VeriSpeech-Repair: Geometry-Constrained Local Repair with Clean-Context Audits

Anonymous authors
Paper under double-blind review

Abstract

A speech editor may recover a missing word while changing audio outside the intended edit. We study this tension with VeriSpeech-Repair: a fixed set of local speech patches, an explicit no-repair choice, and separate tests for transcript recovery and preservation outside the original corruption span. On controlled LibriSpeech corruptions, broad patch selection roughly halves automatic word error rate (WER) but reduces exact outside-span context by an absolute 22.1% among outputs that can be aligned to the original audio. A score-free geometry rule preserves exact context under local splicing and lowers WER by an absolute 2.8% on 782 sources after excluding earlier calibration sources. In a separate comparison on those sources, the configured systems differ in architecture, input information, and proposal count. dots.tts.edit achieves 52.259% automatic transcript pass on corrupted cases, compared with 28.261% for the risk controller. Only five of the editor's 2,346 corrupted outputs align to the original time map; none has exact context, and the other 2,341 cannot be assessed by this fixed-map measure. Transcript recovery and fixed-map preservation therefore give different outcomes for these systems.

1 Introduction

Speech editors may recover a missing word while altering speech outside the requested edit. Such collateral change matters in media correction and accessibility, where preserving the surrounding utterance is part of the task. A local repair policy must therefore test content recovery and preservation separately.

The anonymous source code and toy example are available at the project repository.[1]

Upstream tokenizers, generators, and reward models improve capability (Yang et al., 2025; 2026; Cai et al., 2026; Wang et al., 2026d; Tian et al., 2025), yet fluent audio can lose facts or commitments (Xie et al., 2026). Residual repair needs bounded decisions and an explicit preservation test.

Figure 1 shows our decision contract. Given flawed audio, target text, and localized failure evidence, the system generates a compact proposal family, then either retains the utterance or applies one bounded repair using content scores, recorded edit geometry, or both.

Our question is whether a fixed policy can recover target content without changing speech outside the intended region. Table 2 shows that broad boundary-grid selection improves automatic content, but the clean-reference audit also reveals losses in outside-span context. Table 4 evaluates a score-free geometry rule on a calibration-disjoint subset and tests whether a content gate survives rescoring with frozen decoders. Table 3 asks a separate question: how does the local controller's transcript recovery compare with dots.tts.edit, and how many editor outputs can be assessed with the same fixed source-time map?

Contributions. We make three contributions:

- We formulate local speech repair as a verifier-constrained choice between retaining the input and accepting one bounded proposal.

[1] `https://anonymous.4open.science/r/verispeech-repair-95EE/README.md`

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.

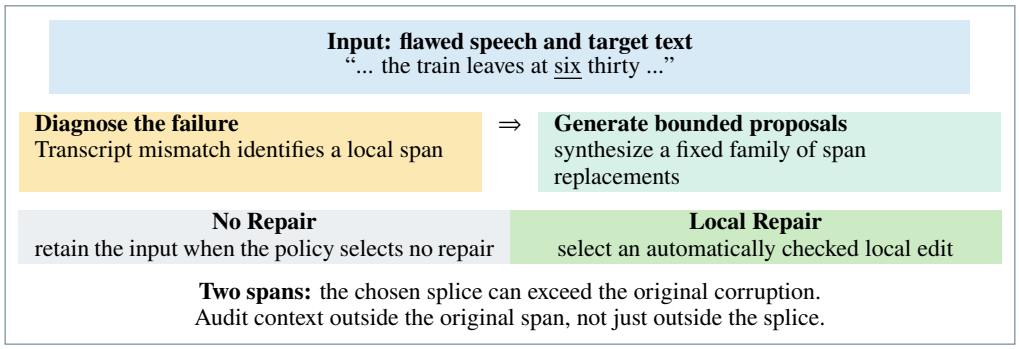


Figure 1: Core local-failure setting for VeriSpeech-Repair. The system starts from flawed speech, proposes bounded repairs, and retains no-repair as an explicit option. Action-local audio checks verify the recorded splice boundary but are not a policy-quality metric.

- We introduce a two-part evaluation that measures target-content recovery and clean-reference context outside the original corruption span, while separating this outcome from action-local splice verification.
- We evaluate score-free geometry and geometry-plus-content rules on a calibration-disjoint source subset and frozen VITS proposals, and compare a direct editor on the same subset. Known-span geometry enforces exact off-span equality under local splicing. dots.tts.edit has stronger automatic transcript recovery, but most outputs fall outside the fixed time map. The incremental score gate does not transfer to two frozen post-selection evaluators.

Secondary decoder, generated-failure, reward, and backend checks probe evaluator and proposal sensitivity.

2 Related Work

Generators and sparse repair. ALMTokenizer, DualSpeechLM, UniAudio 2.0, and Gogo factor speech into semantic, acoustic, or codec-token representations for increasingly unified and controllable generation (Yang et al., 2025; Wang et al., 2026e; Yang et al., 2026; Chen et al., 2026). UALM is the closest foundation-model comparator because its audio-language model includes a self-reflection loop (Tian et al., 2025). It densely rewrites an utterance; our policy retains no-repair and accepts only a bounded patch. A fair direct comparison would require matched failure-conditioned audio, independent evaluators, and a restoration measure not defined by either method's action boundary. We therefore study the post-generation decision layer, not a new generator or a UALM head-to-head.

Speech editing, preservation, and abstention. SpeechSEC, VoiceNoNG, FluentEditor, and SpeechPalette address localized content, consistency, or hallucination control (Liang et al., 2025; Huang et al., 2025; Liu et al., 2024; Wang et al., 2026c); Voicebox, VoiceCraft, and VoiceCraft-X provide masked or token-infill editing, and SpeechEditBench evaluates bilingual multi-attribute edits (Le et al., 2023; Peng et al., 2024; Zheng et al., 2025; Zhang et al., 2026). The closest recent preservation study is dots.tts.edit and its bilingual doteBench suite (Wang et al., 2026b). It trains a 2B-parameter editor from structured transcript instructions and evaluates instruction following, non-edit recognition, aligned duration and pitch, speaker similarity, and predicted audio quality. Our intervention is a reject-option policy over already generated patches: it may retain the input, and its controlled diagnostic compares output with the clean source at the sample level outside the original corruption span. These are different tasks and preservation denominators. We compare dots.tts.edit and our controller on the same controlled inputs and evaluators in Table 3; published doteBench results remain contextual and are not pooled with this experiment.

Local generation, rewards, and memory. Masked Text-to-Audio Flow-Matching combines masked generation with reward feedback (Huang et al., 2026); UniSRM, SupCLAP, and MMSU provide finer speech and audio evaluation signals (Wang et al., 2026d; Luo et al., 2026; Wang et al.,

2026a). We keep the primary acceptance rule transparent and treat learned rewards as secondary checks. VoiceMem further motivates non-regression tests because a local correction should preserve earlier facts and commitments, although we do not evaluate VoiceMem directly (Xie et al., 2026).

