PRECOMPUTED DECODER JACOBIANS FOR FEW-STEP AUDIO GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Training-free guidance steers a pretrained generator with the gradient of a score on its output. In latent audio generators this score sits behind a decoder, so every guidance step pays for differentiating it, and the common remedy trains a latent head that imitates the decoder's gradient. We observe that for attributes such as loudness the decoder's Jacobian changes little from one latent state to another, and propose Mean-Jacobian guidance (MaJa), which replaces it with a single matrix: the Jacobian averaged over a handful of training states, computed once and reused for every state, prompt and step. Paired with a head that predicts only the attribute value, this takes the decoder out of the sampling loop and makes guidance an order of magnitude cheaper. On few-step samplers the shortcut comes without a loss of accuracy in almost every setting. Although its direction has a cosine of only 0.60 to 0.91 to the decoder gradient, MaJa is level with or ahead of differentiating the decoder in fourteen of sixteen settings across two generators and five attributes, and reduces target error by 31% on average relative to learned latent guidance heads. Closer imitations of the gradient do not help: a head that reproduces it at cosine up to 0.99 controls no better, and the exact gradient through the remaining sampler steps does worse. On these samplers, agreement with the decoder gradient is not what makes a guidance direction effective.

1 INTRODUCTION

Guidance steers a pretrained generator toward targets it was never trained for, by following the gradient of a score on its output (Dhariwal & Nichol, 2021; Bansal et al., 2023; Chung et al., 2023). In latent generative models that score sits behind a decoder, so every guidance step pays to run the decoder and differentiate through it. The common remedy is to train a small latent head that predicts the score directly and imitates the decoder's gradient (Novack et al., 2026).

We start from a simple observation. For attributes such as loudness, the decoder's Jacobian changes little from one latent state to another: the direction that makes the output louder is nearly the same wherever the sampler happens to be. If so, there is no need to differentiate the decoder at every step, nor to train a network to imitate it. We propose **Mea**n-**Ja**cobian guidance (MaJa): compute the Jacobian at a handful of states, average it once, and apply this single matrix at every state, prompt and step (Figure 1). Paired with a head that predicts only the attribute's value, MaJa takes the decoder out of the sampling loop altogether.

This shortcut should come at a price. The average discards everything specific to the current state and points noticeably away from the true gradient; by first-order reasoning it is simply the worse direction. Yet on few-step samplers the price never materialises: MaJa guides as well as differentiating the decoder in nearly every setting we test, and better in several (Table 1). The difference shows up after the step rather than at it. With only a handful of steps, the sampler keeps some components of an update and washes out others, so a direction is worth only as much of it as reaches the output, and more of the averaged update does (Section 6.5). Agreement with the local gradient, the yardstick that learned heads implicitly optimise, is therefore the wrong measure on these samplers; even the exact gradient of the final output does not guide better.

Because the matrix is fixed, MaJa is also simple to use. It needs no training beyond the value head, is built once from a handful of states, and adds only a head evaluation and a matrix product to each guidance step. Two attributes are controlled together by adding their matrices, and clips too long

