PREFERENCE-OPTIMISED MUSIC MASTERING UNDER DEGRADATION CHAINS

Anonymous authors
Paper under double-blind review

ABSTRACT

Music reaches listeners through production, re-recording and distribution stages, and degradations accumulate along the way, several at once and in order. SonicMaster, the prior model for joint mastering and restoration, is trained mostly on single degradations and never on more than three, so it can succeed by correcting each artifact in isolation. We instead train on *degradation chains* of one to eight degradations (mean 4.8, against 1.6) from 24 types, most of them drawn from six production scenarios, plus pass-through pairs that teach the model when to leave well-mastered music unchanged. On these chains we train PRISM, an instruction-conditioned rectified flow from degraded to reference-master latents. Matching a reference alone does not ensure high production quality, so we refine it through a single offline round of preference optimization, ranking the model's own samples by predicted production quality (PQ) and mel-SSIM to the reference. On held-out chains, PRISM improves mel-SSIM by $+0.112$ versus $+0.002$ and better preserves already-mastered inputs (0.98 vs. 0.82). Across 574 real low-production-quality releases it raises predicted PQ from 5.36 to 5.87, improving 79.3% of tracks, against $+0.09$ for SonicMaster and $+0.32$ for the same architecture trained further on its original corpus. Ablations show that a model trained on the prior corpus barely improves on our chains, and that our advantage holds at every chain depth tested (one to five)[1].

1 INTRODUCTION

Music post-production modifies an audio through a sequential chain of audio effects (Zölzer et al., 2002; Katz & Katz, 2003). Audio quality may deteriorate throughout this process; for example, a home-mastered track can be over-limited, re-recorded on a mobile device in a reverberant environment, and subsequently uploaded at a low bitrate. Unlike deliberate production choices, this chain is accidental: it is not recorded, and no clean reference is available. We use *mastering* for two parts of the task. *Corrective* processing removes accumulated degradations such as noise or hum; noise reduction often starts a mastering session (Katz & Katz, 2003). *Finishing* produces a technically sound master. We exclude creative decisions such as genre-specific loudness targets or stylistic colouration. Image and speech restoration also model degradation as a chain of stages(Zhang et al., 2021; Wang et al., 2021; Serrà et al., 2022; Zhang et al., 2024a). Order matters. For example, if a track is limited and then lossy-encoded, decoding can produce inter-sample peaks above the limit, which is why mastering uses a true-peak ceiling (ITU-R, 2023; Audio Engineering Society, 2015). Each stage also limits what later stages can recover: once a codec discards high frequencies, no equaliser can restore them. Furthermore, these degradations often co-occur depending on the production context, resulting in poorly produced audio that exhibits multiple simultaneous degradations.

Learning-based methods address individual degradations such as clipping (Záviška et al., 2020), band-limiting (Liu et al., 2024), lossy-codec artifacts (Li & Luo, 2025), and distortion (Imort et al., 2022; Rice et al., 2023), which mastering engineers typically correct before finalization. SonicMaster (Melechovsky et al., 2025) simultaneously addresses restoration and mastering, whereas text-driven systems select effect parameters based on descriptive input (Chu et al., 2025; Doh et al., 2025). Matched degraded and reference versions of the same music cannot be collected at scale, so mastering models, including ours, train on synthetic pairs. The degradation pipeline therefore

[1] Demo, Code

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.

defines the task. Existing music pipelines generate random chains of independently parameterized effects (Rice et al., 2023; Melechovsky et al., 2025). The SonicMaster corpus assigns each clean clip four single, two double, and one triple degradation, so no example has more than three. A model can do well on such data by correcting each artifact on its own. We find that this does not transfer to our chains, at any depth (Section 4.2). Another limitation is that matching a reference master does not guarantee high production quality. Reference-free quality predictors (Tjandra et al., 2025) can assess production quality, but optimizing their scores may satisfy the predictor rather than the listener (Rafailov et al., 2024a). Preference optimization (Rafailov et al., 2023; Wallace et al., 2024) needs only a ranking of candidate outputs, so the score does not have to be differentiable. We rank the model's own outputs by production quality and reference fidelity together, so that gaining on one at the cost of the other is not rewarded.

We present PRISM, an instruction-conditioned model that corrects compound degradations and produces a finished master without a separate repair stage. On held-out chains, it improves mel-SSIM by $+0.112$, compared with $+0.002$ for SonicMaster, while preserving already-mastered inputs at 0.978 versus 0.821. Across 574 real releases, predicted Audiobox production quality (Tjandra et al., 2025) increases by $+0.509$, with 79.3% of tracks improving. In a blind three-way listening study, ten listeners with mixing or mastering experience selected PRISM in 70.7% of judgements across fifteen tracks. On the ten held-out synthetic chains, it received 85% of votes, compared with 8% for SonicMaster and 7% for AudioSR, and was the most-selected system on every item (Section 4.5). Our contributions are threefold: **(1)** a synthesis pipeline with 24 degradation types, including eight absent from SonicMaster, arranged into chains of up to eight degradations through six production scenarios and signal-path ordering, with pass-through pairs teaching preservation; **(2)** mastering preference optimisation in one offline round using fixed candidate pairs and a preferred-sample likelihood anchor (Pang et al., 2024); and **(3)** controlled comparisons using the same architecture that demonstrate gains at every tested chain depth, including single degradations, alongside instruction-based control over whether the model intervenes.

2 RELATED WORK

Restoration and mastering. Music restoration addresses declipping (Záviška et al., 2020), bandwidth extension (Lee & Han, 2021; Mandel et al., 2023; Moliner et al., 2024a; Liu et al., 2024), codec restoration (Li & Luo, 2025), historical recordings (Moliner & Välimäki, 2022; Moliner et al., 2023; 2024b; Lemercier et al., 2024), and effect removal (Imort et al., 2022; Rice et al., 2023; Kandpal et al., 2022), alongside speech restoration (Liu et al., 2021; Koizumi et al., 2023; Richter et al., 2023). Automatic mixing and mastering uses parameter estimation (De Man et al., 2017; Mimilakis et al., 2016), differentiable effects (Engel et al., 2020; Ramírez et al., 2021; Steinmetz et al., 2021; Martínez-Ramírez et al., 2022), reference-based style transfer (Steinmetz et al., 2022; Koo et al., 2023; Vanka et al., 2024; Steinmetz et al., 2024), graph search (Lee et al., 2024), and mastering-specific models (Koo et al., 2022; 2025). Parameter prediction is interpretable but limited by the available effects, while restoration alone does not target a finished master. SonicMaster (Melechovsky et al., 2025) first combined these goals; PRISM extends them to deeper chains.

Text control and degradation pipelines. Language guides effect-parameter optimisation (Chu et al., 2025), parameter prediction (Doh et al., 2025), and effects-chain generation (Doh et al., 2026), supported by conditioning mechanisms from instruction-guided editing (Wang et al., 2023; Han et al., 2023; Zhang et al., 2024b; Manor & Michaeli, 2024) and text-to-audio generation (Liu et al., 2023; Evans et al., 2025; Hung et al., 2026). These approaches realise a described sound; we correct an unobserved degradation chain. We exclude CLAP text–audio similarity (Wu et al., 2023), directly optimised by Text2FX, but retain audio-only CLAP-FAD. For degradation synthesis, speech corpora combine independently sampled noise, reverberation, and codecs (Valentini-Botinhao et al., 2016; Maciejewski et al., 2020; Reddy et al., 2020), while universal enhancement uses long randomised chains (Serrà et al., 2022; Zhang et al., 2024a). In image super-resolution, BSRGAN shuffles degradation stages (Zhang et al., 2021), and Real-ESRGAN repeats a fixed-order pipeline (Wang et al., 2021). We extend music pipelines with deeper, production-ordered chains, additive and time-varying degradations, and pass-through pairs (Section 3.2).

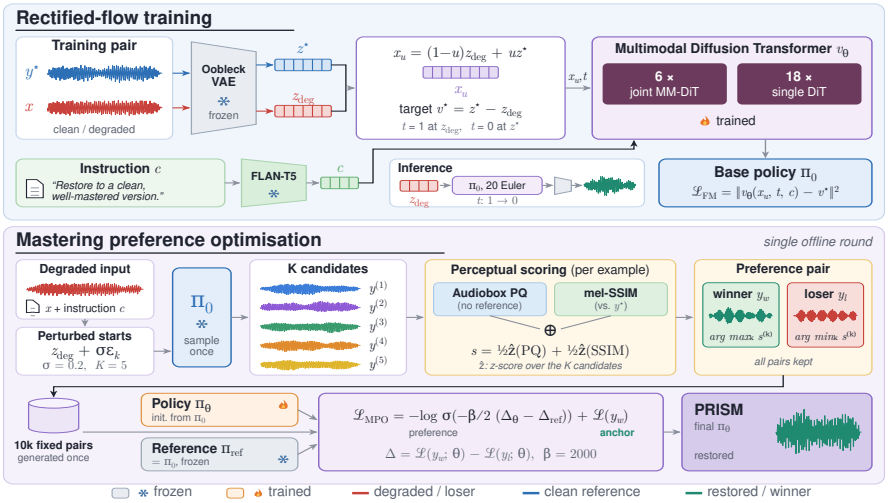


Figure 1: Overview of PRISM. (a) Rectified flow from z_{deg} to $z^\star$ in a frozen autoencoder latent space, conditioned on a frozen FLAN-T5 instruction embedding, giving π_0. (b) Mastering preference optimisation: K candidates from perturbed starts are ranked by production quality and mel-SSIM to the reference; best and worst form a pair, drawn once from π_0, the frozen reference.

Preference optimisation. Direct preference optimisation learns from pairwise preferences without a separate reward model (Rafailov et al., 2023), with extensions to diffusion (Wallace et al., 2024), text-to-audio (Majumder et al., 2024; Liao et al., 2024; Hung et al., 2026), and music generation (Cideron et al., 2024; Novack et al., 2024). Our automatically generated preferences combine no-reference production quality (Tjandra et al., 2025) with reference fidelity. To our knowledge, this is the first application to music mastering. Learned scorers are susceptible to over-optimisation, including in offline alignment (Rafailov et al., 2024a;b). Iterative on-policy methods refresh preference pairs but repeatedly optimise against the same scorer (Guo et al., 2024; Pang et al., 2024).

3 METHOD

3.1 OVERVIEW AND PROBLEM SETTING

Let $y^\star \in \mathbb{R}^{C\times T_s}$ denote an excerpt of a reference master with C channels and T_s samples at sampling rate f_s. A degradation chain is an ordered composition $D = f_T \circ \cdots \circ f_1$ of T parametric operators, each drawn from a bank $\mathcal{F}$ of degradation types. Given the degraded signal $x = D(y^\star)$ and a natural-language instruction c, we seek a mastering model $\hat{y} = M(x, c)$ whose output approximates $y^\star$ (the corrective component) and has high production quality (the finishing component); we score the two separately (Section 4.1). The problem differs from conditional generation in three ways. First, D is unobserved and compound: several degradations are applied in sequence and none is identified at test time. Second, D may be the identity ($T = 0$), so the model must decide whether to intervene before deciding how. Third, $y^\star$ exists only when D is synthesised. Instead of estimating D, we learn a transport from the degraded signal to the reference master in the latent space of a frozen autoencoder $\mathcal{E}$, with $z_{\text{deg}} = \mathcal{E}(x)$ and $z^\star = \mathcal{E}(y^\star)$, both in $\mathbb{R}^{L\times d}$.

3.2 COMPOUND DEGRADATION SYNTHESIS

Our pipeline composes 24 degradation types into ordered chains of one to eight types (mean 4.8). The bank covers 15 *mix and mastering* flaws, 5 *re-recording* degradations, and 4 *transfer and distribution* artefacts (Figure 4, Table 6). Eight types are absent from SonicMaster: background noise, mains hum, automatic gain control, lossy-codec artefacts, dropouts, wow and flutter, harmonic saturation, and over-limiting. Our band-limiting cutoff also varies from 4 to 16 kHz, rather than remaining fixed at 4 kHz. Parameters follow direct, inverse, or severity-independent mappings from a

shared latent $s \sim \mathrm{Beta}(2, 2)$ (Appendix B); sampling s independently per type shows that severity coupling does not explain the gains (Section 4.2).

Of the degraded examples, 70% follow six production scenarios: phone recording at a gig, laptop recording across a room, cassette transfer, low-bitrate internet rip, amateur master, and call capture. Each specifies four to eight types, their order, and optional restrictions on parameter ranges (Table 7). Scenario order can cross stage boundaries: cassette transfer places wow and flutter before hiss and hum; internet rip places codec band-limiting before re-compression and limiting; and call capture places far-end band-limiting before device coloration. These chains are unavailable under SonicMaster's sampling rule. The remaining 30% randomly combine one to five types in mix and mastering $\rightarrow$ re-recording $\rightarrow$ transfer and distribution order. Quality checks reject negligible changes or destroyed content (Appendix D). Pass-through pairs constitute 15% of all examples, teaching preservation by using the reference master as input; SonicMaster contains none.

