THE HALF-LIFE OF REPAIRS: WHY FIXES FAIL UNDER CONTINUED TRAINING AND HOW TO MAKE THEM LAST

Anonymous authors
Paper under double-blind review

ABSTRACT

Post-hoc repairs to a trained differentiable audio synthesizer—a frozen gain, an inserted filter, a reverb reparameterization—are routine and physically verifiable, yet continued training silently undoes them. We formulate repair persistence as an engineering objective and measure it with a pre-registered four-arm paired-continuation protocol and tri-state verdicts: unrepaired training, immediate repair, repair under continuation, and repair with behavioral anchoring. In a controlled cello-vocoder experiment, a $+5.03$ dB gain restores the median loudness drift from -5.03 to -0.0006 dB; after 15,000 continued-training steps the unanchored arm degrades to -5.25 dB ($R = -0.27$, the repair-retention ratio; directional, within the seed-variance band), the benefit halving within 500–1,000 steps (single-seed). Three instruments localize the cancellation to the harmonic and noise branches; output-level gradient pressure falls in all 12 instrumented runs. Anchoring is a fuse, not a retainer—no stable persistence advantage over plain repair (1/3 vs 1/3 seeds)—and the repair-plus-anchor package backfires on both URMP seeds. A 34-run, 13-arm, two-domain matrix shows erasure replicating at the gradient-instrument level; persistence tracks the calibration source (3/3 vs 1/3 seeds; not separable, Fisher $p = 0.20$). Claims cover the loudness-defect domain only. Contributions: a repair-visibility taxonomy; a persistence-time / compensation-path detection protocol; a constructive candidate route to durable repair.

1 INTRODUCTION

Repairing a differentiable audio synthesizer is routine engineering: quiet output gets a gain calibration; a narrowband leak in silent passages gets a filter; metallic reverb ringing gets a structural change. These repairs are usually physically impeccable, and immediate post-repair measurements pass in full. The default engineering assumption is that *a fix that works now will keep working*.

The evidence in this paper shows the assumption fails in specific, reproducible ways—though not universally: verdict-level erasure is conditional (§4.2). Three archived cases: a physically correct low-pass repair was fully bypassed after 1,500 continued-training steps, rebounding 38.51 dB to *worse than before the repair* (§2.1); a structured reverb reparameterization survived (§2.2); and a controlled four-arm experiment upgraded "repair eaten" from observation to a directional verdict under controlled design (single seed; magnitude readouts within the seed-variance band; multi-seed robustness closed by the matrix of §4) while measuring the benefit half-life at no more than 1,000 steps (§3). A 34-run multi-start matrix (§4) is complete with matrix-level conclusions: repair erasure replicates in 12/12 instrumented repair runs at the gradient-instrument level (pressure falling sharply; no unrepaired-arm gradient control run, §8); persistence tracks the calibration-source choice, external-holdout calibration persisting on all seeds (3/3); and "repair + anchoring" transferred to the URMP domain (Li et al., 2019) backfires on both seeds (backfire: $R(\mathrm{end})$ negative, worse than no repair)—persistence covers only the level-defect domain, and self-recorded-domain conclusions do not transfer to URMP. We state the problem as:

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.

A physical constraint that holds at inference need not remain a constraint on the equilibrium of the re-optimized model: the remaining trainable modules may form compensation paths that route around or cancel the repair.

A *compensation path* is an adjustment by trainable parameters other than the repair that offsets it—e.g., after a low-pass is inserted, noise amplitudes and reverb gains regrow a mapping around the filter.

Citation delimitation. The umbrella phenomenon—decay of edits under continued training—is not our discovery (LLM editing decay (Cheng et al., 2026); text-to-image editing robustness (He et al., 2025); systematic study of whether erased concepts are truly removed (Lu et al., 2025); §7). *Our contribution is an acceptance methodology and standard validated on one system and designed for transfer rather than the phenomenon*: a taxonomy organizing risk diagnosis (prospective prediction as future work), lightweight anchoring as an overshoot guard, structural reparameterization as a root-cure candidate, and a four-arm test protocol—a toolbox narrative, not a silver bullet, with delivery accepted only through the unified protocol.

Contributions. (1) **A layered repair toolbox**: diagnosis (repair-visibility taxonomy—formerly *loss visibility*; retention metric, compensation-path detection), overshoot guard (behavioral anchoring), and root cure (structural reparameterization), each with in-domain evidence (§5). (2) **A four-arm test protocol**: same-start pairing, deployment-setting readouts, contemporaneous denominators, pre-registered tri-state verdicts, and a noise floor (§6; one closed loop in-domain; cross-domain porting a design goal). (3) **To our knowledge, the first fully machine-archived case set and multi-start matrix**: C3/R3/the four-arm experiment, every claim's number recomputable from archives, plus 34 training runs (slot accounting in §4.1) extending to multi-start, multi-calibration-source evidence (§3, §4). (4) **Pre-registration and evaluation discipline**: criteria frozen before runs, seed-variance bands, leak-proof reference sets, the single-anchor rule (§6, §8).

Declared scope and claim discipline (enforced throughout). Controlled evidence spans from a single-seed four-arm design to a 34-run matrix; we make no universal claims ("every repair gets eaten" and "reparameterization is never eaten" alike); the surviving case R3 is included as a counterexample (§2.2); anchoring guards overshoot and does not preserve repairs (§3, §4).

Number-disclosure caliber. Readings are reported as recorded in the machine archives; derived differences are computed on unrounded values—recomputing derived quantities from the displayed values can differ by one unit in the last digit; this applies throughout the paper (Appendix L).

2 Motivating Cases: a Repair Eaten and a Repair that Lasted

Two archived cases frame the problem (details and evaluation configurations in Appendix B).

2.1 C3: a physically correct repair eaten

C3 (zero-visibility insertion; observational): a physically correct 13 kHz-cutoff low-pass on the noise branch of a cello differentiable-synthesis chain—its defect a persistent 14 kHz silence-window leak, architectural in root cause—cut the dry-end (pre-reverb signal) leak from -18.82 to -49.67 dB at drop-in, and was fully bypassed after 1,500 continued-training steps: dry-end rebound to -11.17 dB ($+38.51$ dB, *7.65 dB worse than the pre-repair baseline*; both ends worse), while the training score function barely noticed (weighted-spectral + mel loss 0.97 $\rightarrow$ 0.98, $+0.5\%$). Module level, the reverb gain $|H(14\text{k})|$ rose $+0.30$ dB—the first sign of a compensation path. Full readings are in Appendix B.

2.2 R3: a structural reparameterization that lasted

R3 (in-sight reparameterization; constructive existence): the free 144,000-tap reverb IR, squeezed by reconstruction pressure into coherent high-Q ringing ($T_{60}^{\text{est}} = 4.18$ s at 14,594 Hz, Q up to 27,745—the metallic timbre), was reparameterized into 30 bands $\times$ (band gain, decay) = 60 physical knobs under a hard T_{60} bound (ringing physically inexpressible) pinned by a reverb anchor loss; the pathology dropped to 0.12–0.14 s and held through training (author listening checks over three sessions; "init is the endpoint": 0.070 dB extrapolated drift against the 3.69 dB re-fit recovery

gap). Mechanistic reading (hypothesis-level; controlled test in §3): the low-pass lies in no loss term's computation graph (zero gradient anchor, unanchored drift), while the reparameterized knobs remain watched by the anchor loss—visibility, not physical correctness, decides survival. Full parameterization, bounds, identity disclosure (C3 a temporary recording stopgap, not an architectural repair), and the cross-structure boundary (not a same-structure counterexample pair) are in Appendix B.

3 Controlled Verdict: the Four-Arm Paired-Continuation Experiment and the Noise Floor

The two motivating cases are observational (C3: no contemporaneous control) and historical (R3: differing training stages). This section upgrades "repair eaten" to a directional verdict under a pre-registered controlled design (single seed; magnitude readouts within the seed-variance band; robustness boundary in §3.4, §8).

3.1 Design

From one production checkpoint at step 77,000 (identical start fingerprint in all arms), four arms identical except for treatment each continue 15,000 steps (77,000 $\rightarrow$ 92,000) with 9 checkpoints sampled densely early and sparsely late (schedule in Table 3, Appendix C): **A1** unrepaired continuation (contemporaneous baseline drift, the reading denominator); **A2** initial model + fixed repair, no training (immediate effect, reference D_{rep}); **A3** fixed repair in the forward pass + continuation (main "eaten" arm); **A4** = A3 + output-behavior anchoring + continuation (overshoot-guard arm).

Repair target and form. The production chain no longer has the C3-shaped 14 kHz leak (an architectural double soft-gate—noise and harmonic gates; wet-end (post-reverb) silence-window readings below the original recording's floor in 11/11 samples, pre-registered leak gates not triggered; full readings and the pre-run feasibility check in Appendix A), so the target is re-grounded on *underloudness in voiced passages*, $med11 = -5.03$ dB ($med11$: median over 11 fixed validation probes of each probe's signed median over voiced frames; a second readout scale, -5.86 dB on an eight-anchor frame-power subset, is reported side by side and never mixed; Appendix B). The defect substitution preserves the failure mode under study: both defects are scalar deployment-time quantities, both repairs are pure output-stage functions, and in both cases the loss supervises neither the defect nor the repair's retention (C3's filter sits outside the loss graph—zero visibility; the frozen gain sits inside it with weak indirect pull—weak visibility; §5). The repair is a *frozen constant gain on the reverb output stage* $g^* = +5.03$ dB (pure function; no parameters, no randomness). Immediate effect (A2): $med11$ $-5.03 \rightarrow -0.0006$ dB (constructive zeroing, exactly matching theory; frame-level absolute-error median 1.17 dB / p90 4.83 dB—residuals retained as such).

A1/A3/A4 share the start, data, and training seed. Anchoring (A4, pre-registered): an output-level-drift hinge—each training-side reference sample referenced to *its own just-repaired level*, penalized only beyond 3 dB drift (static unevenness unpenalized); reference set = 32 systematically sampled training-side samples (the 11 validation probes excluded—no judge-set leakage); applied every 25 steps to a random subset with an RNG fence around it (full configuration in Appendix L).