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

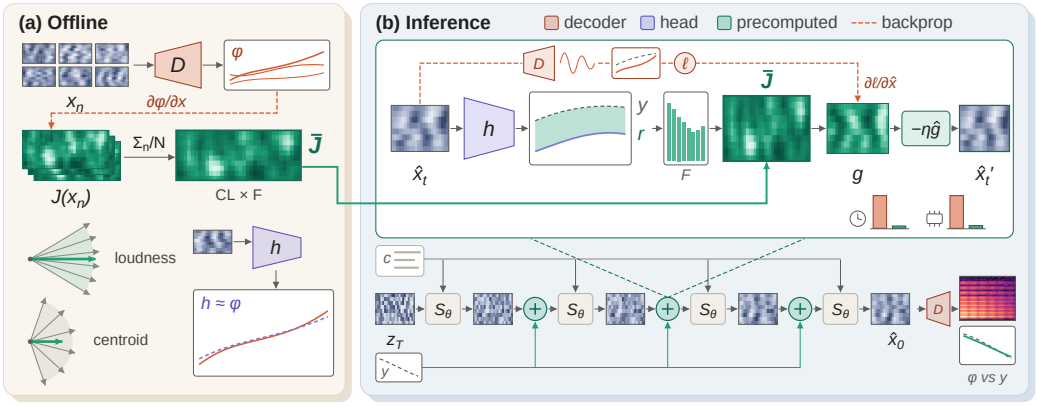

Figure 1: Guidance with a precomputed decoder Jacobian (MaJa). (a) Offline, once per decoder and attribute. The Jacobian of the attribute φ through the decoder D is computed by backpropagation at N states taken from the student's trajectories and averaged into a single matrix $\bar{J}$. A value head h is trained to predict $\varphi(D(x))$ directly from the latent. The fans show per-state gradient directions for two attributes: when they agree (loudness), their mean (teal) keeps nearly full length; when they spread (spectral centroid), it shortens. Our screen measures this spread with three backward passes per state. (b) Inference with a few-step student S_θ conditioned on a prompt c. After every non-terminal step, the head predicts the attribute of the clean estimate $\hat{\mathbf{x}}_t$, the residual r to the target y is mapped through $\bar{J}$ to a latent direction g, and $\hat{\mathbf{x}}_t$ takes a normalized step along it. The standard alternative (dashed, coral) decodes $\hat{\mathbf{x}}_t$ and backpropagates the loss through D at every step. With $\bar{J}$, the decoder leaves the sampling loop and is applied once, to the final output. Bars compare the per-step time and memory of the two paths, drawn to scale (values in Section 6.7).

for the decoder path to fit on a single GPU still fit. Whether an attribute suits averaging can be checked before any guidance is run, with three backward passes per state that measure how much its Jacobian moves. More broadly, few-step generators are now common well beyond audio, and wherever guidance must pass through a decoder the same question arises: which direction a guidance step should follow, and whether imitating the gradient is the right way to find it.

Contributions. (i) We propose MaJa, a state-independent guidance direction given by the decoder Jacobian averaged over a handful of states and computed once, which guides few-step audio generators as well as differentiating the decoder (Section 4, Table 1). (ii) We give a decoder-free form that pairs this matrix with a head predicting only the attribute value, taking the decoder out of the sampling loop and making guidance an order of magnitude cheaper (Sections 4.2 and 6.7). (iii) Experiments on two few-step audio generators and five attributes show that MaJa reduces target error by 31% on average relative to learned latent guidance heads, and by at least 11% in every one of the sixteen settings (Section 6.2).

2 Related Work

Latent audio generation and control. Text-to-audio and text-to-music systems put a diffusion or flow model in the latent space of a neural audio autoencoder, so every waveform-level quantity sits behind a decoder (Liu et al., 2023a; Ghosal et al., 2023; Evans et al., 2024; Tian et al., 2025; Gong et al., 2026)[1], and distillation compresses the sampler to a few steps (Song et al., 2023; Yin et al., 2024b; Saito et al., 2025; Novack et al., 2025b;a). Control is added by training conditional branches (Wu et al., 2024; Devis et al., 2023; Demerlé et al., 2024) or by optimising at inference time against a differentiable objective on the decoded audio (Novack et al., 2024a;b; Koo et al., 2025), which is the pressure that motivates a latent head.

[1]This section is a summary; Appendix A discusses the same work at length, including the lines of work our result does not follow from.

Training-free guidance and the reference it imitates. Guidance adds the gradient of an external score to the sampler update (Dhariwal & Nichol, 2021; Song et al., 2021; Bansal et al., 2023; Yu et al., 2023; Chung et al., 2023), and the loss-based family has been unified (Shen et al., 2024; Ye et al., 2024; Ma et al., 2024). Closest to our setting, LatCH trains latent heads for audio attributes and guides with their input gradients (Novack et al., 2026). Any such head imitates the decoder gradient, and there are already signs that this is not what guidance wants: robust classifiers give better directions than accurate ones (Kawar et al., 2023) and gradient-free schemes reach comparable control (Shenoy et al., 2024). None of this work reports a state-independent baseline.

Linearising the decoder, and what survives distillation. Replacing the decoder by a linear map is familiar from latent inverse problems (Rout et al., 2023; Song et al., 2024; He et al., 2024; Hong et al., 2026; Askari et al., 2025), but there the Jacobian is taken at the current state and the decoder gradient is the reference that linearisation approximates; ours is one matrix frozen before any panel is run and measured against that reference. Separately, distillation is known to break plugged-in control modules (Xiao et al., 2024; Shafi et al., 2026), while sliders and distilled guide modules transfer (Gandikota & Bau, 2025; Hsiao et al., 2024; Shi & Fu, 2025). Sobolev training and Jacobian matching supervise derivatives as well as values (Czarnecki et al., 2017; Srinivas & Fleuret, 2018; Huang et al., 2026); the VJP-matched head we use as a baseline is an instance.