Instructions come from a fixed bank generated once offline by a language model, indexed by scenario or type, severity, and register. Each example receives an engineer-style *directive* and a listener-style *freeform* description. Complete scenario chains receive scenario-level instructions; others name one or two applied degradations. Examples without matching instructions are excluded.

3.3 Model and training

The mastering model v_θ is a MM-DiT (Shin et al., 2025) conditioned on frozen FLAN-T5 instruction embeddings (Chung et al., 2024). Its inherited audio-conditioning branch is zeroed at evaluation (Appendix L). Rectified flow transports degraded latents to reference latents along

$$x_u = (1-u)z_{\mathrm{deg}} + uz^\star, \qquad v^\star = z^\star - z_{\mathrm{deg}}, \tag{1}$$

by minimising $\mathbb{E}\|v_\theta(x_u, 1-u, c) - v^\star\|_2^2$. Starting from z_{deg} anchors generation to the input track. Pass-through pairs have the exact target $v^\star = 0$, which the learned field approximates. We sample u equally from $\mathcal{U}(0,1)$ and $p(u) \propto 1-u$, emphasising the degraded end. Every example includes an instruction and inference uses N Euler steps from z_{deg}, followed by decoding. Paired evaluation uses decoded reference latents to exclude autoencoder reconstruction error (Section 4.1).

Mastering preference optimisation. MPO refines the supervised model in one offline round using ranked outputs from synthetic training chains. Since integration is deterministic, we generate diverse candidates from perturbed starts: $z^{(k)} = \Phi(z_{\mathrm{deg}} + \tau\epsilon_k, c; \theta_0)$, where $\epsilon_k \sim \mathcal{N}(0, I)$, Φ integrates the learned field, $K = 5$, and $\tau = 0.2$ (Appendix M). We rank decoded candidates $y^{(k)}$ by combining reference-free Audiobox production quality (PQ) with reference mel-SSIM:

$$s^{(k)} = \tfrac{1}{2}\hat{z}\Big(\mathrm{PQ}(y^{(k)})\Big) + \tfrac{1}{2}\hat{z}\Big(\mathrm{SSIM}(y^{(k)}, y^\star)\Big), \tag{2}$$

where $\hat{z}$ standardises each metric within the candidate set. The two scores balance production quality and fidelity: PQ alone does not ensure content preservation, while fidelity alone does not ensure perceptual improvement. The highest- and lowest-scoring candidates form the winner–loser pair (z^w, z^l), and all pairs are retained. Appendix M reports raw score gaps, since standardised gaps reflect within-set variation rather than absolute differences.

Objective and offline training. Both candidates share z_{deg}, c, and a sampled u. For candidate z, define $x_u = (1-u)z_{\mathrm{deg}} + uz$, $v^\star = z - z_{\mathrm{deg}}$, and $\mathcal{L}(z;\theta) = \|v_\theta(x_u, 1-u, c) - v^\star\|_2^2$. With $\Delta_\theta = \mathcal{L}(z^w;\theta) - \mathcal{L}(z^l;\theta)$ and Δ_{ref} its frozen-reference counterpart, we minimise

$$\mathcal{L}_{\mathrm{MPO}} = -\log\sigma\Big(-\tfrac{\beta}{2}(\Delta_\theta - \Delta_{\mathrm{ref}})\Big) + \mathcal{L}(z^w;\theta), \tag{3}$$

where σ is the logistic function and $\beta = 2000$. The Diffusion-DPO term (Wallace et al., 2024) rewards relative improvement, while the preferred-sample anchor (Pang et al., 2024) discourages improvement solely by worsening the loser. We generate and rank candidates once for $10{,}000$ training examples under the supervised policy π_0, which also initialises the policy and serves as the frozen reference, $\pi_{\mathrm{ref}} = \pi_0$. Training uses fixed pairs for $2{,}500$ steps with audio conditioning zeroed (Appendix M). This offline setup and anchoring limit, but do not eliminate, scorer overoptimisation (Rafailov et al., 2024b). We report the final checkpoint without selection on WILD. Although K, τ, and β were not chosen on WILD, the learning-rate schedule followed a gentler pilot also evaluated on it (Appendix N).

4 EXPERIMENTS AND RESULTS

Data, splits and baselines. We train on the union of two paired corpora: the synthesis of Section 3.2 applied to JamendoMaxCaps (Roy et al., 2025),and the released SONICMASTER corpus (Melechovsky et al., 2025) at its published split (Appendix G). Reference masters score at least 7.5 on Audiobox Aesthetics production quality, a proxy rather than a claim of professional mastering; the remaining pool filters are in Appendix D. Our corpus is split artist-disjointly, and released-corpus rows whose source track falls in our validation or test split are dropped from training. The released corpus carries no artist identifiers, so artist overlap across the two corpora cannot be excluded. Table 8 lists the five evaluation sets. Four are paired and measure correction and preservation. The fifth, WILD, holds 574 real low-production-quality releases and is the only test of whether what is learned from synthetic degradations transfers to real ones; its tracks were selected at production quality at most 5.5 on a centred 60 s excerpt and scored on a different 30 s one, so some score above 5.5, and a track is included only if its artist is absent from our reference pool, since same-artist tracks often share a room and a mastering chain. WILD is not genre-filtered (Appendix I). We compare against a static DSP mastering preset, Text2FX (Chu et al., 2025), LLM2Fx (Doh et al., 2025), AudioSR (Liu et al., 2024) and the released SonicMaster checkpoint, covering parametric mastering, text-driven parameter estimation and generative restoration. The DSP preset and AudioSR take no instruction; all baseline settings are in Appendix E.

4.1 IMPLEMENTATION DETAILS AND METRICS

Architecture, schedule and inference settings are in Appendix E, metric settings in Appendix F. Paired metrics compare each output with its reference decoded from the same latent, so autoencoder error is excluded and an unchanged pass-through input scores mel-SSIM 1. We report three signal measures, mel-SSIM, log-spectral distance (LSD) and mel-spectral KL, and a paired KL between PANNs Cnn14 AudioSet posteriors (Kong et al., 2020). Distributional measures catch shifts shared across outputs, such as a common compression signature, that paired measures average out: Fréchet distance (Kilgour et al., 2018) over PANNs (FD) and CLAPcembeddings (Wu et al., 2023) (CLAP-FAD), since FAD depends on the embedding (Gui et al., 2024; Tailleur et al., 2024). Fréchet estimates are biased at these sample sizes, so we read only rankings. WILD has no reference, so we report Audiobox PQ (Tjandra et al., 2025) and the fraction of tracks improved, with BS.1770 loudness, true peak and PLR (ITU-R, 2023) in Section 4.2.

4.2 MAIN RESULTS

Mastering quality and correction. Table 1 gives both settings. On WILD, PRISM adds $+0.509$ production quality and improves 79.3% of tracks, against $+0.086$ for the released SonicMaster checkpoint. Part of that gap is continued training: the same architecture trained further on the SonicMaster corpus alone gains $+0.323$ (Section 4.2), leaving $+0.120$ to the new corpus and $+0.066$ to MPO. AudioSR gains a similar $+0.084$ by extending high frequencies. The text-driven baselines do worse than doing nothing: Text2FX loses 0.389 PQ and drops pass-through fidelity to 0.492, because fitting an effects chain to a text embedding alters already-mastered and degraded audio alike, and LLM2Fx leaves mastered input untouched but its predicted EQ and reverb degrade every paired measure on degraded input. On synthetic chains the gap widens: PRISM improves mel-SSIM by $+0.112$ against SonicMaster's $+0.002$, and preserves already-mastered input at 0.978 against 0.821. Appendix H breaks this down by scenario.

Fidelity. Table 2 separates the systems more sharply than mel-SSIM does. PRISM is the only system that improves on the degraded input under all five measures, roughly halving LSD and cutting PANNs FD from 11.234 to 2.519. No baseline improves all three paired measures, and the best baseline FD (9.177, AudioSR) and CLAP-FAD (0.129, SonicMaster) rank well behind PRISM. Appendix J covers where the measures disagree.

Contribution of the corpus: Table 9 (Appendix G) varies only the training data. Training on the SONICMASTER corpus alone gains just $+0.001$ mel-SSIM on COMPOUND validation, despite 11,000 steps against 4,500 for the COMPOUND model, so shallow, independently sampled combinations do not transfer to our chains; adding our corpus raises the mel-SSIM gain from $+0.001$ to

system	WILD		COMPOUND val		COMPOUND test	
	PQ	improved	mel-SSIM	pass-through	mel-SSIM	pass-through
Unprocessed input	5.364	–	0.5955	1.0000	0.5833	1.0000
DSP preset	$5.369_{+0.005}$	46.7%	$0.5937_{-0.0017}$	0.9890	$0.5801_{-0.0031}$	0.9865
Text2FX (Chu et al., 2025)	$4.975_{-0.389}$	29.4%	$0.3919_{-0.2036}$	0.4923	$0.4007_{-0.1826}$	0.5035
LLM2Fx (Doh et al., 2026)	$5.297_{-0.067}$	32.1%	$0.5888_{-0.0067}$	1.0000	$0.5753_{-0.0080}$	1.0000
AudioSR (Liu et al., 2024)	$5.448_{+0.084}$	64.1%	$0.6142_{+0.0187}$	0.9593	$0.5909_{+0.0077}$	0.9478
SonicMaster (Melechovsky et al., 2025)	$5.450_{+0.086}$	52.8%	$0.6005_{+0.0051}$	0.8214	$0.5854_{+0.0021}$	0.8172
PRISM w/o comp. corpus	$5.687_{+0.323}$	68.1%	$0.5967_{+0.0013}$	0.8015	$0.5831_{-0.0002}$	0.8034
PRISM w/o MPO	$5.807_{+0.443}$	76.7%	$0.7177_{+0.1222}$	0.9843	$0.6935_{+0.1102}$	0.9835
PRISM	$\mathbf{5.873}_{+0.509}$	**79.3%**	$\mathbf{0.7183}_{+0.1229}$	0.9779	$\mathbf{0.6949}_{+0.1116}$	0.9770

Table 1: Main results. Subscripts: change from the unprocessed input. WILD: Audiobox PQ on real releases, *improved*: fraction of tracks whose PQ rises. Pass-through: mel-SSIM on already-mastered input. Bold marks the best system, pass-through columns excluded; set sizes in Table 8.

system	paired			distributional	
	LSD ↓	mel-KL ↓	PANNs KL ↓	FD ↓	CLAP-FAD ↓
Degraded input	24.171	0.5219	0.3903	11.234	≈0.157
DSP preset	$23.542_{-0.629}$	$0.5394_{+0.0176}$	$0.3911_{+0.0008}$	$11.240_{+0.006}$	0.1577
Text2FX (Chu et al., 2025)	$25.301_{+1.130}$	$1.3107_{+0.7888}$	$0.9911_{+0.6008}$	$31.026_{+19.792}$	0.4196
LLM2Fx (Doh et al., 2025)	$24.250_{+0.079}$	$0.6149_{+0.0930}$	$0.4114_{+0.0211}$	$12.179_{+0.945}$	0.1688
AudioSR (Liu et al., 2024)	$18.774_{-5.397}$	$0.5257_{+0.0039}$	$0.3712_{-0.0191}$	$9.177_{-2.057}$	0.1748
SonicMaster (Melechovsky et al., 2025)	$25.000_{+0.829}$	$0.4794_{-0.0425}$	$0.3525_{-0.0378}$	$9.428_{-1.806}$	0.1290
PRISM w/o comp. corpus	$25.977_{+1.806}$	$0.4754_{-0.0465}$	$0.3585_{-0.0318}$	$9.570_{-1.664}$	0.1352
PRISM w/o MPO	$13.390_{-10.781}$	$0.1423_{-0.3796}$	$0.1620_{-0.2283}$	$2.586_{-8.648}$	0.0326
PRISM	$\mathbf{13.213}_{-10.958}$	$\mathbf{0.1383}_{-0.3836}$	$\mathbf{0.1597}_{-0.2306}$	$\mathbf{2.519}_{-8.715}$	0.0304

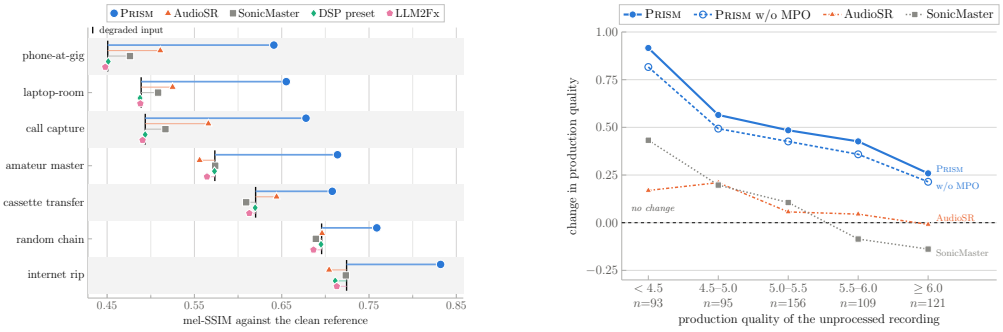
Table 2: Fidelity on COMPOUND test. Subscripts: change from the degraded input. Column groups follow Section 4.1; FD uses PANNs Cnn14 and CLAP-FAD uses CLAP audio embeddings. CLAP-FAD scores a random 10 s window per clip, so it varies across runs, carries no subscript.