3.2 Primary readout: repair retention R(t)

$$R(t) \;=\; \frac{D_t - D_{\mathrm{base}}}{D_{\mathrm{rep}} - D_{\mathrm{base}}}, \tag{1}$$

where D is the $med11$ readout on the *deployment setting* (the scale keeping the repair in its deployed form—A1 read without the gain, A3/A4 with g^*, i.e. real post-launch usage); D_{base} is A1's *contemporaneous per-checkpoint paired value* (common baseline drift cancels inside the numerator); D_{rep} is A2's reading at $t = 0$. $R = 1$: full retention; 0: back to baseline; negative: worse than not repairing. t_{half}: the first step with $R(t) \leq 0.5$ (benefit half-life). Scale disclosure (full working in Appendix J): the denominator stretches with the baseline's contemporaneous reading (inflating $|R|$ by $\approx$22% vs. a fixed t=0 denominator at A1's $+0.91$ dB drift); verdicts use the pre-registered contemporaneous scale, the fixed scale the robustness control, t_{half} unchanged under it. A construct boundary applies to R: it measures the retention of the *repair's contribution*, not the persistence of the *problem state*—if the baseline arm were itself to converge toward the repair target, a low R

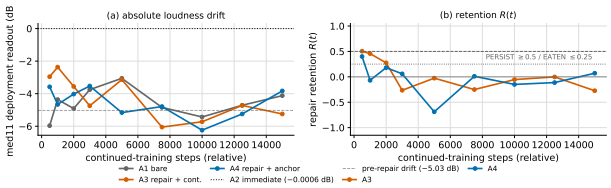

Figure 1: **Four-arm experiment results. Left:** deployment-setting $med11$ per checkpoint (A1/A3/A4; A2's immediate value as horizontal reference; single seed). **Right:** retention $R(t)$ for A3/A4 with the 0.5/0.25 tri-state thresholds; both arms cross early (A4 <500 steps, A3 within 500–1,000). Endpoint verdicts are the mechanical tri-state assignments of $R(\text{end})$; full values in Appendix C.

could coexist with "the problem is naturally resolved"; engineering users care about the latter. In our instance A1 drifts only $+0.91$ dB (no practical effect), but protocol ports should report both.

3.3 Endpoint verdict (checkpoint 92000; pre-registered tri-state, mechanical)

Tri-state: PERSIST $\iff$ $R(\text{end}) \geq 0.5$; EATEN $\iff$ $R(\text{end}) \leq 0.25$; PARTIAL otherwise. The 0.5/0.25 thresholds are engineering decision values written down before the runs (empirically annotated; no calibration anchor). A listening check (the audio gate) is a post-verdict acceptance step for PERSIST and is not part of this verdict layer; the gate for the matrix PERSIST arms has been executed and passed (§4.2).

Endpoint verdicts (checkpoint 92000, mechanical tri-state): A3 $med11 = -5.2471$ dB, $R(\text{end}) = -0.2733$ **EATEN** (overcorrection: 1.126 dB worse than no repair, within-band magnitude); A4 $med11 = -3.8306$ dB, $R(\text{end}) = +0.0705$ **EATEN** (overshoot guard: 1.42 dB above A3, within-band magnitude); A1 contemporaneous reference -4.1211 dB; A2 immediate -0.0006 dB (definitional $R = 1$). Benefit half-life: $t_{\text{half}}(\text{A3}) \in [+500, +1000]$ steps (boundary margins inside the propagated R-noise scale of ± 0.4–0.5, hence a single-seed descriptive reading; exact margins in Appendix C); $t_{\text{half}}(\text{A4}) < +500$ steps. The headline "$\leq$ 1,000 steps" carries the qualitative conclusion "fast failure, on the order of hundreds to a thousand steps" (mechanism, hypothesis-level: the repair lifts the deployed output above the training equilibrium; the supporting pilot and first-checkpoint readings are in Appendix C). The curves are non-monotone; Figure 1 shows the curves; Table 3 (Appendix C) the full checkpoint table (t_{half} is exceptionally defined by intermediate readings and carries its interval upper bound).

3.4 The noise floor, reading rules, and engineering meaning

All between-arm magnitude readings are quoted against the seed-variance band (scale reference; the seed-replication control retrained two arms byte-identical except the seed—a cross-segment proxy, n=2; per-anchor paired-difference median -0.3295 dB, 90% CI crossing zero; endpoints, intervals, and the deepest-anchor split in Appendix I). Reading rules: A3's overcorrection and A4's overshoot-guard gain (magnitudes in §3.3) are within-band magnitude-level readings; no statistical-significance claim is made. The verdict rule itself does not depend on the band for execution (pre-registered, preventing post-hoc cherry-picking of the reading scale). Single-point robustness: A4's endpoint verdict flips into PARTIAL under an R perturbation of $+0.18$ (distance to PERSIST $+0.43$); A3 needs $+0.52$ (0.5233 exactly)—the R noise-propagation scale ± 0.4–0.5 is of threshold order, so label robustness is answered by the multi-seed distribution of §4 (under curve-aggregated scales both verdicts stand: per-checkpoint means and recent-segment means all < 0.25—noise threatens the endpoint single-point label, not the phenomenon). Single-anchor readings are barred from verdicts. A1's $+0.91$ dB drift is of the seed-replication scatter order—the contemporaneous paired definition cancels such common drift by construction.

The four arms upgrade "repair eaten" from observation to a *directional verdict under controlled design* (A3 vs. A1; single seed, magnitude within the seed-variance band; multi-seed robustness closed by §4). A4 vs. A3 is the interventional effect of anchoring—*anchoring does not preserve the repair, but under this design it blocks the "repair makes it worse" overshoot* (fuse semantics

$=$ the A4$-$A3 post-hoc annotation; main-effect/interaction separation in §8). Uncorrected readouts (without g^*) give *directional, quantitative* compensation evidence: A3's endpoint uncorrected reading ≈ -10.28 dB (deployment -5.2471 minus g^* 5.03) is 6.16 dB below A1's uncorrected (-4.12 dB)—training drove the repaired arm's uncorrected output down by 5.25 dB, $\approx 104\%$ of the repair magnitude 5.03 dB, overshooting below the baseline arm's own level (-10.28 vs. -4.12 dB) while A1 rose $+0.91$ dB over the same window (the two arms move in opposite directions—the directed push is not a drift shared by both arms); A4's uncorrected reading ≈ -8.86 dB (4.74 dB lower), a smaller directed push. We treat *compensation against the repair* and *loss-driven re-equilibration of a mis-calibrated constant* as competing descriptions: both predict a downward push, and the 104%-vs-100% excess ($\approx$0.2 dB) sits well inside the seed-replication band. The discriminating observation is the *overshoot beyond the baseline's own level* (-10.28 vs. -4.12 dB) together with the dissipated-pressure pattern (12/12 at matrix level, §4.2); the sham control contributes only the anchoring attribution. The t=0 over-loudness data (§4.2) leave equilibrium proximity a live confound, so both descriptions remain open. The side-effect panel is archived per checkpoint; verdicts use the primary scale (panel definitions in Appendix A).

4 Scale-Up Evidence: The 34-Run Comparison Matrix

§3 is a single-start, single-seed controlled verdict. This section extends the test along three dimensions—seeds, calibration sources, and data domains—via a pre-registered scale-up built for exactly this purpose (the domain dimension contains both successes and failures).

4.1 Design (pre-registered)

34 training runs, 13 arms (13 treatment arms plus one run-level sham control; B4sham is counted as a run, not an arm; per-arm intervention registry in Table 1, Appendix A): 36 slots (self-recorded 10 arms $\times$ 3 seeds; URMP 3 arms $\times$ 2 seeds) minus 3 zero-training B1 measurement slots plus 1 sham-intervention run, entering five waves (7/7/7/6/7) as 26 pre-copied + 8 surgical starts. Arm types: unrepaired continuation (B0); inference-only calibration (B1; zero training—constructive start-checkpoint zeroing of the four-arm A2 reading, the B1 measurement archive); loudness-loss weighting (B2); frozen-gain repair (B3); gain+anchoring (B4; C1 is a same-configuration parallel arm—the multi-seed replication group of the four-arm A4—kept separate to preserve within-arm seed numbering, combined reading 2/6); *structural surgery* (B5 = reverb band-gain lift + re-anchoring; B6 = B5+anchoring); cross-source calibration (C2, C3); URMP unrepaired / repair+anchor / reparameterization (U1/U2/U3). Suite-spec arm codes are unrelated to the low-pass case C3 of §2.1.

Three-source calibration registry: the repair gain has three registered sources—difficult-anchor end 5.03 dB, training-distribution median 1.23 dB, URMP external-holdout measured 3.99 dB—all values single-sourced from the registry; that the three sources give three different values itself exhibits the gain's dependence on the reference-set choice. **Frozen criteria**: the configuration-freeze file passed integrity and tamper-proofing double checks (start-checkpoint fingerprints verified one by one; surgical files complete in the artifact registry; anchor reference samples all-or-none, partial inclusion banned; repair gain single-sourced; wave scheduling exhaustive and non-overlapping), all passed. **Surgical-parameter solving** converged before the runs ($\Delta_s = 4.92$ dB, $\Delta_u = 3.46$ dB; iteration counts, terminal residuals, and the convergence band in Appendix A).

4.2 Matrix verdicts: full R(t) tri-states and the compensation-localization instruments

Training side: all 34 runs completed to 92,000 steps; 306 checkpoint snapshots fingerprint-verified (34×9). **Evaluation**: checkpoint evaluation passed 306/306 (itemized domain counts in Appendix A).

R(t) tri-states (29 repair runs $= 34 -$ unrepaired baselines B0$\times$3 and U1$\times$2; single-source file: the matrix reading file): the matrix's $R(t)$ is isomorphic to §3.2, differing in reading unit and normalization anchor—per-anchor paired-difference median normalized to the run's first checkpoint; the reading scale was pre-registered; CI = anchor-level bootstrap (B=1000, seed 20260918). Totals: **PERSIST (audio gate passed) 15 / EATEN 10 / PARTIAL 4**. Seed-level counts per arm: B2 3/3, C3 3/3, B5 2/3, C2 2/3, B3/B4/C1/B6 1/3, U3 1/2, B4sham 0/1, U2 0/2 (per-seed values in

Appendix A, Table 2; plotted in Appendix E). An objective audio-side check of the rendered PERSIST endpoints is recorded: the loudness-difference direction matches the med11 readings in 26/30 probe pairs, with all four disagreements concentrated on a single healthy-anchor probe—waived under the single-anchor rule (§3.4); the listening verification has since been executed and passed (author check, direction-consistent with the verdicts; archived as the audio-gate record; full check configuration in Appendix L).