3 Preliminaries

3.1 Guidance with a Direction Source

Write $\hat{\mathbf{x}}_t \in \mathbb{R}^{C\times L}$ for the clean latent estimate at sampler time t (C=64 channels, L latent frames), $\phi : \mathbb{R}^{C\times L} \to \mathbb{R}^F$ for an attribute of the decoded audio on F frames, and $y \in \mathbb{R}^F$ for the target. A *direction source* g supplies a direction from a state and a target. All direction sources are compared with one fixed intervention: with $\ell(\hat{\mathbf{x}}) = \frac{1}{20^2 F}\|\phi(\hat{\mathbf{x}}) - y\|^2$ and $d = g(\hat{\mathbf{x}}_t, y)$,

$$\hat{\mathbf{x}}_t \;\leftarrow\; \hat{\mathbf{x}}_t - \eta\,\|\hat{\mathbf{x}}_t\|_{\mathrm{rms}}\,\frac{d}{\|d\|_{\mathrm{rms}}}, \tag{1}$$

and the sampler receives the corresponding velocity correction. The $1/20$ scales decibels to order one and makes the loss dimensionless; it affects no comparison because the direction is normalised. The *decoder gradient* is $g = \nabla_{\hat{\mathbf{x}}}\ell$, a head gives $g = J_h(\hat{\mathbf{x}})^\top \nabla\ell$, and the *mean Jacobian* $\bar{J}$ of Section 4.1 gives $g = \bar{J}\,r$ with $r = \phi(\hat{\mathbf{x}}) - y$ the residual. One complete run of Equation (1) over a prompt set is a *panel* and η is the *step size*. A matrix of $\bar{J}$'s shape built at some other state or from some other source is an *operator*. Few-step samplers are the setting throughout; Section 6.4 shows why the number of remaining steps matters.

3.2 Latent Guidance Heads

A latent guidance head $h(\mathbf{x}, t)$ is a small network trained by regression to predict the decoded attribute from the latent clean estimate and the time step (Novack et al., 2026). Guidance with a head uses $\nabla_{\mathbf{x}}\ell(h(\mathbf{x}, t), y) = J_h(\mathbf{x})^\top\nabla\ell$, the head's Jacobian applied to the loss residual; the reference is $J_\phi(\mathbf{x})^\top\nabla\ell$ with ϕ the decoded attribute. The regression loss constrains the value along the data manifold; at the standard data budget it leaves the derivative in the directions guidance explores largely unconstrained (Section 6.2). Architecture and training data are given in Section 5.3.

4 Method

4.1 Precomputed Mean Decoder Jacobian

The decoder's Jacobian varies little from state to state for the attributes we control. For the constant direction u_0 in loudness space, decoder VJPs at different training states have mean pairwise cosine 0.84, and the VJP at an unseen state has cosine 0.88 to the mean VJP over training states, while the gradient is nearly orthogonal to the state itself ($|\cos| = 0.09$). Guidance with $\bar{J}$ is MaJa.

We therefore precompute

$$\bar{J} = \frac{1}{N} \sum_{n=1}^{N} J_\phi(\mathbf{x}_n)^\top \tag{2}$$

over N stored training states (212 backward passes each) and guide with $g = \bar{J}\, r$ for the residual $r = \phi(\hat{\mathbf{x}}) - y$ at every state, prompt and step. Nothing is trained. We use $N{=}80$; eight states give control indistinguishable from 80 in a paired test at about a quarter of an hour of GPU time (Appendix B.5). The matrix is 12.8 MB, or 6.4 MB in half precision, and one matrix serves all guidance steps (Appendix B.5). The direction $\bar{J}r$ is a descent direction whenever its per-state deviation is smaller than the state's own gradient, and the screen of Section 4.3 estimates this margin (Appendix C).

4.2 Decoder-Free Guidance

Guiding with $\bar{J}$ still reads the residual r from the decoder, so it saves the backward pass but not the forward one. To remove the decoder call, a head supplies the attribute value, giving a residual estimate $\hat{r} = h(\hat{\mathbf{x}}_t, t) - y$, and the direction is $\bar{J}\,\hat{r}$. Nothing touches the decoder at inference. A head is needed for the value because the affine predictor that $\bar{J}$ itself implies, $b + \bar{J}\mathbf{x}$, is a poor one: the Jacobian's direction transfers between states while its scale does not (numbers in Section 6.7).

4.3 Screening Jacobian State-Dependence

The relevant property of an attribute is how much its decoder Jacobian moves with the state, and three backward passes per state measure it. Store a few dozen clean estimates from the sampler to be guided, at the steps where guidance will act; for each of three fixed directions u_k in attribute space take one decoder backward pass per state; and report two numbers: the mean pairwise cosine between the resulting vectors at different states, and the mean cosine between a held-out state's vector and the mean over the others, which is the direction accuracy a state-independent rule can reach.

Practical procedure. For a new attribute and a new latent generator: (1) run the screen above on a few dozen stored clean estimates; (2) if the two numbers are high, compute the full Jacobian at about eight of those states, average them, and guide with the average; one state alone is usually as good but occasionally fails outright, and a handful removes that. (3) if they are low, neither the matrix nor a head helped on the one such attribute we measured; use the decoder gradient with a small step (Appendix E.5). (4) Either way, compare against MaJa and against a random matrix of the same shape. Appendix B.5 gives the construction in detail and Appendix F the ways we saw it fail.