$+0.122$ and WILD PQ by a further $+0.120$.. The SONICMASTER-only model raises WILD PQ but does not improve correction on our chains. Pass-through fidelity rises from 0.802 to 0.984 with our corpus, which, unlike the SONICMASTER corpus, contains pass-through pairs; we do not isolate these pairs from the rest of the corpus. Drawing severity once per degradation type rather than once per chain changes results negligibly (-0.0025 mel-SSIM, -0.0018 pass-through, $+0.015$ PQ).

The advantage does not grow with depth. Scenario chains are both deeper and more severe than random ones, so we compare the SONICMASTER-only and union models of Table 9 on the 738 random-chain examples of COMPOUND validation and test, where depth is drawn independently. The advantage stays between $+0.072$ and $+0.078$ mel-SSIM at every depth from one to five, and between $+0.068$ and $+0.081$ at depths one to four when restricted to degradation types with a SonicMaster counterpart (depth five has only six such examples). Gains rise with depth for both models only because deeper chains degrade more: with input mel-SSIM held fixed, the gain per added degradation is slightly negative (-0.0053, 95% CI $[-0.0081, -0.0024]$, union; -0.0052 $[-0.0092, -0.0014]$, SONICMASTER-only). The corpus therefore helps at every depth, single degradations included.

Quality–preservation trade-off: MPO raises WILD production quality from 5.807 to 5.873 and the share of improved tracks from 76.7% to 79.3%, while correction is unchanged or slightly better ($+0.0006$ mel-SSIM on COMPOUND validation, $+0.0014$ on test). The cost is preservation: pass-through fidelity falls from 0.984 to 0.978 (Table 1). Across the 25 checkpoints of the run, validation mel-SSIM stays within 0.0012 while pass-through fidelity declines steadily, so the checkpoints trace a trade-off between quality and preservation rather than trading correction for quality (Appendix N).

What preference optimisation changes. MPO ranks candidates with the predictor we also report, so its gain could belong to the predictor rather than the listener. Measures outside the ranking argue against that. On COMPOUND test, under an embedding MPO never sees, PANNs FD falls from 2.586 to 2.519 and paired PANNs KL from 0.162 to 0.160; LSD and mel-spectral KL also improve slightly. All four are significant under a paired bootstrap over test rows (Appendix N), so MPO improves every deterministic fidelity measure while ranking only one. The gain tracks input quality,



(a) Mel-SSIM by synthetic scenario; black ticks mark degraded inputs.

(b) Mean production-quality gain by input-quality bin ($n = 93, 95, 156, 109, 121$).

Figure 2: Gain by input quality. (a) COMPOUND validation, scenarios and random chains. (b) WILD, binned by the Audiobox PQ of the evaluated excerpt.

		degraded input			mastered input
model	corpus	gain	sensitivity	share of gain	sensitivity
Text2FX	COMPOUND	−0.2036	+0.0009	–	−0.0550
LLM2Fx	COMPOUND	−0.0067	+0.0020	–	+0.0341
SonicMaster	COMPOUND	+0.0051	+0.0010	–	−0.0275
SONICMASTER-only	COMPOUND	+0.0013	+0.0022	–	−0.0350
PRISM w/o MPO	COMPOUND	+0.1222	+0.0024	2.0%	+0.0264
PRISM	COMPOUND	+0.1229	+0.0021	1.7%	+0.0324
SonicMaster	SONICMASTER	+0.1212	+0.0699	57.7%	–
PRISM	SONICMASTER	+0.1251	+0.0488	39.0%	–

Table 3: Instruction use. *Gain*: mel-SSIM improvement over the degraded input. *Sensitivity*: gain lost when the instruction is swapped for another example's; on already-mastered input, the change under a mismatched instruction. *Share*: sensitivity as a fraction of gain, omitted where the gain is negative or negligible. SONICMASTER-only is PRISM w/o comp. corpus in Table 1.

falling from $+0.100$ on the lowest-quality WILD inputs to $+0.044$ on the highest, and from $+0.061$ on degraded COMPOUND validation input to $+0.013$ on already-mastered input (Table 11). That is consistent with correction, though poor inputs leave more headroom for any predictor-driven change. The cost is preservation ($0.984 \rightarrow 0.978$); the listening study cannot settle this, since PRISM w/o MPO was not a condition.

Quality gains beyond loudness. Matching every output to its input's integrated loudness (BS.1770) before scoring leaves PRISM $+0.483$ [$+0.432$, $+0.534$] of its $+0.509$ gain, with 80.7% of tracks still improving. PRISM raises loudness by only 1.0 LU and lifts the peak-to-loudness ratio from 10.5 to 11.2 dB, the opposite of the heavier limiting a loudness exploit needs, and per track its gain is anticorrelated with its loudness change ($\rho = -0.27$, $n = 574$). SonicMaster instead adds 5.0 LU and drops the ratio to 8.4 dB. Brightness and bandwidth do not explain it either (Appendix J).

4.3 DOES THE MODEL KNOW WHAT IS WRONG?

A model that repairs a master must find what is wrong with it, yet no term in its loss asks for this. We test whether that knowledge is in the hidden states, and whether it depends on the training corpus.

Protocol. We freeze each model and give it an empty instruction. On each of the 24 blocks (double-stream D0–D5, single-stream S0–S17) we time-average the audio-token states and fit a linear probe (Alain & Bengio, 2016), a multi-label logistic regression over SONICMASTER's five degradation groups (EQ, dynamics, reverb, amplitude, stereo). Using its groups rather than our 24 types puts both models on the same labels: 19 of our types map to a group, and the 5 fall outside all five and are labelled negative. Each probe trains on 10k clips from its model's corpus and is

Table 4: Probing block S13 at t=0: AP per group, grey rows are base rates, cAP is chance-corrected AP averaged over groups. S13 is SONICMASTER's best block on chains; every other row is within 0.010 mean AP of its own best block (Appendix P.3).

Model (test corpus)	EQ	Dyn.	Rev.	Amp.	Stereo	mAP	cAP
Chain test set (1,550 clips)							
chance (base rate)	.770	.588	.137	.259	.113	.373	0
Ours (own corpus)	.987	.971	.978	.843	.848	.925	.89
SONICMASTER (other corpus)	.908	.644	.764	.551	.130	.599	.38
SONICMASTER test set (7,000 clips)							
chance (base rate)	.584	.235	.382	.286	.149	.327	0
SONICMASTER (own corpus)	.961	.963	.997	.995	.993	.982	.97
Ours (other corpus)	.814	.579	.956	.594	.874	.764	.64

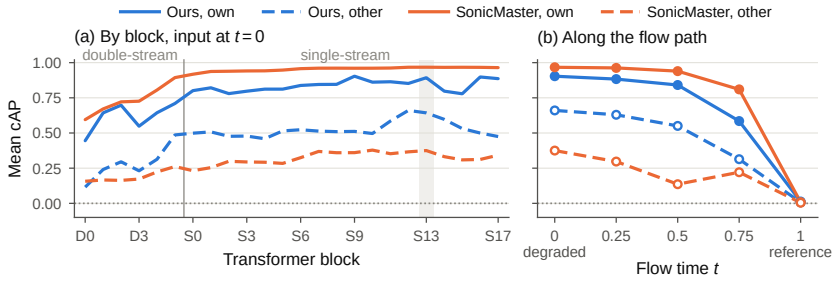


Figure 3: Mean cAP over the five groups (0 = chance); solid: own corpus, dashed: other corpus. (a) By block at t=0; grey band: S13 (Table 4). (b) Along the flow path, each pair at its best block at t=0 (own/other: S9/S12 for ours, S15/S13 for SONICMASTER); at t=1 the input is the clean.

tested on that corpus and on the other. Base rates π differ between the two test sets, so we report chance-corrected AP, $\mathrm{cAP} = (\mathrm{AP} - \pi)/(1 - \pi)$ (Appendix P).

Both models encode the damage, but only one code transfers. With no instruction, a single hidden layer says which groups are present: on its own corpus our model reaches mean AP 0.925 and SONICMASTER 0.982 (Table 4). The asymmetry is in transfer. Our probes still read the SONICMASTER test set, all five groups well above chance (cAP 0.64), whereas SONICMASTER's probes reach cAP 0.38 on our chains, with stereo at chance. What transfers is also built deeper: after the first block each probe reads its own corpus (cAP 0.45 and 0.59) but barely the other (0.12 and 0.16), and cross-corpus readability peaks at S12 or S13 (Figure 3a). The training corpus thus shapes which faults can be read from the hidden states, not only which faults get fixed.

Control and scope. Probing along the flow path, where the input is $(1-t)\, z_{\mathrm{deg}} + t\, z^{\star}$ as in Eq. 1, every probe falls to within 0.04 AP of its base rate at the reference end (Figure 3b). The music is fully present there, so what the probes read at the degraded end comes almost entirely from the degradation. Two limits apply: a probe shows that information is present, not that the model uses it, and each comparison changes the weights and the probe's training data together, both of which follow from the training corpus, the variable we study (Appendix P.4).

4.4 WHAT THE INSTRUCTION CONTROLS

Whether to intervene. On already-mastered input nothing in the signal says whether to act, so the decision must come from the instruction. Under the correct instruction PRISM leaves such input nearly unchanged; a mismatched one more than doubles the departure (1 − mel-SSIM from 0.022 to 0.055), a gap of +0.032 that MPO widens from +0.026 (Table 3). SonicMaster (−0.028), Text2FX (−0.055) and the SONICMASTER-only model trained without pass-through pairs (−0.035) all alter already-mastered audio *less* under the wrong instruction, so none reads it as a cue to hold back; LLM2Fx's positive gap arises only because it processes nothing under the correct one. Even mismatched, PRISM retains a pass-through mel-SSIM of 0.946, above SonicMaster's 0.849 mismatched and 0.821 correct.

		chosen as best (% of judgements)			
subset	judgements	AudioSR	SonicMaster	PRISM	PRISM modal
COMPOUND test	$10 \times 10 = 100$	7.0 [0, 18]	8.0 [0, 21]	**85.0** [68, 97]	10/10
WILD	$5 \times 10 = 50$	16.0 [0, 40]	42.0 [18, 70]	42.0 [16, 68]	2/5
all items	$15 \times 10 = 150$	10.0 [1, 23]	19.3 [7, 35]	**70.7** [53, 87]	12/15

Table 5: Ten listeners with mixing or mastering experience each judged the same 15 trials and picked the version with the best production and mastering quality, so chance is 33.3%. Brackets are 95% percentile intervals from a bootstrap that resamples listeners and items together. PRISM *modal*: items on which PRISM received more votes than either other system. Bold marks a rate whose interval excludes chance. Per-item counts are in Table 12.

Which degradation to correct. On degraded input, the weight of the instruction appears to depend on the corpus. In the corpus where every degradation is named, swapping in another example's instruction removes 39.0% of PRISM's gain and 57.7% of SonicMaster's (a text-blind model would lose nothing); on our chains it removes only 1.7%, and even Text2FX, whose only mechanism is text, changes by just $+0.0009$. Either the audio itself reveals what to correct when several degradations co-occur, or our instructions, which name only one or two of them, carry less information by construction. The instruction thus governs whether the model acts but not where: naming one of several co-occurring degradations does not make the model correct it preferentially (Appendix K).

4.5 LISTENING STUDY

Every result so far comes from a learned predictor or a signal measure, and MPO ranks its own candidates with one of the predictors we then report. Ten listeners with mixing or mastering experience each judged the same 15 trials. A trial plays one input processed by PRISM, SonicMaster and AudioSR as three players marked A, B and C in an order drawn per trial, and asks which version has the best production and mastering quality. There is no reference and no "no preference" option, so chance is 33.3%. Ten inputs come from COMPOUND test and five from WILD, giving 150 judgements over headphones. Protocol, per-item votes and statistics are in Appendix O.

Listeners chose PRISM in 70.7% of judgements and all ten chose it above chance (sign test, $p = 0.001$); the intervals in Table 5 come from a bootstrap resampling listeners and items, since judgements are nested in both. On the ten synthetic chains the margin is large, 85.0% [68, 97] against 8.0%, with PRISM modal on every item. On the five WILD items PRISM and SonicMaster tie at 42.0% and the interval contains chance; five items cannot rank three systems, so the predicted WILD gain of $+0.509$ against $+0.086$ is not corroborated by listeners and we report it as a predictor score rather than a perceptual claim. Stimuli were not loudness-matched, which normally favours the louder system, but on WILD SonicMaster is 5.0 LU louder against PRISM's 1.0 LU (Section 4.2), so level does not explain the preference. Two conditions are absent: PRISM w/o MPO, so the study cannot show that preference optimisation is audible, and the unprocessed input, so it ranks the three systems against each other rather than against leaving a track alone.

