pretraining; this is not held-out generalization evidence.

Method	α	Label gain (pp)	Reference cosine gain
Pretrained FiLM	1	**+11.45**	-0.0094
Pretrained FiLM	1.25	**+14.68**	-0.0153
Pretrained FiLM	1.5	**+17.81**	-0.0225
Pretrained concat	1	+11.33	-0.0096
Pretrained concat	1.25	+14.51	-0.0155
Pretrained concat	1.5	+17.47	-0.0229
Shared residual	1	+6.06	**-0.0029**
Shared residual	1.25	+7.73	**-0.0043**
Shared residual	1.5	+9.35	**-0.0061**

and their measurements averaged, unlike the main experiment's pre-synthesis residual ensemble. The source editor has no target-derived activation request.

The fixed diagnostic subset comprises batches 0–15: 4,096 queries, 3,605 source keys and 90,112 waveforms (Table 15). Each of the four emotions contributes 1,024 queries. All subset queries belong to the original pretraining training split. The subset was selected for inspection after synthesis began; it is not a preregistered sequential significance test. We average seeds within query, queries within source-key/emotion, available emotions within source key, and finally source keys equally. No query is filtered by its measured score.

A positive source response with a reference-proximity trade-off. At $\alpha = 1$, FiLM increases target-label probability by 11.45 percentage points over unedited synthesis, versus 6.06 points for source pooling. At $\alpha = 1.25$ and 1.5, its gains are 14.68 and 17.81 points, versus 7.73 and 9.35 for pooling. Both network architectures show larger mean label gains than pooling at every tested alpha. Within each emotion, the FiLM mean gain also exceeds the pooled mean at the three tested strengths; the sadness response is smaller than the other categories. However, the network produces larger reference-cosine decreases. These observations establish neither superior identity retention nor a direction-only advantage: equal alpha does not equalize predicted residual magnitude. They are descriptive training-domain waveform results, not significant held-out improvements, MOS evidence, or external generalization.

G Complete Text-Driven Development Dose Grid

This earlier two-speaker study reports all tested strengths, not independently selected deployment doses. It is distinct from the 18-speaker panel in Table 5. Mean direction and FiLM variants are our source-trained editing implementations; Unedited CV3 retains the same native interface without the additional edit. Bold marks the best displayed value per metric across all listed methods and doses within each speaker–interface block, including the unedited baseline and ties (larger e2v/P(target)/CAM, smaller error). These are descriptive grid maxima/minima, not significance or validated optimal doses. No cross-language error average is formed.

H Reference- and Emotion-Dependent Operating Boundaries

Emotional response is not sufficient for a usable edit. The preceding waveform study tests whether the frozen interface can enhance emotion. We now examine where that response incurs identity costs. Direction and magnitude both determine the supplied condition; neither is assumed irrelevant or universal. The following historical sweeps characterize variation in measured dose re-

Table 16: ZH development speaker: all methods and doses. Error is mean utterance CER in percent.

Interface	Method	α	e2v	P(target)	CAM	Error
Zero-shot	Unedited CV3	0	0.3111	0.0000	0.8175	2.60
Zero-shot	Mean direction	0.5	0.3163	0.0001	**0.8179**	3.79
Zero-shot	Mean direction	1	0.3359	0.0003	0.8167	3.07
Zero-shot	Mean direction	1.5	0.3515	0.0020	0.8132	3.07
Zero-shot	FiLM	0.5	0.3441	0.0103	0.8158	3.63
Zero-shot	FiLM	1	0.3886	0.0637	0.8029	2.19
Zero-shot	FiLM	1.5	**0.4503**	**0.2168**	0.7802	**2.11**
Zero-shot	FiLM (norm-matched)	0.5	0.3179	0.0002	0.8176	3.47
Zero-shot	FiLM (norm-matched)	1	0.3232	0.0003	0.8169	3.31
Zero-shot	FiLM (norm-matched)	1.5	0.3388	0.0072	0.8159	3.08
Instruct2	Unedited CV3	0	0.5547	0.2535	**0.6552**	4.76
Instruct2	Mean direction	0.5	0.5702	0.2760	0.6497	4.88
Instruct2	Mean direction	1	0.5746	0.2876	0.6441	5.30
Instruct2	Mean direction	1.5	0.5878	0.3412	0.6369	5.04
Instruct2	FiLM	0.5	0.5839	0.2985	0.6413	5.65
Instruct2	FiLM	1	0.6119	0.3533	0.6196	**4.42**
Instruct2	FiLM	1.5	**0.6350**	**0.3816**	0.5842	4.45
Instruct2	FiLM (norm-matched)	0.5	0.5622	0.2571	0.6535	5.30
Instruct2	FiLM (norm-matched)	1	0.5707	0.2754	0.6483	5.27
Instruct2	FiLM (norm-matched)	1.5	0.5797	0.2874	0.6432	5.43

Table 17: EN development speaker: all methods and doses. Error is mean utterance WER in percent.