5 Experimental Setup

5.1 Generators

AudioX (Tian et al., 2025) is a latent rectified-flow (Liu et al., 2023b; Lipman et al., 2023) transformer for text-to-audio generation, sampled natively with 32 Euler steps and classifier-free guidance 6 (Ho & Salimans, 2022); AudioX-Turbo (Tian et al., 2026) is its DMD student (Yin et al., 2024b;a), a stochastic four-step sampler over $t \in \{1, 0.75, 0.5, 0.25\}$ that forms $\hat{\mathbf{x}}_t = \mathbf{z}_t - t\, v_\theta$ at each step and re-noises it. Both decode with the same released VAE; clips are 10 s stereo at 44.1 kHz. ACE-Step 1.5 (Gong et al., 2026) is a 48 kHz stereo music generator whose SFT teacher we run with 48 linear Euler steps and APG guidance 7 and whose Turbo student runs 8 linear Euler steps, so the student's seven non-terminal times are exact grid points of the teacher; both share one Oobleck codec (Kumar et al., 2023; Evans et al., 2025).

5.2 Attributes and Targets

All attributes are frame-wise scalars on one grid (212 frames for AudioX, 497 for ACE): a loudness envelope (RMS in dB), a spectral band contrast ($[1, 8)$ kHz over $[50, 1000)$ Hz power, dB) that we call brightness, a spectral centroid in $20 \log_{10}$Hz, a spectral flux that we call onset strength, and a high-band loudness on $[4, 16)$ kHz (Appendix B.2). Targets are the attribute of the unguided clip

after a signal-processing operation: an exact ± 6 dB gain (*gain*), a linear $\mp 6 \to \pm 6$ dB ramp (*fade*), a per-frame gain imposing the loudness envelope of another clip (*transfer*), a loudness-matched ± 6 dB shelf equaliser (*shelf*), a shelf whose amount is solved by bisection so that the mean centroid of the valid frames moves by the requested amount, capped at ± 18 dB of shelf (*shift*), the transient component of the log spectrogram scaled by $1 + a$ with $a = 2.5$ or $a = -0.95$ and resynthesised with the original phase (*shape*), and a zero-phase ± 6 dB gain on $[4, 16)$ kHz (*band gain*). Signal processing attains every target exactly, so the "signal-processing reference" row is the floor.

5.3 Direction Sources Compared

- **Decoder gradient.** $g = \nabla_{\hat{\mathbf{x}}} \ell$ through the decoder: the reference every head imitates, and the only direction source used to choose the step size.
- **Fixed direction.** Move every state along the mean training VJP of u_0, signed by the mean residual; it tests whether a single direction suffices.
- **Regression head.** A head $h(\mathbf{x}, t)$ is a 6.9M-parameter bidirectional transformer over the latent frames with a Fourier time token, the learned latent guidance head of prior work (the LatCH design in spirit), trained for 10k AdamW updates on *direct* labels: the attribute of the decoded estimate it sees. Training states are 512 trajectories per generator (128 prompts over eight instrument classes, six for training and two held out, four seeds) with $\mathbf{z}_t, \hat{\mathbf{x}}_t$ stored at the four common times.
- **Equal-cost regression head.** The same head given every advantage that does not involve a gradient label: twelve extra generation seeds per training prompt (four times the student trajectories), the precision and batch of the VJP-matched fit, and a wall-clock-matched update budget (Table 46).
- **VJP-matched head.** A head trained to reproduce the decoder gradient as closely as we could, used to test whether reproducing the reference brings better control. The full Jacobian ($F \times CL$, here $212 \times 15{,}104$) is too large to store per state, but guidance only ever applies it to residual vectors, so it suffices to match its action on a few directions. We fix $K+1$ directions in attribute space, $u_0 = F^{-1/2}\mathbf{1}$ and $u_1, \ldots, u_K$ random unit vectors, precompute for every stored training state the decoder VJPs $\mathbf{v}_k(\mathbf{x}) = \partial(u_k \cdot \phi(\mathbf{x}))/\partial \mathbf{x}$, one decoder backward pass each, and train the head with

$$\mathcal{L} = \left\| \tfrac{1}{20}(h(\mathbf{x}, t) - \phi(\mathbf{x})) \right\|^2 + \lambda \, \frac{1}{K+1} \sum_k \left(1 - \cos\left(\frac{\partial \, (u_k \cdot h(\mathbf{x}, t))}{\partial \mathbf{x}}, \, \mathbf{v}_k(\mathbf{x})\right)\right), \quad (3)$$

  where the head's VJP is obtained with a retained graph so that the loss back-propagates into its parameters (Appendix B.1). $\lambda = 1$ and $K = 2$ unless stated. This is Sobolev training (Czarnecki et al., 2017) and Jacobian matching (Srinivas & Fleuret, 2018) applied to the one function whose derivative is what is consumed; the labels are free because the decoder is available at training time. It costs one decoder backward pass per stored state once, about thirteen times a regression fit to train, and nothing at inference.

