Input Embeddings as a Compositional Pronunciation Interface for Frozen LLM-Based Chinese TTS

Anonymous authors
Paper under double-blind review

Abstract

LLM-based text-to-speech systems can misread a character even when its intended pronunciation is known. We study input embeddings as a pronunciation interface for a frozen speech model. A closed-form additive regression maps initials, finals and tones to the embeddings of common characters. To construct a target embedding, we start from real donor rows and add the regression-estimated phonological difference between donor and target. This preserves information in real rows while changing the specified reading; the TTS parameters remain frozen and no recording of the target pronunciation is required. On 120 held-out common-word items, excluding the target's entire initial–final combination in all tones from both regression fitting and donor selection, composition reaches 94% pronunciation accuracy on CosyVoice2 versus 88% for regression alone, with the same ordering on Spark-TTS and GLM-TTS. Under the same exclusion, composition repairs 77% of 193 verified rare-word errors versus 65% for regression alone. In ordinary use, a unified rule that reuses homophones when available repairs 79%, versus 70% for text rewriting. These results connect compositional control to real pronunciation repair, while revealing a trade-off between donor availability and damage to correctly read words.

1 Introduction

A speech model can produce fluent sentences while misreading individual characters: CosyVoice2 reads 香橼 (xiāng yuán, citron) as xiāng yǎn and 铼 (lái, rhenium) as shì. On 1,440 lexicon-mined rare-character items, its verified error rate is 47%, versus 2% for a lexicon-front-end system. Given the intended reading, how can a frozen model pronounce such targets correctly?

Trained pronunciation channels and post-hoc model or conditioning edits require training access, corrected labels or reference audio. Internal homophone rewriting is simpler but requires another character with the desired reading. We instead ask whether input embeddings can express readings assembled from familiar phonological parts.

Common-character embeddings contain linearly decodable initials, finals and tones. We add an additive regression's predicted phonological difference to real donor rows (Figure 1), retaining their residual beyond the prediction. Homophone averaging is the zero-difference case. The constructed vector replaces only the target token or byte-fragment span; the model parameters and original sentence remain unchanged (Figure 3).

We develop this argument through three contributions.

1. **An input-level intervention.** Probes reveal phonological structure in common rows; controlled transplants redirect spoken readings (Section 3).
2. **Composition beyond available syllables.** Withholding the target's initial–final combination in all tones from fitting and donors separates composition from homophone lookup across three backbones (Section 5).
3. **Real repair and a second script.** We test repair under the same withholding, compare ordinary use with rewriting and resampling, and measure damage to correct words (Section 6). A Korean extension composes syllabic blocks from component letters (Section 7).

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.

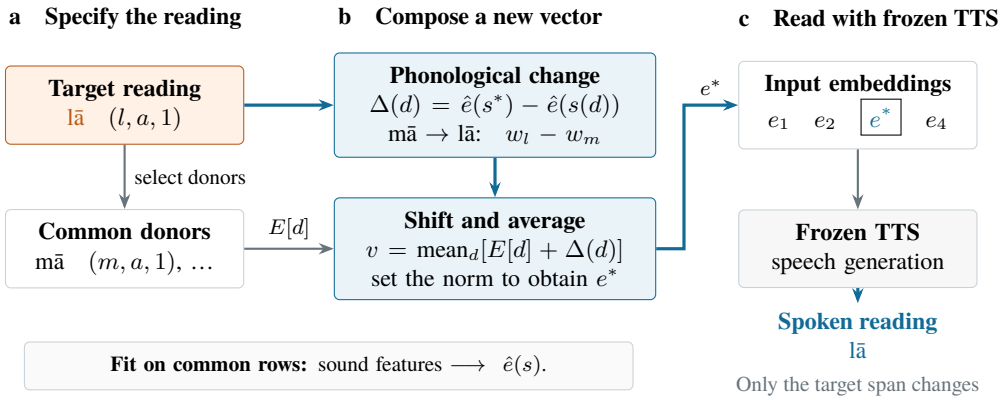


Figure 1: **Pronunciation by composition.** Shift common donor rows by their predicted phonological difference from the target, then average and set the norm. The example changes mā to lā. Regression uses eligible common rows; the TTS model remains frozen.

2 Related Work

Pronunciation control and editing. Recent LLM-based TTS systems include CosyVoice2/3, Spark-TTS, GLM-TTS, IndexTTS2, Llasa and Fish-Speech (Du et al., 2024; 2025; Wang et al., 2025; Cui et al., 2025; Zhou et al., 2026; Ye et al., 2025; Liao et al., 2024). CosyVoice 3 trains pinyin inpainting, and several systems accept pronunciation markup through trained channels. We compare with the channel in the released Fun-CosyVoice3-0.5B checkpoint (FunAudioLLM, 2025). Post-hoc approaches include rank-one weight edits (SonoEdit; Singh et al., 2026a), optimisation of text conditioning from reference audio (FlowEdit; Singh et al., 2026b), activation steering (Lee et al., 2025), LoRA training on corrected labels (UtterTune; Kato, 2025) and prompt-audio splicing (GRAFT; Asonitis et al., 2026). Our setting supplies the intended reading and access to the embedding lookup. Appendix B compares the information and optimisation these approaches require; it is not an empirical accuracy comparison.

Constructing embeddings from known rows. Vocabulary-transfer methods initialise embeddings from sub-tokens or other known rows: FVT uses means (Gee et al., 2022), WECHSEL, FOCUS and OFA use weighted combinations (Minixhofer et al., 2022; Dobler & de Melo, 2023; Liu et al., 2024), and MIMICK predicts embeddings from surface form (Pinter et al., 2017). VocADT learns combination weights with the model frozen (Han et al., 2025); Token Distillation fits a new row to reproduce hidden states of the original segmentation (Dobler et al., 2026). Our homophone-mean case follows the same transfer principle. The additional question is whether a row for a specified reading can be assembled from phonological components and interpreted by a frozen speech decoder. McLaughlin et al. (2025) establish linear phonetic representations and interventions in text LLMs. We test this structure in generated speech under complete phonological-combination withholding. Studies of under-trained tokens (Land & Bartolo, 2024) and frequency-dependent linear structure (Merullo et al., 2025) also motivate our common-versus-rare row analysis.

Measuring pronunciation. ASR may normalise a correct rare reading or infer the intended word from incorrect speech (Park, 2026; Yu & Kang, 2026). Building on global calibration with synthesised controls (Kawajiri, 2026), our protocol calibrates the judge per item and abstains when controls cannot be separated. Human checks and independent resampling assess this measurement process (Appendices A and E).

3 Input Embeddings as a Pronunciation Intervention

Measurement. We place each target in a carrier sentence and compare the generated reading with its dictionary pronunciation. Whisper, SenseVoice and Paraformer propose pronunciation errors (Radford et al., 2023; An et al., 2024; Gao et al., 2022). A forced-choice judge is then calibrated using positive and negative homophone controls synthesised by the model under test; inseparable

controls yield an abstention. This rule agrees with two listeners' consensus on 91% of 157 decided items from five systems ($\kappa = 0.82$). A *stable error core* contains items wrong in every baseline sampling used to define it; a further, held-out sampling measures spontaneous recovery. Appendix A gives the calibration, oracle validation and repeatability analysis.

Rare characters expose an input-level weakness. On the 143-word curated seed set, lexicon-front-end systems err on 12–18% of words and four LLM backbones on 35–55%. On the mined set, CosyVoice2 errors increase as character frequency falls, reaching all 33 byte-fragmented targets in the initial run (32 in a later sampling; Figure 2a). Only 6% of its wrong readings follow the character's phonetic component, compared with 29% for the lexicon system. Appendix A specifies the challenge-set construction and the rates' denominators.

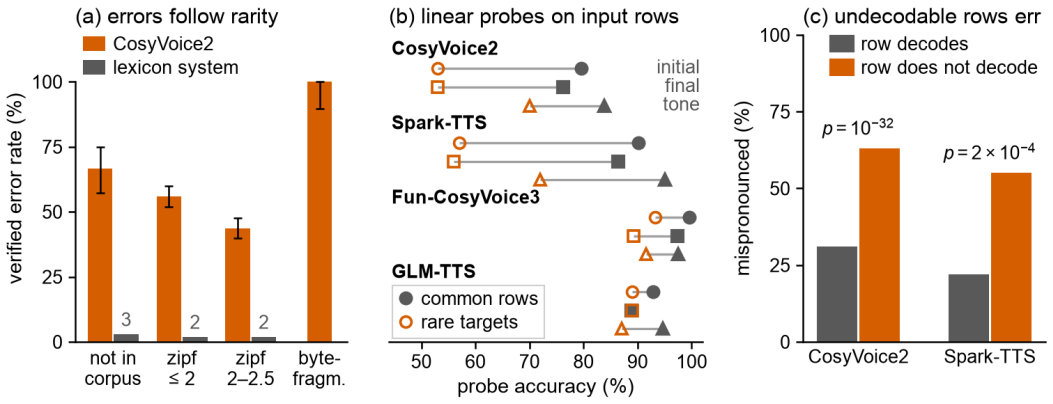


Figure 2: **Phonological structure and pronunciation errors.** (a) Verified error rates by target-character rarity. (b) Linear probes trained on common-character input rows (5-fold, filled) and applied to rare targets (hollow). (c) Mispronunciation is more frequent when a row's base syllable cannot be decoded.