Interface	Method	α	e2v	P(target)	CAM	Error
Zero-shot	Unedited CV3	0	0.3829	0.0000	0.7894	3.44
Zero-shot	Mean direction	0.5	0.3849	0.0000	0.7891	**2.99**
Zero-shot	Mean direction	1	0.3890	0.0001	0.7865	3.44
Zero-shot	Mean direction	1.5	0.3949	0.0069	0.7880	3.44
Zero-shot	FiLM	0.5	0.3903	0.0072	0.7889	3.44
Zero-shot	FiLM	1	0.4020	0.0169	0.7848	3.44
Zero-shot	FiLM	1.5	**0.4173**	**0.0377**	0.7728	3.44
Zero-shot	FiLM (norm-matched)	0.5	0.3846	0.0001	0.7890	3.44
Zero-shot	FiLM (norm-matched)	1	0.3904	0.0044	**0.7903**	3.44
Zero-shot	FiLM (norm-matched)	1.5	0.3955	0.0075	0.7876	3.44
Instruct2	Unedited CV3	0	0.5066	0.1591	**0.7930**	3.20
Instruct2	Mean direction	0.5	0.5116	0.1766	0.7923	2.94
Instruct2	Mean direction	1	0.5122	0.1815	0.7845	**2.41**
Instruct2	Mean direction	1.5	0.5143	0.1963	0.7799	2.87
Instruct2	FiLM	0.5	0.5224	0.2134	0.7821	2.60
Instruct2	FiLM	1	0.5308	0.2101	0.7683	2.54
Instruct2	FiLM	1.5	**0.5410**	**0.2319**	0.7513	2.68
Instruct2	FiLM (norm-matched)	0.5	0.5160	0.1883	0.7879	2.71
Instruct2	FiLM (norm-matched)	1	0.5186	0.2152	0.7844	2.95
Instruct2	FiLM (norm-matched)	1.5	0.5261	0.2160	0.7774	3.21

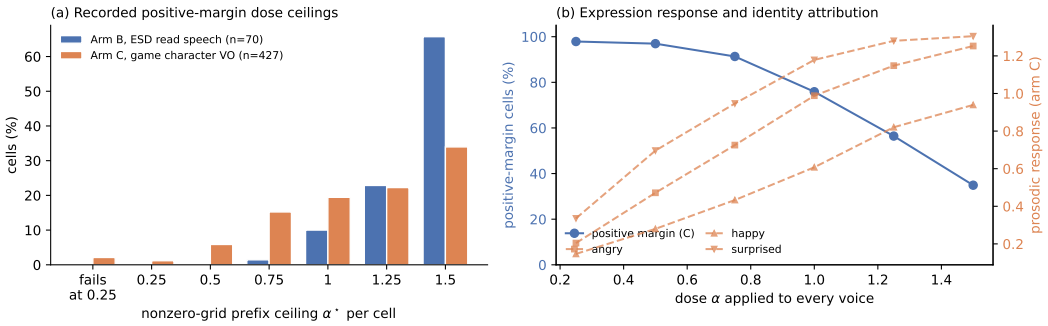


Figure 6: Historical angular-direction dose curves, not the source-trained FiLM dose study. Recorded positive-margin prefix ceilings use 70 ESD and 427 baseline-gated character-voice cells. The character positive-margin rate decreases as angry, happy and surprised acoustic projections increase. The sadness response is described in the affect-specific analysis below. Different competitor pools and dose definitions preclude direct comparison with the main experiment.

sponses, not a learned dose-selection policy. In particular, alpha values in different representations and editing rules are not calibrated to a common perceptual strength.

Historical angular dose rule. These B/C sweeps do not use the neural editor's additive dose directly. Let d_a be a pooled emotion direction, $u_{s,a} = \mathcal{N}(d_a - (e_s^\top d_a)e_s)$ its unit tangent direction at the enrollment vector, and β_s the reference angle stored in the experiment contract. The historical condition follows the great-circle step $e_s(\alpha) = \cos(\alpha\beta_s)e_s + \sin(\alpha\beta_s)u_{s,a}$. A control direction uses the same angular displacement. This geometric calibration controls perturbation angle, not perceptual intensity or emotion purity. The network experiment instead uses Eq. 1; alpha is not a common dose scale between these protocols.

Read speech and character voices. Arm B uses ESD: 18 speakers, leave-one-speaker-out directions, and 70 speaker–affect cells. Arm C draws on 253 qualifying characters voiced by 53 actors. Directions are estimated from within-character neutral/affect aggregates with *actor*-level donor exclusion; the donor audit flags no leakage among 547 planned cells. Of 522 scored character–affect cells, 427 pass the pitch-based baseline precondition G_0 and form the primary response cohort; the full cohort supplies sensitivity results. Both sweeps retain fixed speech tokens while varying the speaker condition over $\alpha \in \{0, 0.25, \ldots, 1.5\}$. A positive identity margin means that the output is closer to its own neutral centroid than to the nearest competing centroid in a fixed affect-specific pool. Pools contain 16–17 competitors in B and 79–161 in C. Pool composition can change both margin size and sign; cross-arm differences are descriptive, not an isolated causal effect of corpus type. These are role-level outcomes in C, not independent samples of 253 people.