3 Problem Setting

Each repair case contains target text, optional prompt or reference audio, a candidate audio file, and optional labeled edit spans. The goal is to output either the original candidate or a locally repaired candidate. This study evaluates automatic content improvement using automatic speech recognition (ASR) and records the locality of each implemented action. It does not treat action locality as evidence of restoration, speaker, or perceptual quality.

Formally, let a case be $(x, a^{\text{cand}}, a^{\text{ref}}, s)$, where x is target text, a^{cand} is candidate audio, a^{ref} is optional prompt or reference audio, and s denotes optional span supervision. A repair backend proposes a set of local candidates $C = \{a_0, a_1, \ldots, a_K\}$ that always includes the no-repair choice $a_0 = a^{\text{cand}}$. A verifier records automatic content outcomes for each candidate. A fixed selection policy may rank those outcomes, restrict actions by recorded geometry, or combine both; it returns a_0 whenever no repair is selected.

The recorded action space has two outcomes: repair and no repair. The no-repair record does not distinguish an already acceptable input, insufficient automatic evidence, or the absence of an eligible proposal.

This formulation separates generator quality from decision quality. A strong generator can still produce a poor local edit, and a proposal from a weaker generator can be useful if the verifier selects it under the stated automatic contract.

4 Method

4.1 Candidate Generation

We study two frozen backend families. VITS supplies lightweight span patches (Kim et al., 2021); SpeechT5 repeats the same interface (Ao et al., 2022). Let t be the normalized target tokens and h the primary automatic speech recognition (ASR) hypothesis. The first Levenshtein difference gives the target patch tokens and a hypothesis-side boundary. The nine local proposals comprise five timing rules plus four delete-boundary shifts. The five rules are: the ASR word-time interval; the target-difference token interval scaled uniformly over the controlled reference duration; that duration centered on the ASR interval; for deletion, the reference–candidate duration gap placed at the ASR start; and the same gap placed after the hypothesis-side difference. The last two fall back to the ASR interval for other operators. Four additional deletion proposals shift the final insertion point by -0.20, -0.10, $+0.10$, or $+0.20$ seconds.

Each backend synthesizes only the target patch text. We linearly resample the patch to the proposed action length, match its root-mean-square level to 0.20 seconds of adjacent input with a 0.98 peak cap, and concatenate samples without cross-fading: mute replaces the interval, delete inserts at its start, and duplicate removes the second copy. The controlled reference duration and recorded corruption span help construct the proposal grid; the controller receives neither the clean waveform nor a clean-reference score. "Failure-conditioned" means that ASR word diagnostics localize proposals, not that natural faults are localized online. Two frozen backends test interface transfer but do not establish generator independence across speech editors.

4.2 Verification and Selection

Each candidate is transcribed with Transformers `openai/whisper-small.en` (Whisper-small) and scored against target text (Radford et al., 2022). Text normalization applies Unicode NFKC, lowercasing, punctuation removal except apostrophes, underscore replacement, and whitespace collapse. Per-case WER is word-level edit distance divided by the number of target words; reported WER is its arithmetic mean over cases. The score is $\max(0, 1 - \text{WER})$, and pass requires a score of at least 0.995. We report pass, WER, and mean score; normalized exact transcript equality is retained in the

machine-readable audit but is not a separate selection objective. The original boundary-grid selector ranks candidates lexicographically by score, pass, and negative word-error count. No-repair is included in every candidate set.

The headline selector uses Whisper-small. Post-selection checks use larger Whisper and Wav2Vec 2.0 connectionist temporal classification (CTC) decoding (Baevski et al., 2020) on frozen outputs; neither enters generation or selection. These LibriSpeech-domain checks assess evaluator sensitivity.

4.3 Diagnostics

We distinguish action-local equality from clean-reference context. Let clean waveform r have original corruption interval $[b, e)$ and length $L = e - b$. Its context is $C(r) = r[: b] \oplus r[e :]$. For an output y, the operation-aware map removes $[b, e)$ for identity, mute, or clean-duration output; uses all of y for a source-duration deletion; and removes $[b, e + L)$ for a source-duration duplicate. Any other operator–duration combination is unaligned. Among aligned outputs, exact context is $\mathbb{1}[C(y) = C(r)]$, including equal length and sample values. We separately report the unconditional exact-context rate, which counts unaligned outputs as failures on the full case denominator. Context mean squared error (MSE) zero-pads the shorter mapped sequence before averaging squared sample differences and is defined only for aligned outputs. The unaligned rate always uses all cases. Unalignment means that this fixed map cannot assess off-span samples; it does not establish that those samples changed.

For example, if clean audio is $u \oplus v \oplus w$ and deletion produces $u \oplus w$, inserting a length-$|v|$ patch gives $u \oplus \hat{v} \oplus w$ and both contexts are $u \oplus w$; an output of any other duration is unaligned. More generally, if the corrupted input equals r outside $[b, e)$, the chosen action is contained in that interval, output duration matches the operator-specific clean duration, and assembly copies every off-span sample, then $C(y) = C(r)$ by prefix–suffix identity. This is the construction invariant tested by the audit. A separate action-local audit compares input and output outside the chosen splice interval and verifies that no-repair returns the input bit for bit. This implementation check cannot rank exact-splice policies or establish clean restoration. Content intervals use paired source-cluster bootstrap resampling, keeping all four variants from one utterance in one unit.

5 Experimental Setup

The n128 and n512 names count source utterances, not repair examples. Each source yields four variants, so n128 contains 512 examples and each n512 split contains 2,048 examples. We fix all selector settings before scoring the source-disjoint n512 data. For paired uncertainty, we resample source utterances so all variants derived from one source remain in the same bootstrap unit. The headline analysis excludes identity controls, retains 3,072 corrupted cases from 1,024 sources, and uses 10,000 paired bootstrap draws with a fixed seed and percentile 95% intervals. These intervals quantify source variation conditional on one frozen proposal realization; they do not include synthesis-draw variation.

5.1 Data and Evaluation Settings

Experiments use cached LibriSpeech clean-test utterances (Panayotov et al., 2015). The n128 setting contains 512 repair examples. The main n512 setting contains 2,048 examples from source utterances 0–511, and the second n512 setting contains another 2,048 examples from source-disjoint utterances 512–1023. The two n512 settings share one speaker, so they are not speaker-disjoint. Each source yields identity, mute-middle, delete-middle, and duplicate-middle variants. Identity examples measure no-op behavior and source-ASR limits; corrupted examples test content recovery.

We use 128 FLEURS en-US utterances, expanded into 512 four-variant examples, only for decoder calibration under the same audio-backed protocol (Conneau et al., 2022). No FLEURS repair outcome enters the quantitative results or the paper's conclusions. Finally, SpeechT5 generated-failure subsets test selective repair on flawed full-utterance outputs. The 71-source n128 and 319-source first-n512 subsets share 64 sources. All 71 n128 failures and 76 first-n512 failures also occur in calibration; 206 first-n512 failures occur in the 782-source controller subset. These overlapping diagnostics are not independent replications.