Common rows contain usable phonological structure. Linear probes trained on 3,134 single-token characters decode initial, final and tone with 80%, 76% and 84% accuracy on CosyVoice2. On rare target rows, accuracy falls to 53%, 53% and 70%. Spark-TTS shows the same decline (Figure 2b). A CosyVoice2 target whose row does not decode to its base syllable is mispronounced in 63% of cases, compared with 31% when it does; the corresponding Spark-TTS rates are 55% and 22%. Rare CosyVoice2 rows also have a lower median norm (0.36 versus 1.05). Frequency, decodability and row norm covary, so these observations alone do not establish causation.

Changing the row changes the spoken reading. We replace the first character's row in common words using the same placeholder interface as the method. A common donor with a different reading redirects the output to its own reading in 113 of 120 cases on CosyVoice2, 106 on Spark-TTS and 113 on GLM-TTS. Replacing it with a rare *homophone*'s own row preserves the intended reading in only 59% of 91 cases on CosyVoice2 (60% on Spark-TTS), with survival increasing with donor frequency (Appendix C). These controls establish the row as an effective intervention point and motivate using reliable common rows as donors.

This account covers an important source of repairable errors. Other errors require choosing a non-dominant reading of a common character, and Fun-CosyVoice3 retains well-decodable rare rows despite pronunciation errors. Throughout, the intended reading is supplied explicitly; the method controls its realisation rather than inferring it from context.

4 Composing a Pronunciation Embedding

Let E be the speech LLM's input-embedding table and C the eligible common single-token characters with dominant readings $s(u) = (\text{initial}, \text{final}, \text{tone})$. We construct e^* for the specified target reading s^* from rows in C.

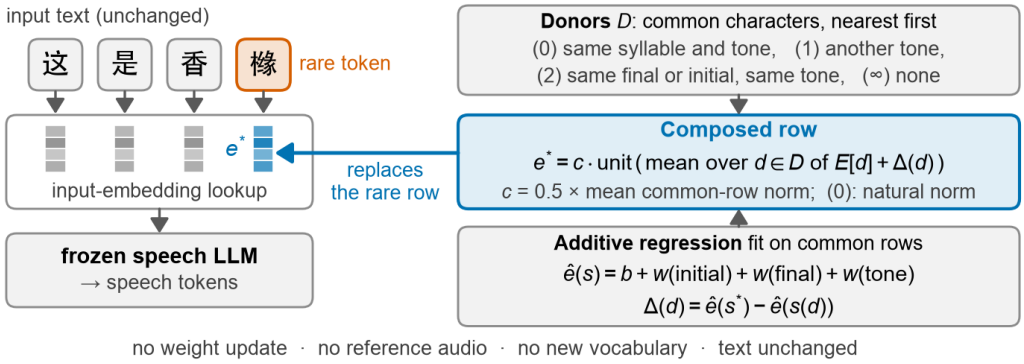


Figure 3: **A concrete intervention: 这是香橼.** Only the input span for 橼 is replaced by e^*; other tokens retain their rows. The diagram's c denotes the full written norm ($c\bar{r}$ in the text); homophone means keep their natural norm. "No reference audio" means no recording of the *target pronunciation*; a speaker prompt may still be used.

An additive phonological map. Ridge regression maps one-hot phonological features to common-character rows:

$$\hat{e}(s) = b + w_{\text{init}(s)} + w_{\text{fin}(s)} + w_{\text{tone}(s)}. \tag{1}$$

This closed-form map is the only fitted object. Additivity gives a shared direction for each component change, including the same tone difference across syllables.

Real donors plus the predicted difference. We select donors at the closest available phonological distance: (0) the same syllable and tone; (1) the same initial and final at another tone; or (2) the same final or initial with the same tone. We take up to 12 donors, ordered by frequency. For each donor d, the offset $\hat{e}(s^*) - \hat{e}(s(d))$ changes the predicted phonological components. Their average is

$$\begin{aligned} v(s^*) &= \frac{1}{|D|} \sum_{d \in D} [E[d] + \hat{e}(s^*) - \hat{e}(s(d))] \\ &= \hat{e}(s^*) + \frac{1}{|D|} \sum_{d \in D} [E[d] - \hat{e}(s(d))]\,. \end{aligned} \tag{2}$$

The target prediction thus gains the donors' mean residual. Homophones have zero offset, yielding their mean; without eligible donors, $v = \hat{e}(s^*)$ recovers regression alone.

Norm and inference. Let $\bar{r}$ be the mean norm of common-character rows. Homophone means keep their natural norm; other constructed vectors are written as $e^* = c\bar{r}\, v/\|v\|$, with default $c = 0.5$. This convention also applies to the regression-only fallback. Section 5 distinguishes comparisons at this shared norm from norm sensitivity; Appendix C reports the scan that motivated the default.

A request-specific placeholder supplies e^* for the target token or entire byte-fragment span (Figure 3). Targets are rare or fragmented characters, or common characters with a specified non-dominant reading. Vectors can be cached by reading and donor pool, adding one lookup per selected position. Appendix B specifies the implementation.

5 Composition with Held-Out Phonological Combinations

We first test composition on common characters the model can ordinarily pronounce. Replacing their embeddings while withholding their readings from the construction separates the ability to assemble a pronunciation from the difficulty of repairing a naturally rare word.

Design. For each backbone, we select 160 frequent two-character words, one per base syllable, with dominant readings and no tone sandhi or neutral tone. Eligible syllables have at least three common characters, and their initial and final each occur in at least four other syllables. Forty items are reserved for tuning; the remaining 120 form the test set. We replace the first character's row in the carrier 这是 X。. Two tiers withhold characters from *both* the regression fit and donor selection: **T2**

excludes the target's toned syllable; **T3** excludes its complete initial–final combination in all tones. The TTS parameters are unchanged, and the withholding does not apply to its pretraining.

Controls use the same placeholder: the character's own row, regression alone, the same donors without the phonological difference, the nearest remaining real row to the regression prediction, and a random vector. A reading is correct when at least two of the three open judges recover the first syllable with its tone (the *fast protocol*). The unmodified models score 97–98% on these common words. The own row uses its natural norm; other constructed rows use the shared default $c = 0.5$, selected on the seed set. GLM-TTS additionally compares $c \in \{0.35, 0.5, 0.7\}$ on its reserved items and retains 0.5. These are comparisons at a common norm, with sensitivity examined below.

The composed row, regression alone, own row and T3 nearest-real-row control are sampled three times; secondary controls are sampled once. We report means and word-bootstrap 95% intervals, with word-level sign-flip tests. The CosyVoice2 T3 composition-versus-regression comparison is one of three confirmatory comparisons fixed before the repeats; all three survive Holm correction (Appendix B). Full results and additional controls are in Appendix C.

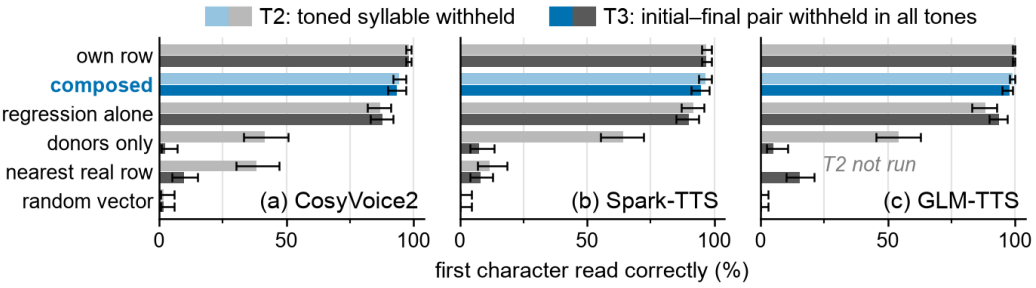


Figure 4: **A composed row can express a withheld reading.** First-character accuracy on 120 test items per backbone (119 for GLM-TTS T3 composition). Own row, composition, regression alone and the T3 nearest row: three-sampling mean and word-bootstrap 95% interval. Other bars: one sampling and Wilson interval. GLM-TTS has no T2 nearest-row result. Numerical values are in Table 3.