Figure 6 summarizes the primary cohort. In the full 522-cell sensitivity cohort, prefix-ceiling bin shares are 2.7/1.1/6.7/15.1/19.9/21.1/33.3%; the unequal competitor pools remain unchanged across doses.

Descriptive dose ceilings. For continuity with the recorded analysis, $\alpha^\star$ denotes the largest nonzero grid value with positive margins at that value and all preceding *nonzero* values; it is coded as zero when the first dose fails. This statistic does not check the zero-dose baseline and is not a safe deployment policy. On B, 65.7% of cells have $\alpha^\star = 1.5$ under CAM++; on the C primary cohort, 2.1% fail at the first dose, a further 7.0% have $0 < \alpha^\star \leq 0.5$, and 34.0% reach 1.5. Already at zero dose, only 419/427 C cells have positive CAM++ margins. Consequently, returning to zero cannot be assumed to provide 100% identity retention, and no oracle safety guarantee is reported.

A measured emotion–identity trade-off. At $\alpha = 0.5$, 97.0% of C primary-cohort cells retain a positive CAM++ margin. Angry, happy, and surprised prosodic projection scores are respectively 41.1%, 34.1%, and 54.4% of their values at 1.25 ($+0.4718$ versus $+1.1480$, $+0.2799$ versus $+0.8204$, and $+0.6962$ versus $+1.2801$). At 1.25, the positive-margin rate falls to 56.4%. These

Table 18: Positive identity-margin rates at measured doses. C uses the 427-cell G_0-passed cohort; B uses all 70 cells. WeSp. denotes the additional WeSpeaker verifier (Section H.1). These are cell-weighted descriptive percentages, not actor-level confidence intervals, perceptual identity scores, or proof of a safe dose-selection policy. The zero-dose row exposes pre-existing attribution failures.

	C: character voices		B: ESD	
Dose α	CAM++	WeSp.	CAM++	WeSp.
0 (unedited)	98.1%	82.0%	100.0%	100.0%
0.5	97.0%	78.5%	100.0%	100.0%
1.0	75.9%	55.0%	98.6%	92.9%
1.25	56.4%	37.9%	88.6%	71.4%
1.5	34.9%	19.2%	65.7%	47.1%

are responses at measured doses, not fractions of an established maximum achievable emotion or a calibrated cross-system intensity scale. Table 18 reports both verifiers; choosing a deployment point additionally requires explicit identity, content, and quality tolerances.

H.1 Identity costs under an additional speaker verifier

We re-encode the existing B/C waveforms with a verifier outside the edited CAM++ representation, rebuilding reference centroids in the same new space. This changes the identity readout, not the synthesis intervention. The first evaluated model, WavLM-base-plus-sv, gives a negative C mean margin on both real neutral recordings (-0.0221) and unedited synthesis (-0.0238). This limits its usefulness for this role-attribution criterion, but does not invalidate adverse results or identify whether the cause is the model, implementation, or data. After inspecting that result, we set a real-recording screening rule for subsequent candidates: mean margin at least 0.08 and at least 90% of identities with positive mean margins. The rule is calibrated from CAM++ on the same development panel, not an independent or universally calibrated verifier standard (Appendix I).

WeSpeaker CN-Celeb ResNet34, with 256-dimensional embeddings, meets that rule. The checkpoint identifier is `Wespeaker/wespeaker-cnceleb-resnet34-LM`; the evaluated WavLM identifier is `microsoft/wavlm-base-plus-sv`. For WeSpeaker, the real-recording results are: C real-recording mean margin is $+0.1039$ with 95.3% of identities positive, versus CAM++ $+0.1082$ and 96.0%; B gives $+0.3545$ and 100.0%. This is a different encoder architecture and checkpoint; we have not established that its training identities are disjoint from CAM++ or the evaluation corpora. The recorded nonzero-grid prefix ceiling has mean 1.275 on B and 0.821 on the C primary cohort. Those are descriptive dose summaries, not learned-policy results. The primary C cohort already has only 350/427 positive margins at zero dose (82.0%), falling to 335/427 at 0.5 (78.5%) and 162/427 at 1.25 (37.9%). Thus baseline attribution errors and additional editing costs both matter; a positive aggregate baseline does not remove the former.

In the separate full-coverage G4 reports, B passes all four emotion-specific mean-margin interval checks at $\alpha = 1$. C passes none at $\alpha = 1.25$, with mean margins from -0.0103 to -0.0386 under WeSpeaker, compared with two passes under CAM++ in the original role-level report. Thus, evidence for positive emotion-specific mean attribution margins at this C dose is not consistent across verifiers. Both results are retained as evidence of readout sensitivity. Automatic margins do not establish perceptual identity, and the historical actor-weighted analysis remains separate from these role-level gate reports.