Table 1 separates evaluation settings from calibration and diagnostics.

Table 1: Evaluation settings. Each setting isolates a distinct part of the evidence.

Setting	Cases	Evaluation role
LibriSpeech n128	512	Span calibration; selector ablations
Main LibriSpeech n512	2,048	Primary boundary-grid evaluation
Second LibriSpeech n512	2,048	Source-disjoint validation
SpeechT5 failures n128	71	Gated repair on flawed outputs
SpeechT5 failures n512	319	Larger selective-repair stress test
FLEURS en-US	512	Decoder calibration only

Both n512 runs reuse n128 settings: Whisper-small scoring, the English Massively Multilingual Speech (MMS) VITS checkpoint, 0.20 s context, a 0.98 peak limit, and the $\{-0.20, -0.10, +0.10, +0.20\}$ boundary shifts.

Each boundary-grid case evaluates ten candidates: the no-repair action plus nine local variants. Each n512 split therefore produces up to 18,432 synthesized patch candidates and scores 20,480 candidate instances including no-repair. This offline budget excludes online latency.

5.2 Frozen Clean-Reference-Free Context Controller

The clean reference is excluded from every controller input. We sort source IDs by a SHA-256 key and reserve the first 128 sources (512 four-variant cases) as a fixed development partition; the remaining 896 sources (3,584 cases) form the original controller evaluation. A later provenance audit found that 114 of these 896 sources occur in the earlier n128 calibration artifacts. We therefore use the fixed 782-source, 3,128-case subset obtained by removing every n128 source for the primary controller comparison, and retain the 896-source analysis only as a transparent sensitivity in Appendix A.4. The geometric rules are specified directly from the corruption map and are not fitted on either partition. Source exclusion does not make their design independent of earlier inspection of the original 896-source cohort. No context target, reference audio, or reference-context metric is used to tune a decision. Score-using controllers receive the existing Whisper-small verifier pass, score, and word-error fields together with the corruption operator, recorded original failure span, recorded repair span, and candidate and corrupted-input waveform durations. Geometry-only receives just the operator, spans, durations, and candidate label. The operator and original span are known synthetic-failure metadata, so this is a controlled geometry test and not a system for localizing faults in natural audio. Thus the calibration-disjoint controller audit separates clean-reference evaluation from selection. Score-using policies remain coupled to the automatic content evaluator; geometry-only removes score access but shares the same frozen candidate grid.

Let n_0 be corrupted-input length, L the original failure-span length, and n_c candidate length. The operator-specific expected length is $n_0 + L$ for delete, $n_0 - L$ for duplicate, and n_0 otherwise; duration consistency requires an error of at most two samples. Hard geometry also requires the recorded repair action $[p, q)$ to satisfy $b \le p \le q \le e$. We define outside-span displacement as $d = \max(0, b - p, q - e)$, so every contained proposal has $d = 0$. Contained proposals therefore often tie on geometry.

The implemented policies use the following rules:

- Geometry-only minimizes $(d, |n_c - n_{\text{exp}}|, \text{candidate label})$ among hard-geometry proposals, otherwise returning no repair.
- Hard feasibility requires hard geometry and a lexicographically greater (pass, score, −WER, −word errors, candidate ID) tuple than no repair, then maximizes that tuple.
- The risk controller permits duration-consistent proposals with $d \le L$ and the same lexicographic eligibility, then minimizes displacement, duration error, the negatives of the first three content fields, and label.

The final ID participates in eligibility: numerically tied proposals can replace no repair. These are non-regressing content rules, not strict numeric-improvement gates. Geometry-only never reads con-

tent outcomes or clean references. Under exact local-splice assembly, known-span containment, and duration alignment, off-span equality follows from copied context. The comparison measures coverage and automatic content at the same nine-proposal budget; clean-reference context is evaluated only after decisions are frozen.

5.3 Direct Editor Comparison

We also compare the risk controller with dots.tts.edit (Wang et al., 2026b) on the same fixed 782-source subset. It contains three corrupted variants per source (2,346 cases) and one unchanged identity control (782 cases). The sources come from the same 1,024-source headline pool: 333 carry earlier Main/development labels and 449 carry Second labels. Removing n128 calibration sources does not make an untouched test set. The editor receives the corrupted audio and a word-difference instruction built deterministically from its input ASR transcript and target text; insertions, deletions, and substitutions are marked in that instruction. It receives no original corruption-span coordinates. It produces one output per corrupted case; identity controls copy the input. The risk controller selects from nine VITS local proposals and no repair using known synthetic-failure spans. Its permitted off-span displacement removes the geometry-only exact-context guarantee. The comparison mixes input information, model architecture, and proposal budget. We score all three arms with shared text normalization and the same Whisper-small pass threshold and original-span context map, and assess transcript sensitivity with frozen larger-Whisper and Wav2Vec2 CTC decoders. Paired intervals use 10,000 bootstrap draws over 782 source utterances, keeping each source's four cases together. The context auditor uses the clean source only after generation and selection. Its fixed-map alignment requirement cannot assess time-warped preservation or perceptual quality.

5.4 Execution

The boundary-grid headline selection and its input baseline use Transformers `openai/whisper-small.en`; each selected record retains the transcript and score of its chosen candidate. A separate Faster-Whisper `small.en` run, using the converted `Systran/faster-whisper-small.en` snapshot `d1d751a5f827`, rescores the frozen input and selected waveforms in Appendix Table 6. The earlier headline run records identify the Transformers backend and model but do not retain its exact runtime revision. The editor comparison uses `openai/whisper-small.en` at revision `e8727524f962`, with Transformers 4.57.6 transcription and word timestamps; VITS uses the English MMS checkpoint for patch synthesis, and SpeechT5 supplies the transfer backend with its text-to-speech and HiFi-GAN components. A cached Faster-Whisper large-v3-turbo model and the fixed `facebook/wav2vec2-base-960h` CTC checkpoint with greedy decoding (revision `22aad52d435e`) (Baevski et al., 2020) rescore frozen no-repair, geometry-only, and hard-feasibility outputs after all decisions are fixed. The editor rescore uses Faster-Whisper 1.2.1 and CTranslate2 4.4.0, float16 and word timestamps, with the remaining transcription arguments at library defaults. The editor comparison uses a separate frozen decoding run of the same cases: three larger-Whisper no-repair transcripts differ from the controller audit, slightly changing WER with unchanged pass rate. The older controller decoding protocol is incompletely recorded. These evaluators do not alter synthesis, policy choice, or controller geometry. The source allocation, controller definitions, metric normalization, and evaluator checkpoints are specified here; anonymous code and generated audio are not provided in this manuscript package. The matched dense baseline used in the reward-reranking audit (Appendix Table 13) reuses the same generated-failure inputs and non-regression checks as sparse repair. A cached large-v3-turbo audit scores all 4,096 raw four-variant n512 examples before boundary-grid selection; matched selected outputs are scored separately. The headline analysis uses 3,072 corrupted cases; its 1,024 identity controls are reported separately. Frozen evaluator checks do not make candidate generation or selection evaluator-independent.