5 CONCLUSION

A mastering model can only correct the damage its training data contains, and prior corpora contain at most three independent effects. Training on ordered chains of up to eight, with pass-through pairs and one offline round of preference refinement, closes a gap that further training on the old corpus does not: the prior corpus transfers almost nothing to our chains at any depth, and only the chain-trained model's hidden states encode faults that a probe can read across corpora. Listeners confirm the correction result on held-out chains. On real releases the improvement is so far supported only by a learned predictor, and the five real items in our listening test cannot separate the systems. Degradation-chain synthesis and preference refinement should carry over to mixing and to speech enhancement.

REPRODUCIBILITY STATEMENT

The problem, degradation synthesis, model and preference optimisation are specified in Section 3. All 24 degradation types, with parameter ranges and severity modes, are listed in Table 6, the six production scenarios with their order of application in Table 7, and the severity mapping in Appendix B. Corpus construction, quality control and WILD selection are in Appendix D; architecture, training schedules and the DSP baseline in Appendix E; metric settings in Appendix F; and the MPO pair construction, including the sampling seed and raw winner–loser gaps, in Appendix M. The instruction bank is generated once, offline, and is fixed, so the corpus is reproducible without network access. All data and models are public: JamendoMaxCaps (Roy et al., 2025), the released SonicMaster corpus and checkpoint (Melechovsky et al., 2025), the Stable Audio Open autoencoder (Evans et al., 2025), FLAN-T5 (Chung et al., 2024) and Audiobox Aesthetics (Tjandra et al., 2025). Code, the taxonomy configuration, the instruction bank, checkpoints and per-track outputs will be released upon publication.

ETHICS STATEMENT

This work uses publicly released, Creative Commons-licensed music from MTG-Jamendo and the released SonicMaster corpus, and involves no personal data. PRISM alters recorded music and makes no creative choices; because WILD includes genres in which rough production can be intentional (Appendix I), a higher predicted production quality is not necessarily an improvement from the artist's perspective, and the model should not be applied to an artist's work without their consent. Production-quality scores come from a learned predictor and should not be read as listener judgements.

LLM USAGE STATEMENT

A large language model (Claude) generated the fixed instruction bank once, offline (Section 3.2). The authors also used an AI assistant to help paraphrase sentences and carry out grammatical checks. All research ideas, experiments and conclusions are the authors', who take full responsibility for the content.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Audio Engineering Society. Recommendation for loudness of audio streaming and network file playback. Technical Report AES TD1004.1.15-10, Audio Engineering Society, 2015.

Annie Chu, Patrick O'Reilly, Julia Barnett, and Bryan Pardo. Text2fx: Harnessing clap embeddings for text-guided audio effects. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of machine learning research*, 25(70):1–53, 2024.

Geoffrey Cideron, Sertan Girgin, Mauro Verzetti, Damien Vincent, Matej Kastelic, Zalán Borsos, Brian McWilliams, Victor Ungureanu, Olivier Bachem, Olivier Pietquin, et al. Musicrl: Aligning music generation to human preferences. *arXiv preprint arXiv:2402.04229*, 2024.

Brecht De Man, Joshua Reiss, and Ryan Stables. Ten years of automatic mixing. In *3rd Workshop on Intelligent Music Production (WIMP 2017)*, 2017.

Seungheon Doh, Junghyun Koo, Marco A Martínez-Ramírez, Wei-Hsiang Liao, Juhan Nam, and Yuki Mitsufuji. Can large language models predict audio effects parameters from natural language? In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5. IEEE, 2025.

Seungheon Doh, Junghyun Tony Koo, Marco Martínez-Ramírez, Woosung Choi, WeiHsiang Liao, Qiyu Wu, Juhan Nam, and Yuki Mitsufuji. Llm2fx-tools: Tool calling for music post-production. In *International Conference on Learning Representations*, volume 2026, pp. 100979–100999, 2026.

Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting frechet audio distance for generative music evaluation. In *ICASSP 2024-2024 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 1331–1335. IEEE, 2024.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Bing Han, Junyu Dai, Weituo Hao, Xinyan He, Dong Guo, Jitong Chen, Yuxuan Wang, Yanmin Qian, and Xuchen Song. Instructme: An instruction guided music edit and remix framework with latent diffusion models. *arXiv preprint arXiv:2308.14360*, 2023.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pp. 2733–2743, 2019.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. In *International Conference on Learning Representations*, volume 2026, pp. 150793–150816, 2026.

Johannes Imort, Giorgio Fabbro, Marco A Martínez Ramírez, Stefan Uhlich, Yuichiro Koyama, and Yuki Mitsufuji. Distortion audio effects: Learning how to recover the clean signal. *arXiv preprint arXiv:2202.01664*, 2022.

ITU-R. Algorithms to measure audio programme loudness and true-peak audio level. Technical Report Recommendation ITU-R BS.1770-5, International Telecommunication Union, 2023.

Nikhil Kandpal, Oriol Nieto, and Zeyu Jin. Music enhancement via image translation and vocoding. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3124–3128. IEEE, 2022.

Bob Katz and Robert A Katz. *Mastering audio: the art and the science*. Butterworth-Heinemann, 2003.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *arXiv preprint arXiv:1812.08466*, 2018.

Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Yu Zhang, Wei Han, Ankur Bapna, and Michiel Bacchiani. Miipher: A robust speech restoration model integrating self-supervised speech and text representations. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5. IEEE, 2023.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

Junghyun Koo, Seungryeol Paik, and Kyogu Lee. End-to-end music remastering system using self-supervised and adversarial training. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4608–4612. IEEE, 2022.

Junghyun Koo, Marco A Martínez-Ramírez, Wei-Hsiang Liao, Stefan Uhlich, Kyogu Lee, and Yuki Mitsufuji. Music mixing style transfer: A contrastive learning approach to disentangle audio effects. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Junghyun Koo, Marco A Martinez-Ramirez, Wei-Hsiang Liao, Giorgio Fabbro, Michele Mancusi, and Yuki Mitsufuji. Ito-master: Inference-time optimization for audio effects modeling of music mastering processors. *arXiv preprint arXiv:2506.16889*, 2025.

Junhyeok Lee and Seungu Han. Nu-wave: A diffusion probabilistic model for neural audio upsampling. *arXiv preprint arXiv:2104.02321*, 2021.

Sungho Lee, Marco A Martínez-Ramírez, Wei-Hsiang Liao, Stefan Uhlich, Giorgio Fabbro, Kyogu Lee, and Yuki Mitsufuji. Searching for music mixing graphs: A pruning approach. *arXiv preprint arXiv:2406.01049*, 2024.

Jean-Marie Lemercier, Julius Richter, Simon Welker, Eloi Moliner, Vesa Välimäki, and Timo Gerkmann. Diffusion models for audio restoration. *arXiv preprint arXiv:2402.09821*, 2024.

Kai Li and Yi Luo. Apollo: Band-sequence modeling for high-quality audio restoration. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li. Baton: Aligning text-to-audio model with human preference feedback. *arXiv preprint arXiv:2402.00744*, 2024.

Haohe Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. Voicefixer: Toward general speech restoration with neural vocoder. *arXiv preprint arXiv:2109.13731*, 2021.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Haohe Liu, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D Plumbley. Audiosr: Versatile audio super-resolution at scale. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1076–1080. IEEE, 2024.

Matthew Maciejewski, Gordon Wichern, Emmett McQuinn, and Jonathan Le Roux. Whamr!: Noisy and reverberant single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 696–700. IEEE, 2020.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572, 2024.

Moshe Mandel, Or Tal, and Yossi Adi. Aero: Audio super resolution in the spectral domain. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using ddpm inversion. *arXiv preprint arXiv:2402.10009*, 2024.

Marco A Martínez-Ramírez, Wei-Hsiang Liao, Giorgio Fabbro, Stefan Uhlich, Chihiro Nagashima, and Yuki Mitsufuji. Automatic music mixing with deep learning and out-of-domain data. *arXiv preprint arXiv:2208.11428*, 2022.

Jan Melechovsky, Ambuj Mehrish, Abhinaba Roy, and Dorien Herremans. Sonicmaster: Towards controllable all-in-one music restoration and mastering. *arXiv preprint arXiv:2508.03448*, 2025.

Stylianos Ioannis Mimilakis, Konstantinos Drossos, Tuomas Virtanen, and Gerald Schuller. Deep neural networks for dynamic range compression in mastering applications. In *Audio Engineering Society Convention 140*, 2016.

Eloi Moliner and Vesa Välimäki. Behm-gan: Bandwidth extension of historical music using generative adversarial networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:943–956, 2022.

Eloi Moliner, Jaakko Lehtinen, and Vesa Välimäki. Solving audio inverse problems with a diffusion model. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Eloi Moliner, Filip Elvander, and Vesa Välimäki. Blind audio bandwidth extension: A diffusion-based zero-shot approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:5092–5105, 2024a.

Eloi Moliner, Maija Turunen, Filip Elvander, and Vesa Välimäki. A diffusion-based generative equalizer for music restoration. *arXiv preprint arXiv:2403.18636*, 2024b.

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J Bryan. Ditto: Diffusion inference-time t-optimization for music generation. *arXiv preprint arXiv:2401.12179*, 2024.

Richard Yuanzhe Pang, Weizhe Yuan, Kyunghyun Cho, He He, Sainbayar Sukhbaatar, and Jason Weston. Iterative reasoning preference optimization, 2024. *URL https://arxiv.org/abs/2404.19733*, 3, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37:126207–126242, 2024a.

Rafael Rafailov, Yaswanth Chittepu, Ryan Park, Harshit Sikchi, Joey Hejna, W Bradley Knox, Chelsea Finn, and Scott Niekum. Scaling laws for reward model overoptimization in direct alignment algorithms. *Advances in Neural Information Processing Systems*, 37:126207–126242, 2024b.

Marco A Martínez Ramírez, Oliver Wang, Paris Smaragdis, and Nicholas J Bryan. Differentiable signal processing with black-box audio effects. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 66–70. IEEE, 2021.

Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al. The interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. *arXiv preprint arXiv:2005.13981*, 2020.

Matthew Rice, Christian J Steinmetz, George Fazekas, and Joshua D Reiss. General purpose audio effect removal. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5. IEEE, 2023.

Julius Richter, Simon Welker, Jean-Marie Lemercier, Bunlong Lay, and Timo Gerkmann. Speech enhancement and dereverberation with diffusion-based generative models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2351–2364, 2023.

Abhinaba Roy, Renhang Liu, Tongyu Lu, and Dorien Herremans. Jamendomaxcaps: A large scale music-caption dataset with imputed metadata. In *2025 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2025.

Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065*, 2022.

Joonghyuk Shin, Alchan Hwang, Yujin Kim, Daneul Kim, and Jaesik Park. Exploring multimodal diffusion transformers for enhanced prompt-based image editing. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19492–19502. IEEE, 2025.

Christian J Steinmetz, Jordi Pons, Santiago Pascual, and Joan Serrà. Automatic multitrack mixing with a differentiable mixing console of neural audio effects. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 71–75. IEEE, 2021.

Christian J Steinmetz, Nicholas J Bryan, and Joshua D Reiss. Style transfer of audio effects with differentiable signal processing. *arXiv preprint arXiv:2207.08759*, 2022.

Christian J Steinmetz, Shubhr Singh, Marco Comunità, Ilias Ibnyahya, Shanxin Yuan, Emmanouil Benetos, and Joshua D Reiss. St-ito: Controlling audio effects for style transfer with inference-time optimization. *arXiv preprint arXiv:2410.21233*, 2024.

Modan Tailleur, Junwon Lee, Mathieu Lagrange, Keunwoo Choi, Laurie M Heller, Keisuke Imoto, and Yuki Okamoto. Correlation of fréchet audio distance with human perception of environmental audio is embedding dependent. In *2024 32nd European Signal Processing Conference (EUSIPCO)*, pp. 56–60. IEEE, 2024.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi. Investigating rnn-based speech enhancement methods for noise-robust text-to-speech. In *Proc. SSW 2016*, pp. 146–152, 2016.

Soumya Sai Vanka, Christian Steinmetz, Jean-Baptiste Rolland, Joshua Reiss, and George Fazekas. Diff-mst: Differentiable mixing style transfer. *arXiv preprint arXiv:2407.08889*, 2024.

Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8228–8238. IEEE, 2024.

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1905–1914. IEEE, 2021.

Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36:71340–71357, 2023.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Pavel Záviška, Pavel Rajmic, Alexey Ozerov, and Lucas Rencker. A survey and an extensive evaluation of popular audio declipping methods. *IEEE Journal of Selected Topics in Signal Processing*, 15(1):5–24, 2020.