Magnitude matters, but adaptation remains to be demonstrated. The norm-matched ablation in Section B reduces the learned editor's mean emotion-cosine gains from 0.0278–0.1043 to 0.0033–0.0283. This shows sensitivity to residual magnitude, not that the network has learned an optimal per-reference dose. Together with the pooled comparator's higher joint-success rate in the separate calibrated ESD extension (Table 14), this motivates evaluating direction and magnitude jointly under explicit identity and quality constraints.

Affect-specific acoustic responses. On C, the sad-direction prosodic projection is slightly positive at $\alpha = 0.25$ ($+0.0104$) and negative at the tested doses from 0.5 to 1.5 (-0.0287 at 0.5

and -0.1147 at 1.25; dose rank correlation -0.9643). Angry, happy, and surprised have positive target-minus-control prosodic contrasts at 1.25 ($+0.7947$, $+0.3843$, $+0.9637$). A post hoc descriptor decomposition shows positive sad-direction projections in voice quality ($+1.1188$) and envelope ($+2.3064$) at that dose despite the negative prosodic projection. Thus acoustic effects differ across descriptor blocks; this does not establish a disentangled representation, a universal failure of speaker-independent directions, or successful perceived sadness. Sections 6 and A.1 delimit these observations.

I Verifier Admission: Both Candidates in Full

The real-recording diagnostic uses leave-one-recording-out scoring: a neutral clip is compared with its own remaining-reference centroid and the closest competing centroid within the affect block. CAM++ is evaluated on the same file lists. Reference-set size, text, duration, and probe multiplicity differ from the synthetic-baseline comparison, so this is not a controlled intervention on audio domain. Clips shared across affect blocks contribute repeated probe records; 5,157 C records are not 5,157 distinct recordings or independent speakers. After the WavLM evaluation, thresholds were fixed for screening subsequent candidates using C real recordings: separation ≥ 0.08 and $\geq 90\%$ of identities with a positive mean margin. They are development heuristics derived from CAM++, not a universal or cross-encoder calibrated standard. The WavLM rows in Table 19 show how its recorded values compare with that later rule, not a prospectively screened experiment.

Table 19: Verifier-admission measurements on real recordings. Separation is mean cosine to the own reference minus mean cosine to the nearest competing reference. Pos. ids is the share of identities whose mean margin is positive. The text below separately compares WavLM real-recording and unedited-synthesis measurements; the synthetic probes are not rows in this table.

Arm	Encoder	Probes	Ids	Separation	Pos. ids	Admitted
C	WeSpeaker CN-Celeb ResNet34	5,157	253	$+0.1039$	95.3%	yes
C	CAM++ (positive control)	5,157	253	$+0.1082$	96.0%	—
C	WavLM-base-plus-sv	5,157	253	-0.0221	10.7%	no
B	WeSpeaker CN-Celeb ResNet34	280	18	$+0.3545$	100.0%	yes
B	CAM++ (positive control)	280	18	$+0.3733$	100.0%	—
B	WavLM-base-plus-sv	280	18	$+0.0439$	100.0%	no

What the diagnostic establishes. For the evaluated WavLM pipeline, C mean margin is -0.0221 on real recordings and -0.0238 on unedited synthesis; B gives $+0.0439$ and $+0.0411$. Poor C attribution is therefore not confined to the synthetic probes. These aggregate similarities do not establish equivalence between domains or rule out a domain effect. They also do not diagnose a root cause: front-end and inference behavior, recording variation, and the definition of role identity remain possible contributors. WavLM-base-plus-sv is a speaker-verification model built on a self-supervised backbone, not an unadapted general-purpose speech representation.

Interpretation limits. CAM++ also gives smaller C than B real-recording margins ($+0.1082$ versus $+0.3733$), but a ratio of margins is not a ratio of corpus difficulty, and margins from different encoders are not calibrated to a shared scale. Competitor pools, actor/role structure, languages, and recording conditions differ. The diagnostic motivates reporting multiple identity measures. The unfavorable verifier result remains part of the evidence, and these comparisons do not isolate corpus characteristics as the cause of the tighter boundary. Section H.1 reports the editing outcomes with these limitations.

J From an Editable Interface to a Learned Editor

Earlier direction-editing experiments examine expressive response, closed-set speaker attribution and audibility. They motivate the source-trained main editor but evaluate different editing rules and data populations. Table 4 separates them from the main model and the calibrated ESD extension.