Compensation-localization instruments (the compensation-localization archive):

(i) *gradient cross-check* (paired before/after-repair comparison within each run): output-level pressure falls sharply in all 12 instrumented repair runs (the 12 = the paired-comparison run set—frozen-gain arms B3/B4 and structural-surgery arms B5/B6, 3 seeds each; no unrepaired-arm gradient control yet, §8); the output-level channel moves t_0 $+14.50$ to $+15.36$ $\rightarrow$ t_{15k} -4.35 to $+5.07$, with 11/12 reversing sign negative (the B5_s0 residual $+5.07$ an expected exception; reading dimension and exception analysis in Appendix F); (ii) *internal differencing*: cancellation lives in the harmonic (-5.08 dB) and noise (-5.96 dB) branches, not deep layers (dense $+0.33$ dB; n=1 run pair: B3_s0 vs. B0_s0, same seed, endpoint checkpoint 92,000, 11 validation anchors, median of per-anchor uncorrected RMS dB differences); (iii) *compensation decomposition* is an accounting identity (Appendix F); (iv) *sham-intervention control*: B4sham$-$B3 shows no significant difference (-0.73 dB, CI crossing zero) $\rightarrow$ the B4$-$B3 difference is attributable to anchor gradients rather than anchor-forward side effects, verified on two channels (the paired-difference CI test—registry pre-noted as a scale reference—and the directly measured anchor gradient $+1.74$; the two lines of evidence converge on the gradient pathway; dimension detail in Appendix F). Figure 3 in Appendix F.

Finding 1 (matrix level; three-instrument cross): repair erasure is a real, replicable risk—verdict-level erasure is conditional rather than typical (15 of 29 persist, frozen-gain majority 8/15 driven by C3 3/3 and C2 2/3; 14 EATEN or PARTIAL)—while at the gradient-instrument level the repair is erased by continued training: pressure dissipated in 12/12, branch-level cancellation (the gradient-instrument evidence limited to the 12 instrumented repair runs; no B0 gradient control—a domain-wide relaxation explanation is not excluded, §8); verdicts track the calibration source (Finding 2). The pressure fall is read as an accompanying form of post-repair re-equilibration; its correspondence with verdict-level outcomes (persist/eaten) awaits separation by the B0 gradient control.

Finding 2 (matrix level): *persistence tracks the calibration source*—C3 (external-holdout calibration) persists 3/3 seeds (vs. C1 difficult-anchor 1/3, C2 training-distribution 2/3); external-holdout calibration has the highest observed persistence count within the self-recorded domain (tied with B2 at 3/3). Two boundaries on this observation: at three seeds per arm, 3/3 vs. 1/3 is not statistically separable (Fisher exact, one-sided $p = 0.20$); we report it as a matrix-level observation pending seed replication, not a separable arm ranking. And the t=0 over-loudness data (below) show the arms differ in equilibrium proximity at t=0, so calibration-source effects and initial-mismatch effects are confounded; separating them would require arms matched on initial over-loudness.

Finding 3 (matrix level; construct correction): *on the URMP domain, "repair + anchoring" (U2) backfires on both seeds* ($R(\mathrm{end}) = -1.26$ / -1.13, CIs entirely negative)—U2's $g^* = 3.99$ was measured on URMP itself (not a cross-domain transplant), so the failure happens on the "URMP in-domain calibration + anchoring" route—*the repair-plus-anchoring route fails on URMP under in-domain calibration; no arm directly tests cross-domain transfer* (the matrix's strongest failure signal). Attribution caveat: URMP has no repair-without-anchor arm (U1 unrepaired / U2 repair+anchor / U3 reparam), so repair$\times$domain, anchor$\times$domain, and anchor$\times$repair are confounded—the backfire is attributable to the package, not separately to anchoring. Calibration-source coupling disclosed: C3 and U2 share the $g^* = 3.99$ source, so C3's success and U2's failure cannot be attributed to the calibration value.

Four-arm side-by-side (seed replication of the single-seed verdict): the archived four-arm A3/A4 EATEN verdicts *do not remain stable* when replicated along the seed axis—same-arm B3 seeds span -12.05 to $+2.35$, B4 spans -0.92 to $+2.86$; the four-arm experiment readings fall within their arm's seed distributions, not as outliers. *A single-seed EATEN verdict is one realization of the seed distribution; the matrix-level multi-seed distribution is the unit of analysis* (this closes the single-seed robustness question of §3.4). Scale caveat: the four-arm experiment and matrix R values use different reading units and normalization anchors (§4.1); the side-by-side reading is therefore ordinal only.

Calibration-source over-loudness side effect (t=0, training-distribution 32-sample reference scale): post-repair over-loudness median C1 $+3.80$ / C2 -0.00 (constructive zeroing holds) / C3 $+2.76$ dB—"how loud the repair lands" is tied directly to the calibration-source choice.

The B2 route (training-objective repair). B2 re-weights the loudness term of the training loss (a training-objective change; value and side-effect profile registered in Appendix A). Its 3/3 has a natural reading: the repair target becomes part of the optimization objective itself—the limiting case of moving a behavior into the loss's line of sight—so there is no unsupervised output-stage repair for continuation to eat. Boundaries: spectral-quality side effects are not separately reported; 3/3 at three seeds is not a separable ranking; and unlike the other routes B2 falls outside the P/Ψ formalism of §6.1.

Claim stratification: established = repair erasure (three-instrument cross), C3 highest observed persistence count (3/3), U2 double backfire, sham attribution closed; *under current evidence, anchoring (B4/C1) has no stable headline-level advantage over plain repair (B3) (B4 1/3, C1 1/3 vs. B3 1/3; large between-seed variance; more seeds or longer training needed)*; open = two out-of-band validation probes unattributed.

Sensitivity analysis (CI criterion downgraded to a note): it reads PERSIST 14 / EATEN 2 / PARTIAL 13 and flips 9 run-level verdicts relative to the registered threshold criterion—but all three headline findings survive under both criteria: their supporting evidence is criterion-invariant (gradient-pressure dissipation in 12/12 runs for Finding 1; C3's 3/3 seed persistence for Finding 2; U2's entirely negative CIs for Finding 3), not carried by the CI-versus-threshold verdict split.

Reporting discipline: the single-source reading documents (the matrix reading file and its interpretation report, both released with the artifact) are authoritative for the matrix, recomputable from the released reading files; the construct of R is the loudness-defect domain (A-weighted loudness under/over, not the timbre domain)—persistence conclusions cover level defects only (§8).

5 A Layered Repair Toolbox

Three layers, each with in-domain evidence (details in Appendix G); no universality is promised. **Layer A — diagnosis**: the *repair-visibility taxonomy* (formerly *loss visibility*) grades repairs on two axes—*inside the loss gradient-backprop graph* $\times$ *reparameterized*—with four evidenced cells (the two $\pm$anchor cells are control arms within one four-arm experiment): *zero-visibility insertion* (C3, observational)—eaten, fast rebound; *weak visibility, no anchor* (A3, controlled)—eaten with overcorrection, half-life $\approx$1k steps; *weak visibility + lightweight anchor* (A4)—still eaten, overshoot guarded (fuse); *in-sight reparameterization* (R3, constructive existence)—survives. Empty cells are unevidenced; predictive power is bounded by the evidenced cells. $R(t)/t_{\mathrm{half}}$ (§3.2) and the localization instruments (§4.2) complete the diagnostic layer, with the pre-registered gate test rejecting the insufficient-optimization-pressure hypothesis (Appendix B). **Layer B — overshoot guard**: the hinge anchor is a fuse, not a retainer (A4: $+1.42$ dB over A3, a within-band magnitude-level reading, still EATEN; matrix: no stable persistence advantage over plain repair, B4 1/3 vs. B3 1/3 seed-level; the sham-control attribution, the observed lower envelope (B4 worst seed -0.92 vs. B3 worst -12.05), and the single-frozen-configuration boundary in §4.2, Appendix G). **Layer C — root cure**: structural reparameterization seals the pathological solution space and is pinned by an anchor loss (R3; a *joint* artifact whose net effect cannot yet be separated, and whose survival is "survival under continuous anchor supervision"; matrix evidence B5 2/3, B6 1/3, U3 1/2 seed-level). Layers B and C form the cost–persistence spectrum; the engineering rules (hard cuts/masking ship with anchoring; inference-end repairs—e.g. per-sample loudness calibration, constructively zeroing $med11$ $-5.03 \rightarrow -0.0006$ dB at zero training cost—with efficacy claims *limited to frozen deployment*; durable delivery $\rightarrow$ evaluate reparameterization first) are stated in Appendix G.

6 The Four-Arm Test Protocol: an Engineering Acceptance Standard

The protocol is specified on three interfaces—*output behavior* (the deployed-output quantity a repair targets and the readout measures), *intervention site* (where the repair enters the computation—an

output transform or a parameter re-expression), and *continuation protocol* (how training proceeds after the repair—same-start paired continuation with pre-registered checkpoints and verdicts).

6.1 FORMALIZATION

Let c be an acoustic condition, y a reference waveform, and $F_\theta(c)$ the model. A repair is either a fixed output transform P (gain, filtering) or a parameter re-expression $\theta = \Psi(\phi)$ (structural reparameterization). For each recording define a target behavior b_j and tolerance τ_j, with defect quantity in hinge form $q_j = \max\{0, |b_j - b_j^*| - \tau_j\}$ (or one-sided; tolerances, reference sources, and evaluation spans registered before the task). On a frozen evaluation set define the immediate benefit and the contemporaneous net benefit:

$$B_0 = q(F_{\theta_0}) - q(P \cdot F_{\theta_0}), \qquad B_t = q(F_{\theta_t^0}) - q(P \cdot F_{\theta_t^P}), \tag{2}$$

where B_t pairs the unrepaired continuation trajectory θ_t^0 against the repaired θ_t^P at equal steps (B_0/B_t are benefit-quantity symbols, unrelated to the arm code B0 of §4; B_t is the contemporaneous net benefit). Joint acceptance requires all of: (i) the absolute defect within its registered cap (or at its own best improvement); (ii) the scheduled final segment and key checkpoints all passing; (iii) worst-case degradation over the run and the passing-checkpoint fraction reported together, both not worse than the registered thresholds; (iv) side-effect criteria not worse than the pre-set allowances (in the present instances the $med11$ level reading corresponds to the signed-median form of q, and the normalization denominator of $R(t)$ is the deployment-scale choice—mapping in §3.2). Two estimator notes (the hinge-vs-signed-median porting mapping and the two repair classes P/Ψ vs. training-objective interventions) are recorded in Appendix M.