Kai Zhang, Jingyun Liang, Luc Van Gool, and Radu Timofte. Designing a practical degradation model for deep blind image super-resolution. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4771–4780. IEEE, 2021.

Wangyou Zhang, Robin Scheibler, Kohei Saijo, Samuele Cornell, Chenda Li, Zhaoheng Ni, Anurag Kumar, Jan Pirklbauer, Marvin Sach, Shinji Watanabe, et al. Urgent challenge: Universality, robustness, and generalizability for speech enhancement. *arXiv preprint arXiv:2406.04660*, 2024a.

Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco A Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Musicmagus: Zero-shot text-to-music editing via diffusion models. *arXiv preprint arXiv:2402.06178*, 2024b.

Udo Zölzer, Xavier Amatriain, Daniel Arfib, Jordi Bonada, Giovanni De Poli, Pierre Dutilleux, Gianpaolo Evangelista, Florian Keiler, Alex Loscos, Davide Rocchesso, et al. *DAFX-Digital audio effects*. John Wiley & Sons, 2002.

A LIMITATIONS

First, the real-audio evaluation employs a learned quality predictor, which also guides the selection of reference masters (PQ ≥ 7.5) and the ranking of MPO candidates. As a result, WILD PQ is not entirely independent of training. The listening study (Section 4.5) provides complementary human judgements, and it supports the result on synthetic chains, but its five WILD items are too few to separate the systems on real audio. The real-audio gain therefore rests on the predictor alone, and a larger loudness-matched study that also includes the unprocessed input and the model before preference optimisation is the obvious next step. Second, our focus is on the corrective and finishing aspects of mastering. Nine of the 24 types are re-recording and transfer artifacts that an engineer would typically address during pre-processing. Creative choices, such as genre-specific loudness or the use of a reference track, are beyond our scope; therefore, we do not compare against reference-based mastering approaches such as Matchering (Koo et al., 2025). WILD also emphasizes experimental, ambient, and noise genres, where a rough sound may be intentional. Third, we leave to future work isolating the contributions of chain ordering and pass-through pairs, as well as testing whether the severity wording of an instruction influences the strength of the model's intervention.

B SEVERITY MAPPING

Each parameter p has a range $[a, b]$, a linear or logarithmic scale, and a response to severity:

$$p = \begin{cases} a + \tilde{s}\,(b-a) & \textit{direct}\text{: the degradation increases with severity,} \\ b - \tilde{s}\,(b-a) & \textit{inverse}\text{: the parameter value decreases as severity rises,} \\ a + \xi\,(b-a), \quad \xi \sim \mathcal{U}(0,1) & \textit{none}\text{: independent of severity,} \end{cases} \tag{4}$$

where $\tilde{s} = \mathrm{clip}_{[0,1]}(s+\eta)$ and the per-parameter jitter $\eta \sim \mathcal{U}(-\delta, \delta)$, with $\delta = 0.15$, keeps degradations loosely coupled rather than moving in lockstep. Examples of inverse parameters are the SNR of background noise ($35 \to 5$ dB) and codec bitrate ($192 \to 32$ kbps); peaking-EQ gain ($0 \to 12$ dB) is direct. On logarithmic scales the interpolation is geometric; otherwise a resonance sampled uniformly over 800–4000 Hz would put most of its mass above 2 kHz. The *none* mode covers parameters such as the base frequency of mains hum, where no value is worse than another. Drawing s once per degradation type instead of once per chain leaves each family's $\mathrm{Beta}(2,2)$ marginal unchanged and alters only whether severities move together within a chain.

C DEGRADATION TAXONOMY AND PRODUCTION SCENARIOS

Table 6 lists the 24 degradation types by stage, with their parameters, severity modes (Appendix B) and SonicMaster counterparts. Mode *d* is direct, *i* inverse and *n* independent of severity; ranges marked log are interpolated geometrically. Table 7 gives the order in which each production scenario applies its degradations.

D CORPUS CONSTRUCTION DETAILS

Quality control. A degraded pair is kept only if the reference excerpt exceeds -45 dBFS RMS, the difference signal $x - y^\star$ is at least -35 dB relative to $y^\star$, and no more than 98% of frames are left untouched. Pass-through pairs bypass these checks, which would otherwise reject all of them. Each kept example records the ordered chain with every parameter value, the severity latent s, the scenario name and a frame-level mask of untouched regions, so the corpus can be audited.

Instructions. Severity bands are mild ($s < 0.34$), moderate ($0.34 \leq s < 0.67$) and severe ($s \geq 0.67$). The two registers are drawn independently, which lets paraphrase robustness be tested separately from instruction sensitivity. When no complete scenario chain applies, the instruction names one applied degradation type, or two with probability 0.4, sampled from all types present rather than the last one. The last type in a random chain belongs to the latest stage present, usually transfer or distribution, so keying on it would over-describe codec artefacts and rarely describe mix flaws. Because the bank is fixed, the corpus is reproducible and building it requires no network access.

stage	degradation type	parameters (range, mode)	SM
Mix & mastering	peaking-EQ imbalance	1–4 bands of alternating sign (d); gain 0–12 dB (d); Q 0.4–6 (n)	✓
	over-compression	threshold -40 to -8 dB (i); ratio 2–20 (d); attack 0.5–40 ms (i); release 30–800 ms (n)	✓
	digital clipping	ceiling 2–9 dB below the clip's peak (i); oversampling $\{1, 4\}$	✓
	stereo-width error	width 0–1.8, narrowing or over-widening (n)	✓
	low-mid build-up	boost 2–10 dB at 200–500 Hz (d)	✓
	loss of air	high shelf at 10 kHz, cut 3–12 dB (d)	✓
	dull top end	high shelf at 6 kHz, cut 3–12 dB (d)	✓
	excess brightness	high shelf at 6 kHz, boost 3–12 dB (d)	✓
	thin low-mids	low shelf at 400 Hz, cut 3–12 dB (d)	✓
	upper-range roll-off	4 kHz low-pass, order 2–8 (d)	✓
	transient softening	attack 1–20 ms (i); release 20–200 ms (d); look-ahead 1–8 ms (n); reduction 3–18 dB (d)	✓
	recessed vocal	band-stop 350–3500 Hz, attenuation 3–20 dB (d)	✓
	low programme level	amplitude 0.05–0.7, log (i)	✓
	harmonic saturation	drive 2–24 dB (d); wet mix 0.3–1.0 (d)	–
	over-limiting	threshold -18 to -1 dB (i); attack 5–100 ms (i); release 5–100 ms (n)	–
Re-recording	transducer coloration	low cut 80–500 Hz, log (d); high cut 3.5–12 kHz, log (i); resonance 0.8–4 kHz, log (n), 0–9 dB (d); tilt -6 to 6 dB (n)	✓
	room reverberation	RT60 0.2–1.4 s (d); distance 0.3–6 m (d); absorption 0.15–0.75 (i)	✓
	automatic gain control	target -24 to -12 LUFS (n); attack 5–120 ms (i); release 80–1200 ms (i); max. gain 6–24 dB (d)	–
	background noise	crowd, hiss, street, room tone or Gaussian (n); SNR 5–35 dB (i)	–
	mains hum	50/60 Hz (n); harmonics 1–6 (d); level -48 to -28 dB (d)	–
Transfer & distribution	lossy-codec artefacts	Opus, AAC or MP3 (n); bitrate 32–192 kbps, log (i); generations 1–3 (d)	–
	band-limiting	cutoff 4–16 kHz, log (i); order 4–10 (d)	partial
	dropouts	rate 1–30/min (d); duration 20–400 ms (d); depth -60 to -6 dB (i)	–
	wow & flutter	wow 0.3–2 Hz (n); flutter 6–14 Hz (n); depth 3–45 cents (d)	–

Table 6: The 24 degradation types. *SM*: counterpart in the SonicMaster bank. SonicMaster models transducer coloration with measured microphone impulse responses where ours is parametric, and its band-limiting is a fixed 4 kHz low-pass. Instructions for *excess brightness* keep SonicMaster's correction label "darkness".

scenario	degradations, in the order applied
phone at a gig	background noise → transducer → AGC → codec → band-limiting
laptop across a room	room → transducer → AGC → codec → band-limiting
cassette transfer	wow & flutter → background noise → band-limiting → mains hum
internet rip	codec → band-limiting → over-compression → over-limiting → digital clipping
amateur master	low-mid build-up → loss of air → dull top end → transient softening → over-compression → over-limiting → digital clipping → stereo-width error
call capture	band-limiting → transducer → AGC → mains hum → codec → dropouts

Table 7: Production scenarios. Scenario parameter overrides are clamped to each type's range, so a scenario narrows but never widens it. Cassette transfer, internet rip and call capture depart from the stage order used for random chains.

E Implementation details

Audio is processed as 30-second stereo excerpts at 44.1 kHz ($C = 2$, $T = 30 f_s$) and encoded using the frozen Oobleck autoencoder from Stable Audio Open (Evans et al., 2025), resulting in latents of shape $L \times d = 645 \times 64$. The MM-DiT architecture comprises 6 joint and 18 single blocks, 8 attention heads with dimension 128, and a joint attention dimension of 1024. Instructions are encoded using a frozen FLAN-T5-large model (Chung et al., 2024). Training begins from the public SonicMaster checkpoint, proceeding for 10 epochs on the released corpus and then on the union dataset (approximately 192,000 examples, effective batch size 128) for 14,000 steps, the checkpoint behind every reported result, both at a learning rate of 2×10^{-5} with 200 warmup steps. MPO (Section 3.3) trains on its 10,000 fixed pairs for 4 epochs at 5×10^{-6} with a per-GPU batch size of 4 (effective batch size 16), concluding at 2,500 steps. All experiments utilize AdamW optimization on four NVIDIA A100 GPUs. At inference we use $N = 20$ uniform Euler steps with guidance 1.0, and audio longer than 30 s is processed as consecutive 30 s excerpts. The DSP mastering preset is a fixed six-band parametric EQ (low shelf at 50 Hz, -2 dB, $Q = 0.7$; peak at 300 Hz, -2 dB, $Q = 1.0$; peak at 3 kHz, $+1$ dB, $Q = 0.8$; high shelf at 10 kHz, $+2$ dB, $Q = 0.7$; two bands at 0 dB)

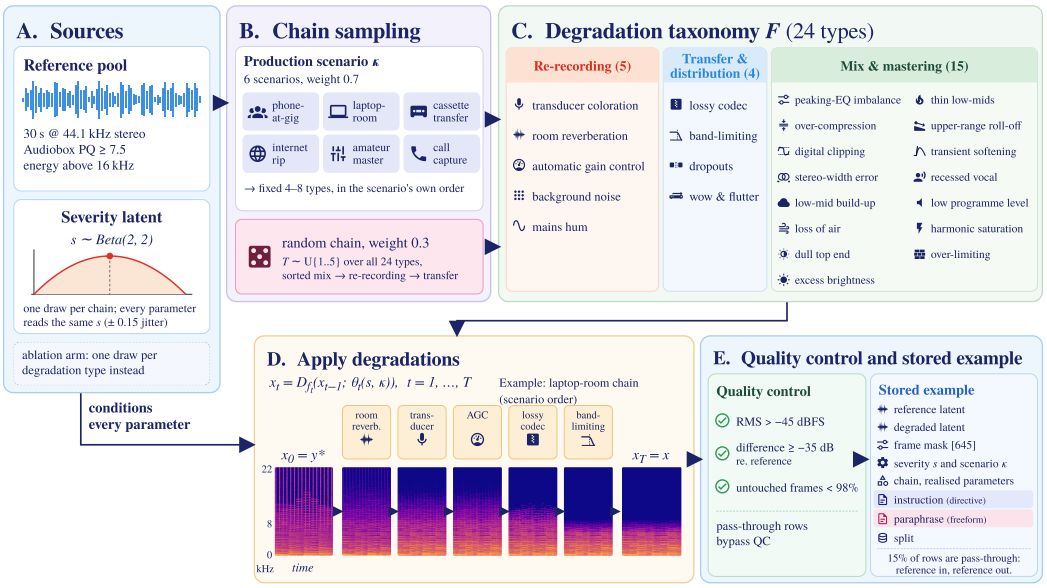


Figure 4: Degradation-chain synthesis. A reference-master excerpt and one severity latent $s \sim \mathrm{Beta}(2, 2)$ per chain (A) feed chain sampling (B): a production scenario with probability 0.7, otherwise a random chain of one to five degradation types ordered mix and mastering → re-recording → transfer and distribution. Types are drawn from the 24-type taxonomy (C; Table 6) and applied in order (D); each pair then passes the quality check and is stored with its full chain record (E).

		degradations	examples		reference
set	degradations	per example	degraded	pass-through	master
SONICMASTER-val	released corpus	1–3	3,500	–	✓
SONICMASTER-test	released corpus	1–3	7,000	–	✓
COMPOUND-val	compound chains, ours	1–8 (4.8)	1,253	229	✓
COMPOUND-test	compound chains, ours	1–8	1,289	261	✓
WILD	real, not synthesised	unknown	574	–	–

Table 8: Evaluation sets. SONICMASTER refers to the released corpus of Melechovsky et al. (2025) at its published split, containing one to three effects per example and at most one per category. COMPOUND consists of ordered chains of one to eight degradations from the 24 types (mean 4.8); no scenario chain is reachable under this rule, and about 18% of random chains (133 of 738 in validation and test, about 5% of degraded examples) have a composition that is reachable under this rule, ignoring parameter ranges and order. Pass-through examples take the reference master as input and are evaluated separately. WILD comprises real low-production-quality releases without a reference master, so only reference-free metrics apply.

followed by a memoryless $\tanh$ soft clipper above 0.7 (-3.1 dBFS), after which the output is scaled to the input's peak level. It has no compressor, reverb or loudness normalisation, applies the same processing to every track, and takes no instruction.