Reaching a reading absent from the construction. Under T3, composition reaches 94% on CosyVoice2, compared with 88% for regression alone ($+5.8$ points, 95% interval $[+1, +11]$, $p = 0.02$). Same-initial donors give a similar result to same-final donors. The ordering replicates on Spark-TTS (95% versus 90%) and GLM-TTS (98% versus 94%). Their own trained rows score 98%, 97% and 100%, respectively. Thus composition recovers much of the trained row's pronunciation accuracy while using neither the target combination in the fit nor a donor with that combination.

Both the donor and the difference contribute. On CosyVoice2 T3, the nearest remaining real row reaches 10%, and donors without the difference reach 2%. The large separation from these controls shows that retrieval or averaging alone does not explain the composed output. T2 isolates tone: composition reaches 94% versus 87% for regression alone. Without the difference, the judges score the donor mean at 42% with tone and 89% for the base syllable. The majority of three independent listeners hears the donor's tone in 21 of 24 checked clips, including 9 of the 12 accepted by the automatic judges. The control therefore retains the donor's tone, and the automatic score overestimates its tone accuracy (Appendix E).

Sensitivity and remaining errors. Composition retains 90–96% accuracy across $c = 0.35$–1.0 in a single-sampling scan; regression alone varies from 92% to 77%. In a 25-character carrier, composition reaches 95% and regression alone 98% (one sampling). The default-norm advantage therefore coexists with settings where regression performs as well or better. Syllables written with *u* for /y/ after j, q, x and y are a concentrated weakness on CosyVoice2 and Spark-TTS; recoding them with *ü* does not remove it. Appendix C gives the per-subset results, norm scan and donor-count sensitivity.

Table 1: **CosyVoice2 repair under restricted and ordinary donor access.** Mean of three samplings; calibrated labels; all verified baseline errors in each column. The pooled column deduplicates the seed, homophone and fragmented sets. ‡171 words: the fragmented subset was not run in every resampling. Additional conditions, including T2, are in Appendix D.

Condition	homophone (100)	fragmented (32)	pooled (193)
Target initial–final pair withheld in all tones			
composed row	78%	77%	**77%**
regression alone	63%	67%	65%
Ordinary donor access			
unified method	80%	80%	**79%**
homophone mean	80%	73%	77%
internal text rewrite	73%	71%	70%
held-out resampling	5%	4%	8%‡

6 From Composition to Rare-Word Repair

6.1 Setup

We test CosyVoice2-0.5B, Fun-CosyVoice3-0.5B, Spark-TTS-0.5B and GLM-TTS-1.5B. The first three use Qwen-family text models; GLM-TTS supplies another backbone and tokenizer. CosyVoice2 and Spark-TTS share token ids but have different embedding tables (Appendix B). The main analysis pools the seed set with mined homophone and byte-fragmented subsets, yielding 193 distinct verified CosyVoice2 errors after deduplication. Spark-TTS has 59 verified seed errors. A separate 50-item set has no common homophone for the intended reading.

All repair conditions use per-item calibration. *Repaired* is the share of baseline errors corrected, counting inapplicable or unjudgeable repairs as unsuccessful; *broken* is the share of baseline-correct words that become errors. Both are compared with fresh resampling. Main CosyVoice2 and Spark-TTS conditions have three samplings, each with its own baseline and calibration controls. The initial error set is fixed, words are counted once, and intervals and tests operate over words. Appendix B details the sampling protocol, confirmatory comparisons and Spark-TTS decoding configuration; Appendix D gives the full seed results and additional subset comparisons.

6.2 Repair with and without access to homophones

The held-out construction repairs real errors. To connect Section 5 to practical repair, we rebuild each target's row after excluding its entire initial–final combination in all tones. On the 193 verified errors, composition repairs 77% versus 65% for regression alone under the same exclusion ($+12.3$ points, 95% interval $[+7, +17]$, $p < 10^{-4}$; Table 1). Homophone averaging with unrestricted donors has the same point estimate (77%). Withholding increases damage to baseline-correct words: 26%, compared with 13% for the unrestricted unified method and 20% for resampling. On Spark-TTS, held-out composition repairs 81% versus 77% for held-out regression, with an interval that includes zero ($+4.0$ points $[-6, +13]$).

Using the closest available donors. The unified method uses homophones when available and composes from more distant donors otherwise. It repairs 79% of CosyVoice2 errors, compared with 70% for rewriting the text internally with the same homophone ($+9.0$ points $[+4, +14]$, $p = 0.001$). The homophone mean alone reaches 77% and is not separated from the unified method. This ordinary-access result mainly establishes the usefulness of transferring reliable rows; the withholding experiment above isolates composition beyond homophones. On the 80 baseline-correct words, the unified method breaks 13% versus 29% for rewriting (-16 points $[-27, -6]$; exploratory comparison). On Spark-TTS, the unified method, homophone mean and text rewrite repair 80%, 81% and 79%, with no clear separation.

Human evaluation without common homophones. Three independent native Chinese speakers rate stored clips for the 50 selected words. Baseline is correct on 14/50 (28%) and half-norm regression on 36/50 (72%). Other-tone donors plus the predicted tone difference are correct on 36/45 available clips (80%); five words have no clip in this condition, giving confirmed correct coverage

of 36/50 over the selected cohort. On the same 45 words, regression scores 32/45 (71%); the methods are not separated by a paired exact test ($p = 0.39$). Both improve over their matched baselines ($p < 0.001$, including Holm correction across the three exploratory comparisons). These single-sampling listening results support pronunciation repair on this subset without establishing an advantage of composition over regression. Appendix E reports the denominators, conditional repair rates and side effects.

6.3 Preservation, other backbones and listening evidence

Preserving ordinary speech. On 300 Seed-TTS-eval test-zh sentences (ByteDance Speech Team, 2024; Anastassiou et al., 2024), replacing one random common character with its homophone mean changes Paraformer CER by -0.03 ± 0.47 points on CosyVoice2 and $+0.01 \pm 0.40$ on Fun-CosyVoice3; speaker-similarity changes are $+0.002$ and -0.002. This preservation check concerns the homophone-mean intervention. Natural 20–30 character carriers place seed targets at the start, middle and end and give repair rates of 76–85% with target-only homophone transfer (Appendix D).

Additional backbones. In single-sampling runs, homophone-mean transfer repairs 82% of Fun-CosyVoice3 errors and 77% of GLM-TTS errors. Fun-CosyVoice3's trained pinyin channel repairs 70%, without a significant paired difference ($p = 0.21$). These comparisons and side-effect rates are in Appendix D.

Listeners confirm the interface and expose judge errors. A blind A/B test with three listeners covers 98 CosyVoice2 pairs using whole-word homophone-mean transfer: the edited side is correct on 67 of 68 baseline-error pairs and breaks 4 of 30 controls, with naturalness responses reported in Appendix E. Agreement with the calibrated protocol is $\kappa = 0.78$.

Three independent listeners also rate 165 clips spanning held-out composition and composition on real errors. By majority vote, all 40 automatically accepted composed clips are correct, and 10 of 21 rejected composed clips are also correct. On real errors under complete withholding, listeners reject 2 of 25 clips labelled repaired and accept 8 of 15 labelled not repaired. Pairwise agreement is $\kappa = 0.92$, 0.92 and 1.00. These checks support the constructed readings and indicate conservative automatic labels in the sampled strata. Appendix E reports the sampling strata, tone-specific checks and estimates adjusted for the stratum sizes.

7 Extension to Korean Syllabic Blocks

Hangul provides a second setting for constructing a pronunciation from component features. We replace Chinese initial, final and tone with a block's initial, medial and optional final letter. Common-block rows supply donors, and the predicted letter differences produce the target vector. The intervention can also replace the full byte-fragment span of a block.

CosyVoice2 and Fun-CosyVoice3 read blocks in the frame 이것은 X 입니다. The evaluation restricts the eligible codas and accounts for liaison and specified vowel mergers. We retain items whose target reading is recovered by at least one of two ASR judges from a reference-engine rendering. This keeps 22% of the zero-frequency blocks and 17% of fragmented blocks, yielding 232 rare or fragmented items; results apply to this filtered subset (Appendix F).

The stable error cores contain 198 items on CosyVoice2 and 195 on Fun-CosyVoice3. Composition repairs 48% and 62%, compared with 1% and 5% for held-out resampling and 0% for random vectors. Regression alone at half norm reaches 47% and 63%, so the donor residual does not add a consistent gain here. The experiment supports the feasibility of constructing usable rows from letter features in a second script. Its labels remain automatic, and its evidence is limited to the reference-transcribable blocks, one frame and one prompt voice. Additional results, including frequent-block withholding and a stricter two-judge criterion, are in Appendix F.