5.4 Protocol and Statistics

The *calibrated multi-step* probe applies Equation (1) at every non-terminal step with a step size chosen from $\{1, 2, 3, 5\}$ % by the decoder gradient alone on eight development prompts and then frozen (AudioX loudness 3 %, brightness 2 %; ACE 3 %). Calibrating on the decoder favours the decoder; no other direction source gets its own step size. No panel prompt appears in head training. For AudioX the *new-prompt* panel has sixteen prompts, two per instrument class over all eight classes of the training manifest, so four of them are on the two classes held out from head training, with two unseen generation seeds each, and the *unseen-instrument* panel has eight prompts on eight instruments that appear nowhere in the training manifest, one seed each; the eight development prompts only choose the step size. For ACE the new-prompt panel has twelve prompts with two seeds each and the unseen-class panel eight prompts with one seed each. Every run asserts bit-exact reproduction of the unguided output under a zero-strength hook and of guided outputs under repetition, and every panel is re-scored from saved waveforms by independent NumPy/SciPy code (Appendix B.1). Comparisons are paired by prompt after averaging target directions and seeds within a prompt, with a 95 % cluster bootstrap over instrument clusters (10,000 draws, seed fixed in advance); new-prompt panels resample eight (AudioX) or six (ACE) instrument clusters and unseen-instrument panels

resample each prompt as its own cluster, so intervals on the latter are wide for a structural reason. A comparison whose interval contains zero is called level or a tie throughout; Appendix E.2 adds a 48-prompt panel with a pre-registered ± 0.2 dB equivalence test. We apply no correction for multiple comparisons (Appendix G).

6 RESULTS

6.1 MAIN RESULTS

Table 1 gives every direction source's error on the sixteen calibrated panels and, in its last three columns, the paired difference to the decoder gradient; Figure 3 in the appendix plots the errors. The fixed direction suffices for a uniform gain target (2.03 dB, the decoder gradient's own error) and comes close on the brightness shelf (1.96 against 1.87), but fails once the target varies along the clip (transfer 5.71, fade 4.18 against 3.26 unguided) and is behind MaJa on every other panel. MaJa fails nowhere, reaching 1.64 dB on gain where the decoder gradient reaches 2.03, 3.64 on envelope transfer against 3.71, 2.17 on fade against 2.64, 0.97 on unseen instruments against 0.93, and 1.87 on brightness against 2.38, with a cosine of only 0.60 to 0.91 to that gradient across all panels. Four of the paired differences have intervals that exclude zero in its favour: gain on new prompts ($+0.39$ dB), the fade target ($+0.47$) and, marginally, envelope transfer ($+0.07$, CI $[+0.00, +0.15]$) and high-band loudness on new prompts ($+0.16$, CI $[+0.00, +0.36]$). Two have intervals that exclude zero against it, both ACE new-prompt panels: loudness by 0.14 dB, marginally (CI $[-0.25, -0.00]$), and brightness by 0.34. On the remaining ten, including brightness ($+0.51$), the centroid ($+0.05$) and every unseen-instrument panel, the interval contains zero, so a precomputed matrix matches differentiating the decoder there. At 48 prompts the brightness tie becomes a lead of 0.42 dB $[0.16, 0.71]$ and the loudness lead is 0.28 $[0.20, 0.36]$ (Appendix E.2).

6.2 COMPARISON WITH LEARNED HEADS

Table 14 reports the mean cosine between a head's input gradient and the decoder gradient at stored t=0.75 states. A loudness head trained on student states predicts the envelope of those states to 0.55 dB and points at cosine 0.78; the teacher head on teacher states predicts to 0.74 dB and points at 0.63; for brightness the own-domain cosines are 0.49 to 0.60. Prediction accuracy and direction accuracy come apart: refitting the brightness head on the student's states improves its prediction five-fold ($4.05 \rightarrow 0.8$ dB) while the cosine moves only $0.49 \rightarrow 0.60$, and inside the sampler the refit head controls no better than the teacher's.

Against these heads MaJa is ahead on every one of the sixteen panels, by 0.19 to 2.77 dB (Table 1). Averaged over the sixteen settings, MaJa reduces target error by 31% relative to the regression head, and by no less than 11% in any setting. On the eleven panels where the comparison is paired the interval excludes zero every time and at most one prompt per panel favours the head (Table 11).