F METRIC SETTINGS

All signal-based measures use mono audio, obtained by averaging channels, at 44.1 kHz. Each paired measure compares the output with the reference decoded from its latent representation.

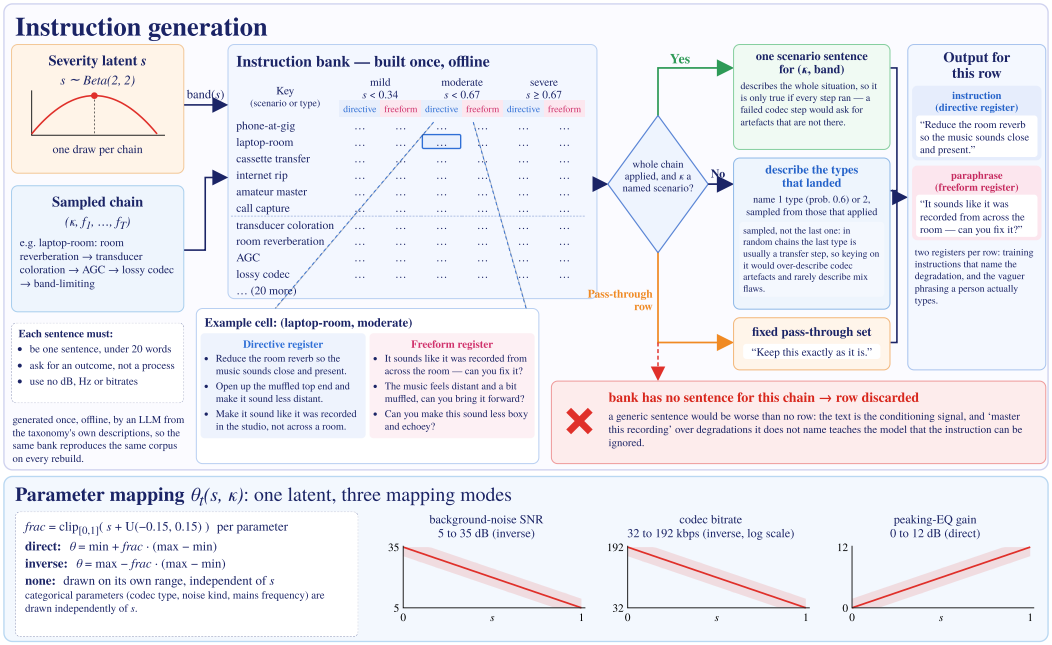


Figure 5: Instruction generation and parameter mapping. Top: each example draws a directive and a freeform instruction from a fixed, offline bank indexed by scenario or degradation type and severity band; a complete scenario chain receives its scenario sentence, otherwise one or two applied types are described, pass-through rows receive a fixed instruction, and chains the bank cannot describe are discarded. Bottom: the severity latent sets every parameter through a direct, inverse or severity-independent mapping (Appendix B).

Mel-SSIM. We compute a 128-band power mel spectrogram using a 2048-point FFT, a 2048-sample Hann window, and a hop size of 512 samples. Each spectrogram is converted to decibels relative to its maximum, floored at -80 dB, and min–max scaled to $[0, 1]$. We then compute SSIM using a 7×7 uniform window with $K_1 = 0.01$ and $K_2 = 0.03$.

Mel-spectral KL. We average the same power mel spectrogram over frames to obtain a long-term spectrum, then normalise it to sum to one across the 128 bands. We compute $\mathrm{KL}(\text{output} \,\|\, \text{reference})$ in nats, using $\epsilon = 10^{-10}$ for numerical stability.

Log-spectral distance (LSD). We compute the STFT using a 2048-point FFT, a 2048-sample Hann window, and a hop size of 512 samples. For each frame, we take the root mean square across frequency bins of the difference between output and reference magnitudes expressed as $20 \log_{10}$ values. We average these distances over frames and report the result in dB.

G Training-data composition

Training-data composition influences adaptation to the new corpus and performance on the original domain. Training primarily on the new corpus with 15% rehearsal gives the highest scores on the new validation set (0.7230) and Wild (5.866). However, it scores 0.7581 on SonicMaster validation, falling 0.0158 below the SonicMaster-only model (0.7739). Training on both corpora nearly eliminates this gap, reaching 0.7734, with small decreases on the new validation set (0.7177) and Wild (5.807). Joint training thus better preserves performance on the original domain while retaining most of the gains on the new data.

	WILD	COMPOUND val		SONICMASTER val
training corpus	PQ	mel-SSIM	pass-through	mel-SSIM
none (input)	5.364	0.5955	1.0000	0.6466
SONICMASTER only	$5.687_{+0.323}$	$0.5967_{+0.0013}$	0.8015	$0.7739_{+0.127}$
COMPOUND†	$5.866_{+0.502}$	$0.7230_{+0.1276}$	0.9822	$0.7581_{+0.112}$
severity per type	$5.881_{+0.517}$	$0.7205_{+0.1250}$	0.9804	–
union of both	$5.807_{+0.443}$	$0.7177_{+0.1222}$	0.9843	$0.7734_{+0.127}$

Table 9: Training-data ablation before MPO. Subscripts: change from the unprocessed input. All rows start from the released SonicMaster checkpoint; steps: 11,000 (SONICMASTER), 4,500 (COMPOUND rows), 14,000 (union). †With 15% SONICMASTER rehearsal. *Severity per type*: one latent per degradation type rather than per chain. Union = PRISM w/o MPO (Table 1).

H CORRECTION BY SCENARIO

Figure 2a reports COMPOUND validation results by scenario, ordered by input quality. PRISM generally achieves larger gains on lower-quality inputs. Across the six scenarios and random chains, mel-SSIM gain is negatively correlated with input quality ($r = -0.90$, Spearman $\rho = -0.86$; descriptive statistics, $n = 7$). The three lowest-quality scenarios, phone at a gig (0.451), laptop across a room (0.489), and call capture (0.494), improve by $+0.190$, $+0.166$, and $+0.184$, respectively. Gains are smaller for cassette transfer (0.620 input; $+0.088$) and random chains (0.696; $+0.063$). Internet rip departs from this trend: despite having the highest input quality, it improves by $+0.108$.

SonicMaster produces small changes across scenarios (-0.011 to $+0.026$), with declines in three cases and improvements on only 23.5% of cassette transfers. Trained on independently sampled combinations of up to three effects, it shows limited gains across the evaluated scenarios. However, deeper chains also involve more severe degradation in this corpus, so these results do not isolate the effect of chain depth. Section 4.2 examines depth within random chains to address this confound.

On WILD (Figure 2b), PRISM's gain decreases from $+0.916$ in the lowest-quality bin to $+0.259$ in the highest, with at least 74% of tracks improving in every bin. The model before MPO shows the same ordering, indicating that this pattern is already present before preference optimisation. SonicMaster's average change becomes negative above an input quality of 5.5, indicating reduced predicted quality for better-produced tracks.

I WILD BY GENRE

WILD is not filtered by genre and is skewed toward genres in which rough production can be intentional: by MTG-Jamendo genre tags, 33% of its tracks are experimental, 27% electronic and 25% ambient, and 43 are tagged noise, 18 drone and 9 lo-fi; tags overlap. Table 10 breaks the result down by tag.

PRISM improves predicted production quality across all genres, although the gains vary substantially. The largest gains occur in conventionally produced music, including rock ($+0.812$) and techno ($+0.716$), with nearly nine in ten tracks improving. Gains are smallest ($+0.17$ to $+0.30$) for dark ambient, electronic, experimental, ambient and noise, a group with substantial tag overlap in which rough textures may be intentional. SonicMaster reduces predicted quality in all five genres, with the largest decline in dark ambient (-0.380). Excluding tracks tagged as noise, lo-fi, drone or dark ambient increases PRISM's gain from $+0.509$ to $+0.554$ ($n = 466$; 81% of tracks improved), while its advantage over SonicMaster remains substantial ($+0.147$ on this subset).

PRISM also outperforms its counterpart before preference optimisation in every row, with additional gains ranging from $+0.049$ for electronic to $+0.104$ for pop. These gains are therefore not limited to genres in which the supervised model already performs well. However, MPO ranks candidates partly using the same predictor used for this evaluation. Consistent gains across genres are thus expected and do not provide independent evidence of perceptual improvement. The listening study in Section 4.5 includes only five WILD items, too few to test whether these gains are audible; on those five, listeners split evenly between PRISM and SonicMaster.

genre tag	n	PRISM	w/o MPO	SonicMaster	AudioSR
all tracks	574	$+0.509$ (79%)	$+0.443$	$+0.086$	$+0.084$
rock	75	$+0.812$ (89%)	$+0.723$	$+0.399$	$+0.024$
techno	45	$+0.716$ (89%)	$+0.628$	$+0.340$	$+0.059$
pop	24	$+0.693$ (92%)	$+0.589$	$+0.350$	$+0.135$
metal	52	$+0.590$ (85%)	$+0.525$	-0.068	-0.062
industrial	73	$+0.498$ (77%)	$+0.436$	$+0.023$	$+0.126$
alternative	83	$+0.470$ (78%)	$+0.408$	$+0.092$	$+0.121$
ambient	143	$+0.297$ (73%)	$+0.240$	-0.120	$+0.135$
noise	43	$+0.297$ (70%)	$+0.245$	-0.155	$+0.157$
experimental	189	$+0.269$ (68%)	$+0.212$	-0.145	$+0.096$
electronic	155	$+0.262$ (67%)	$+0.213$	-0.206	$+0.056$
dark ambient	47	$+0.174$ (62%)	$+0.116$	-0.380	-0.020
excl. noise, lo-fi, drone, dark ambient	466	$+0.554$ (81%)	$+0.487$	$+0.147$	$+0.088$

Table 10: Change in predicted production quality (Audiobox PQ) on WILD by MTG-Jamendo genre tag. Tags overlap, so a track counts toward every genre it carries; lo-fi ($n = 9$) and drone ($n = 18$) have no row of their own but are removed in the last row. Parentheses give the fraction of tracks PRISM improves. Rows with fewer than about 50 tracks are noisy.

J DISAGREEMENT BETWEEN FIDELITY METRICS

Multiple fidelity measures are needed because they capture different aspects of audio quality, as illustrated by two contrasting results in Table 2. AudioSR achieves the largest reduction in LSD among the baselines (-5.397) and improves PANNs FD (-2.057), yet leaves mel-spectral KL nearly unchanged ($+0.004$). Its CLAP-FAD (0.175) is also worse than that of the unprocessed input. These results suggest that high-frequency extension improves band-wise spectral fidelity without correcting tonal balance, defined as the distribution of energy across frequency bands. The improvement in PANNs FD also contrasts with the greater divergence from reference masters measured by CLAP-FAD. Conversely, the model trained exclusively on SONICMASTER yields the largest increase in LSD ($+1.806$), despite improving both KL divergences. Together, these findings show that improvements in band-wise spectral fidelity need not coincide with improvements in tonal balance or learned representations, and that the two encoders can assess the same outputs differently. Brightness and bandwidth do not explain the production-quality gains. Text2FX raises the spectral centroid by 2.2 kHz and the roll-off frequency by 4.6 kHz yet loses 0.39 in production quality, and AudioSR extends bandwidth by 5.9 kHz for a gain of 0.08, whereas PRISM moves the centroid by 0.11 kHz and the roll-off by 0.54 kHz.

K DEGRADATION-SPECIFIC INSTRUCTION FOLLOWING

We test whether naming a degradation in the instruction causes the model to prioritise its correction. We map fifteen degradation types to four signal descriptors: high-frequency roll-off (band-limiting, lossy codec, loss of air), spectral centroid (dull top end, excess brightness, low-mid build-up, upper-range roll-off, thin low-mids, transducer coloration), crest factor (over-compression, over-limiting, transient softening, digital clipping, harmonic saturation), and RMS level (low programme level). We exclude background noise, mains hum, room reverberation, dropouts, and stereo-width error.