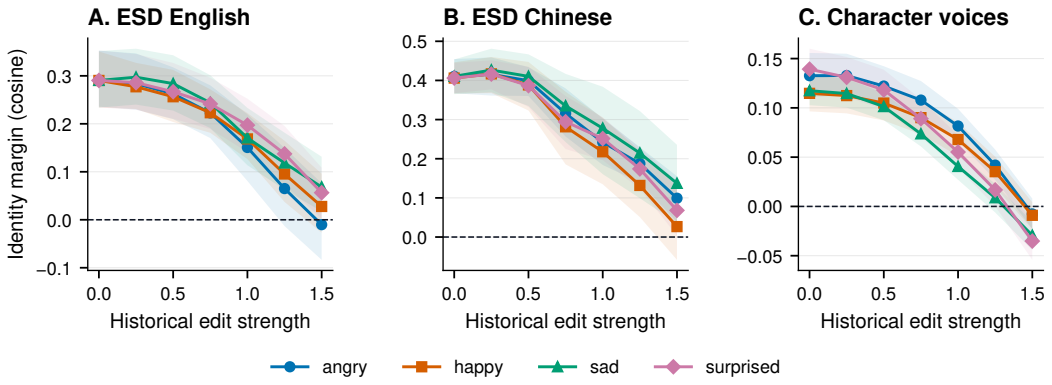


Figure 7: **Historical direction edits exhibit dose-dependent attribution margins.** Each curve shows mean own-reference cosine minus nearest-competitor cosine. A positive mean does not require every output to have positive margin. Shading is a pointwise 95% cluster-bootstrap interval, not a simultaneous band. Panels A/B use G0-passed ESD targets and speaker clustering; panel C uses G0-passed character voices and actor clustering. Competitor pools differ across panels, so absolute heights are not corpus rankings. These are historical direction edits, not outputs of either neural editor.

J.1 Data lineage across the experimental arms

B uses ESD real parallel emotional recordings, with English and Chinese as language strata of one corpus. C uses within-role neutral/affect aggregates with actor-level donor exclusion, not same-sentence emotional pairs. B/C directions exclude the target speaker/actor from donor estimation. D constructs synthetic neutral/angry pairs with dots and then tests their directions in frozen CV3. Its source recordings are drawn from Emilia, so D is not an independent real-emotional-speech corpus. The earlier Emilia analysis (A) is excluded from the reported result aggregation. The 576-dimensional calibrated ESD extension is a B-derived construction, distinct from D and from the 192-dimensional source-trained main editor.

D's angry execution study contains 39 queries and 351 CV3 waveforms. It uses the query's dots neutral embedding as the editing origin and target real-speech tokens held fixed across conditions. The wrong-direction control is another speaker's angry direction, not another emotion. The reported G1–G4 checks pass, but the personalized-over-pooled comparison P0 does not establish a personalized synthesis advantage. Equivalent execution evidence is not available for happy and surprise, and the related sadness experiment did not establish the intended response.

J.2 Intervention logic and the identity boundary

Earlier direction-editing experiments hold the speech-token plan H fixed while changing only the realization condition U. The B/C target-affect and different-affect directions are compared at the same angular distance to distinguish direction-dependent response from a generic displacement effect; D instead uses the same-emotion, different-speaker control described above. This controls the intervention geometry, not perceived strength. Acoustic projection progress, emotion-model cosine, and identity margin are separate endpoints: one cannot stand in for another.

Figure 7 carries forward the audited identity analysis on ESD and character voices. The primary historical population retains targets passing the pitch-based baseline reliability gate (G_0), not an identity gate; the competitor pool is left unchanged rather than simplified after filtering. ESD uncertainty is clustered by speaker. In the character-voice population, roles are first averaged within actor and uncertainty is clustered by actor. We do not treat multiple roles of one actor as independent people.

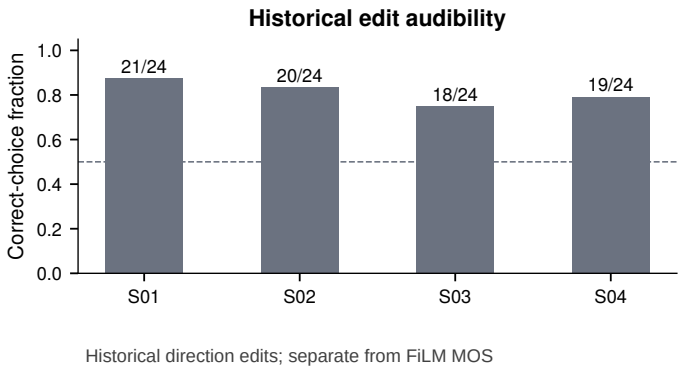


Figure 8: Historical direction-editing audibility, not the main FiLM MOS study. Four listeners each completed the same 24 trials, with correct counts 21, 20, 18 and 19. The dashed line denotes chance; listener-level one-sided sign-flip $p = 0.0625$.

At historical strength $\alpha = 1$, all four affect-specific mean-margin intervals are positive in both ESD languages. English margins range from 0.1507 to 0.1972; Chinese margins range from 0.2178 to 0.2778. Depending on affect, the G0 population contains nine English speakers and seven or eight Chinese speakers. These results identify a tested operating region with positive affect-specific mean CAM++ attribution margins, not perceptual identity preservation or positive margins for every output.

Character voices show smaller margins under their own fixed competitor pool at the tested doses; this is not a controlled corpus comparison. At $\alpha = 1.25$, actor-weighted margins are 0.04189 for angry and 0.03523 for happy, both with positive intervals. Sad is 0.00878 with interval [-0.00613, 0.02298]. Surprise is 0.01673 with lower bound 0.000119 in the G0 population, but its all-target sensitivity interval crosses zero. Thus the higher-strength result is population-sensitive and affect-dependent. The four affect groups contain 37–48 actors; they are not identical populations. Their newer actor-level identity results cannot be combined with older role-level acoustic intervals into an unqualified all-gates-passed claim.