Protocol elements (extracted from the four-arm experiment; portable piece by piece). Nine elements—four arms from one start (checkpoint-fingerprint whitelist); treatment matrix = baseline-no-repair / repair-only / repair+continuation / repair+anchor+continuation; primary readout $R(t)$ (Eq. 1) with the deployment-setting scale pinned + contemporaneous paired denominator; t_{half} + dense-early/sparse-late checkpoints; pre-registered tri-state endpoint verdict; seed-variance-band noise floor; leak-proof reference sets; side-effect panel; single-anchor readings barred from verdicts—are enumerated with their sub-clauses in Appendix M.

6.2 SCOPE

The protocol's interfaces are designed to be architecture-agnostic, with cross-architecture validation as future work; the implementation blueprint is in Appendix D (the taxonomy's prospective predictive power untested—rankings must be issued before continuation outcomes are known—and the blueprint a design claim).

7 RELATED WORK

Editing-decay neighborhood (citation delimitation first): LLM-edit decay under fine-tuning (Cheng et al., 2026), text-to-image editing robustness (He et al., 2025), and systematic study of whether erased concepts are truly removed (Lu et al., 2025) establish and quantify the phenomenon outside audio. Model-editing persistence (Mitchell et al., 2022) and catastrophic forgetting (McCloskey & Cohen, 1989) form the known neighborhood: this paper operates on physical constraints as the intervention object. *Our contribution is an acceptance methodology and standard validated on one system and designed for transfer rather than the phenomenon.* Relative to continual-learning regularizers (EWC/MAS—memory aware synapses—(Kirkpatrick et al., 2017; Aljundi et al., 2018), LwF (Li & Hoiem, 2018)), our anchor targets a scalar deployment-time behavior, is hinge-shaped, references each sample's own post-repair level, and bounds off-target side effects rather than retaining the repair—not claimed as a novel regularizer; retention metrics (BWT/forgetting (Lopez-Paz & Ranzato, 2017)) share ancestry with $R(t)$, differing in the deployment-setting scale and contemporaneous paired denominator. The differentiable digital signal processing (DDSP) lineage (Engel et al., 2020; Renault et al., 2023) and vocoder families (Siuzdak, 2024; Du et al., 2024; Lee et al., 2023) define the adaptation surfaces of the studied system; the band-gain×exponential-decay reverb of §2.2 belongs to the standard DDSP decay-reverb family (our contribution is the survival evidence and acceptance test, not the form). Literature search (2026-09): no prior work on audio

repair/calibration persistence under continued training; the nearest relatives (Jain et al., 2024; Lee et al., 2025) measure task performance, not repair persistence. Extended discussion in Appendix H.

8 LIMITATIONS

The claim domain is repair persistence for level defects, with pre-registered objective criteria as the evidence standard; listening tests (MOS/preference) and a public audio page are outside it and listed as post-acceptance verification (no verdict depends on listening scores). Evidence spans one system family and the loudness-defect construct ($med11$/A-weighted; not timbre); cross-domain generalization covers only URMP (U2 double backfire)—self-recorded conclusions do not transfer. Single-seed EATEN verdicts are unstable when more seeds are added (gain-family arm seeds (B3/B4) jointly spanning -12 to $+2.9$); C3 and R3 remain single cross-structure cases demoted to motivation (no contemporaneous control). Tri-state thresholds are uncalibrated engineering values with boundary-near readings; the two R scales admit ordinal comparison only, and the matrix normalization at $+500$ steps is blind to the sub-500 transient motivating the title quantity. The taxonomy's three empty cells are unevidenced; the four-arm design lacks a "no-repair + anchor" control arm (A4$-$A3 mixes the anchor main effect with the anchor$\times$repair interaction; the matrix's B4sham fills only the *gradient-attribution* gap, §4.2); and the unrepaired-baseline arm (B0) has no contemporaneous gradient control, so the alternative explanation—training pressure relaxing domain-wide even without a repair—is not excluded (§4.2). The four-arm endpoint numbers are authoritative from the machine archive, recomputable from the released reading files. Extended statement in Appendix K.

9 CONCLUSION

This paper organizes repair persistence as a measurable, comparable engineering objective: the taxonomy organizes risk diagnosis before surgery; $R(t)$ and the half-life turn the failure process into a recomputable curve; the four-arm protocol turns "will it survive continued training" into a controlled acceptance test rather than waiting for production incidents. The controlled experiment and the 34-run matrix yield two directional observations: lightweight anchoring has *no stable persistence advantage* over plain repair (B4 1/3 vs. B3 1/3 seed-level)—fuse semantics holds at matrix level (blocks overshoot: B4 worst seed -0.92 vs. B3 worst -12.05; does not retain repairs, the sham control attributing its effect to anchor gradients); and *persistence tracks the calibration source* (C3 3/3 vs. difficult-anchor 1/3; not separable at three seeds—one-sided Fisher exact $p = 0.20$, t=0 initial-mismatch confounding not separated). Two routes persist on all seeds (audio gate passed; not separable at three seeds; certified on the $+500$-normalized scale, window caveat in §8): training-objective repair (B2—a natural first candidate when the defect stems from an objective blind spot and the loss can be changed; its re-weighting value and side-effect profile registered in the artifact) and holdout-calibrated frozen gain (C3). Structural reparameterization is a constructive candidate demonstrated once under anchor supervision (R3), with matrix-level seed evidence below the two routes (B5 2/3, B6 1/3, U3 1/2; B6 = B5 + anchoring, reading below B5—1/3 vs. 2/3, not separable at three seeds—consistent with anchoring's lack of a stable gain). Cross-domain generalization (URMP U2 double backfire) is the main open boundary. Absolute quality, contemporaneous relative benefit (Eq. 2), and acoustic side effects jointly decide whether a repair is fit for delivery—continued training itself becomes part of repair acceptance.

Data and materials availability. All numbers reside in machine archives (archive components enumerated in Table 4); the artifact is released at submission with an anonymized repository and de-anonymized upon acceptance (code, starting checkpoints, data cards, timestamped pre-registration files, evaluation scripts, per-anchor raw matrices; public DOIs replacing the internal component names of Appendix L). Licensing/de-identification of the self-recorded cello data is in progress; if it cannot be released, a formal unavailability statement will be provided and the URMP public arms will serve as the externally runnable subset.

REFERENCES

Rahaf Aljundi, Francesca Babiloni, Mohamed Elhoseiny, Marcus Rohrbach, and Tinne Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *European Conference on Computer Vision*, 2018. arXiv:1711.09601.

Yinjie Cheng, Paul Youssef, Christin Seifert, Jörg Schlötterer, and Zhixue Zhao. Can fine-tuning erase edits? On the fragile coexistence of knowledge editing and fine-tuning. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 604–615, 2026. arXiv:2511.05852.

DDSP-SVC contributors. DDSP-SVC: Real-time end-to-end singing voice conversion system based on DDSP. `https://github.com/yxlllc/DDSP-SVC`. Open-source software; accessed 2026-09-24.

Hui-Peng Du, Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. APNet2: High-quality and high-efficiency neural vocoder with direct prediction of amplitude and phase spectra. In *Man-Machine Speech Communication*, pp. 66–80, 2024. arXiv:2311.11545.

Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. DDSP: Differentiable digital signal processing. In *International Conference on Learning Representations*, 2020. arXiv:2001.04643.

Jessica Z. Forde and Michela Paganini. The scientific method in the science of machine learning. *arXiv preprint arXiv:1904.10922*, 2019. ICLR 2019 Debugging Machine Learning Models Workshop.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting Fréchet audio distance for generative music evaluation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1331–1335, 2024. arXiv:2311.01616.

Feng He, Zhenyang Liu, Marco Valentino, and Zhixue Zhao. How robust is model editing after fine-tuning? An empirical study on text-to-image diffusion models. *arXiv preprint arXiv:2506.18428*, 2025.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018. arXiv:1709.06560.

Sarthak Jain, Orchid Chetia Phukan, Swarup Ranjan Behera, Arun Balaji Buduru, and Rajesh Sharma. SeQuiFi: Mitigating catastrophic forgetting in speech emotion recognition with sequential class-finetuning. *arXiv preprint arXiv:2410.12567*, 2024.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, pp. 2350–2354, 2019. arXiv:1812.08466.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017. arXiv:1612.00796.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR–half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019. arXiv:1811.02508.

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. arXiv:2506.00832.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. BigVGAN: A universal neural vocoder with large-scale training. In *International Conference on Learning Representations*, 2023. arXiv:2206.04658.

Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a multitrack classical music performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Transactions on Multimedia*, 21(2):522–535, 2019. arXiv:1612.08727.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2018. arXiv:1606.09282.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, 2017. arXiv:1706.08840.

Kevin Lu, Nicky Kriplani, Rohit Gandikota, Minh Pham, David Bau, Chinmay Hegde, and Niv Cohen. When are concepts erased from diffusion models? In *Advances in Neural Information Processing Systems*, 2025. arXiv:2505.17013.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In Gregory H. Bower (ed.), *The Psychology of Learning and Motivation*, volume 24, pp. 109–165. Academic Press, 1989.

Thomas McGrath, Matthew Rahtz, Janos Kramár, Vladimir Mikulik, and Shane Legg. The Hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D. Manning. Fast model editing at scale. In *International Conference on Learning Representations*, 2022. arXiv:2110.11309.

Lenny Renault, Rémi Mignot, and Axel Roebel. DDSP-Piano: A neural sound synthesizer informed by instrument knowledge. *Journal of the Audio Engineering Society*, 71(9):552–565, 2023.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and Fourier-based neural vocoders for high-quality audio synthesis. In *International Conference on Learning Representations*, 2024. arXiv:2306.00814.

Da-Yi Wu, Wen-Yi Hsiao, Fu-Rong Yang, Oscar Friedman, Warren Jackson, Scott Bruzenak, Yi-Wen Liu, and Yi-Hsuan Yang. DDSP-based singing vocoders: A new subtractive-based synthesizer and a comprehensive evaluation. In *International Society for Music Information Retrieval Conference*, 2022. arXiv:2208.04756; the system is named SawSing.