For a descriptor g, compliance is

$$\mathrm{Comp}_g = 1 - \frac{|g(\hat{y}) - g(y^\star)|}{|g(x) - g(y^\star)|}, \tag{5}$$

the fraction of the input's deviation from the reference removed by the model. We name one degradation in each chain and compare its compliance with that of unnamed degradations in the same chain.

Across $1{,}721$ matched pairs, PRISM achieves median compliance of $+0.483$ for named degradations and $+0.478$ for unnamed degradations. Named degradations score higher in 742 pairs and unnamed

degradations in 726, with 253 ties, mostly due to the roll-off descriptor's discrete FFT-bin resolution. Excluding ties, named degradations score higher in 50.5% of pairs for PRISM, 50.2% for the supervised model (745 versus 739; 237 ties), and 50.7% for the released SonicMaster checkpoint (849 versus 825; 47 ties). These results provide no evidence that naming a degradation selectively improves its correction under the evaluated descriptors.

Classifier-free guidance does not improve selectivity. As the guidance scale increases from 1.5 to 5.0 for the supervised model, median compliance remains equal for named and unnamed degradations and falls from $+0.452$ to $+0.209$. This is consistent with the model not having been trained with text conditioning omitted.

L MODEL DETAILS

Instruction tokens enter the joint blocks as a sequence. They are also mean-pooled using the token mask and projected, together with a learned embedding of excerpt duration, to form the text component of the pooled modulation vector. This component is concatenated with a pooled audio projection and passed through a final fusion layer.

During supervised training, the audio branch receives the first third of the reference master and is dropped with probability 0.75. During preference optimisation and all evaluation, it receives zeros. Supervised training therefore provides reference information that is unavailable at evaluation. The SonicMaster checkpoint uses the same arrangement, so this conditioning setup is shared by both models. We retain the architecture and conditioning scheme of the initial checkpoint; differences between that checkpoint and PRISM arise from the training data and schedule.

M PREFERENCE OPTIMISATION DETAILS

We selected $K = 5$ and $\tau = 0.2$ by examining candidate-score variation across values of τ on a held-out training probe. Small values of τ produce nearly indistinguishable candidates, whereas large values produce outputs that no longer resemble alternative masters of the same input. MPO uses $10{,}000$ examples from a pool of $27{,}013$ examples drawn exclusively from our training corpus, with no examples from the released corpus. Approximately 15% are pass-through examples. We select the first $10{,}000$ entries from a round-robin ordering over (scenario, severity band) groups to cover different scenarios and severities. Candidates are generated once using the supervised model, and preference pairs are fixed before optimisation.

The standardised winner–loser gap in the combined score has a mean of 1.95, a median of 2.02, a tenth percentile of 1.14, and a minimum of 0.108. No examples are discarded. Because scores are standardised within each candidate set, these gaps reflect differences relative to the within-set variation, rather than large absolute differences.

In raw units, preferred candidates exceed rejected candidates by an average of 0.098 in production quality (median 0.063) and 0.0096 in mel-SSIM (median 0.0074). The pairs therefore capture fine-grained preferences between alternative masters of the same input rather than contrasts with gross failures. Although the production-quality gaps are small relative to the predictor's range, they are large relative to the variation within each candidate set.

N FRONTIER DETAILS

Checkpoints at 800, 1,500, and 2,500 steps achieve (pass-through fidelity, WILD PQ) values of $(0.9805, 5.859)$, $(0.9785, 5.870)$, and $(0.9779, 5.873)$, respectively, against $(0.9843, 5.807)$ before MPO. By step 800, about three-quarters of the gain and over half of the preservation cost are realised. A gentler pilot schedule (5×10^{-7} for 2 epochs instead of 5×10^{-6} for 4), run first and also scored on WILD before the final schedule was chosen, achieves $(0.9834, 5.819)$, intermediate between the supervised model and PRISM, consistent with the learning rate setting the operating point along the same trade-off. Across all 25 saved checkpoints, validation mel-SSIM stays between 0.7175 and 0.7187, and pass-through fidelity declines from 0.983 to 0.978. Audiobox content enjoyment and content usefulness, which MPO does not rank but which share the PQ predictor's backbone,

input	input PQ	n	PQ added by MPO
WILD, PQ < 4.5	4.13	93	$+0.100$ $[+0.083, +0.118]$
WILD, PQ 4.5–5.0	4.77	95	$+0.072$ $[+0.057, +0.088]$
WILD, PQ 5.0–5.5	5.25	156	$+0.059$ $[+0.048, +0.069]$
WILD, PQ 5.5–6.0	5.70	109	$+0.068$ $[+0.058, +0.078]$
WILD, PQ ≥ 6.0	6.63	121	$+0.044$ $[+0.037, +0.052]$
COMPOUND val, degraded	6.84	1,253	$+0.061$
COMPOUND val, pass-through	8.19	229	$+0.013$

Table 11: Production quality added by preference optimisation (PRISM minus the pre-MPO model, on the same item), by the quality of the input, measured on the evaluated excerpt (WILD was selected on a different, centred 60 s excerpt; Section 4). Brackets are 95% paired bootstrap intervals over tracks. Per track on WILD, input quality and added quality are anticorrelated (Spearman $\rho = -0.22$, $p < 10^{-7}$, $n = 574$).

increase on both evaluation sets: by $+0.045$ and $+0.042$ on COMPOUND validation, and by $+0.060$ and $+0.088$ on WILD. Audiobox production complexity, which measures the number of audio components in the scene rather than the amount of processing, changes in opposite directions on the two sets ($+0.191 \rightarrow +0.168$ on COMPOUND validation, $+0.348 \rightarrow +0.362$ on WILD); we do not interpret it as evidence about mastering. All four paired changes from MPO are significant under a paired bootstrap over COMPOUND test rows: LSD -0.177 $[-0.211, -0.144]$, mel-spectral KL -0.0040 $[-0.0062, -0.0018]$, PANNs KL -0.0023 $[-0.0043, -0.0003]$ and mel-SSIM $+0.0014$ $[+0.0008, +0.0019]$. CLAP-FAD moves from 0.033 to 0.030, within its window noise.

O LISTENING STUDY DETAILS

Protocol. Fifteen trials in a fixed order, no training phase and no reference. Each trial plays the same input track processed by PRISM, SonicMaster and AudioSR as three players marked A, B and C, and asks which version has the best production and mastering quality. Exactly one answer is required. Participants were asked to use headphones and could replay each version as often as they liked. Ten participants, all with mixing or mastering experience, answered all 15 trials. The median session took 18 minutes, with a range of 4 to 140 minutes; the form recorded no demographic data. Stimuli are the raw system outputs and are not loudness-matched.

Items. Ten items were drawn from COMPOUND test and five from WILD, before any responses were collected (Table 12). The COMPOUND items carry synthetic chains for which a reference master exists, although listeners never heard it. The WILD items are real low-production-quality releases with no reference.

Label assignment. The map from system to label was drawn once per trial and held fixed across listeners, so label position is a property of the item and not of the listener. The draw was not balanced across systems: over the 15 trials PRISM appeared as A, B and C on 3, 6 and 6 trials, SonicMaster on 8, 0 and 7, and AudioSR on 4, 9 and 2. Two checks suggest this does not drive the result. Listeners showed no preference among the three labels (52 A, 44 B, 54 C; $\chi^2 = 1.12$, $p = 0.57$), and on COMPOUND test PRISM was chosen in 90.0%, 90.0% and 80.0% of judgements in the trials where it sat at A, B and C. A balanced assignment, randomised per listener, would be better and we would use one in a larger study.

Statistics. Each judgement is nested in a listener and in an item, so a binomial test over the 150 judgements would overstate precision. We report 95% percentile intervals from a bootstrap that resamples listeners and items together, with 50,000 replicates. Bootstrap p-values for PRISM against chance ($1/3$) are $p < 10^{-4}$ over all items, $p < 10^{-4}$ on COMPOUND test and $p = 0.26$ on WILD. A listener-level sign test agrees: all ten listeners chose PRISM above chance overall and on COMPOUND test ($p = 0.001$, one-sided), and seven of ten did so on WILD ($p = 0.17$). Dropping AudioSR and counting only head-to-head judgements, PRISM beats SonicMaster on 106 of 135 overall (78.5%), 85 of 93 on COMPOUND test (91.4%) and 21 of 42 on WILD (50.0%).

trial	source item	votes (of 10) AudioSR	SonicMaster	PRISM
COMPOUND test				
t01	04_279604_c1_v1	0	1	9
t02	07_884407_c1_v1	0	0	10
t03	17_1095717_c0_v1	1	2	7
t04	21_1216121_c0_v1	1	3	6
t05	26_75226_c1_v0	0	0	10
t06	27_1216127_c1_v1	2	0	8
t07	27_303227_c1_v1	0	0	10
t08	28_1216128_c2_v0	1	1	8
t09	28_944928_c1_v0	0	1	9
t10	34_457234_c2_v0	2	0	8
WILD				
t11	04_1399804	1	4	5
t12	12_1033912	2	6	2
t13	27_785627	2	4	4
t14	42_1097842	2	6	2
t15	45_1012345	1	1	8
total		15	29	106

Table 12: Per-item votes in the listening study. Each row is one input track processed by the three systems; ten listeners each chose one version. PRISM is the modal choice on all ten COMPOUND test items and on two of the five WILD items, and is tied with SonicMaster on a third (t13).

listener	1	2	3	4	5	6	7	8	9	10	mean
PRISM (of 15)	12	12	7	10	11	13	9	14	7	11	10.6
SonicMaster	3	3	1	5	4	2	6	0	4	1	2.9
AudioSR	0	0	7	0	0	0	0	1	4	3	1.5

Table 13: Per-listener choices over the 15 trials. Chance is 5 of 15. Every listener chose PRISM more often than chance. Listeners 1 and 2 are the identical pair discussed above. Listeners 3 and 9 account for 11 of the 15 AudioSR votes.

Duplicate response vectors. Two of the ten response vectors agree on all 15 trials, one submitted from a phone and one from a desktop browser. We cannot tell whether this is one participant submitting twice or two participants who happened to agree, and we report all ten. Dropping one of the pair changes little: PRISM is chosen in 69.6% of judgements overall, 83.3% on COMPOUND test and 42.2% on WILD.

P PROBING ANALYSIS: DETAILS

P.1 SETUP

Models and features. Both models use the SONICMASTER backbone: a flow-matching transformer with a FLAN-T5-large text encoder, six double-stream blocks (D0 to D5) followed by 18 single-stream blocks (S0 to S17), hidden size 1024 (8 heads of size 128), 645 audio tokens per clip and a 64-channel latent input. Ours is the final checkpoint after preference optimization; SONICMASTER is the released checkpoint. All weights are frozen. A forward hook on each block stores the hidden states of the audio tokens, and we average them over the 645 tokens. This gives one 1024-dimensional vector per clip and block. The instruction is left empty, because it names the degradations, and chunk-continuity conditioning is not used.

Flow time. We read the features at $t \in \{0, 0.25, 0.5, 0.75, 1\}$. As in training, the input at t is $(1-t)\,\mathbf{z}_{\mathrm{deg}} + t\,\mathbf{z}_{\mathrm{ref}}$, where $\mathbf{z}_{\mathrm{deg}}$ and $\mathbf{z}_{\mathrm{ref}}$ are the latents of the degraded clip and of its reference

master. So t=0 is the degraded input and t=1 is the reference itself. We fit a separate probe at each t.

Probes. For each block (and each t) we fit one multi-label logistic regression with 1024 inputs and five sigmoid outputs, one per group. Features are standardized. We train with full-batch gradient descent for 400 epochs, learning rate 0.5, weight decay 10^{-3} and seed 0. The settings are the same for every model, block and t.

Data. Each model's probes are trained on 10,000 clips sampled (seed 0) from that model's training corpus. A further 100 clips are used only for the check in Appendix P.2. We test on the SONICMASTER test split (7,000 clips) and on our chain test split (1,550 clips).

Labels. We use the five groups of SONICMASTER: EQ, dynamics, reverb, amplitude and stereo. A clip can belong to several groups. For chains, 19 of our 24 degradation types map to one of these groups. The other five (wow/flutter, hum, codec, dropouts and noise) fall outside all five groups and count as negatives; the three further types with no counterpart in the SonicMaster bank (automatic gain control, over-limiting and harmonic saturation; Section 3.2) do map to a group.; 1.36% of the chain training clips contain only such types and so have no positive label. For SONICMASTER, its fine-grained tags map to the same five groups. Its data pipeline also applies a clipping pass that the tags do not record; we add it to the amplitude label so that these clips are not marked negative by mistake.

Metrics. We report average precision (AP) per group; our implementation matches scikit-learn's `average_precision_score`. The base rate π of a group is close to the AP of a random ranking, so every AP must be read against the base rate of its own test set (Table 14). To compare across test sets we use chance-corrected AP, $\mathrm{cAP} = (\mathrm{AP} - \pi)/(1 - \pi)$, averaged over the five groups.

Table 14: Base rates (fraction of clips that contain each group) in the two test sets.