6 Results

Table 2 reports Transformers Whisper-small automatic content outcomes on corrupted cases only. On the pooled 1,024-source benchmark, the fixed boundary-grid selector changes pass rate from 12.0% to 30.8%, WER from 27.0% to 13.2%, and mean verifier score from 0.7569 to 0.8683. For selected minus input, source-cluster paired 95% intervals for the absolute pass-rate difference (%)

and WER difference (%) are [+17.3%, +20.3%] and [-16.3%, -11.7%]; the score difference has interval [+0.1066, +0.1164]. The two source-disjoint n512 portions have the same direction. A separate Faster-Whisper small.en rescore of the frozen waveforms also improves pass and WER (Appendix Table 6). Identity controls are reported separately. Their small score difference arises from six selections whose candidate labels point to the input waveform but carry different stored scores (Appendix Section A.2). These are automatic outcomes, not independent evidence of naturalness, speaker identity, or deployment safety.

Table 2: Automatic content outcomes under the fixed Transformers Whisper-small selector. The first three settings contain corrupted cases only; identity cases are a no-repair control. Pass is a percentage and larger is better; WER is a percentage and lower is better; score is unitless and larger is better. Shaded cells identify selected local repairs. The original scorer runtime revision is unavailable; Appendix Table 6 rescores these waveforms with a separate decoder.

Data	Input or policy	Pass↑ (%)	WER↓ (%)	Score↑
All LibriSpeech corrupted sources (1,024)	Corrupted input	12.0	27.0	0.7569
	Boundary-grid selector	**30.8**	**13.2**	**0.8683**
Main corrupted sources	Corrupted input	11.9	29.3	0.7533
	Boundary-grid selector	**28.5**	**14.0**	**0.8616**
Second n512 sources	Corrupted input	12.0	24.6	0.7604
	Boundary-grid selector	**33.1**	**12.5**	**0.8750**
Identity controls (1,024)	Corrupted input	59.4	4.8	0.9518
	Boundary-grid selector	59.6	4.8	0.9524

Across all 4,096 selected n512 outputs, the post-repair waveform is exact outside the recorded splice action: every audit record has zero padded mean squared error and zero duration difference. This verifies the splice implementation, not a performance difference among policies or restoration fidelity against clean speech.

The fixed original-span audit covers all 4,096 outputs, including identity controls. Of 3,635 aligned outputs, 2,832 selected outputs have exact context (77.9%) versus all 3,635 inputs: an absolute -22.1% difference (95% source-cluster interval [-23.2%, -20.9%]). Thus 69.1% of all selected outputs are both aligned and exact. Mean padded context MSE rises from 0 to 0.000163; 461 delete outputs (11.3%) are unaligned because their durations are unsupported. Mute and alignable-delete outputs account for the exact-context loss; duplicate and identity controls remain exact.

Direct editor contrast. On the same 2,346 corrupted cases, dots.tts.edit achieves 52.259% Whisper-small pass and 10.561% WER, versus 28.261% and 15.663% for the risk controller and 11.807% and 27.772% for no repair. The editor is the stronger automatic transcript-recovery system here. Table 3 reports context over all 3,128 cases, including 782 identity controls. Only 787 editor outputs align to the fixed map: 782 identity controls and five corrupted outputs, none of which has exact context. The controller aligns throughout and has 64.514% unconditional exact context; the editor's 25.000% exact rate comes solely from identity controls. Unaligned outputs cannot be assessed for physical off-span change by this map. Appendix Tables 8 and 9 separate denominators and post-selection decoders. For editor minus controller over all cases, pass differs by an absolute +18.00% (95% source-cluster interval [+16.14%, +19.85%]) and WER by -3.83% ([-5.38%, -1.85%]), seed 20260919. These configured systems differ in inputs, architecture, and proposal budgets.

Table 3: Direct comparison on 3,128 matched cases from 782 sources. Pass and WER use Whisper-small; alignment and exact context use the original corruption-span time map on all cases. Unaligned outputs fail the unconditional exact-context endpoint. Identity controls are unchanged in every arm. Higher pass, alignment, and exact context are preferable; lower WER is preferable. The editor has one generated output per corrupted case; the controller chooses from nine VITS patches plus no repair.

System	Pass↑ (%)	WER↓ (%)	Aligned↑ (%)	Exact↑ (%)
No repair	23.881	22.002	100.000	100.000
dots.tts.edit	54.220	9.095	25.160	25.000
Risk controller	**36.221**	**12.921**	**100.000**	**64.514**

Table 4 evaluates constrained policies on the calibration-disjoint subset of the same source pool. Rules are fixed without fitting the 128-source controller development partition; the primary analysis removes every earlier n128 calibration source and contains 782 source clusters. Geometry-only repairs 29.6% of cases, raises Whisper-small pass rate by an absolute 7.3% (95% interval [+6.4%, +8.2%]), lowers WER by an absolute 2.8% ([-4.4%, -0.7%]), and retains exact context and duration alignment throughout. Among hard feasibility's 892 repairs, 456 (51.1%) tie no repair on all four numeric content fields; the risk controller has 891 such ties among 2,003 repairs (44.5%). Hard feasibility repairs 28.5% of cases, raises pass rate by an absolute 7.6%, and lowers WER by an absolute 3.8% versus no repair. Relative to geometry-only, its Whisper-small ranker changes pass by an absolute +0.35% (95% interval [+0.16%, +0.58%]) and WER by -0.97% ([-2.60%, -0.12%]); both policies retain exact context. Table 5 shows why this is an in-contract ranking effect, not an evaluator-independent increment.

Table 4: Calibration-disjoint controller audit on 3,128 cases from 782 source clusters, including identity controls. The controller uses known synthetic-failure metadata but no clean waveform or clean-reference metric at selection. Exact context is over aligned outputs; unaligned uses all cases. All reported metrics are percentages. Shaded rows have exact clean context.

Policy	Coverage↑ (%)	Pass↑ (%)	WER↓ (%)	Exact context↑ (%)	Unaligned↓ (%)
No repair	0.0	23.9	22.0	100.0	0.0
Geometry-only	**29.6**	**31.2**	**19.2**	**100.0**	**0.0**
Hard feasibility	28.5	31.5	18.2	100.0	0.0
Risk controller	64.0	36.2	12.9	64.5	0.0

Table 5 tests evaluator sensitivity without retuning a policy on the calibration-disjoint 782-source subset. The two evaluators score frozen audio and have no input to synthesis, policy choice, or controller geometry.

Table 5: Fixed post-selection evaluator checks on 3,128 calibration-disjoint cases. Each evaluator scores frozen input and selected audio without altering synthesis or policy selection. Faster-Whisper uses large-v3-turbo; Wav2Vec 2.0 (W2V2) uses fixed greedy CTC decoding (Baevski et al., 2020). Both are LibriSpeech-domain automatic rescores. Coverage, pass, and WER are percentages. Shaded rows identify geometry-preserving policies.