A Experimental Substrate: System, Data, Metrics, and Intervention Registry

This appendix documents the experimental substrate at the resolution available in the machine archives; items marked † ship verbatim in the anonymized artifact (training and continuation code, starting checkpoints, configuration files, and dataset cards) rather than being fully restated here, in keeping with the page budget.

A.1 System. The studied synthesizer is a differentiable digital-signal-processing (DDSP) chain for solo cello at 48 kHz with three audio-generating branches—a harmonic (sinusoidal-sum) branch, a filtered-noise branch (Engel et al., 2020) (bow noise and breath-type components; the module whose lack of a silence gate underlies the C3 leak, §2.1), and a convolutional reverb branch (the module reparameterized in R3, §2.2)—plus conditioning encoders and a final output level. Training uses a reconstruction objective family (weighted multi-scale spectral L_1 + mel loss) with a multi-scale parameter-domain adversarial discriminator, and stage-dependent curriculum weights; the exact architecture dimensions, discriminator configuration, optimizer, learning-rate schedule, batch size, and clipping thresholds† are frozen in the archived launch configurations. The production checkpoint at step 77,000 is the shared starting point of the four-arm and matrix experiments; the C3 and R3 cases reside on an earlier line of the same family with the same branch inventory. The production chain no longer has the C3-shaped 14 kHz leak: an architectural double soft-gate (noise and harmonic gates), wet-end (post-reverb) silence-window readings -5.92 dB relative to the original recording's floor, 11/11 samples below it, cross-confirmed under the measured scale and the

archived residual scale, pre-registered leak gates not triggered; the pre-registered feasibility check was completed before the runs (§3.1).

A.2 Data. Two domains. *Self-recorded cello*: studio-condition 48 kHz solo-cello recordings of étude repertoire selected for technique coverage (long bow, pizzicato, harmonics), one performer per instrument; training segments, the 11 fixed validation probes (anchor-level reading units; §3.1), and the 32 training-side reference samples (the 32-sample reference set; anchor references) are disjoint, with the selection rules† archived. *URMP* (Li et al., 2019): the public multi-instrument ensemble dataset; the subset, resampling, and the cascade evaluation definition used for the URMP arms (U1/U2/U3)† are archived with the dataset card. Licensing/de-identification of the self-recorded domain is in progress (§9); the URMP arms constitute the externally runnable subset.

A.3 Metrics. $med11$: for each of the 11 validation probes, the signed median over voiced frames of the A-weighted loudness difference between the deployment-setting render and its reference; $med11$ is the median of the 11 probe values. Voiced-frame detection, the A-weighting operator, and the frame rate† are fixed in the evaluation harness. The deployment-setting readout keeps the repair in its deployed form (A1 read without the gain; repaired arms read with g^*); the uncorrected readout removes it. The propagated R-noise scale of ± 0.4–0.5 derives from the seed-replication control seed-variance band (§3.4); $med8$ (a secondary eight-anchor aggregate) and the side-effect panels (band shares, discriminator state) are defined in the archived evaluation scripts.

A.4 Intervention registry (per arm; 13 treatment arms plus one run-level sham control—B4sham is counted as a run, not an arm).

Arm	Intervention	g^* source	Seeds	Start
B0	unrepaired continuation	—	3	pre-copied
B1	inference-only calibration (zero training)	—	3	attached to B0
B2	loudness-loss weighting (training-objective change)	n/a	3	pre-copied
B3	frozen-gain repair	difficult-anchor 5.03 dB	3	pre-copied
B4	frozen gain + behavioral anchor	difficult-anchor 5.03 dB	3	pre-copied
C1	g^*=5.03 replication of the four-arm configuration	difficult-anchor 5.03 dB	3	pre-copied
C2	cross-source calibration	training-distribution median 1.23 dB	3	pre-copied
C3	cross-source calibration	external-holdout measured 3.99 dB	3	pre-copied
B5	structural surgery: reverb band-gain lift Δ_s=4.92 dB + re-anchoring	—	3	surgical
B6	B5 + behavioral anchor	—	3	surgical
B4sham	sham intervention (identical to B4 with the anchor score identically zero)	difficult-anchor 5.03 dB	1	pre-copied
U1	URMP unrepaired continuation	—	2	pre-copied
U2	URMP frozen gain + anchor	URMP in-domain measured 3.99 dB	2	pre-copied
U3	URMP structural surgery Δ_u=3.46 dB	—	2	surgical

Table 1: Intervention registry. B3, B4, and C1 share the difficult-anchor gain (5.03 dB) and differ in anchoring and replication provenance; the B2 re-weighting value, the anchor-loss weight, the surgical parameterization details, and the exact seed values† are registered in the artifact. "Pre-copied" starts are frozen copies of the shared checkpoint; "surgical" starts apply the reparameterized correction before continuation (§4.1).

A.4b Surgical-parameter solving. Converged before the runs: $\Delta_s = 4.92$ dB—six iterations, terminal residual $+0.033$ dB, within the ± 0.05 dB convergence band; $\Delta_u = 3.46$ dB—eight iterations, residual -0.0047 dB, within the band; full-precision values recorded in the intervention-registry archive (§4.1).

A.5 Per-seed endpoint values (Figure 2 in numbers).

A.6 Compute and artifact contents. Compute: the matrix alone is 34 training runs $\times$ 15,000 steps (plus the four-arm experiment and the seed-replication control runs) with 306 checkpoint evaluation passes—checkpoint evaluation passed 306/306 (self-recorded domain: med11 primary + med8 secondary (the eight-anchor secondary readout) + side-effect panels, 252 items; URMP domain: cascade, 54 items; structural arms: g_{db} drift, 72 readings); single-run hardware allocation and wall-clock† are archived. The artifact, released at submission with an anonymized repository (de-anonymized upon acceptance), contains: training and continuation code with launch configurations; the starting checkpoints (including the shared production checkpoint at step 77,000); dataset cards with split and selection rules; the timestamped pre-registration files (four-arm spec, suite spec and plan); the evaluation harness (med11/med8, side-effect panels, cascade); the anchor-loss implementation; and all reading files enumerated in Appendix L.

Arm	$R(\text{end})$ per seed	Tri-state per seed
B2	$+0.62$ / $+7.42$ / $+2.11$	P / P / P
C3	$+1.08$ / $+2.60$ / $+1.76$	P / P / P
B5	$+14.88$ / -0.80 / $+8.15$	P / E / P
C2	-0.30 / $+1.96$ / $+5.21$	E / P / P
B3	-12.05 / $+0.30$ / $+2.35$	E / Pa / P
B4	-0.92 / -0.39 / $+2.86$	E / E / P
C1	$+0.26$ / -0.22 / $+4.67$	Pa / E / P
B6	$+0.08$ / $+1.00$ / $+0.34$	E / P / Pa
B4sham	-0.12	E
U3	$+0.26$ / $+0.58$	Pa / P
U2	-1.26 / -1.13	E / E

Table 2: Matrix endpoint retention per run (single-source file: the matrix reading file; P = PERSIST, E = EATEN, Pa = PARTIAL). Values outside $[0, 1]$ reflect near-zero first-checkpoint denominators on the $+500$-normalized scale (§8).

B Mechanistic Background and Historical Assets

B.1 Case C3 (full text): a physically correct repair eaten in 1,500 steps

Defect, repair, and the eating. From a cello differentiable-synthesis production chain: a persistent leak near 14 kHz during silent passages, architectural in root cause (the filtered-noise module has no silence gate—a standing leak, not learned contamination). The physically straightforward repair: a 13 kHz-cutoff low-pass at the noise-branch output, warranted by the absence of true sounding content above 13 kHz. A drop-in probe passed all listed readings: dry-end (pre-reverb) leakage in the 14 kHz silence window (bins 295–306; 8 validation samples) fell from -18.82 to -49.67 dB (30.86 dB); the product end (the final output) fell $-55.15 \rightarrow -71.85$ dB; the 4–13 kHz bow-noise share moved only -0.074 pp (0.3 pp safety threshold). After 1,500 further training steps (checkpoint 40000 $\rightarrow$ 41500), the dry-end leak rebounded to -11.17 dB—a $+38.51$ dB rebound from drop-in and *7.65 dB worse than the pre-repair baseline* (steady-state criterion ≤ -46.67 dB: fail); the product end rebounded to -36.74 dB (both ends *worse* than their pre-repair baselines: dry $+7.65$, product $+18.41$ dB; the two ends are reported separately, never merged). Side effect: the 4–13 kHz share fell 5.885 pp (true bow noise damaged). The training score function barely noticed: weighted-spectral + mel loss 0.97 $\rightarrow$ 0.98, $+0.5\%$ relative. Module level, the reverb gain $|H(14\text{k})|$ rose $+0.30$ dB—the reverb no longer needed its notch cancellation, a re-adaptation within mechanistic expectation (five-item steady-state summary in the archival-details subsection below).

Mechanistic reading (hypothesis-level; controlled test in §3) and disclosure. The inserted low-pass lies in no loss term's computation graph—zero gradient through it, the noise-branch amplitude near 14 kHz unanchored by any gradient, and continuation drifts it back (two nuances below: loss-value invariance vs. nonzero training gradient, half/full-derivative ratio 94.45; and the above-baseline rebound's tension with "zero gradient $\Rightarrow$ no directed drift," suggesting an unlocalized systematic pull—candidates: nonzero effective loss weight at 14 kHz; parameter coupling). In the taxonomy (§5) the case is classified under *zero-visibility insertion* (post hoc). Identity disclosed: C3 is a temporary recording stopgap, not an architectural repair; we recruit it as a motivating exhibit with that status stated (evaluation configuration below).

B.2 CASE R3 (FULL TEXT): WHY A STRUCTURAL REPARAMETERIZATION WAS NOT EATEN

Defect. Earlier, on the same synthesis line: the free reverb IR (all 144,000 sample taps independently trainable) was squeezed by reconstruction pressure near 14 kHz into high-Q narrow peaks: measured decay $T_{60}^{\text{est}} = 4.18$ s at 14,594 Hz—roughly an order of magnitude longer than typical room reverberation ($\sim$0.5 s); peaks exceed neighbors by $+17.92$ dB, Q up to 27,745. Coherent high-Q ringing is the source of the metallic timbre; the root cause is structural in the solution space—a high-Q narrow peak is the minimum-energy way to satisfy a 14 kHz band-energy match.