J.3 Audibility as foundational, not network-level evidence

An earlier blinded study used the same 24 trials with four listeners. Correct choices were 21, 20, 18, and 19, totaling 78/96 (81.25%). Higher-dose trials yielded 43/56 (76.79%); direction-specificity trials yielded 35/40 (87.5%). These counts replace an earlier withdrawn listening batch and are not combined with it. The listener-level sign-flip test gives one-sided $p = 0.0625$, so the descriptive consistency is not a listener-population confirmation at the 0.05 level. Figure 8 shows listeners separately rather than presenting 96 judgments as 96 independent listeners.

This evidence motivates waveform evaluation of the speaker interface: edits can produce audible differences in the historical materials. It does not establish naturalness gains or perceptual superiority of either current neural editor. Acoustic quality changes are downstream effects of the intervention, not pre-treatment confounders. We report audibility of these materials without requiring expression to act independently of its acoustic realization.

J.4 Why prediction quality alone is insufficient

Earlier paired-embedding prediction experiments provide a second motivation. In the neutral-calibrated protocol, additive prediction reaches residual-direction cosine 0.3635 for leave-one-speaker-out evaluation, 0.3440 for English-to-Chinese transfer, and 0.2574 for Chinese-to-English transfer; pooled directions reach 0.1639, 0.1370, and 0.0868. The three planned additive-versus-pooled tests survive their correction. However, label increments over embedding-only prediction are small and do not survive correction, and a negative-input baseline exceeds additive prediction in Chinese-to-English transfer. These are residual-prediction scores, not generated-speech scores.

These observations motivated testing the entire prediction-to-synthesis chain. The calibrated extension in Appendix E and the separate main source-trained editor ask whether a better embedding prediction becomes a useful input edit after normalization and frozen synthesis. Their waveform results must be measured rather than inferred from prediction scores.

K WAV2VEC2-BERT BRANCH EDITING

The shared development panel contains 18 speakers, 72 queries and three reference conditions per query. Directions are added to every frame of the standardized layer-17 representation. Speaker and emotion reference calls are edited separately; CAM++ and reference mel are unchanged. Strengths are descriptive dose comparisons, not test-selected optima. Means average references, emotions and speakers hierarchically. Random controls use one deterministic norm-matched draw per case. Dependencies between leave-one-speaker-out folds preclude treating these observations as independent population evidence.

Table 20: Changes relative to unedited IndexTTS2 within each native-control setting. The mean-direction edits are our representation-specific implementations; random directions are diagnostic controls. Arm prefixes denote emotion-reference (emo), speaker-reference (spk), or both branches; a denotes strength. Error changes are percentage points: CER for ZH, WER for EN. CAM is reference similarity, not perceptual identity. Bold marks the best displayed value per column among the listed arm–strength combinations within each language–control block (larger e2v/CAM, smaller error), including ties. This descriptive comparison includes random controls, excludes the zero-change baseline, and does not establish a deployment optimum.

Language	Control	Arm	Δe2v	ΔCAM	ΔError
en	default	both_a1	+0.1980	-0.0116	**-0.37**
en	default	emo_a0.5	+0.0864	-0.0147	+0.00
en	default	emo_a1	+0.2074	-0.0240	+0.00
en	default	emo_a1.5	**+0.2846**	-0.0445	+0.21
en	default	emo_random_a1	+0.0036	-0.0087	+0.00
en	default	spk_a1	-0.0018	**+0.0019**	**-0.37**
en	emotion_vector	both_a1	+0.0508	-0.0125	+0.74
en	emotion_vector	emo_a0.5	+0.0303	-0.0142	+1.20
en	emotion_vector	emo_a1	+0.0555	-0.0254	+0.66
en	emotion_vector	emo_a1.5	**+0.0556**	-0.0207	+0.31
en	emotion_vector	emo_random_a1	-0.0315	**+0.0018**	**+0.00**
en	emotion_vector	spk_a1	-0.0139	-0.0011	+0.15
zh	default	both_a1	+0.1621	-0.0413	-0.28
zh	default	emo_a0.5	+0.0539	-0.0103	**-0.35**
zh	default	emo_a1	+0.1655	-0.0470	-0.26
zh	default	emo_a1.5	**+0.3663**	-0.1150	+0.59
zh	default	emo_random_a1	+0.0088	**+0.0000**	-0.26
zh	default	spk_a1	+0.0036	-0.0022	+0.05
zh	emotion_vector	both_a1	-0.0033	-0.0204	**+0.47**
zh	emotion_vector	emo_a0.5	+0.0046	-0.0123	+1.27
zh	emotion_vector	emo_a1	+0.0090	-0.0265	+1.73
zh	emotion_vector	emo_a1.5	**+0.0128**	-0.0376	+0.99
zh	emotion_vector	emo_random_a1	-0.0036	-0.0003	+0.84
zh	emotion_vector	spk_a1	-0.0198	**+0.0075**	+1.27

This representation-specific experiment reports automatic evaluation only. Naturalness, perceptual speaker similarity and requested-emotion correspondence require human ratings; no MOS, human preference or perceptual non-inferiority result is reported for this branch.