A head trained on the decoder's VJPs closes the alignment gap. Guidance-time alignment rises to 0.89 to 0.99 against 0.38 to 0.84 for regression heads (Table 15), out of sample on prompts and seeds no head saw; control matches the decoder gradient to within 0.1 dB on loudness and closes most of the brightness gap; and the head stops caring which generator's states it was trained on, the teacher-trained VJP head controlling the student within 0.02 to 0.08 dB of the student-trained one where the regression gap is 0.5 to 1.3 dB. It does not control better than MaJa: MaJa is ahead of it on brightness (0.69 dB), high-band loudness (0.60 and 0.67), onset strength on new prompts (0.12) and ACE brightness on unseen classes (0.63), and level with it elsewhere (Tables 11 and 13). On 48 prompts it is equivalent to the decoder gradient within 0.2 dB on loudness (Appendix E.2); the gradient through the remaining sampler steps, the other candidate reference, is worse than both (2.37 dB on gain, 2.96 on fade) at its own development-optimal step size (Appendix E.1).

The equal-cost regression head recovers most of the alignment without a gradient label (0.95 to 0.97 at guidance time) and matches the VJP head's control on loudness and brightness to within 0.2 dB (Table 46). The VJP loss buys sample efficiency, not a capability the regression loss lacks. Against MaJa this head is behind on ten of fourteen panels, by 0.10 (onset strength) to 2.03 dB (ACE brightness, unseen classes), and level on the other four (Table 12).

Table 1: Main result: target error after the calibrated multi-step probe, **lower is better**. Units are dB except for the spectral centroid ($20\log_{10}$Hz). Rows are attribute and target family; † marks the unseen-instrument panels (eight prompts on eight instruments absent from the training manifest), the other rows are the sixteen new prompts over all eight classes of the training manifest. Columns are direction sources: no guidance; the decoder gradient, which alone chooses the step size on eight development prompts; the fixed direction; MaJa, the mean Jacobian $\bar{J}$; the regression head; the regression head given four times the trajectories at matched compute; and the VJP-matched head. The last three columns pair MaJa with the decoder gradient by prompt: the mean per-prompt benefit of MaJa (positive favours MaJa), the prompts with a valid score for both and how many of them favour MaJa strictly, and a 95 % cluster bootstrap interval (10,000 draws; Section 5.4). Every other paired comparison is in Tables 11 to 13.

attribute, target	unguided	decoder	fixed dir.	MaJa	reg.	reg. 4×	VJP	benefit	prompts	95 % CI
AudioX-Turbo										
loudness, gain	5.59	2.03	2.03	1.64	2.35	2.02	2.09	+0.39	13/16	[+0.27,+0.50]
loudness, gain†	6.00	0.93	1.50	0.97	1.73	1.05	0.96	-0.05	3/8	[-0.14,+0.04]
loudness, transfer	9.64	3.71	5.71	3.64	4.52	3.64	3.67	+0.07	9/16	[+0.00,+0.15]
loudness, fade	3.26	2.64	4.18	2.17	2.45	2.47	2.71	+0.47	16/16	[+0.38,+0.56]
brightness, shelf	4.82	2.38	1.96	1.87	3.02	2.40	2.57	+0.51	6/9	[-0.04,+1.32]
brightness, shelf†	4.68	1.40	1.59	1.55	2.79	1.80	1.69	-0.15	2/7	[-0.44,+0.24]
centroid, shift	3.08	1.31	1.48	1.25	1.58	1.31	1.34	+0.05	5/9	[-0.05,+0.14]
centroid, shift†	2.91	1.12	1.54	1.35	1.55	1.27	1.30	-0.22	2/8	[-0.46,+0.06]
onset, shape	1.29	0.88	1.28	0.87	1.09	0.97	0.99	+0.01	5/9	[-0.02,+0.05]
onset, shape†	1.62	1.01	1.60	1.11	1.30	1.14	1.15	-0.10	5/8	[-0.35,+0.03]
high band, band gain	5.99	1.62	2.09	1.47	2.39	2.09	2.06	+0.16	6/9	[+0.00,+0.36]
high band, band gain†	6.00	1.85	2.59	1.80	2.86	2.48	2.46	+0.05	6/8	[-0.23,+0.27]
ACE-Step Turbo										
loudness, gain	6.00	2.54	3.29	2.67	3.68	3.44	2.59	-0.14	4/12	[-0.25,-0.00]
loudness, gain†	6.00	3.06	4.00	2.92	3.76	3.71	3.07	+0.15	4/8	[-0.03,+0.33]
brightness, shelf	5.17	1.46	2.48	1.80	3.97	3.58	2.02	-0.34	2/12	[-0.57,-0.12]
brightness, shelf†	4.98	1.65	2.66	1.61	4.38	3.64	2.24	+0.04	3/8	[-0.32,+0.38]