Repair. Not a filter but a *reparameterization* (changing the parameterization): 144,000 free taps $\rightarrow$ 30 bands $\times$ (band gain g_b + decay time T_b) $= 60$ physical knobs, each band's IR a fixed band-limited noise times an exponential decay, $h(t) = \sum_b g_b\, n_b(t)\, \exp(-\ln(1000)\, t/T_b)/N_b$ with $T_b \in (0.05, 1.5)$ s (N_b normalizes each band's fixed band-limited noise n_b; the $\ln(1000)$ exponent makes the amplitude at $t = T_b$ exactly one-thousandth $= -60$ dB, so T_b directly reads as the band's T_{60} decay time). The hard T_{60} bound caps the equivalent Q at 14.5 kHz (conversion: $T_{60} \leq 1.5$ s gives a bandwidth floor $2.2/T_{60} = 2.2/1.5 \approx 1.4667$ Hz, so the equivalent Q cap is $14{,}500/1.4667 \approx 9{,}886$, recorded as 9,890 by tens-digit rounding)—coherent high-Q ringing is physically inexpressible under the new parameterization. This eliminates the pathological solution space, not the reconstruction pressure: the pressure remains, and the only legal outlet becomes wideband band-level gain (natural reverberation). A reverb anchor loss (band-gain hinge, weak all-band L_2, log-domain decay prior) pins the 60 knobs.

Survival evidence and "init is the endpoint." Post-training retention: T_{60} pathology 4.18 s $\rightarrow$ 0.12–0.14 s (scale: the cross-band range of the 30 parameterized band T_{60} values; 0.1185–0.1440 s measured at the acceptance checkpoint 43000); author listening checks over three sessions (init / checkpoint 43000 / 48000: ringing gone, closer to the original recording, bow noise retained, natural tails) all passed; checkpoint 48000 became the new base checkpoint of this model line (best validation loss 17.229 at checkpoint 48000 on that line's validation-loss scale, lower is better); the long-training health gate raised no alarms on three keys ($T_{60}^{\max}$ drifted mildly 0.1445 $\rightarrow$ 0.1683 s over the 43000$\rightarrow$48000 segment—the 0.1445 is the measured $T_{60}^{\max}$ machine reading, close to but on a different scale from the 0.1440 parameterized band-range upper end at the acceptance checkpoint, both originals preserved in the archive—within $9\times$ margin of the 1.5 s bound); the line completed delivery; disclosed: an independent output-shape check carries one minor failing item (not affecting the main-line conclusion; minor blemishes listed with the artifact). "Init is the endpoint," empirically: extrapolating the parameter-drift slope of 50 real Adam steps over 3,000 steps gives 0.070 dB, far below the 3.69 dB re-fit recovery gap (scale: the 14 kHz band-energy difference of the reparameterized initial state vs. the old base, cross-sample mean)—no rebound potential to release (extrapolation assumption disclosed: validity bounded by the measured slope; no fast transient observed under anchor supervision).

Mechanistic reading and boundary. In taxonomy language the repair sits *inside the loss's line of sight*—the reparameterized parameters remain watched by both the anchor loss and reconstruction pressure while pathological solutions are structurally sealed; hence survival. R3 is the constructive existence proof for the root-cure layer (§5). Boundaries: C3 is an inference-end insertion on the noise branch, R3 a reparameterization of the reverb branch—a cross-structure contrast, not a same-structure counterexample pair; the training scenarios also differ (mid-continuation insertion vs. fresh retrain), so a third variable is not separated. R3 proves the class "in-sight reparameterization survives" constructively; it does *not* prove "reparameterization is never eaten."

B.3 C3 ARCHIVAL DETAILS

The five-item steady-state summary after the 1,500-step rebound (§2.1; $\times$ fail / $\triangle$ partial warning / ✓ pass): rebound $\times$, mid-band side effect $\triangle$, IR no longer notching ✓, cancellation anomaly $\triangle$, loss stable ✓. The product-end readings: $-55.15 \rightarrow -71.85$ dB at drop-in, rebound to -36.74 dB ($+18.41$ dB above the pre-repair baseline, as in §2.1). Loss decomposition: weighted-spectral $+1.8\%$, mel $+0.0\%$. Evaluation configuration: 8 validation samples, seed 20260806, STFT n_{fft}=1024, hop $= 256$, 48 kHz.

B.4 CONDITIONAL NUMERICAL INVARIANCE: A LOSS VALUE CAN BE EXACTLY BLIND TO A CLASS OF TRANSFORMS

When three conditions hold (a shared positive scalar gain applied to both prediction and target; identical loudness operator, fixed frames, fixed masks, fixed weights; the loudness/power floors and clipping never triggered), the forward value of a dB-domain L_1 loudness loss is invariant to the gain: the losses at $\delta = +2$ and $+6$ are exactly equal, and their difference from $\delta = 0$ is one float32 unit in the last place ($1\,\mathrm{ULP} = 9.54 \times 10^{-7}$). Under a shared gain the squared error is $g^2(\hat{y} - y)^2$ with a nonzero derivative in g—invariance of the value does not mean a zero gradient. We emphasize that the causal statement "C3 was eaten because of this invariance" is not supported: the subject of C3 is the continuation dynamics and the compensation paths, and the causal chain is strengthened by the controlled experiment of §3.

B.5 HALF-GRADIENT / FULL-DERIVATIVE SEPARATION

On the same probe, the norm ratio between the training-returned backpropagated gradient (target side cut by stop-gradient) and the mathematical full derivative is **94.45** (three conditions: a single synthetic input; stop-gradient structure as the main factor; prediction side touching the 0.4% floor). "The loss value does not move" and "the gradient is zero" are two different statements; all visibility discussion is split by units. The band-average net push is -0.996 (14 of 30 components negative $=$ 46.7%; the bands do not push uniformly—we avoid "all bands push up" formulations; true cosine -0.148, energy share 2.19%).

B.6 PRE-REGISTERED REJECTION OF THE SUFFICIENT-PRESSURE HYPOTHESIS

A pre-registered gate test of the "insufficient optimization pressure in the volume direction" hypothesis: across three checkpoint generations the gradient-magnitude ratio of the volume virtual knob to the noise/reverb reference knobs is $\hat{\rho} = 1.3160/1.4888/1.4616$ (bootstrap 90% CI lower bound at least 1.2816), against a criterion requiring $\hat{\rho} \leq 0.20$ and CI upper bound ≤ 0.35—the measured values sit far above the threshold cap, consistent across the three checkpoint generations; verdict **not passed**: the pressure is not lacking but 32%–49% larger than the reference directions, while the under-loudness persists. **Background reading**: the presence of optimization pressure is no survival guarantee—resonant with the A4 verdict (artificially reinforced visibility by anchoring still failed to retain the repair). Instrument credibility: two-sided blind 54/54, sensitivity positive control 54/54, controlled ± 8 dB scenarios 17/17 all passed; shape control 34/54 (20 mismatches adjudicated as data-shape differences, not instrument failure; the verdict is unaffected).

B.7 UNDER-LOUDNESS READINGS AND THE VARIANCE BAND

The systematic under-loudness has two readout scales reported side by side (never mixed): -5.86 dB (eight-anchor subset, frame-power domain, six qualifiers archived with the reading) and -5.03 dB (the shared production checkpoint at step 77,000, anchor-11, A-weighted); dry/wet split $-18.64/+12.85$ dB (strictly additive per anchor, not additive at the median layer; residual 0.0636 dB). The seed-replication variance band (§3.4) gives the *scale reference* for training-induced level-change readings in this domain (cross-segment proxy, $n{=}2$): *the direction of the phenomenon is real (both arms under-loud) while the headline number has high between-anchor heterogeneity*—single-point readings must not be read in isolation, single anchors are barred from verdicts, and reports must carry the variance band. The cause remains open and is not claimed in this paper.

B.8 POSITIONING OF INFERENCE-TIME SELF-CALIBRATION

Per-sample loudness calibration at inference (checking the actual output against a loudness target specified by input features and correcting the difference; consumes no reference audio, zero training cost): $med11$ $-5.03 \rightarrow -0.0006$ dB (within three conditions: ± 12 dB clipping limit, 100-voiced-frame gate, floors untouched; the zeroing is constructively achieved, not model self-healing; frame-level residual 1.17 dB). **Taxonomy position = the class that never faces the training test**: efficacy claims limited to "frozen-deployment" scenarios—a zero-training emergency item in the toolbox narrative, not occupying layers B/C.

B.9 HISTORICAL ASSETS (INCLUSION SCOPE: ITEMS DIRECTLY SUPPORTING THE ARGUMENT)

The 2×2 loss-design interaction at -12.46 percentage points (L1/L2 $\times$ clamp/log1p, four cells, independent recomputation agrees); a branch loudness-loss local improvement of -0.31 dB (median of 15 improvements over 16 samples, lacking a blank control—a demonstration that neither confirms nor denies); UV phase freezing pressing non-zero DC down by 29–34 dB (re-verification of an existing conclusion, not claimed as new); a naive replication-control endpoint difference of $+1.80$ dB vs. a same-anchor paired median of $+0.38$ dB (ratio ≈ 4.7; the anchor-subset uncertainty band $[-0.75, +0.95]$ dB written separately from the seed variance); an oracle ceiling reference: BigVGAN's (Lee et al., 2023) near-ceiling reconstruction error 0.44 percentage points, correlation 0.965 (the route motivation that "the ceiling lives in the parametric synthesizer"). The R3 four-way inference-interchange and loss-design reversal history are retained in the artifact (not included in this paper).

C FULL R(T) CHECKPOINT TABLE (FOUR-ARM EXPERIMENT)

Steps	$+500$	$+1$k	$+2$k	$+3$k	$+5$k	$+7.5$k	$+10$k	$+12.5$k	$+15$k
R(A3)	0.5044	0.4574	0.2756	-0.2645	-0.0262	-0.2503	-0.0541	-0.0019	-0.2733
R(A4)	0.4003	-0.0685	0.1805	0.0594	-0.6885	0.0106	-0.1493	-0.1131	0.0705

Table 3: Pre-registered-checkpoint $R(t)$ values for the four-arm experiment arms (the curve shapes are carried by Figure 1; this table is the data record).

Boundary margins and fast-transient support (§3.3). The $+500$ point sits 0.0044 above the threshold, the $+1000$ point 0.0426 below—*both margins lie inside the paper's own propagated R-noise scale of ± 0.4–0.5, hence a single-seed descriptive reading*. The headline "$\leq$ 1,000 steps" is supported by a 200-step pilot with $R \approx 0.20$ and first-checkpoint absolutes A1 -5.97 / A3 -2.96 / A4 -3.58 dB—a real fast transient in the repaired arms.