8 Limitations

The method requires a specified reading and access to the input-embedding lookup. It controls pronunciation without resolving contextual polyphone ambiguity. The evidence establishes an effective

intervention point, not a unique explanation of rare-character errors. All evaluated backbones are 0.5–1.5B, and three share a tokenizer.

Performance depends on donor availability, vector norm and context. Main composition-versus-regression comparisons use a shared norm; the scans are not an equal-budget search for each method's optimal configuration. Complete withholding increases collateral errors, and the ordinary-access advantage over text rewriting is clear on CosyVoice2 but not on Spark-TTS. Secondary controls and repair on Fun-CosyVoice3 and GLM-TTS use one sampling. No post-hoc editing method was run as an empirical baseline.

The natural no-common-homophone listening study covers 50 words from one stored CosyVoice2 sampling, with donor-based outputs available for 45; it does not separate composition from regression. The controlled-composition listening check covers selected strata, and Korean results remain automatic and filtered by reference transcribability. Natural-carrier repair and ordinary-sentence preservation primarily test homophone transfer. English embedding repair has not been evaluated; a separate measurement pilot on lexicon/G2P systems is retained in the supplementary experiment archive.

9 Conclusion

Input embeddings provide a compositional pronunciation interface for frozen LLM-based TTS. Adding a predicted phonological difference to real donor rows creates vectors that express readings excluded from the construction's fit and donor pool. Controlled tests on three backbones connect this representation-level operation to spoken output, and the same construction repairs verified rare-word errors under complete phonological-combination withholding. Homophone transfer is the zero-difference case, enabling one rule to use whichever donors are available. The results support pronunciation control through the existing input interface, with donor availability, norm and evaluation reliability defining its current limits.

AI use statement

Generative AI tools were used primarily to assist with writing and debugging experimental and analysis code, and with drafting and editing the manuscript. They also assisted with feedback on the methodology and drafts, literature search and summaries, interpretation of results, the method diagram, LaTeX preparation and reference formatting. The author reviewed and revised AI-assisted text and code and checked reported results against the underlying experiment and listening records.

For the new 193-clip listening analysis, all ratings came from human listeners. Answer IDs were matched to a fixed sample manifest, and source audio copies were checked against the export manifest. Majority counts were independently recomputed with set operations; paired exact tests were independently checked with a statistical library. The reproducible scoring code preserves missing votes and includes checks of the joint-vote rule and paired-test edge cases. No human ratings were generated or imputed by an AI model. The author takes responsibility for the final content of the paper.

Ethics statement

Our evaluation involved a human listening study of synthesized speech. All participants voluntarily participated in the study and provided informed consent. No personally identifiable or sensitive information was collected. Aggregate results are reported in the paper, and anonymized item-level ratings are included in the supplementary materials for reproducibility.

Reproducibility statement

Appendix A specifies measurement and data construction; Appendix B specifies serving, statistical procedures and record locations. All TTS backbones are public checkpoints, and the only fitted object in the intervention is a closed-form ridge map on input embeddings. The experimental supplementary archive contains code, word lists, the available generation and verification records, statistical reports and four listening pages with embedded audio. Its update adds the 193-clip study's anonymised

answer snapshots, manifest, per-item votes, coverage audit and standard-library scoring script. The new listening results can be recomputed directly from these answer files; Appendix E states the scoring rule and denominators.

References

Keyu An, Qian Chen, Chong Deng, Zhihao Du, Changfeng Gao, Zhifu Gao, Yue Gu, Ting He, Hangrui Hu, Kai Hu, Shengpeng Ji, Yabin Li, Zerui Li, Heng Lu, Haoneng Luo, Xiang Lv, Bin Ma, Ziyang Ma, Chongjia Ni, Changhe Song, Jiaqi Shi, Xian Shi, Hao Wang, Wen Wang, Yuxuan Wang, Zhangyu Xiao, Zhijie Yan, Yexin Yang, Bin Zhang, Qinglin Zhang, Shiliang Zhang, Nan Zhao, and Siqi Zheng. FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs. *arXiv preprint arXiv:2407.04051*, 2024. Technical report describing SenseVoice. https://arxiv.org/abs/2407.04051.

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, Mingqing Gong, Peisong Huang, Qingqing Huang, Zhiying Huang, Yuanyuan Huo, Dongya Jia, Chumin Li, Feiya Li, Hui Li, Jiaxin Li, Xiaoyang Li, Xingxing Li, Lin Liu, Shouda Liu, Sichao Liu, Xudong Liu, Yuchen Liu, Zhengxi Liu, Lu Lu, Junjie Pan, Xin Wang, Yuping Wang, Yuxuan Wang, Zhen Wei, Jian Wu, Chao Yao, Yifeng Yang, Yuanhao Yi, Junteng Zhang, Qidi Zhang, Shuo Zhang, Wenjie Zhang, Yang Zhang, Zilin Zhao, Dejian Zhong, and Xiaobin Zhuang. Seed-TTS: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024. https://arxiv.org/abs/2406.02430.

Antonis Asonitis, Francesco Verdini, Aref Farhadipour, Vijeta Avijeet, Pierre-Edouard Honnet, Marzieh Razavi, and Juan Pablo Zuluaga Gomez. GRAFT: Grafted reference audio for fine-grained pronunciation in zero-shot text-to-speech. *arXiv preprint arXiv:2607.02633*, 2026. https://arxiv.org/abs/2607.02633.

ByteDance Speech Team. seed-tts-eval. https://github.com/BytedanceSpeech/seed-tts-eval, 2024. Test set and scoring scripts released with Seed-TTS.

CC-CEDICT contributors. CC-CEDICT. https://cc-cedict.org/wiki/, 2026. Community-maintained, Creative Commons-licensed Chinese-English dictionary; a continuation of Paul Andrew Denisowski's 1997 CEDICT project. Downloads: https://www.mdbg.net/chinese/dictionary?page=cc-cedict. Accessed 2026-09-17.

Yafeng Chen, Siqi Zheng, Hui Wang, Luyao Cheng, Qian Chen, and Jiajun Qi. An enhanced Res2Net with local and global feature fusion for speaker verification. In *Interspeech 2023*, pp. 2228–2232, 2023. doi: 10.21437/Interspeech.2023-1294. URL https://www.isca-archive.org/interspeech_2023/chen23o_interspeech.html.

cjkvi project. cjkvi-ids: CJK ideographic description sequence data. https://github.com/cjkvi/cjkvi-ids, 2026. Collection of Ideographic Description Sequence (IDS) data describing the compositional structure of CJK Unified Ideographs. Accessed 2026-09-17.

Jiayan Cui, Zhihan Yang, Naihan Li, Jiankun Tian, Xingyu Ma, Yi Zhang, Guangyu Chen, Runxuan Yang, Yuqing Cheng, Yizhi Zhou, Guochen Yu, Xiaotao Gu, and Jie Tang. GLM-TTS technical report. *arXiv preprint arXiv:2512.14291*, 2025. https://arxiv.org/abs/2512.14291.

Konstantin Dobler and Gerard de Melo. FOCUS: Effective embedding initialization for monolingual specialization of multilingual models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13440–13454, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.829. URL https://aclanthology.org/2023.emnlp-main.829/.

Konstantin Dobler, Desmond Elliott, and Gerard de Melo. Token distillation: Attention-aware input embeddings for new tokens. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=n20ml5nGEo. arXiv:2505.20133.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. CosyVoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024. https://arxiv.org/abs/2412.10117.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. CosyVoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*, 2025. https://arxiv.org/abs/2505.17589.

FunAudioLLM. Fun-CosyVoice3-0.5B-2512. https://huggingface.co/FunAudioLLM/Fun-CosyVoice3-0.5B-2512, 2025. Model card and open-weight release; see also https://github.com/QwenAudio/CosyVoice.

Zhifu Gao, Shiliang Zhang, Ian McLoughlin, and Zhijie Yan. Paraformer: Fast and accurate parallel transformer for non-autoregressive end-to-end speech recognition. In *Interspeech 2022*, pp. 2063–2067, 2022. doi: 10.21437/Interspeech.2022-9996. URL https://www.isca-archive.org/interspeech_2022/gao22b_interspeech.html.

Leonidas Gee, Andrea Zugarini, Leonardo Rigutini, and Paolo Torroni. Fast vocabulary transfer for language model compression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 409–416, Abu Dhabi, UAE, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-industry.41. URL https://aclanthology.org/2022.emnlp-industry.41/.