Evaluator	Frozen policy	Coverage↑ (%)	Pass↑ (%)	WER↓ (%)
Faster-Whisper large-v3-turbo	No repair	0.0	31.1	16.6
	Geometry-only	29.6	36.2	14.8
	Hard feasibility	28.5	36.2	14.8
W2V2 CTC	No repair	0.0	14.5	22.8
	Geometry-only	29.6	28.7	16.9
	Hard feasibility	28.5	28.4	17.0

Geometry-only improves both post-selection decoders over no repair (Table 5). The hard-feasibility increment does not transfer: versus geometry-only, large-v3-turbo pass changes by an absolute -

0.064% (95% interval [-0.192%, +0.064%]) and WER by +0.006% ([-0.042%, +0.052%]); W2V2 pass changes by -0.256% ([-0.448%, -0.096%]) and WER by +0.096% ([+0.038%, +0.157%]). Removing the 456 candidate-ID tie repairs lowers pass under both decoders (Appendix Table 7). These frozen LibriSpeech rescores test decoder sensitivity; candidate generation and policy selection remain coupled to Whisper-small. Appendix Table 12 separates corruption types.

In the overlapping SpeechT5 failure sets, ungated timestamp repairs reduce first-n512 pass, while a matched large-v3-turbo verifier accepts 12/71 n128 and 103/319 first-n512 cases (Appendix Table 14). These sets overlap calibration and controller sources. The 319-item UniSRM check has no SupCLAP reranking comparison because it tests one candidate per case (Appendix Table 13).

7 Discussion

Broad selection lowers Whisper-small WER, but only 77.9% of aligned outputs retain exact context (Table 2). On 782 calibration-disjoint sources, geometry-only retains context at 29.6% coverage. The content gate's small Whisper-small gain vanishes under large-v3-turbo and reverses pass under W2V2; removing 456 numeric-tie repairs worsens both auxiliary results (Appendix Table 7). Geometry-only has stronger support for decoder robustness. dots.tts.edit recovers more transcripts, but only five corrupted outputs align to the fixed map, leaving preservation mostly unmeasured (Table 3). Table 4 reports the disjoint comparison; Appendix Figure 2 shows the overlapping cohort.

8 Limitations

Controlled LibriSpeech corruptions provide known spans; FLEURS supports calibration only. Natural-failure localization and listener judgments of identity or quality remain untested. Splice equality does not establish restoration, and failure diagnostics overlap calibration. The editor differs in architecture, input, and proposal budget; the fixed map cannot assess most outputs. Scorer runtime records are incomplete; score conflicts affect identity controls, not corrupted cases (Appendix Section A.1). Latency and cost are unmeasured.

9 Conclusion

On 782 calibration-disjoint sources, known-span geometry retains exact context at 29.6% VITS repair coverage. The direct editor recovers more transcripts, but the fixed map cannot assess most outputs. Natural-failure localization and perceptual quality remain untested.

AI Use Statement

Generative AI tools helped polish the writing and find related work. They also helped review analysis code and discuss recorded results. The paper describes these uses and their role.

Reproducibility Statement

The Methods and Experimental Setup sections specify the proposal rules, data splits, and evaluation measures. The supplementary MIT-licensed source archive contains text-only decision helpers and a toy example; it does not contain the audio, model weights, predictions, or benchmark runner needed to reproduce the audio experiments.

References

Junyi Ao, Rui Wang, Long Zhou, Shujie Wang, Yao Ren, Yu Wu, Chengyi Zhang, Ming Zhou, and Zhoujun Li. SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5723–5738. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.acl-long.393/.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Yunrui Cai, Xu Li, Yucheng Zhou, Jinchao Li, Dingdong Wang, Dongchao Yang, Xixin Wu, Chen Zhang, Zhiyong Wu, Pengfei Wan, and Helen Meng. Sonicweave: Chunk-routed mixture-of-experts for unified audio scene generation, 2026. URL https://arxiv.org/abs/2608.09571.

Weidong Chen, Xixin Wu, and Helen M. Meng. Gogo: Group-wise granularity-ordered codec for stable and efficient speech generation, 2026. URL https://openreview.net/forum?id=JbLmIoWwDC. ICLR 2026 poster.

Alexis Conneau et al. FLEURS: Few-shot learning evaluation of universal representations of speech, 2022. URL https://arxiv.org/abs/2205.12446.

Rongjie Huang, Dongchao Yang, Wenxiang Guo, Huadai Liu, Xize Cheng, Zehan Wang, Zhou Zhao, Xixin Wu, and Helen M. Meng. Masked text-to-audio flow-matching and reward feedback optimization. In *Findings of the Association for Computational Linguistics: ACL 2026*. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1891. URL https://aclanthology.org/2026.findings-acl.1891/.

Sung-Feng Huang, Heng-Cheng Kuo, Zhehuai Chen, Xuesong Yang, Pin-Jui Ku, Ante Jukic, Huck Yang, Yu Tsao, Yu-Chiang Frank Wang, Hung-yi Lee, and Szu-Wei Fu. VoiceNoNG: Robust High-Quality Speech Editing Model without Hallucinations. In *Interspeech 2025*, pp. 3469–3473, 2025. doi: 10.21437/Interspeech.2025-431.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5530–5540. PMLR, 2021. URL https://proceedings.mlr.press/v139/kim21f.html.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale, 2023. URL https://arxiv.org/abs/2306.15687.

Liming Liang, Dongchao Yang, Xianwei Zhuang, Yuxin Xie, Luo Chen, Yuehan Jin, and Yuexian Zou. SpeechSEC: A Unified Multi-Task Framework for Speech Synthesis, Editing, and Continuation. In *Interspeech 2025*, pp. 3464–3468, 2025. doi: 10.21437/Interspeech.2025-1776.

Rui Liu, Jiatian Xi, Ziyue Jiang, and Haizhou Li. FluentEditor: Text-based speech editing by considering acoustic and prosody consistency. In *Interspeech 2024*, 2024. doi: 10.21437/Interspeech.2024-2320. URL https://doi.org/10.21437/Interspeech.2024-2320.

Jiehui Luo, Yuguo Yin, Yuxin Xie, Jinghan Ru, Xianwei Zhuang, Minghua He, Aofan Liu, Zihan Xiong, and Dongchao Yang. Supclap: Controlling optimization trajectory drift in audio-text contrastive learning with support vector regularization, 2026. URL https://arxiv.org/abs/2509.21033.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. VoiceCraft: Zero-shot speech editing and text-to-speech in the wild, 2024. URL https://arxiv.org/abs/2403.16973.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv.org/abs/2212.04356.

Jinchuan Tian, Sang-gil Lee, Zhifeng Kong, Sreyan Ghosh, Arushi Goel, Chao-Han Huck Yang, Wenliang Dai, Zihan Liu, Hanrong Ye, Shinji Watanabe, Mohammad Shoeybi, Bryan Catanzaro, Rafael Valle, and Wei Ping. UALM: Unified audio language model for understanding, generation and reasoning, 2025. URL https://arxiv.org/abs/2510.12000. ICLR 2026 oral.