Test set	Clips	EQ	Dyn.	Rev.	Amp.	Stereo	Mean
Chain test set	1,550	.770	.588	.137	.259	.113	.373
SONICMASTER test set	7,000	.584	.235	.382	.286	.149	.327

P.2 CHECKS

Probes against shuffled labels. A probe can score above the base rate by fitting noise, and taking the best of 24 blocks adds a further upward bias. We therefore refit every probe on shuffled labels (five shuffles), in the spirit of control tasks (Hewitt & Liang, 2019). On the 100 held-out clips, we compare each group's best AP over the 24 blocks with the same statistic under shuffled labels, and call a group decodable if the gap is at least 0.15 (Table 15). Nine of the ten tests pass. The exception is EQ for our model, where the test cannot pass: 82 of these 100 clips carry EQ, so shuffled-label probes already reach 0.947, and no AP can exceed that by 0.15. On the full chain test set the same group reaches 0.987 at S13, against a base rate of 0.770 (Table 4).

Stereo on chains is at chance. SONICMASTER's stereo AP on the chain test set is 0.130, against a base rate of 0.113. To see whether this differs from chance, we scored 20,000 random rankings of a test set with the same size and base rate. Their AP is 0.117 $\pm$ 0.009 (mean $\pm$ s.d.), and 8% of them reach 0.130 or more, so this probe does no better than a random ranking. Its dynamics AP (0.644, base rate 0.588) is higher than all 20,000 random rankings, so dynamics is above chance, but only slightly (cAP 0.14).

Choice of block. Table 4 reports every pair at the same block, S13, so that no block is chosen per row on test data. S13 is the best block for SONICMASTER on chains, so this choice gives the baseline its best case in the cross-corpus comparison. The best block per pair (bold in Tables 16 and 17) is S9 for ours on its own corpus (mean AP 0.935), S12 for ours on the other corpus (0.772), S15 for SONICMASTER on its own corpus (0.982, tied with S12 and S13 at three decimals) and S13 for SONICMASTER on the other corpus (0.599). S13 is within 0.010 mean AP of each.

Table 15: Decodability check on the 100 held-out clips. Base: base rate in these clips. Shuffled: best AP over the 24 blocks for probes fit on shuffled labels. Best: best AP over the 24 blocks with the true labels, reached at the given block. Decodable: Best − Shuffled ≥ 0.15.

Model	Group	Base	Shuffled	Best	Block	Gap	Decodable
Ours	EQ	.820	.947	.992	S16	.045	no
Ours	Dynamics	.590	.764	.973	S17	.209	yes
Ours	Reverb	.090	.309	1.000	D4	.691	yes
Ours	Amplitude	.270	.469	.929	S5	.460	yes
Ours	Stereo	.120	.261	.860	S4	.599	yes
SONICMASTER	EQ	.640	.690	.980	S8	.290	yes
SONICMASTER	Dynamics	.180	.369	.997	S10	.628	yes
SONICMASTER	Reverb	.340	.468	.999	S13	.531	yes
SONICMASTER	Amplitude	.300	.436	.999	S17	.563	yes
SONICMASTER	Stereo	.100	.290	.973	S8	.683	yes

P.3 FULL RESULTS

Depth. Tables 16 and 17 give AP for every block at t=0. On its own corpus, each model's probe is well above chance from the first block and comes within 0.03 mean AP of its best block early, by S0 for SONICMASTER and by S6 for ours. On the other corpus, both probes start near chance and peak later, at S12 (ours) and S13 (SONICMASTER). The per-group columns show where the cross-corpus gap comes from: SONICMASTER's stereo AP on chains never exceeds 0.212 at any block (base rate 0.113), while our model's stereo AP on the SONICMASTER test set reaches 0.874 at S13 (base rate 0.149).

Table 16: AP by block on the chain test set (t=0). Left: our model, probe trained on its own corpus. Right: SONICMASTER, probe trained on its own corpus (so tested on the other corpus). Bold: best block by mean AP.

	Ours (own corpus)						SONICMASTER (other corpus)					
Block	EQ	Dyn.	Rev.	Amp.	Ste.	mAP	EQ	Dyn.	Rev.	Amp.	Ste.	mAP
D0	.723	.614	.843	.816	.820	.763	.940	.460	.175	.516	.083	.435
D1	.978	.646	.839	.699	.790	.790	.937	.463	.153	.577	.078	.442
D2	.980	.673	.845	.820	.815	.827	.934	.469	.135	.583	.075	.439
D3	.824	.678	.841	.738	.846	.785	.936	.466	.141	.608	.080	.446
D4	.897	.731	.906	.770	.769	.815	.940	.520	.192	.641	.080	.475
D5	.899	.792	.945	.785	.864	.857	.949	.509	.419	.570	.090	.507
S0	.985	.797	.954	.821	.874	.886	.902	.498	.452	.582	.117	.510
S1	.925	.957	.951	.799	.880	.902	.872	.532	.531	.624	.124	.536
S2	.933	.813	.953	.876	.882	.891	.900	.542	.566	.619	.164	.558
S3	.945	.820	.953	.884	.887	.898	.896	.542	.567	.574	.212	.558
S4	.949	.847	.951	.885	.879	.902	.901	.548	.582	.571	.159	.552
S5	.960	.871	.966	.828	.838	.893	.895	.561	.610	.558	.104	.546
S6	.961	.886	.962	.888	.850	.910	.898	.628	.613	.579	.101	.564
S7	.969	.900	.969	.837	.874	.910	.907	.665	.707	.566	.102	.589
S8	.970	.903	.973	.852	.845	.909	.899	.657	.695	.590	.093	.587
S9	.986	.967	.968	.891	.861	**.935**	.901	.649	.709	.576	.107	.588
S10	.971	.904	.969	.893	.863	.920	.909	.659	.725	.585	.108	.597
S11	.988	.906	.968	.839	.872	.914	.907	.632	.725	.532	.123	.584
S12	.988	.896	.970	.834	.844	.906	.911	.631	.737	.557	.127	.593
S13	.987	.971	.978	.843	.848	.925	.908	.644	.764	.551	.130	**.599**
S14	.864	.971	.978	.857	.882	.910	.886	.603	.755	.553	.118	.583
S15	.833	.971	.978	.893	.876	.910	.870	.589	.748	.551	.118	.575
S16	.987	.972	.978	.839	.874	.930	.864	.605	.769	.554	.095	.578
S17	.987	.971	.980	.805	.862	.921	.899	.610	.756	.554	.097	.583
chance	*.770*	*.588*	*.137*	*.259*	*.113*	*.373*	*.770*	*.588*	*.137*	*.259*	*.113*	*.373*

Flow time. Table 18 fixes each pair at its best block at t=0 and moves along the flow path. For all four pairs AP is highest at t=0. For the two own-corpus pairs it changes little up to t=0.5 and drops mostly between t=0.75 and t=1. Our cross-corpus pair falls more steadily, and SONICMASTER's cross-corpus pair is not monotone (0.462 at t=0.5, 0.464 at t=0.75). At t=1 the input is the

Table 17: AP by block on the SONICMASTER test set (t=0). Left: SONICMASTER, probe trained on its own corpus. Right: our model, probe trained on its own corpus (so tested on the other corpus). Bold: best block by mean AP (three blocks tie for SONICMASTER at three decimals).

	SONICMASTER (own corpus)						Ours (other corpus)					
Block	EQ	Dyn.	Rev.	Amp.	Ste.	mAP	EQ	Dyn.	Rev.	Amp.	Ste.	mAP
D0	.666	.443	.784	.912	.981	.757	.643	.196	.424	.532	.212	.401
D1	.737	.495	.841	.948	.981	.800	.766	.205	.505	.469	.446	.479
D2	.782	.572	.868	.947	.981	.830	.771	.216	.529	.651	.407	.515
D3	.789	.562	.875	.949	.986	.832	.714	.217	.553	.475	.427	.477
D4	.818	.701	.950	.960	.993	.884	.756	.241	.727	.638	.228	.518
D5	.860	.883	.992	.985	.993	.943	.775	.398	.844	.627	.603	.649
S0	.903	.886	.996	.990	.992	.953	.795	.366	.869	.620	.625	.655
S1	.926	.918	.996	.990	.993	.965	.765	.411	.893	.578	.692	.668
S2	.931	.916	.996	.990	.992	.965	.732	.381	.896	.640	.582	.646
S3	.935	.916	.996	.990	.993	.966	.748	.368	.893	.644	.570	.644
S4	.937	.916	.997	.988	.993	.966	.782	.408	.905	.643	.360	.620
S5	.940	.932	.998	.986	.993	.970	.775	.546	.906	.600	.507	.667
S6	.948	.955	.998	.986	.994	.976	.782	.545	.917	.653	.451	.670
S7	.954	.959	.998	.985	.993	.978	.762	.554	.930	.646	.431	.665
S8	.955	.959	.998	.983	.994	.978	.740	.547	.941	.680	.412	.664
S9	.955	.959	.998	.982	.993	.978	.796	.540	.938	.678	.319	.654
S10	.957	.958	.998	.981	.992	.977	.761	.527	.944	.651	.365	.650
S11	.956	.962	.997	.983	.994	.979	.804	.545	.955	.622	.654	.716
S12	.960	.964	.997	.995	.993	**.982**	.832	.539	.956	.683	.850	**.772**
S13	.961	.963	.997	.995	.993	**.982**	.814	.579	.956	.594	.874	.764
S14	.960	.962	.997	.994	.993	.981	.770	.558	.956	.631	.747	.733
S15	.960	.961	.997	.997	.993	**.982**	.764	.496	.949	.584	.615	.682
S16	.960	.959	.997	.997	.993	.981	.810	.515	.924	.528	.467	.649
S17	.957	.958	.997	.997	.992	.981	.805	.458	.924	.509	.457	.630
chance	*.584*	*.235*	*.382*	*.286*	*.149*	*.327*	*.584*	*.235*	*.382*	*.286*	*.149*	*.327*

reference, and all 20 group APs are within 0.04 of their base rates. The largest gap is stereo on the SONICMASTER test set: 0.186 for SONICMASTER's probe and 0.180 for ours, against a base rate of 0.149. This gap is small but real, since 20,000 random rankings of that test set give 0.150 $\pm$ 0.004. The input at t=1 depends only on the reference, so in that test set the clips with a stereo label differ from the others even in their reference audio. On the chain test set, both models' stereo probes stay at chance at t=1 (0.114 for ours and 0.119 for SONICMASTER's, base rate 0.113).

P.4 LIMITATIONS OF THE PROBING ANALYSIS

First, a probe shows that information can be read linearly, not that the model relies on it; showing use would need an intervention, for example removing the probe direction and measuring the change in the output. Second, each pair changes the model weights, the probe's training data and the label map together, so the comparison is between training corpora, the variable this paper changes; fitting probes for both models on both corpora would separate the weights from the probe data. Third, each probe is fit once (seed 0), so we report no error bars. Fourth, features are averaged over the clip, so the probes see whether a degradation is present, not when. Fifth, five of our 24 degradation types have no SONICMASTER group and are left out of the labels.

Table 18: AP along the flow path, each pair fixed at its best block at t=0. The input at t is $(1-t)\,\mathbf{z}_{\mathrm{deg}} + t\,\mathbf{z}_{\mathrm{ref}}$, so at t=1 it is the reference master.

t	EQ	Dyn.	Rev.	Amp.	Stereo	mAP	cAP
Ours, own (chain test), block S9							
0	.986	.967	.968	.891	.861	.935	.90
0.25	.968	.959	.970	.879	.869	.929	.88
0.5	.956	.937	.944	.822	.871	.906	.84
0.75	.886	.725	.778	.754	.713	.771	.58
1	.777	.592	.149	.266	.114	.380	.01
chance	.770	.588	.137	.259	.113	.373	0
SONICMASTER, other (chain test), block S13							
0	.908	.644	.764	.551	.130	.599	.38
0.25	.887	.596	.713	.503	.078	.556	.30
0.5	.804	.618	.441	.362	.085	.462	.14
0.75	.933	.631	.349	.303	.102	.464	.22
1	.778	.582	.134	.258	.119	.374	.00
chance	.770	.588	.137	.259	.113	.373	0
SONICMASTER, own (SONICMASTER test), block S15							
0	.960	.961	.997	.997	.993	.982	.97
0.25	.953	.953	.998	.994	.999	.979	.96
0.5	.930	.929	.997	.977	.997	.966	.94
0.75	.832	.739	.985	.885	.983	.885	.81
1	.597	.246	.372	.263	.186	.333	.01
chance	.584	.235	.382	.286	.149	.327	0
Ours, other (SONICMASTER test), block S12							
0	.832	.539	.956	.683	.850	.772	.66
0.25	.830	.415	.961	.697	.837	.748	.63
0.5	.829	.313	.898	.694	.706	.688	.55
0.75	.739	.267	.792	.569	.230	.520	.31
1	.584	.238	.381	.284	.180	.333	.01
chance	.584	.235	.382	.286	.149	.327	0