HyoJung Han, Akiko Eriguchi, Haoran Xu, Hieu Hoang, Marine Carpuat, and Huda Khayrallah. Adapters for altering LLM vocabularies: What languages benefit the most? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KxQRHOre9D. arXiv:2410.09644.

Shuhei Kato. UtterTune: LoRA-based target-language pronunciation edit and control in multilingual text-to-speech. *arXiv preprint arXiv:2508.09767*, 2025. https://arxiv.org/abs/2508.09767.

Ryoma Kawajiri. ja-tts-g2p-bench: Japanese TTS Kanji-reading (g2p) benchmark. https://github.com/filmapp/ja-tts-g2p-bench, 2026. Software, v0.1.0.

Sander Land and Max Bartolo. Fishing for Magikarp: Automatically detecting under-trained tokens in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11631–11646, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.649. URL https://aclanthology.org/2024.emnlp-main.649/.

Kyowoon Lee, Artyom Stitsyuk, Gunu Jho, Inchul Hwang, and Jaesik Choi. Counterfactual activation editing for post-hoc prosody and mispronunciation correction in TTS models. In *Interspeech 2025*, pp. 434–438, 2025. doi: 10.21437/Interspeech.2025-723. URL https://www.isca-archive.org/interspeech_2025/lee25f_interspeech.html.

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

Yihong Liu, Peiqin Lin, Mingyang Wang, and Hinrich Schuetze. OFA: A framework of initializing unseen subword embeddings for efficient large-scale multilingual continued pretraining. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 1067–1097, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.68. URL https://aclanthology.org/2024.findings-naacl.68/.

Oliver McLaughlin, Arjun Khurana, and Jack Merullo. I have no mouth, and I must rhyme: Uncovering internal phonetic representations in LLaMA 3.2. *arXiv preprint arXiv:2508.02527*, 2025. https://arxiv.org/abs/2508.02527.

Jack Merullo, Noah A. Smith, Sarah Wiegreffe, and Yanai Elazar. On linear representations and pretraining data frequency in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EDoD3DgivF. arXiv:2504.12459.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3992–4006, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.293. URL https://aclanthology.org/2022.naacl-main.293/.

Seongjin Park. Intent vs. surface: Recovering acoustic realization from modern ASR for pronunciation training. In Ekaterina Kochmar, Bashar Alhafni, Stefano Bannò, Marie Bexte, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Anais Tack, Victoria Yaneva, and Zheng Yuan (eds.), *Proceedings of the 21st Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2026)*, pp. 336–346, San Diego, California, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-409-5. doi: 10.18653/v1/2026.bea-1.23. URL https://aclanthology.org/2026.bea-1.23/.

Yuval Pinter, Robert Guthrie, and Jacob Eisenstein. Mimicking word embeddings using subword RNNs. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 102–112, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1010. URL https://aclanthology.org/D17-1010/.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. https://arxiv.org/abs/2412.15115.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 886–890, 2022. doi: 10.1109/ICASSP43922.2022.9746108. URL https://doi.org/10.1109/ICASSP43922.2022.9746108.

Ayush Pratap Singh, Harshit Singh, Nityanand Mathur, Akshat Mandloi, and Sudarshan Kamath. SonoEdit: Null-space constrained knowledge editing for pronunciation correction in LLM-based TTS. In *Conference on Parsimony and Learning*, volume 328 of *Proceedings of Machine Learning Research*, pp. 1090–1100. PMLR, 2026a. URL https://proceedings.mlr.press/v328/singh26a.html.

Harshit Singh, Ayush Pratap Singh, and Nityanand Mathur. FlowEdit: Associative memory for lifelong pronunciation adaptation in flow-matching TTS. *arXiv preprint arXiv:2606.20518*, 2026b. https://arxiv.org/abs/2606.20518.

Unicode Consortium. Unihan Database. https://www.unicode.org/reports/tr38/, 2026. Unicode Standard Annex #38, "Unicode Han Database (Unihan)." Data files: https://www.unicode.org/Public/UCD/latest/ucd/Unihan.zip. Accessed 2026-09-17.

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-TTS: An efficient LLM-based text-to-speech model with single-stream decoupled speech tokens. *arXiv preprint arXiv:2503.01710*, 2025. https://arxiv.org/abs/2503.01710.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. https://arxiv.org/abs/2407.10671.

Zhen Ye, Xinfa Zhu, Chi-Min Chan, Xinsheng Wang, Xu Tan, Jiahe Lei, Yi Peng, Haohe Liu, Yizhu Jin, Zheqi Dai, Hongzhan Lin, Jianyi Chen, Xingjian Du, Liumeng Xue, Yunlin Chen, Zhifei Li, Lei Xie, Qiuqiang Kong, Yike Guo, and Wei Xue. Llasa: Scaling train-time and inference-time compute for Llama-based speech synthesis. *arXiv preprint arXiv:2502.04128*, 2025. https://arxiv.org/abs/2502.04128.

Taehyung Yu and Seongjae Kang. Best-of-N TTS evaluation is confounded by ASR family alignment. *arXiv preprint arXiv:2607.08256*, 2026. Accepted at the ICML 2026 Workshop on Machine Learning for Audio. https://arxiv.org/abs/2607.08256.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. IndexTTS2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):35139–35148, 2026. doi: 10.1609/aaai.v40i41.40820. URL https://ojs.aaai.org/index.php/AAAI/article/view/40820.

A Data and Pronunciation Measurement

Chinese test sets. The seed set contains 143 words: rare characters (45), place names (26), surnames with a non-default reading (22), chemical elements (21), context-dependent polyphones (14), traditional-medicine terms (10) and polyphonic medicine terms (5). There are 49 one-character words, 93 two-character words and one four-character word. Each item provides its dictionary reading in toned pinyin, a carrier of 2–8 characters and a likely wrong reading; 73 carriers are of the form 这是 X。.

The mined set uses cjkvi-ids decompositions, Unihan, CC-CEDICT and wordfreq (cjkvi project, 2026; Unicode Consortium, 2026; CC-CEDICT contributors, 2026). Candidate characters are in GB2312 with zipf frequency at most 2.5. We decompose them two levels deep, remove generic components appearing in more than 60 characters, and collect more frequent sibling characters sharing the remaining components. Siblings with a different Mandarin reading supply distractors. A CC-CEDICT entry supplies the target word and reading, always in 这是 X。. The 1,527 eligible characters yield 1,440 items. Twenty-five seed words overlap the mined set, giving 1,558 distinct words and 2,800 word–character slots across the two sets. Frequencies are wordfreq zipf values, the base-10 log frequency per billion words.

The repair subsets contain 120 mined words with a common homophone, 33 byte-fragmented words and 50 words without a common homophone. The later three-sampling runs begin with 100 and 32 verified errors in the first two subsets; pooling them with the seed errors and deduplicating gives 193 words. The separate 50-item set is excluded from that pooled comparison. Controlled composition instead uses the frequent-word test/tuning split of Section 5.

Pronunciation labels. Whisper large-v3-turbo (faster-whisper), SenseVoiceSmall and Paraformer-zh transcribe each utterance (Radford et al., 2023; An et al., 2024; Gao et al., 2022). We align transcript and reference at the pinyin level, accepting alternative readings of polyphonic transcript characters and applying tone sandhi to the reference. Labels concern the target span: correct, tone error, segment error or missing. The fast protocol on common-word tests requires agreement by at least two of the three judges on the target syllable and tone.

For calibrated labels, teacher-forced Whisper scores common-homophone proxies of the intended and competing readings. Proxy scores are pooled by log-sum-exp (Kawajiri, 2026); the margin m is the log-probability difference between the intended reading and its strongest competitor. Competitors come from dictionary alternatives, phonetic components, sibling characters and open-judge transcripts. Thus the wrong-reading candidate is partly judge-derived.

For each flagged item, the TTS under test synthesises positive and negative controls in the same carrier using common characters with the intended and suspected readings, never the tested character. The calibrated score is

$$q = \frac{m_{\text{orig}} - m_{\text{neg}}}{m_{\text{pos}} - m_{\text{neg}}}. \tag{3}$$

An item is a verified error when $q < 0.5$ and a cleared judge artefact otherwise; the rule abstains when $m_{\text{pos}} - m_{\text{neg}} < 2$ nats. The threshold and separability gate were fixed before repair experiments.

Table 2: Initial diagnostic runs. Verified errors are divided by all synthesised items; abstentions are also shown. Later repair runs use their own baseline error sets. †: lexicon/G2P front end.