Dingdong Wang, Junan Li, Jincenzi Wu, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. Mmsu: A massive multi-task spoken language understanding and reasoning benchmark, 2026a. URL https://arxiv.org/abs/2506.04779.

Hankun Wang, Bohan Li, Shi Lian, Xiaoyu Gu, Jing Peng, Da Zheng, Yiwei Guo, Colin Zhang, and Kai Yu. dots.tts.edit: Precisely controlled speech editing with a continuous autoregressive model, 2026b. URL https://arxiv.org/abs/2608.02673.

Tao Wang, Jiangyan Yi, Ruibo Fu, Chunyu Qiang, Dading Chong, Chao Wang, Dongyang Dai, Zhengqi Wen, and Jianhua Tao. SpeechPalette: A comprehensive speech editing method for text-based speech editing, one-shot TTS and attributes editing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026c. doi: 10.1109/TPAMI.2025.3626793. URL https://doi.org/10.1109/TPAMI.2025.3626793.

Yuanyuan Wang, Dongchao Yang, Yayue Deng, Zhiyong Wu, Steven Y. Guo, Helen M. Meng, and Xixin Wu. UniSRM: A unified speech reward model for reasoning-based fine-grained assessment. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2026d. doi: 10.18653/v1/2026.acl-long.2150. URL https://aclanthology.org/2026.acl-long.2150/.

Yuanyuan Wang, Dongchao Yang, Yiwen Shao, Hangting Chen, Jiankun Zhao, Zhiyong Wu, Helen Meng, and Xixin Wu. DualSpeechLM: Towards unified speech understanding and generation via dual speech token modeling with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026e. doi: 10.1609/aaai.v40i40.40663. URL https://ojs.aaai.org/index.php/AAAI/article/view/40663.

Zhifei Xie, Jiaqi Lang, Ze An, Yifan Zhao, Dongchao Yang, Kai Li, Ziyang Ma, Mingbao Lin, Chunyan Miao, and Shuicheng Yan. Voicemem: Streaming dual-brain memory for real-time interaction, 2026. URL https://arxiv.org/abs/2608.26005.

Dongchao Yang, Songxiang Liu, Haohan Guo, Jiankun Zhao, Yuanyuan Wang, Helin Wang, Zeqian Ju, Xubo Liu, Xueyuan Chen, Xu Tan, Xixin Wu, and Helen M. Meng. ALMTokenizer: A low-bitrate and semantic-rich audio codec tokenizer for audio language modeling. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70850–70872. PMLR, 2025. URL https://proceedings.mlr.press/v267/yang25q.html.

Dongchao Yang, Yuanyuan Wang, Dading Chong, Songxiang Liu, Xixin Wu, and Helen Meng. Uniaudio 2.0: A unified audio language model with text-aligned factorized audio tokenization, 2026. URL https://arxiv.org/abs/2602.04683.

Hanlin Zhang, Daxin Tan, Dehua Tao, Xiao Chen, Haochen Tan, and Linqi Song. SpeechEditBench: A bilingual multi-attribute benchmark for instruction-guided speech editing, 2026. URL https://arxiv.org/abs/2606.01804.

Zhisheng Zheng, Puyuan Peng, Anuj Diwan, Cong Phuoc Huynh, Xiaohang Sun, Zhu Liu, Vimal Bhat, and David Harwath. VoiceCraft-X: Unifying multilingual, voice-cloning speech synthesis and speech editing. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. doi: 10.18653/v1/2025.emnlp-main.137. URL https://doi.org/10.18653/v1/2025.emnlp-main.137.

A Supplementary Results

The direct editor analysis below first separates corrupted cases from unchanged controls and then tests automatic transcript recovery with two additional decoders. The remaining tables examine tie breaking, corruption type, reward selection, and generated-failure gating. Their populations and decision rules are distinct; none is a human listening test.

A.1 Additional Measurement Limits

The direct editor comparison uses one dots.tts.edit output against nine VITS proposals plus no repair, mixing architecture, input information, search budget, and training-data differences. Its unaligned outputs need time-warp-aware or human-assessed preservation tests. SpeechSEC, Voicebox, VoiceCraft, VoiceNoNG, and SpeechEditBench have no interface-matched comparisons here. Frozen larger-Whisper and Wav2Vec2 rescoring tests decoder sensitivity on LibriSpeech, while original generation and selection remain coupled to Whisper-small. The earlier headline Transformers records identify the model but lack an exact runtime revision. Twelve identity-control cases have inconsistent scores across labels pointing to the same waveform; their exact scorer-run cause remains unresolved. The earlier controller run's full scorer-runtime lineage has not been reconstructed.

A.2 Headline Scorer and Selection Lineage

Table 6 scores the same frozen input and boundary-grid-selected waveforms with a separate Faster-Whisper small.en decoder. The original headline selector and Table 2 instead use Transformers Whisper-small.en. Each of the 4,096 selected records matches its stated no-repair or proposal-source record on waveform, transcript, and verification result; recomputing the frozen selection key yields no label mismatches. The chosen waveform path appears under multiple candidate labels in 1,049 cases. Twelve of these, all identity controls, have different stored Transformers scores for that same path. Six select a nonbaseline label with the same input waveform and a higher score. Canonicalizing each waveform path to its first scored occurrence changes no selected waveform, but these inconsistent stored scores account for the small identity-control metric differences. These 12 records do not enter the corrupted-case headline aggregate; the duplicate-path check cannot rule out scoring inconsistencies on other waveforms. The separate Faster-Whisper pass and WER direction agrees with the headline result, but its numeric values are decoder-specific.

Table 6: Separate Faster-Whisper small.en rescore of the frozen headline waveforms on 3,072 corrupted cases from 1,024 LibriSpeech sources. The selected audio is unchanged from Table 2; this decoder did not choose repairs. Pass and WER are percentages.

Waveform	**Pass↑ (%)**	**WER↓ (%)**
Corrupted input	11.523	24.266
Boundary-grid selected	**27.930**	**14.906**

A.3 Numeric-Tie Admission in the Context Controller

Table 7 removes one implementation detail from hard feasibility: the candidate ID may win eligibility even when all four numeric Whisper-small fields tie no repair. On the 782-source controller subset, 456 of 892 hard-feasibility repairs enter through such ties. A strict numeric gate returns the original waveform for those 456 cases while retaining the other 436 repairs; it does not regenerate or rescore audio. The primary Whisper-small pass and WER are unchanged by construction. On the independent decoders, however, removing ties lowers pass and raises WER relative to the implemented policy. Most removed repairs are duplicate-middle cases (391/456). This establishes sensitivity to the frozen tie rule, not a causal benefit of candidate IDs.

Table 7: Frozen-audio numeric-tie ablation on 3,128 cases from 782 calibration-disjoint sources. Hard feasibility admits numeric ties through candidate-ID ordering; strict numeric requires improvement in a content field. All policies preserve exact fixed-map context on every case. Pass, WER, and coverage are percentages.