D CROSS-ARCHITECTURE IMPLEMENTATION BLUEPRINT (DESIGN; NOT EXECUTED)

Adaptation surfaces for three classes of generative structure: explicit harmonic/noise/reverb structures (this system—branch, filter, reverb, and final-level interfaces); explicit spectral generation (Vocos / APNet2-class—magnitude, phase, and frequency-grid interventions); time-domain waveform generation (BigVGAN-class—output-transform and behavioral-loss interfaces). A cross-architecture main comparison should include at least one shared final-output repair task (one target definition, one repair-operator family, one calibration rule, one success condition); architecture-specific interventions are auxiliary only, matched on intervention strength/bandwidth/group delay and equipped with empty-intervention controls. The taxonomy's prospective predictive power must be tested by issuing rankings before continuation outcomes are known and scoring them on held-out tasks—that test has not been executed, and this appendix is a design claim. Three construct-matched cross-domain pre-registered validation targets are also registered (image white balance, point-cloud alignment, contrastive-learning augmentation; anchored to the comparison-suite spec), pending execution.

E MATRIX ENDPOINTS FIGURE

F COMPENSATION-LOCALIZATION FIGURE

Reading dimension (gradient cross-check, §4.2). Loss-per-dB signed half-gradient: the directional derivative of the total loss with respect to a 1-dB turn of the output-level knob, in training-backprop units—positive = the loss increases as the output level rises, i.e. optimization pressure points toward suppressing the output level; negative = pressure reversed, pointing toward raising

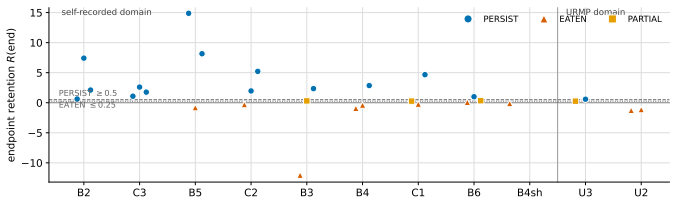

Figure 2: **34-run matrix endpoints.** Endpoint retention $R(\mathrm{end})$ for each arm–seed run (29 repair runs; tri-state coloring and marker shape as shown; 0.5/0.25 thresholds dotted). Self-recorded domain (left of the divider) and URMP domain (right) shown together; cross-scale readings are ordinal only (§4.2).

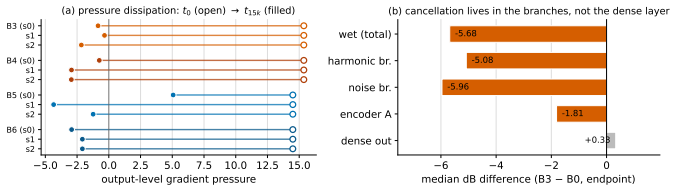

Figure 3: **Compensation localization. (a)** Output-level gradient pressure at t_0 vs. t_{15k} for the 12 instrumented runs (one line per run): pressure dissipates in 12/12 (11/12 flip negative; the B5_s0 residual $+5.07$ reflects structural adaptation—the surgery embeds the gain into the starting checkpoint's parameters). **(b)** Internal differencing—cancellation carried by the harmonic and noise branches, not the dense layer (median dB differences, B3 minus B0; n=1 run pair, §4.2).

it (the anchor-gradient channel of instrument (iv) uses the same dimension; positive = pressure to suppress). The B5_s0 residual $+5.07$ is an expected exception: the structural surgery embeds the gain into the starting checkpoint's parameters (not an external g^*), and the residual positive value reflects the structural parameters' own adaptation process—not in conflict with the pressure-reversal reading of the external-gain arms. **Compensation decomposition (instrument (iii)).** An accounting identity: with-g^* minus the uncorrected reading $\equiv g^*$; the corrected reading takes cancellation from uncorrected-read drift plus branch differencing. Instrument (iv)'s two verification channels: the paired-difference CI test excluding the anchor-forward side-effect explanation, and the directly measured anchor-backpropagated gradient effect ($+1.74$)—the two lines of evidence converge on the gradient pathway.

G Extended Toolbox Layers

Risk unclear → diagnose first (§5); structure untouchable → add the overshoot guard (in the self-recorded domain it prevents backfire; on URMP the package backfired, §5); durable delivery → evaluate the root cure (§5); before delivery → accept through the four-arm protocol (§6). No universality is promised.

G.1 Layer A — diagnosis: prediction and measurement

Repair-visibility taxonomy (formerly *loss visibility*). Repair visibility = the training score function's sensitivity to the repaired behavior (visible = watched by a gradient). Two dimensions—*inside the loss gradient-backprop graph* × *reparameterized*—give a four-cell spectrum, each cell evidenced (the two ±anchor cells are control arms within one four-arm experiment, not cross-experiment replications): *zero-visibility insertion* (C3, observational)—eaten, fast rebound within 1,500 steps; *weak visibility, no anchor* (A3, controlled)—eaten with overcorrection, half-life ≈1k steps, ≈6× faster than the slow-path prior (≈5.9k steps from the registered historical drift rate -0.43 dB/1k steps); *weak visibility + lightweight anchor* (A4 = A3 ±anchor control)—still eaten, overshoot guarded (fuse); *in-sight reparameterization* (R3, constructive existence)—survives.

Boundary: C3 and A3 are cross-chain cases in different cells (annotated as such); the taxonomy's empty cells (zero-visibility+anchor / in-sight+no-anchor / weak-visibility+reparameterized) are unevidenced, and predictive power is bounded by the evidenced cells. **Retention metric.** $R(t)$ and t_{half} (§3.2) turn "how long the repair lasts" into a recomputable curve and a single number; a fast transient is itself a high-risk signal. **Compensation-path detection.** The gradient-decomposition probe suite records module-level readings before and after the repair (module×loss-term matrix; the C3 IR behavior of §2.1 is the prototype), and the matrix-level localization instruments are now in place (§4.2)—*compensation forming against the repair during continuation is cross-confirmed across seeds*. A related pre-registered gate test additionally *rejected* the insufficient-optimization-pressure hypothesis for this defect family: volume-direction gradient pressure ran 32%–49% *above* the reference directions—far above the gate's rejection threshold—across three checkpoint generations (details in Appendix B)—pressure being present is no survival guarantee.

G.2 Layer B — overshoot guard: lightweight anchoring as a fuse

The hinge anchor (§3.1) is a fuse, not a retainer. A4 evidence: prevents backfire ($+1.42$ dB over A3—a within-band magnitude-level reading; no significance claim) but is not sufficient ($R = 0.071$, still EATEN). Matrix-level (§4.2): no stable persistence advantage over plain repair (B4 1/3 vs. B3 1/3 seed-level), but the sham control confirms the effect comes from anchor gradients (B4sham 0/1, no significant endpoint difference from B3)—fuse semantics consistent with the observed lower envelope (B4 worst seed -0.92 vs. B3 worst -12.05; an n=3 extreme-value comparison, directional only). Applicability: minimum protection when the parameterization cannot be restructured; the fuse conclusion rests on a single hinge-anchor configuration frozen before the runs, untested against stronger anchors or EWC/LwF-class regularizers (at three seeds any such contrast would be non-separable, §4.2).

G.3 Layer C — root cure: structural reparameterization

Seal the pathological solution space physically and pin it with an anchor loss (R3, §2.2: 144,000 → 60 knobs; high-Q ringing inexpressible; T_{60} 4.18 s → 0.12–0.14 s; "init is the endpoint" 0.070 dB against a 3.69 dB gap). R3 is a *joint reparameterization + anchor-loss* artifact: the layer's net effect cannot yet be separated (the "reparameterized + unanchored" cell is empty), and its survival is "survival under continuous anchor supervision," weaker than "surviving unprotected continuation." Cost: architecture change and retraining. Matrix evidence: B5 2/3, B6 1/3, U3 1/2 seed-level PERSIST (§4.2).

G.4 The cost–persistence spectrum and engineering rules

Layer B (low cost, guards backfire without retaining) and layer C (architecture change, retention) form the ends of a cost–persistence spectrum. Three rules: (i) hard cuts/masking ship with anchoring; (ii) inference-end repairs (e.g. per-sample loudness calibration, constructively zeroing $med11$ $-5.03 \rightarrow -0.0006$ dB at zero training cost) have efficacy claims *limited to frozen deployment*—they never face continued training; (iii) durable delivery → evaluate reparameterization first.

H Extended Related Work

Intervention-decay neighborhood (citation delimitation first). Cheng et al. (2026) quantify LLM-edit decay under fine-tuning; He et al. (2025) study text-to-image editing robustness under fine-tuning; Lu et al. (2025) examine whether erased concepts are truly removed. Model-editing persistence (Mitchell et al., 2022) and catastrophic forgetting (McCloskey & Cohen, 1989) form the known neighborhood: this paper operates on physical constraints as the intervention object, contributing a layered repair toolbox, module-level compensation measurement, and the repair-visibility taxonomy—the differentiation is on the solution axis. None of these address differentiable audio synthesis, repair acceptance protocols, or compensation-path localization; *relative to all of them, our contribution is an acceptance methodology and standard validated on one system and designed for transfer*—a repair-visibility-framed layered toolbox (diagnosis / overshoot guard / root-cure candidate), a portable four-arm protocol, and (to our knowledge) the first fully machine-archived evidence in differentiable audio synthesis plus constructive durable-repair routes.

Continual-learning regularization, delimited. Behavioral anchoring is related to anti-forgetting regularizers but differs in goal and form: EWC/MAS (Kirkpatrick et al., 2017; Aljundi et al., 2018) anchor the *weight space* (Fisher-weighted quadratic penalties), LwF (Li & Hoiem, 2018) anchors the *full output distribution* (KL distillation), both against old-task forgetting; our anchor targets a *scalar deployment-time behavior* (voiced-frame loudness drift), is *hinge-shaped* (penalized only beyond 3 dB), references *each sample's own post-repair level* (carrying repair-target information), and aims to *bound off-target side effects* rather than retain the repair. We do not claim a novel regularizer; the contribution is the specific form and the controlled failure boundary "guards overshoot, does not retain repairs" (same-cell $\pm$anchor four-arm contrast).