Set	System	items	verified errors	unjudgeable
seed	Tingting†	143	26 (18%)	8
	Edge†	142	17 (12%)	8
	Kokoro†	143	26 (18%)	23
	CosyVoice2	143	73 (51%)	12
	Fun-CosyVoice3	143	50 (35%)	15
	Spark-TTS	143	62 (43%)	9
	GLM-TTS	143	79 (55%)	9
mined	Edge†	1,440	28 (2%)	119
	CosyVoice2	1,439	676 (47%)	114

Validation and sensitivity. On the deterministic Tingting engine, calibration reaches 31 items with an exact waveform-oracle verdict. It correctly labels 13 errors and 11 correct readings and abstains on seven correct readings. Depending on the system, 10–66% of naive flags reaching calibration are judge artefacts. Counting all unjudgeable mined items as errors changes the Edge–CosyVoice2 contrast from 2% versus 47% to at most 10% versus 55%. Human validation is reported in Appendix E.

Relabelling stored margins over thresholds 0.35–0.65 and gates 1–3 nats changes CosyVoice2 seed errors from the default 73 to 66–77 and mined errors from 676 to 584–722; Fun-CosyVoice3 seed errors vary from 50 to 42–58. Across baseline samplings, per-word label agreement is only moderate ($\kappa = 0.53$–0.84). We therefore report both all-error repair rates and a stable error core, always compared with resampling that was not used to define that core. The main comparison uses the two-baseline core described below; earlier exploration records retain their separate three-baseline cores in the experiment archive.

B Implementation and Statistical Analysis

Embedding and serving interface. Each backbone uses its own embedding table and an HTTP wrapper accepting text plus a map of placeholders to vectors. The wrapper registers 16 or 64 placeholder strings as special tokens and substitutes the requested vectors in the embedding module's forward pass. A byte-fragmented target is replaced by one placeholder spanning all fragments. CosyVoice2, Fun-CosyVoice3, Spark-TTS and GLM-TTS use their zero-shot interfaces with fixed repository example prompts; GLM-TTS text normalisation is bypassed for vector-bearing requests because it would rewrite placeholders. TTS weights are frozen. The default vector norm and donor rule are specified in Section 4.

CosyVoice2's released text checkpoint carries Qwen2-0.5B rows, while Spark-TTS descends from Qwen2.5-0.5B (Yang et al., 2024; Qwen et al., 2024). Over the first 20,000 token ids, mean row-wise cosine is 1.000 between the released CosyVoice2 text checkpoint and Qwen2-0.5B, 0.56 between trained CosyVoice2 rows and that initial checkpoint, 0.21 between Spark-TTS and Qwen2.5-0.5B, and 0.000 between the two trained TTS tables. Shared token ids therefore do not imply shared embeddings.

Spark-TTS decoding. Initial runs returned empty audio on about 13% of short carriers. All reported Spark-TTS results use the same correction for baseline and intervention: forbid both end tokens for the first six semantic tokens per CJK character and resample if the output is still empty. Empty outputs are not counted as errors because the first runs did not record which items failed. This convention limits comparison with unconstrained decoding and is distinct from counting an inapplicable or unjudgeable repair as unsuccessful.

Repeated comparisons. Main CosyVoice2 and Spark-TTS repair conditions have three samplings. The baseline-error set is fixed from the initial sampling; stable cores are wrong in two baseline samplings (63 of 72 CosyVoice2 errors and 50 of 59 Spark-TTS errors). An additional baseline sampling supplies the held-out comparator. Each repeat generates its own baseline and positive/negative calibration controls. Rates are computed per sampling and averaged; overlapping words are counted once, and each comparison uses conditions available across its samplings. The pooled resampling row covers 171 words because the fragmented subset was not run in every sampling.

Intervals use 10,000 word-bootstrap resamples of per-word means. Differences use two-sided sign-flip tests on per-word differences with 20,000 permutations (minimum attainable $p = 5 \times 10^{-5}$). Three comparisons were fixed before the repeats: CosyVoice2 T3 composition versus regression on common words ($p = 0.02$), the same comparison on real errors under complete withholding ($p = 5 \times 10^{-5}$), and the unified method versus text rewriting ($p = 0.0013$). All survive Holm correction. The broken-word comparison was exploratory. On the two repeats alone, unified repair is 77.5% versus 69.9% for rewriting ($+7.5$ points, 95% interval $[+1, +14]$, $p = 0.02$).

Conditions within a sampling were run in one session on one machine. Repeats were run in a later session on another machine with two RTX 3090 GPUs; their vector tables, baselines and controls were regenerated there. Secondary single-sampling comparisons use paired discordant counts and exact binomial tests; their rates are not pooled with the repeated primary analyses.

Information supplied to comparison methods. Internal text rewriting uses the intended reading via a common homophone. Our method additionally needs embedding-lookup access. SonoEdit uses a corrected label and gradient-based weight editing; FlowEdit uses reference audio and conditioning optimisation; activation steering uses contrastive label pairs; UtterTune trains a LoRA from corrected labels; GRAFT splices reference audio into the prompt. Trained pronunciation channels require prior training support in the backbone. These distinctions describe the settings of the cited methods (Section 2); no empirical comparison with the post-hoc editing methods was run.

Record files. Per-item `results.jsonl`, verification files and paired-analysis dumps back the reported rates. Common-word exploration runs are in `out/gpu3/x1–x3`; initial calibrated repairs in `out/gpu4/v1_*`, `out/gpu3/v1m_*` and `out/gpu4/v1sp_*`; repeats, sensitivities and GLM-TTS controlled tests in `out/gpu5/`. Aggregate summaries include `v1x3_*.md`, `x3x3_test.md`, `spslot_x3_test.md` and `glm_slot_x3_test.md`. Korean records are in `out/gpu3/ko_*`; listening records in `out/human_*`. The source archive preserves the full earlier appendix, including exploratory runs, all-160-item summaries and the English measurement pilot; these are not additional primary comparisons.

C Held-Out Composition and Sensitivity

Table 3 gives the numerical results behind Figure 4. The 120 test items exclude the 40 reserved tuning items. T2 excludes all characters sharing the target's toned syllable from both fitting and donors; T3 excludes all tones of its initial–final pair. The TTS model itself is unchanged. Own rows use their natural norm; the CosyVoice2 own-row control rescaled to the composed norm scores 97%. One GLM-TTS test item has no same-final T3 donor, leaving 119 items for that composed condition.

Table 3: Held-out common-word accuracy (T2 / T3). Upper block: means of three samplings. †: one sampling; $*$: 119 rather than 120 items at T3.

Row written into the slot	CosyVoice2	Spark-TTS	GLM-TTS
own row	98%	97%	100%
composed: donors + difference	**94 / 94%**	**97 / 95%**	**99 / 98%***
regression alone	87 / 88%	92 / 90%	88 / 94%
nearest remaining real row to the prediction (T3)	10%	8%	15%
donors without the difference†	42 / 2%	64 / 8%	54 / 5%
random vector†	2%	1%	0%

Intervention controls. Common donors with another reading redirect 113/120 CosyVoice2, 106/120 Spark-TTS and 113/120 GLM-TTS outputs to the donor reading. Rare-homophone donors preserve the intended reading on 54/91 CosyVoice2 and 55/91 Spark-TTS items. For donor zipf strata < 2, 2–3, 3–4 and ≥ 4, the preserved counts are 28/55, 20/29, 5/6 and 1/1 on CosyVoice2 and 29/55, 20/29, 5/6 and 1/1 on Spark-TTS (Spearman $\rho = 0.24$, $p = 0.023$ and $\rho = 0.29$, $p = 0.0053$). In the first CosyVoice2 sampling, adding a random difference to the donors yields 43% at T2 and 2% at T3, close to the donors-without-difference control (42% and 2%).

Norm. Table 4 separates controlled common-word accuracy from the fast-protocol seed scan. The latter uses 62 stable errors, distinct from the calibrated cores. Homophone means remain at their natural length: stretching them to the full common-row norm lowers seed repair from 68% to 56% in the fast-protocol run, while their repair rate shows no consistent trend within the natural range. The calibrated full-versus-half-norm regression comparisons are retained in Table 5.

Donors and context. Using 3, 12 or 24 donors gives 88%, 93% and 93% T3 accuracy in one sampling. In a 25-character carrier, the unmodified model scores 99%, its own row through the placeholder 100%, composition 95%, regression alone 98% and the nearest remaining real row 5%. Thus the advantage over regression at the shared default norm is not uniform across contexts.

The /y/ subset. On nine affected CosyVoice2 test items, T3 composition, regression and own rows score 52%, 56% and 85%; on Spark-TTS they score 63%, 59% and 100%. The corresponding seven

Table 4: CosyVoice2 norm sensitivity, one sampling per cell. Common words: T3 fast-protocol accuracy on 120 items. Seed: fast-protocol repair on 62 stable errors and damage to baseline-correct words. Entries are percentages.