Capacity does not help. Heads of 27.6M and 41.3M parameters reach 2.70 and 2.69 dB where the 6.9M head reaches 2.35 and $\bar{J}$ 1.64, with the small head ahead of both large ones on every prompt (Table 31). Regression heads are also fragile to distribution shift: a teacher-trained loudness head goes from 0.74 to 2.76 dB error on the student's states, and from 0.24 to 3.22 on ACE, because distillation moves the clean estimates while leaving the noisy states nearly in place (Fréchet distances 7.22 against 1.37 at t=0.75; Figure 4, Table 17). VJP-matched heads generalise to cotangent directions never used in training (0.95 to 0.96 on unseen-instrument states, 0.87 on five-frame impulses; Table 16), and three initialisations and $\lambda \in \{0.1, 1, 10\}$ change control by at most 0.1 dB (Table 47). Across all of it, no learned head is ahead of $\bar{J}$ by an interval that excludes zero on any of the sixteen panels.

6.3 Ablations on the Mean Jacobian

Three controls say what the matrix has to be (Table 18). A random Gaussian matrix with $\bar{J}$'s row norms does nothing (5.63 dB against 5.59 unguided, cosine -0.004). The brightness operator applied to a loudness target is worse than not guiding (6.11), and the full grid over four attributes shows the same for every other mismatched pair except one: brightness and the spectral centroid measure the same spectral tilt and their operators are interchangeable (Appendix D.3). So the matrix must belong to the direction in signal space the attribute measures. A *linear head trained by regression*, which has exactly $\bar{J}$'s functional form and the regression heads' data, reaches only 2.42 dB at cosine 0.34, losing to $\bar{J}$ by 0.78 dB: the Jacobian label, not state-independence, is the active ingredient, and supervising a head with the same Jacobians is the other way to use it (Section 5.3). Operators built at an unseen instrument's state or at a teacher state work as well as the training-state average (1.68 and 1.62 to 1.78 against 1.64). Latents encoded from real recordings are less reliable: the median over eight recordings is 1.93 dB, but two exceed 2.3 and one fails at 4.50, so generated states are the safer source. Seven of eight single generated states match the 80-state average while one fails outright, so averaging costs 0.05 dB and buys reliability (Table 19).

6.4 Conditions for the Advantage

Table 31 varies one thing at a time. The advantage needs a few-step sampler. Inside the 32-step teacher every direction source lands within 0.12 dB of every other, a correction is washed out ($5.69 \rightarrow 5.39$ dB even with the decoder gradient, against $6.00 \rightarrow 3.25$ for the student), and the teacher is indifferent to which head produced it ($|\Delta| \leq 0.06$ dB), so MaJa neither wins nor loses where later steps can re-project the perturbation. It needs the direction repeated and repeated *late*: with guidance at one or two of the three non-terminal steps MaJa is level with the decoder gradient, and restricting it to the two late steps gives it 0.52 dB, where restricting it to the two early steps gives nothing. It decays as the target moves away from the clip, gaining 0.36, 0.39 and 0.06 dB at ± 3, ± 6 and ± 12 dB, which is what a fixed linearisation should do. And a cheaper decode is not an alternative route: decoding half or a quarter of the clip destroys the direction (5.03 and 5.21 dB), and bf16 reproduces the decoder gradient while saving only $42\,\%$ of its time.

Sensitivity to step-size calibration. The step size is calibrated on the decoder gradient, and MaJa prefers a larger one. On brightness the decoder gradient peaks at $2\,\%$ and degrades to 3.10 dB at $3\,\%$ while $\bar{J}$ improves to 1.49; on the centroid $\bar{J}$ still improves at $5\,\%$ where the gradient has turned. At each method's own best step size the gaps are 1.61 and 0.44 dB rather than 0.51 and 0.05. Table 1 uses the shared step size throughout; these optima are a post hoc observation. Letting every source choose its own step size, or replacing the fixed step by a DPS-style update, leaves the ordering unchanged (Appendices E.3 and E.4).

6.5 Mechanism Analysis

What is consistent across direction sources is what reaches the output: Table 38 projects the injected displacement onto the difference between guided and unguided final latents and finds 1.63 to 1.80 of it surviving for $\bar{J}$ against 1.29 to 1.60 for the decoder gradient. The decoder gradient is also more concentrated in latent space (kurtosis 112 to 170 against 58 to 85), but that is not the lever. Clipping, low-passing, whitening or taking the sign of it leaves control unchanged or worse (2.02, 2.07, 2.11 and 2.96 dB against 2.03; Table 21). Nor is the lever averaging as such: smoothing the decoder gradient in state space, by averaging it over two or eight small random perturbations of the clean estimate, gives 2.05 and 2.06 dB, no closer to $\bar{J}$'s 1.64 than the unsmoothed gradient is (Table 40). $\bar{J}$'s direction is therefore not a smoothed decoder gradient, in coordinate space or in state space. Across the four panels where we measured survival, MaJa has both the highest survival and the lowest error every time, but survival does not order the other direction sources, so we read it as a necessary condition for the advantage rather than an explanation of it (Figure 6). Appendix C makes the condition precise: if the remaining steps discard the part of the decoder gradient orthogonal to $\bar{J}r$, a normalised step along $\bar{J}r$ has $1/\cos\theta$ times the effect at the output; the condition is sufficient, not an explanation, since by the same first-order picture the end-to-end gradient would be best and it is the worst.