Decoder	Policy	Coverage↑ (%)	Pass↑ (%)	WER↓ (%)
Whisper-small	Geometry-only	29.6	31.17	19.22
	Hard feasibility	28.5	31.52	18.24
	Strict numeric	13.9	31.52	18.24
Large-v3-turbo	Geometry-only	29.6	36.22	14.80
	Hard feasibility	28.5	36.16	14.80
	Strict numeric	13.9	34.91	15.19
Wav2Vec2 CTC	Geometry-only	29.6	28.68	16.92
	Hard feasibility	28.5	28.42	17.01
	Strict numeric	13.9	21.20	19.88

Relative to geometry-only, the strict numeric gate changes large-v3-turbo pass by an absolute -1.31% (95% source-cluster interval [-1.76%, -0.90%]) and WER by +0.39% ([+0.27%, +0.53%]); CTC pass changes by -7.48% ([-8.28%, -6.68%]) and WER by +2.96% ([+2.73%, +3.20%]). These 10,000 source-cluster bootstrap draws use seed 20260926 and keep each source's four variants together.

A.4 Original Controller Cohort Sensitivity

In the original 896-source controller cohort, 939 of 1,055 eligible cases tie on the best displacement and duration pair. Reversing the label tie order changes all 939 choices but changes pass by an absolute -0.22% (95% interval [-0.42%, -0.06%]) and WER by +0.01% ([-0.11%, +0.11%]); seeded random eligible routing changes pass by -0.08% ([-0.20%, 0.00%]) and WER by -0.04% ([-0.15%, +0.02%]). All three retain exact context (Appendix Tables 10–11). Thus preservation follows the fixed geometry conditions, while score-dependent routing is a distinct mechanism over frozen proposals.

Frozen SpeechT5 patches retain exact context at 27.2% coverage on the original controller cohort. This is a secondary back-end check, because its evaluated source cohort also overlaps the earlier n128 calibration artifacts. The primary calibration-disjoint result is the VITS geometry comparison in Table 4.

Of the original cohort's 1,011 hard-feasibility repairs, 528 tie no repair on all numeric content fields; the risk controller has 1,031 ties among 2,285 repairs. Candidate IDs admit these ties. The tie-routing sensitivity below also uses this original cohort.

Figure 2 displays the original 896-source controller frontier transparently. It visualizes the context loss incurred by the higher-coverage score-driven policies, while the calibration-disjoint VITS comparison in Table 4 is the primary policy evidence.

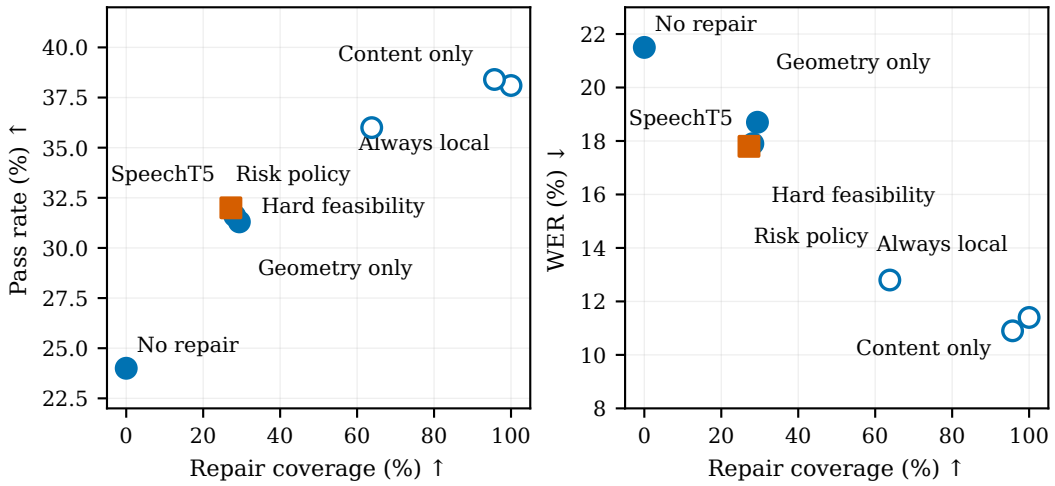


Figure 2: Original controller frontier on 3,584 cases from 896 source clusters, retained for transparency because this cohort overlaps earlier n128 calibration artifacts. Each point uses frozen selections; the two panels report automatic pass rate and WER against repair coverage. Filled blue circles have exact clean-reference context for VITS candidates, while hollow markers identify policies with non-exact context. The orange square is the frozen SpeechT5 hard-feasibility check. "Always local" chooses the highest content-ranked patch without a no-repair option; "content only" includes no repair. Geometry-only selects geometry-qualified candidates by displacement, duration error, and label without automatic-content scores; hard feasibility adds lexicographic content eligibility.

A.5 Editor: Content, Time Map, and Decoder Sensitivity

The main comparison pools four cases from each source. Table 8 asks whether unchanged controls hide failures on cases that actually require repair. On the 2,346 corrupted cases, dots.tts.edit has the highest Whisper-small pass rate (52.259%) and lowest WER (10.561%). The risk controller improves on no repair but has lower pass (28.261%) than the editor. All three arms are identical on the 782 identity controls, where no generation occurs and pass is 60.102%. Pooling these controls changes the denominator and must not be read as additional repair success.

The fixed-map diagnostic reaches a different conclusion on the corrupted subset. The editor aligns on 5/2,346 outputs (0.213%); none of those five retains exactly the original off-span samples. Its other 2,341 outputs are unaligned, so their physical off-span change is unmeasured under this map. The controller aligns all 2,346, with unconditional exact context of 52.685%. No repair is exact on this context measure by construction, despite its poor transcript recovery. For the full set, the editor's 787 aligned outputs comprise these five plus 782 identity controls, and its 25.000% exact-context rate comes solely from controls. Aligned-only context MSE for the editor is 0.00964657 on the five corrupted cases and 0.00006129 on all 787 aligned cases; neither mean describes the unaligned outputs. The controller's aligned-only MSE is 0.00017130 on corrupted cases. The main comparison in Table 3 reports unconditional rates to prevent selecting only alignable editor outputs.

Table 8: Whisper-small transcript outcomes and fixed-map context by corruption stratum. Each source contributes three corrupted cases and one unchanged identity control. Alignment is a count; exact context is an unconditional percentage over the stated stratum, counting unaligned outputs as failures. There is no repair-quality interpretation of no repair's 100% exact context.