	Common words: correct		Seed: regression alone	
Norm c	composed	regression	repaired	broken
0.35	96	92	61	13
0.50	93	87	63	11
0.70	95	77	45	24
1.00	90	78	44	28

GLM-TTS items score 95%, 90% and 100%. Recoding pinyin *u* after j, q, x and y as *ü* gives 94.2% versus 94.8% under the original coding on the 115 shared CosyVoice2 test items (three samplings; -0.6 points $[-3, +2]$, $p = 0.83$). The recoding does not resolve the concentrated weakness.

D Repair Results and Preservation

Repeated seed comparisons. Table 5 reports all main seed conditions, including the stable-core and broken-word columns. The pooled and mined-subset comparisons appear in Table 1. With only the target's toned syllable withheld (T2), composition repairs 71% of the 100-error mined homophone subset and 71% of the 171 pooled errors on which it was run; the fragmented subset has no repeated T2 condition. On those same 171 words, unrestricted homophone averaging repairs 78%. All rates use the calibrated protocol.

Table 5: Seed-set repair: three-sampling means (†: one sampling). Denominators for CosyVoice2/Spark-TTS are 72/59 baseline errors, 63/50 stable-core errors and 63/71 baseline-correct words. "All tones withheld" excludes the target's initial–final pair from fitting and donors.

	CosyVoice2			Spark-TTS		
Condition	repaired	core	broken	repaired	core	broken
held-out resampling	11%	4%	16%	14%	7%	17%
regression alone, full norm	62%	62%	33%	55%	61%	41%
regression alone, half norm	69%	71%	23%	80%	83%	19%
single homophone's row†	62%	62%	24%	81%	82%	10%
internal text rewrite, same homophone	66%	68%	27%	79%	79%	12%
homophone mean (distance 0 only)	76%	77%	13%	81%	81%	7%
unified (nearest donors + difference)	**76%**	**76%**	**11%**	80%	83%	15%
homophones withheld: composed	72%	71%	20%	–	–	–
all tones withheld: composed	78%	78%	24%	81%	81%	20%
all tones withheld: regression alone	68%	68%	25%	77%	81%	23%

Additional backbones and controls. Table 6 reports the single-sampling comparisons supporting Section 6.3. Fun-CosyVoice3 homophone transfer and trained pinyin tokens are not separated ($p = 0.21$ on paired judgeable errors). GLM-TTS has no held-out fourth baseline sampling; its second and third baselines repair 5/79 and 4/79 errors. These exploratory runs use their own denominators and are not averaged into the later repeated comparisons.

Table 6: Additional-backbone seed results, one sampling. Repair is over all baseline errors; broken counts are over baseline-correct words still judgeable after intervention.

Backbone	Intervention	repaired	broken
Fun-CosyVoice3	trained pinyin tokens	35/50 (70%)	10/74 (14%)
	homophone mean, target only	41/50 (82%)	8/68 (12%)
GLM-TTS	homophone mean, target only	61/79 (77%)	6/48 (12%)

On an exploratory CosyVoice2 seed run, random vectors repair 9/73 errors versus 11/73 for resampling, make 46/143 items unjudgeable and break 29/37 judgeable baseline-correct words. Removing

the tone feature from full-norm regression reduces repair from 46/73 to 24/73, with 27 versus 5 discordant errors ($p = 10^{-4}$), and breaks 32/52 correct words. The full exploration tables and pairwise contrasts are retained in the experiment archive.

Ordinary sentences. The preservation study uses 300 Seed-TTS-eval test-zh sentences (ByteDance Speech Team, 2024; Anastassiou et al., 2024). A sparse intervention replaces one random common character per sentence with its homophone mean. A stress test replaces every eligible character: 5,293 of 6,038 positions (88%). Each design has its own baseline sampling. Table 7 reports paired changes in Paraformer CER and ERes2Net speaker similarity (Chen et al., 2023). Sparse changes have intervals including zero. Widespread replacement increases CER; including SenseVoice, the increase ranges from 2.5 to 4.7 points. DNSMOS overall rises by 0.02 on every backbone in that stress test (Reddy et al., 2022); it was not measured for sparse replacement.

Table 7: Preservation on 300 sentences. Paired changes with 95% intervals; CER in percentage points, speaker similarity as cosine difference. Sparse: one character per sentence. Widespread: every eligible character.

	Sparse replacement		Widespread replacement	
Backbone	ΔCER	ΔSIM	ΔCER	ΔSIM
CosyVoice2	-0.03 ± 0.47	$+0.002 \pm 0.004$	$+4.25 \pm 0.96$	$+0.009 \pm 0.004$
Fun-CosyVoice3	$+0.01 \pm 0.40$	-0.002 ± 0.005	$+2.52 \pm 0.72$	$+0.014 \pm 0.004$
Spark-TTS	-0.41 ± 0.74	$+0.004 \pm 0.005$	$+2.86 \pm 0.73$	-0.002 ± 0.003

Natural carriers. We also place the CosyVoice2 seed targets at the start, middle or end of 20–30 character sentences. Target-only homophone transfer repairs 76–85% (Table 8). Baseline labels agree with the short-carrier run on 83%, 77% and 82% of words, close to baseline repeatability in short carriers (84–86%). This check concerns homophone transfer, whereas the long-carrier composition comparison is in Appendix C.

Table 8: CosyVoice2 natural carriers, one sampling. Baseline errors are over judgeable words; repair is over all baseline errors, and broken counts over judgeable baseline-correct words after intervention.

Target position	baseline errors	repaired	broken
sentence start	75/133 (56%)	57/75 (76%)	6/53 (11%)
sentence middle	66/130 (51%)	56/66 (85%)	6/57 (11%)
sentence end	69/129 (53%)	55/69 (80%)	6/54 (11%)
short carrier	73/131 (56%)	50/73 (68%)	8/50 (16%)

E Listening Studies

Blind A/B transfer study. Three native Mandarin speakers (one author, two volunteers) judge 98 CosyVoice2 sentence pairs: unmodified audio and whole-word homophone-mean transfer, with side order randomised and the condition hidden. Each screen asks which side pronounces the target correctly (A, B, both or neither) and which sounds more natural. There are 68 baseline-error pairs and 30 baseline-correct controls. By majority vote, the edited side is correct on 67/68 error pairs (99%, Wilson 95% interval $[92, 100]$), versus 5/68 for baseline, and breaks 4/30 controls (13% $[5, 30]$). Fleiss' κ on the four-way answer is 0.61. Agreement between human majority and the calibrated judge is 90% over 192 judgeable sides ($\kappa = 0.78$).

Two raters prefer the edit's naturalness in 32 and 23 pairs, prefer baseline in 13 and 7, and otherwise report no difference. The third gives the same naturalness and correctness response on 90/98 pairs. These responses are reported separately from pronunciation correctness.

Detection study. Two listeners, including one author, label 188 stratified utterances from Tingting, Edge, Kokoro, CosyVoice2 and Fun-CosyVoice3 as correct, wrong or unsure, seeing the intended reading but no judge output. The strata cover errors, tone errors, cleared flags, unflagged items and abstentions. Listener agreement is 93% ($\kappa = 0.87$). On the 157 items where both listeners agree on a

decided answer and the pipeline also decides, pipeline agreement is 91% ($\kappa = 0.82$), with error-call precision 90% and recall 92%. Of 47 cleared flags, 41 are heard as correct; none of 29 unflagged items is heard as wrong. Tone calls have the weakest agreement (80%).

Composition study. Three independent listeners rate 165 CosyVoice2 clips for initial/final correctness, tone, neighbouring-word damage, naturalness and the pinyin heard when wrong. Clips are shuffled, with condition and automatic label hidden. A clip is correct when the majority accepts both segments and tone. Table 9 reports every sampling stratum. Controlled-test clips come from its second sampling, except the donor-only clips, which come from the first.

Table 9: Composition listening check, 165 clips; majority of three independent listeners. Brackets are Wilson 95% intervals. Automatic labels use the fast protocol on common words and calibration on real errors.

Row written into the slot	judges	clips	right for the listeners
composed, T2 (other-tone donors + tone dir.)	accept	20	20 (100%)
	reject (all such clips)	11	7 (64% [35, 85])
composed, T3 (same-final donors + difference)	accept	20	20 (100%)
	reject (all such clips)	10	3 (30% [11, 60])
regression alone, T3	accept	10	9
	reject	10	2
own row	accept	10	10
nearest real row, T3	reject	10	0
donors only, T2 (sampling 1)	accept with tone	12	3 tone (12 initial+final)
	accept base only	12	0 tone (12 initial+final)
all withheld, real errors	repaired	25	23 (92% [75, 98])
	not repaired	15	8 (53% [30, 75])

For composed common-word rows, all 40 checked accepted clips are correct for listeners, as are 10/21 rejected clips. Weighting the T2 strata by 109 accepted and 11 rejected test items gives 96.7% listener-adjusted accuracy versus 90.8% automatically; weighting T3 by 110 and 10 gives 94.2% versus 91.7%. The analogous T3 regression estimate is 80.7% versus 86.7%. These are stratum-weighted estimates for that sampling, not a full-set listening test.