L Planning–Realization Separation and Native-Interface Reuse

L.1 Conceptual organization and control pathways

The motivation in Figure 2 distinguishes the organization of an utterance from its acoustic realization. Communicative context and function shape prosodic delivery; observed acoustics also depend on speaker traits and expression-related phonation (Scherer, 2003). This is a conceptual organization, not a fitted causal model of human speech. In particular, the figure does not identify identity and emotion as statistically independent latent factors, and its contextual variables are not additional inputs to the editor.

CV3 supplies a useful computational counterpart: the language model produces a speech-token plan H, while the acoustic model also consumes a reference-derived condition U (Du et al., 2025). The plan contains linguistic and prosodic information; the CAM++ condition need not contain only identity (Wang et al., 2023). Neither variable is assumed to be emotion free. Their separate input paths allow a controlled comparison without requiring a disentangled representation.

Control applied during token generation can change H itself and therefore does not isolate acoustic realization at a fixed plan. Our comparison reuses H and changes only U. Token-level control can revise timing and prosodic choices that a fixed-plan editor cannot, whereas condition editing tests realization under an unchanged token sequence. These complementary roles depend on the synthesizer's interfaces; the present experiments evaluate condition editing within CV3, with no intervention on token generation.

L.2 The controlled comparison and its limits

For case i, native interface m, and measured outcome q, the intended paired contrast is

$$\begin{aligned}\Delta^{q}_{i,m}(\alpha) = q(G(H_{i,m}, U_{i,\alpha}; p_i, \epsilon_{i,m})) \\ - q(G(H_{i,m}, e_i; p_i, \epsilon_{i,m}))\,,\end{aligned} \tag{4}$$

where $U_{i,\alpha} = \mathcal{N}(e_i + \alpha\hat{r}(e_i, a_i))$. Text, reference, requested emotion, cached tokens, acoustic prompt and decoding state are matched. The notation $\mathrm{do}(U = U_{i,\alpha})$ denotes this explicit assignment at the synthesis interface. Although the input is assigned within a case, its value remains a function of the reference and requested emotion across cases; it is not independent of them.

The comparison isolates the implemented input change, not a pure emotional factor. Because U is entangled, editing it can also change phonation, reference similarity or intelligibility. These are distinct downstream outcomes of the input assignment. Fixed tokens do not imply unchanged perceived rhythm, content accuracy or identity after nonlinear decoding. The contrast concerns the synthesizer's behavior, not the natural causal effect of emotion in observational recordings.

We compare arms within zero-shot and instruct2 separately. These two interfaces are not the neutral and target cells of a two-path factorial experiment, and we do not estimate a planning–realization interaction. The main objective panel uses text-generated tokens; privileged target-derived tokens in other studies remain separately labeled. Interpretation is also subject to the objective cache-audit scope (Appendix D) and incomplete original waveform-to-checkpoint provenance of the listening material (Appendix C). The reported analyses remain exploratory.

L.3 Adaptation scope and interface reuse

CV3 already consumes a normalized 192-D speaker condition through a learned projection into Flow. Our editor changes the value at this interface without adding a new decoder branch or a new conditioning dimension. Training on cached embedding pairs updates only the editor: no gradient passes through the token LM, Flow or vocoder. The method is therefore lightweight in its added trainable module and adaptation scope, not training free.

For a new neutral reference in the same CV3 setting, the encoder and fitted editor are reused without per-speaker optimization. The residual depends on reference and emotion rather than the synthesis text, so it can be cached for that pair; changing the scalar strength does not require refitting. The three editor predictions used by the main ensemble occur outside the token-generation and diffusion-

step loops. These structural properties do not establish measured latency, memory, energy or total data-generation savings.

Reuse is bounded by the label set and encoder space. The four-way emotion input does not support arbitrary new emotions. Changing the encoder, backbone or condition semantics may require new paired embeddings, a refitted editor and renewed evaluation. Equal vector dimensions do not imply compatible coordinates. The IndexTTS2 extension therefore demonstrates a separate representation-specific edit, not unchanged transfer of the main CAM++ network. We make no formal sample-complexity or cross-backbone efficiency claim.

The native interface accepts edits constructed in its own coordinate space. Our source pooled-residual arm holds the synthesis interface and training split fixed while replacing reference-conditioned prediction with an emotion-level mean. Its role is to test the added reference dependence; the norm-matched arm separately probes residual scaling. These are mechanism controls within the present system.

L.4 Endpoint learning and the geometry of editing strength

Both shared and learned editing start from a reference. Their difference is whether the increment is a fixed emotion-level vector or a prediction conditioned on that reference. A rule $s(e, a)v_a$ is an intermediate possibility: it changes the length of a shared direction using reference information. A full residual predictor permits both length and direction to vary; its architecture does not establish that both freedoms are necessary.