Metrics and interpretability. Continual-learning retention metrics (BWT/forgetting (Lopez-Paz & Ranzato, 2017)) and editing-decay metrics share ancestry with $R(t)$; our differences are the deployment-setting scale, contemporaneous paired denominator, and joint acceptance over absolute quality, relative benefit, and side effects. Mechanistic-interpretability self-repair (post-ablation downstream compensation, the Hydra effect (McGrath et al., 2023)) is closely related to compensation paths; our differences: the compensation forms against a repair during continuation, with directional, quantitative uncorrected-readout evidence.

Differentiable reverb and synthesizer lineage. DDSP (Engel et al., 2020) provides the differentiable DSP interface; DDSP-Piano (Renault et al., 2023) discusses reverb absorbing other modules' duties and constraining it—we introduce continued-training acceptance on the same module-modeling problem. Delimitation: the band-gain$\times$exponential-decay construction of §2.2 belongs to the standard DDSP-lineage decay-reverb family (Engel et al., 2020; Renault et al., 2023); the root-cure layer's contribution is the survival evidence and the acceptance test under continued training, not the parameterization form itself. Vocoder lineages (Siuzdak, 2024; Du et al., 2024; Lee et al., 2023) define the adaptation surfaces of Appendix D; SawSing (Wu et al., 2022) and DDSP-SVC (DDSP-SVC contributors) are infrastructure of the studied system.

Evaluation critique and seed variance. FAD (Fréchet Audio Distance; metric source Kilgour et al., 2019) style metric critiques (Gui et al., 2024) and deep-learning seed variance (Henderson et al., 2018; Forde & Paganini, 2019) are complementary: we provide signal-domain training dynamics specific to our domain and (to our knowledge) its first measured seed-variance-band anchor. Scale-invariant-SNR design critique (Le Roux et al., 2019) is adjacent in theme but a different object.

Literature search. For "audio repair/calibration failing or persisting under continued training," searches over audio/speech/TTS/vocoder/model-surgery keyword families found no prior work; the nearest relatives are catastrophic-forgetting studies of sequential speech-task fine-tuning (Jain et al., 2024) and inference-time activation editing for TTS (Lee et al., 2025)—both measure task performance, not repair persistence.

I Sample Heterogeneity (Seed-Replication Control)

The two seed trajectories' endpoint med11 values are -5.7511 and -3.9876 dB; the difference of the two aggregate medians is -1.7635 dB (scale = subtraction of the two medians; the machine archive has no single archived entry for this—it is a recomputed value); the per-anchor paired-difference median is -0.3295 dB with paired bootstrap 90% interval $[-2.0209, +0.2582]$ dB (at n=11 the nominal 90% bootstrap interval achieves $\approx$85% actual coverage, a discrete-coverage deficit; the secondary scale also crosses zero). The control retrained two arms byte-identical except the seed (R0/R1, 15k steps; a two-segment continuation with a mid-segment topology change; same-segment multi-seed evidence in §4); the deepest-anchor split is amplified to -2.9/-3.7 dB (recorded in the artifact). The two numbers correspond to different aggregation orders and do not contradict each other—the protocol reports both the aggregate change and the per-recording distribution, preserving the heterogeneity of difficult samples.

J Denominator-Scaling Disclosure

The pre-registered contemporaneous denominator $(D_{\mathrm{rep}} - D_{\mathrm{base}})$ moves with the baseline arm's contemporaneous reading. "Common drift cancels internally" holds only for the numerator's shared component. A1 ending $+0.91$ dB louder than its start shrinks the denominator, inflating $|R|$ by about

22% relative to a fixed t=0 denominator (A3's endpoint is $R = -0.224$ under the fixed-denominator scale vs. -0.273 contemporaneous); part of the curve's non-monotone jitter may be transmitted from A1 reading fluctuation. The t_{half} conclusion is unchanged under the fixed-denominator scale (A3 first point $0.598 > 0.5$, second $0.397 < 0.5$; A4 first point $0.475 < 0.5$). This paper's verdicts use the pre-registered contemporaneous scale, with the fixed scale reported as a robustness control.

K Extended Limitations

Evaluation-standard statement. The claim domain is repair persistence for level defects, with pre-registered objective criteria as the evidence standard (R(t) tri-states + gradient cross-check + internal differencing + sham-intervention control). Human listening tests (MOS/preference) and a public audio page are outside the evidence standard and listed as post-acceptance verification—no verdict depends on listening scores; R3's listening evidence is author verification over three sessions, engineering-observational in nature. FAD-style generation-quality metrics are a different object from defect-repair persistence and are not used as criteria.

Evidence scope. Controlled evidence spans from a single-seed four-arm design to the 34-run matrix (§4); but the matrix itself shows *single-seed EATEN verdicts are unstable under seed replication (gain-family arm seeds (B3/B4) jointly spanning -12 to $+2.9$)*—the §3 verdicts are one realization of the seed distribution, not final. C3 and R3 remain single cases (n=1+1, cross-structure). *Persistence conclusions cover only the level-defect domain (med11/A-weighted loudness)—R's construct is not the timbre domain*; audio-domain generalization covers only URMP, and U2 backfires on both seeds—self-recorded-domain conclusions do not transfer to URMP.

Taxonomy empty cells (zero-visibility+anchor / in-sight+no-anchor / weak-visibility+reparameterized) have untested predictive power; cross-model-family replication is future work. **Cross-domain porting** of the protocol is a design claim (not replicated in LLM/T2I/other synthesizer domains).

Verdict thresholds (0.5/0.25) are engineering decision values without calibration anchors; results crossing a boundary carry boundary uncertainty—A3's first checkpoint $R = 0.5044$ and A4's endpoint 0.0705 both sit near thresholds (the R noise-propagation scale is of threshold order, §3.4); label robustness is answered by the matrix's multi-seed distribution.

The four-arm design lacks a "no-repair + anchor" control arm: A4$-$A3 mixes the anchor main effect with the anchor$\times$repair interaction (A4's anchor reference is the post-repair level, carrying repair-target information); the matrix's B4sham sham control fills the *gradient-attribution* gap (gradient vs. forward side effects: B4sham 0/1, no significant endpoint difference from B3—anchoring effects attributed to anchor gradients); separating the anchor main effect from the interaction still requires a "no-repair + anchor" arm.

Gradient-control arm coverage: the matrix-level gradient comparison covers only the 12 instrumented repair runs (B3/B4/B5/B6 $\times$ 3 seeds); the unrepaired-baseline arm (B0) has no contemporaneous gradient comparison, so the alternative explanation—training pressure relaxing universally on unrepaired arms as well (not repair-specific)—is not excluded (§4.2).

First-crossing times are bounded by the sampling interval (A4's t_{half} is an upper bound). **Scale commensurability**: the matrix $R(t)$ normalizes at the $+500$-step first checkpoint, so the scale-up instrument is blind to the sub-500-step fast transient that motivates the title quantity; and its reading unit (per-anchor paired median) differs from the four-arm experiment $R(t)$ (med11 aggregate), so cross-scale statements are ordinal only. **Tail behavior**: med11 is a median statistic; worst-anchor behavior does not enter verdicts (two out-of-band validation probes remain unattributed). **Historical-case control gaps**: C3 has no contemporaneous unrepaired-continuation control; R3's training stage and evaluation definitions differ—both are demoted to the motivation layer; controlled conclusions come only from §3. The four-arm experiment's endpoint numbers and verdicts are authoritative from the machine archive (the four-arm reading archive), recomputable from the released reading files.

L MACHINE-VERIFIED RECORDS (ANONYMIZED)

Every number in this paper is backed by a machine archive; the artifact (released at submission with an anonymized repository; de-anonymized upon acceptance) contains, among others:

Audio-gate objective check configuration (§4.2). Deployment-setting renders of all 15 PERSIST runs against their unrepaired baselines; 30 = 15 PERSIST runs × 2 root validation probes—a healthy anchor and a deep-under-loudness anchor; pass rule: the audio-side repair-vs-baseline level difference and the med11 endpoint difference share sign; the objective audio-check archive in the artifact. **A4 anchoring detail (§3.1).** Applied every 25 steps to a random 8/32 subset with the global RNG state saved and restored around it (an RNG fence).

Content	Artifact component
Four-arm per-checkpoint med11 / R(t) / t_{half}	four-arm reading archive
Four-arm treatments, thresholds, anchor parameters (pre-registered)	four-arm spec (gain g^*=+5.03 single-source; the 32-sample reference set; 8 samples / 25 steps; 3.0 dB margin)
A4 anchor-loss implementation (32-sample reference + frozen baseline + RNG fence)	anchor-loss code
C3 low-pass three-timepoint steady-state report	C3 steady-state report (long-term archive)
R3 structured-reverb design spec and closure	R3 design spec and closure log
pressure-gate hypothesis test (not-passed verdict)	gate adjudication document
seed-replication control: two trajectories and anchor-level intervals	the seed-variance archive; suite report
Matrix design, frozen criteria, wave schedule	comparison-suite spec, plan, progress ledger; the B1 measurement archive
Matrix R(t) (single source of readings)	matrix readings (JSON)
Compensation-path localization	localization readings (JSON)
Audio gate (objective check + record)	audio-gate archive (JSON); listening record

Table 4: Machine-archive components backing the paper's numbers (anonymized paths).

M FOUR-ARM PROTOCOL ELEMENTS (PORTABLE CHECKLIST)

Protocol elements (extracted from the four-arm experiment; portable piece by piece). (1) four arms from one start (checkpoint-fingerprint whitelist); (2) treatment matrix = baseline-no-repair / repair-only / repair+continuation / repair+anchor+continuation; (3) primary readout $R(t)$ (Eq. 1)—deployment-setting scale pinned + contemporaneous paired denominator (dual-scale reporting recommended, §3.2); (4) t_{half} + dense-early/sparse-late checkpoints; (5) pre-registered tri-state endpoint verdict (thresholds frozen before the run); (6) noise floor = seed-variance band (true seed-replication primary; cross-segment proxies as scale references only; anchor-level bootstrap fallback); (7) leak-proof reference sets (anchor references training-side; the verdict set never enters the training graph); (8) side-effect panel; (9) single-anchor readings barred from verdicts.

Estimator notes (§6.1). (i) The hinge form of q_j is the general porting template (a deadband around tolerance), whereas the verdicts of this paper use the *signed* per-probe median ($med11$) with no deadband—a verbatim port computes a tolerance-suppressed quantity and should state the mapping explicitly; (ii) the formalism admits exactly the two repair classes P and Ψ, so training-objective interventions (arm B2) fall outside it and are scored as a separate route.