Listeners hear the donor's tone in 21/24 donor-only clips, including 9/12 automatically accepted with tone. On real errors under complete withholding, listeners reject 2/25 clips labelled repaired and accept 8/15 labelled not repaired. Pairwise right/wrong agreement is $\kappa = 0.92$, 0.92 and 1.00. At least one listener marks a damaged neighbour in 9/165 clips across all tested conditions. The 165 clips are distinct from the additional study below.

Natural no-common-homophone study. Three independent native Chinese speakers, none an author, rate 193 interleaved CosyVoice2 clips with condition labels hidden. They are informed about the task and consent before participation. The sample contains 50 baseline clips, 50 half-norm regression clips and 45 clips using other-tone donors plus the predicted difference for the 50-word no-common-homophone cohort, plus four conditions on 12 matched common words (48 clips). This 193-clip sample is separate from both the 165-clip study above and the 193-word pooled repair analysis.

The interface asks separately about segments, tone, neighbouring-word damage and naturalness (1: unnatural; 2: intermediate; 3: natural), with optional problem tags and heard pinyin. A listener's correctness vote requires both segments and tone to be correct; the clip label requires at least two such votes. The first listener leaves the four rating fields blank on two baseline clips (`s136`, 锘; `s179`, 瞍), marking omission/unintelligibility instead. We do not impute these votes; the other two listeners agree that each is wrong. Every clip therefore has a decided majority label. Pairwise Cohen's κ is 0.940, 0.882 and 0.875 on 191, 191 and 193 jointly rated clips, respectively.

Table 10: Human labels on the natural no-common-homophone cohort, one stored sampling. "Correct" is joint segment-and-tone majority. Repair and breakage are conditional on the same words' human-rated baseline labels. The composition row covers 45 available outputs, not all 50 selected words.

Condition	target correct	baseline errors repaired	baseline-correct broken
unmodified baseline	14/50 (28%)	–	–
regression, half norm	36/50 (72%)	29/36 (81%)	7/14 (50%)
other-tone donors + difference	36/45 (80%)	26/32 (81%)	3/13 (23%)

The Wilson 95% intervals for target accuracy are $[17, 42]$, $[58, 83]$ and $[66, 89]\%$ for baseline, regression and available composed outputs. No composed clip is present for 发扬踔厉, 搦战, 蒽醌, 蔫儿 or 锘. Counting those five as not confirmed correct gives 36/50 (72%) confirmed coverage; it is not an observed 50-clip accuracy. On the 45 shared words, composition scores 36/45 against regression's 32/45, with 8:4 discordant pairs (two-sided exact $p = 0.388$; $+8.9$ points, word-bootstrap 95% interval $[-6.7, +24.4]$). Against baseline, regression has 29:7 discordant pairs on 50 words and composition 26:3 on 45 words; raw exact p values are 3.13×10^{-4} and 1.52×10^{-5}, and Holm-adjusted values within these three exploratory comparisons are 6.25×10^{-4} and 4.57×10^{-5}. These tests do not replace the earlier confirmatory comparisons.

The 14 correct baseline clips in this stored batch make accuracy different from a repair rate. Table 10 therefore reports both, along with damage to baseline-correct words. Neighbour damage by majority vote is observed on 3/50 baseline, 1/50 regression and 0/45 composed clips. The common-word controls are correct on 12/12 own-row, 12/12 T2-composed, 11/12 T3-composed and 10/12 T3-regression clips; two T2-composed clips have majority-rated neighbour damage. These 12 words belong to the existing test partition; the checks are descriptive and do not replace the 120-item repeated experiment. The new labels apply only to the presented stored clips and are not propagated to other samplings. An anonymised manifest, answer snapshots, per-item votes and a standard-library scoring script are included in the accompanying analysis files.

F Korean Evaluation

Inventory and construction. The Qwen2.5 tokenizer shared by CosyVoice2 and Fun-CosyVoice3 represents 2,512 of the 11,172 precomposed Hangul blocks as single tokens and 8,660 as byte fragments. Frequency is the sum of frequencies of words containing the block in a 30,000-word Korean list. There are 1,615 occurring single-token blocks and 897 zero-frequency single-token blocks. Probes decode initial, medial and final letters with 85/78/83% accuracy on the occurring rows of CosyVoice2 and 92/81/88% on Fun-CosyVoice3 (5-fold); on zero-frequency rows these fall to 22/16/20% and 25/19/25%.

A ridge map predicts rows from one-hot initial, medial and final letters, withholding every block in the tested set. Composition uses up to 12 frequent training blocks sharing two of the three letters, adds each target-minus-donor prediction, averages and writes the result at half the mean common-row norm. The regression-only, donor-only and random controls use the same half norm; full-norm regression is also reported below.

Items and pronunciation matching. The main tests draw 200 frequent eligible blocks from the top-500 single-token group, all 347 eligible zero-frequency blocks and 900 eligible fragmented blocks. The pilot blocks that fixed the design are excluded. Targets are open syllables or have final ㄱ, ㄴ, ㄹ, ㅁ, ㅂ or ㅇ, avoiding coda clusters and the selected frame's palatalisation/neutralisation cases. Both backbones read 이것은 X 입니다 with a fixed synthetic Korean prompt voice.

Whisper large-v3-turbo and SenseVoice transcribe in Korean mode. Transcript and target are decomposed into letters, dropping silent initial ㅇ to handle liaison. The comparison folds ㅐ/ㅔ, ㅒ/ㅖ, ㅙ/ㅞ/ㅚ, post-consonant ㅢ/ㅣ, and drops the glide after ㅈ, ㅉ and ㅊ. A target is heard when the vowel and final of 은, the target letters and the vowel of 입 occur contiguously. The default accepts either judge; the strict criterion requires both.

Each sentence is also synthesised by the macOS Yuna reference voice. A test item is retained only if the corresponding criterion accepts this reference rendering. The default retains 177/200

frequent, 78/347 zero-frequency and 154/900 fragmented items; the strict criterion retains 131, 19 and 58. Hence the rare/fragmented evaluation covers 232 items by default and 77 under the strict criterion. The reference filter keeps only 22% and 17% of the two rare groups; this is a selected, ASR-transcribable subset, with no Korean listening study.

Rare and fragmented blocks. The native rows (or fragments) are sampled three times. The stable core is wrong in samplings 1 and 2; sampling 3 is held out. Every intervention is one sampling, and failed or empty synthesis counts as wrong. Table 11 combines the two criteria while retaining the regression and donor-only controls. On the default cores, composition versus resampling has 94:0 discordant pairs on CosyVoice2 and 114:3 on Fun-CosyVoice3; the strict cores give 38:0 and 44:0. Tests are paired exact binomial tests. Composition improves over resampling, but its advantage over regression is not consistent across backbones and criteria.

Table 11: Korean rare/fragmented stable-core repair, percentages. Both criteria filter the reference and evaluate model audio by the same rule. Default and strict criteria have different retained items and cores. Except where stated, constructed vectors use $c = 0.5$.

	Either judge (default)		Both judges (strict)	
	CosyVoice2	Fun-CosyVoice3	CosyVoice2	Fun-CosyVoice3
core / retained items	198/232	195/232	64/77	62/77
held-out resampling	1	5	0	2
regression, full norm	46	53	55	65
regression, half norm	47	63	50	68
composed row	48	62	59	73
donors without difference	18	23	22	27
random vector	0	0	0	0

Held-out frequent blocks. On 177 default-filtered frequent blocks, native-row accuracy across three samplings is 74–80% on CosyVoice2 and 81–86% on Fun-CosyVoice3. Composition gives 69% and 82%, and half-norm regression 71% and 66%. Under the strict criterion on 131 blocks, composition is 53% versus 69–70% for native CosyVoice2 rows (14:36 discordant pairs, $p = 0.003$), while Fun-CosyVoice3 composition is 77% versus 77–80% for native rows. Intermediate-frequency results and per-group tables are preserved in the experiment archive. Records are `out/gpu3/ko_eval_{cv2,cv3}_{big,mid}.md` and their `_both` counterparts, with reference transcripts under `ko_yuna_{big,mid}`.