Cases	System	n	Pass↑ (%)	WER↓ (%)	Aligned	Exact↑ (%)
Corrupted	No repair	2,346	11.807	27.772	2,346	100.000
	dots.tts.edit	2,346	52.259	10.561	5	0.000
	Risk controller	2,346	**28.261**	**15.663**	**2,346**	**52.685**
Identity	No repair	782	60.102	4.695	782	100.000
	dots.tts.edit	782	60.102	4.695	782	100.000
	Risk controller	782	60.102	4.695	782	100.000

Table 9 tests whether the editor's content lead is specific to Whisper-small. The decoder settings are specified in Section 5.4. All three decoders score the same selected waveforms; neither larger-Whisper nor Wav2Vec2 CTC chooses an edit. dots.tts.edit has lower WER than the controller with both additional decoders (6.618% versus 11.589%, and 9.742% versus 16.255%). The pattern supports a content advantage in this LibriSpeech-domain comparison, not perceptual quality or generalization to natural editing failures. For editor minus controller on all cases, large-v3-turbo changes pass by +16.40% (95% interval [+14.48%, +18.29%]) and WER by -4.97% ([-5.64%, -4.30%]); CTC changes pass by +14.23% ([+12.56%, +15.89%]) and WER by -6.51% ([-7.16%, -5.88%]). These are absolute metric differences, using 10,000 paired source-cluster draws with seed 20260919 and unrounded operands. The Whisper-small stratum reanalysis uses seed 20260924 separately.

Table 9: Automatic content on all 3,128 matched cases, including unchanged controls. Larger Whisper (large-v3-turbo) and Wav2Vec2 CTC are post-selection checks on the same audio. Pass uses the score threshold in Section 4.2, which coincides with normalized transcript equality in these outputs; WER is the mean per-case error rate.

Decoder	**System**	**Pass↑ (%)**	**WER↓ (%)**
Large Whisper	No repair	31.138	16.676
	dots.tts.edit	59.527	6.618
	Risk controller	43.127	11.589
Wav2Vec2 CTC	No repair	14.514	22.788
	dots.tts.edit	43.382	9.742
	Risk controller	29.156	16.255

A.6 Geometry Tie-Break Sensitivity

Tables 10 and 11 report the original 896-source controller sensitivity cited in the main results. The ascending and descending policies differ only in the label order used after displacement and duration tie; the seeded policy samples uniformly from all geometry-eligible repairs. Center and word denote the two ASR-span proposals; after and gap denote delete-after-difference and delete-gap proposals; signed columns denote boundary shifts in seconds.

Table 10 asks whether the geometry finding depends on an arbitrary label order when 939 eligible choices tie on displacement and duration. Descending order changes all 939 of those choices but leaves coverage at 29.4%, changes pass from 31.3% to 31.0%, and changes WER from 18.7% to 18.8%. Seeded random routing gives 31.2% pass and 18.7% WER; all three policies retain exact context. These are alternate selections from the same frozen proposals, not independent synthesis runs.

Table 10: Geometry-only routing sensitivity on 3,584 cases from the original 896-source controller cohort. Switches count decisions that differ from ascending-label routing. All metrics use frozen candidates; pass, WER, coverage, and exact context are percentages.

Routing policy	**Switches**	**Coverage↑ (%)**	**Pass↑ (%)**	**WER↓ (%)**	**Exact context↑ (%)**
Ascending label	**0**	**29.4**	**31.3**	**18.7**	**100.0**
Descending label	939	29.4	31.0	18.8	100.0
Seeded random eligible	825	29.4	31.2	18.7	100.0

Table 11: Selected proposal-family counts for each routing policy on the same 3,584 cases. Each policy column sums to 3,584.

Proposal family	Ascending label	Descending label	Seeded random eligible
No repair	2,529	2,529	2,529
ASR word-time interval	950	22	132
ASR word-time interval shifted uniformly	9	0	96
Delete-after-difference	1	1	122
Delete-gap	2	1	108
Boundary shift −0.10 s	4	4	103
Boundary shift −0.20 s	2	2	112
Boundary shift +0.10 s	0	0	90
Boundary shift +0.20 s	2	18	82
Uniform-duration proposal	85	1,007	210

Table 11 explains how such a large choice change can have a small aggregate effect. Ascending routing selects the center label 950 times and uniform duration 85 times; descending routing selects those labels 22 and 1,007 times. Both retain 2,529 no-repair choices. The distribution exposes label sensitivity but cannot show that the alternate patches sound alike or restore the same speaker characteristics.

Table 12: Corruption-type breakdown for the fixed content boundary-grid selector. Entries are verifier pass rates (%).

Corruption	Main input	Main selected	Second n512 input	Second n512 selected
Delete-middle	1.2	18.2	0.6	25.8
Duplicate-middle	32.4	58.4	34.2	63.7
Identity	56.4	56.6	62.3	62.5
Mute-middle	2.1	8.8	1.4	10.0

Table 12 tests whether pooled content gains hide failure types. On main n512, delete-middle pass rises from 1.2% to 18.2%, duplicate-middle from 32.4% to 58.4%, and mute-middle from 2.1% to 8.8%. The source-disjoint second n512 has the same direction, with 25.8%, 63.7%, and 10.0% selected pass. Identity shifts only from 56.4% to 56.6% and from 62.3% to 62.5%, so these controls are not evidence of repair success. The scores remain automatic transcript checks under controlled corruptions.

Table 13: Auxiliary UniSRM likelihood reranking over two whole-utterance candidate families. The deterministic verifier remains the reported acceptance metric.

Policy	Cases	Pass↑ (%)	WER↓ (%)	Score↑	Selected mix
Baseline: pre-enhancement	319	10.0	23.8	0.7630	–
UniSRM reranking	319	**22.9**	**17.2**	**0.8277**	**227 enhancer / 92 pre**

Table 13 asks whether an auxiliary likelihood signal can choose between two whole-utterance candidates on 319 generated-failure items. UniSRM selection uses 227 enhancer outputs and 92 pre-enhancement outputs, and the fixed verifier's pass changes from 10.0% to 22.9%, with WER changing from 23.8% to 17.2%. This is a separate whole-utterance reranking question; it does not validate the local controller's off-span context or perceptual quality. SupCLAP's single-candidate check has no alternative to select and therefore cannot demonstrate a reranking gain.

Table 14: Real SpeechT5 generated-failure repairs under matched large-v3-turbo scoring. The verifier gate accepts a repair only when score does not decrease, WER does not increase, and at least one content metric improves. The relative speaker gate additionally permits at most a 0.02 speaker-similarity decrease.

Subset	Cases	Gate	Accepted	Pass↑ (%)	WER↓ (%)
n128	71	Failure input	0	14.1	14.8
n128	71	Verifier	12	23.9	12.9
n128	71	Verifier + speaker	11	22.5	13.1
first-n512	319	Failure input	0	19.4	20.8
first-n512	319	Verifier	103	23.8	15.4
first-n512	319	Verifier + speaker	95	23.5	15.7

Table 14 tests selective acceptance when the input is an actual flawed SpeechT5 utterance. On 71 n128 failures, the verifier accepts 12 repairs and changes pass from 14.1% to 23.9%, while the additional automatic speaker gate accepts 11 and gives 22.5% pass. On 319 first-n512 failures, the corresponding counts are 103 and 95, and pass is 23.8% and 23.5% against 19.4% before repair. The speaker threshold protects only this measured embedding signal; it does not establish listener-rated identity or safe online fault localization.