For unit paired endpoints x, y, $\|y - x\|_2^2 = 2(1 - x^\top y)$. Raw-residual supervision preserves this pairwise endpoint-distance information, which is removed when each difference is individually normalized before averaging. It is not a calibrated label for perceived emotional intensity. Content, recording and synthesis variation can also contribute to the training residual.

For a unit reference e, decompose a predicted residual as $r = ce + r_\perp$, where $c = e^\top r$ and $e^\top r_\perp = 0$. For $\alpha \geq 0$ and $e + \alpha r \neq 0$, the angle between e and the final normalized condition is

$$\theta = \operatorname{atan2}(\alpha\|r_\perp\|_2,\, 1 + \alpha c)\,. \tag{5}$$

Residuals with the same length can have different parallel and orthogonal components, and hence different final angles. Equal angle still does not imply equal emotion or identity cost after synthesis. The norm-matched control thus tests residual scaling without fully matching effective output cost. Its results constrain a direction-specific superiority claim; they neither establish that all reference conditioning is unnecessary nor prove that learned directions are better.

L.5 Reference dependence, intensity and perceptual evidence

EmoKnob (Chen et al., 2024) and the present editor use different rules for the editing increment. For a fixed demonstration-derived direction and user strength, EmoKnob adds the same increment across target references. The reference still determines the starting embedding, and the decoder can produce different responses for different speakers. Thus a shared increment does not imply identical audio or an absence of speaker information. Our network instead conditions the increment itself on the target reference and emotion, learning from paired endpoints. This is a modeling distinction, not by itself proof of better personalization.

Intensity control and dose response. EmoKnob's Emotion Strength Test asks which of two versions conveys more of the requested emotion, and its ablation varies both strength and the number of demonstration pairs. It explicitly discusses deterioration of speaker similarity and transcription accuracy as strength increases. The present study combines intensity control with reference-conditioned prediction and multidimensional perceptual evaluation. Neither the scalar nor our intensity rating is a calibrated cross-speaker emotional unit, and the editor does not automatically choose a perceptually matched or identity-safe dose.

Perceptual evaluation dimensions. Binary judgments assess comparative emotion expression or recognition. EmoKnob's reported listening endpoints do not provide separate MOS for the four dimensions in our study. Our paired design measures expression gains alongside naturalness and

Table 21: Modeling and evaluation scope. The EmoKnob column describes its published method and experiments (Sections 3.2, 4.1 and 5; Tables 2–3). The present-study column refers to the main editor and bilingual dose study, not the separate historical sweeps. Entries compare questions and measurements, not performance across different models or datasets.

Aspect	EmoKnob	Present study
Reference dependence	Reference supplies the starting embedding; the demonstrated emotion direction is shared	Reference and emotion jointly predict the residual; shared and norm-matched arms test design choices
Intensity control	User-selected scalar, pairwise Emotion Strength Test, and sample-count/strength sweeps of WER and SIM	Scalar editing strength; paired ratings of requested-emotion intensity and appropriateness at low and high doses
Human outcomes	Binary selection, enhancement, discrimination, identification and strength judgments	Separate 1–5 ratings for emotion appropriateness, naturalness, speaker similarity and intensity
Quality and identity	Automatic WER and speaker-encoder SIM; reported means and standard deviations	CER/WER and CAM++ diagnostics, plus paired naturalness and similarity ratings with conditional non-inferiority analysis

similarity costs, separating perceived enhancement from voice-cloning fidelity. Additional rating dimensions broaden the evaluation but do not guarantee lower bias or greater precision: our allocation is unbalanced, and uncertainty remains conditional on the collected participants and recordings (Appendix C).

Uncertainty and preservation. EmoKnob's Table 3 reports WER and SIM with standard deviations and describes scores as similar within uncertainty. Standard deviations describe dispersion; they do not themselves constitute paired confidence intervals for a change or a non-inferiority test. Our listening analysis instead states a margin and estimates paired changes, distinguishing failure to establish non-inferiority from evidence of a loss beyond that margin. The adverse high-dose similarity result illustrates why expression gain and preservation require separate inferences, even when the same edited samples are used.

L.6 Relationship to the unpublished concurrent manuscript

The unpublished PATHOS-TTS manuscript (Anonymous, 2026) develops coordinated instruction-guided planning and native-condition transport. The conceptual planning–realization organization and normalized residual-transport form are shared with that work, rather than claimed here as independently new. This manuscript studies the width-256, two-block reference-conditioned editor as a standalone module with a frozen planner and Flow, and separates paired endpoint prediction, real-reference synthesis and listening outcomes.

No results, pretrained-path efficiency claims or two-path interaction estimates from that manuscript are imported as evidence for the present editor. Its distinct model configurations are not substitutes for this paper's experiments. Conversely, the present calibrated and historical studies are not replications of the main editor. Cross-manuscript sample overlap has not been audited here, so different study definitions do not establish disjoint underlying data. The reference is identified as an unpublished manuscript, not a public preprint or an accepted publication.