6.6 Screen Results

Table 42 has two blocks that must not be mixed: the lower covers five attributes on one common set of 48 AudioX student states at $t{=}0.75$, which is what a practitioner would measure; the upper compares model families using the stored VJP labels across three guidance times and both generators. In the lower block, frame-wise loudness is almost a fixed linear map (0.84 pairwise, 0.92 to the mean); high-band loudness, brightness, the spectral centroid and onset strength follow in that order, down to 0.20 and 0.43 for onset strength. In the upper block the families separate: AudioX loudness 0.92/0.95, AudioX brightness 0.80/0.88, ACE loudness 0.72/0.81. MaJa stays level with or ahead of the decoder gradient at brightness, which the lower block puts at 0.59/0.78 and where it is ahead by 0.5 dB, and at the spectral centroid (0.35/0.56), where it still edges it on new prompts. What the screen tracks, within one generator, is a *trend*: the advantage decays along it and turns slightly negative on unseen instruments for the gain and shelf targets, though not for the three time-varying targets of Appendix D.4. On the five signal-processing attributes no head wins anywhere. The screen does identify where the matrix fails: a learned CLAP-contrast attribute screens at 0.005/0.02, and there MaJa, the heads and a random matrix all leave the error near its unguided value while the decoder gradient at a $1\,\%$ step still works (Appendix E.5).

6.7 Decoder-Free Guidance, Cost, and Composition

With a VJP-matched head's value and $\bar{J}$'s direction the decoder-free rule reaches 1.65 dB on new prompts, indistinguishable from the 1.64 that $\bar{J}$ achieves with the exact decoder value and better than the decoder gradient (2.03) or either head alone (2.02 to 2.09); with an equal-cost regression head's value, 1.71 (Table 44). On unseen instruments the ordering reverses slightly (1.01 against the head's 0.96), and the reason is visible in the same table: there the head's *value* is least accurate, and a good direction cannot repair a wrong residual. The affine predictor $b + \bar{J}\mathbf{x}$ confirms that a head is needed for the value: it reaches 4.51 dB on loudness, 5.40 on brightness and 3.75 on the centroid, against 0.54 to 1.16 for the heads (Appendix D).

Cost and guards. One guidance gradient costs the head 33 ms and 0.9 GB against the decoder's 465 ms and 11 GB on a 10-s AudioX clip, plus a 12.8 MB matrix; on ACE the decoder path needs 27 GB at 20 s and does not fit a 48-GB GPU at 40 s, while the head stays at 6.6 GB up to 120 s (Table 45). Table 49 guards against the explanation that a direction source wins by damaging the audio more. Gain-invariant waveform correlation to the unguided clip stays above 0.99 for every direction source on every panel, and the Fréchet distance (Kilgour et al., 2019) between CLAP (Elizalde et al., 2023) embeddings of the guided and unguided sets agrees to three decimals across the decoder gradient, the VJP heads and MaJa on each panel. One number is not reassuring: on the brightness panel every guided set sits at 0.345 against 0.020 for the signal-processing reference, identical across the four direction sources, so it affects no comparison but means we cannot call those interventions artefact-free (Appendix G). We make no perceptual claim in this paper.

Two attributes at once. Two precomputed matrices add (Table 48). Asking for a ± 6 dB gain and a ± 6 dB shelf together, the summed matrices reach 1.48 dB on loudness and 1.67 on brightness where differentiating the decoder for both terms reaches 1.68 and 1.56, and the two regression heads' summed gradients reach 2.52 and 2.86. On unseen instruments the pattern repeats (1.41/1.75 against 1.54/1.48 and 2.56/2.84). Neither attribute is bought at the other's expense, and the summed rule still needs no decoder backward pass.

7 Limitations and Conclusion

Two teacher/student pairs, one modality, five frame-level scalar attributes, minimal interventions fixed in advance, and effects measured by attribute error and a distribution guard rather than by listening tests (Appendix G). The mean Jacobian is built from the student's own training states for one decoder at one clip length; a different codec needs its own, a few GPU-hours once. The mechanism is only partly characterised (Section 6.5).

We do not claim learned heads are useless. We claim that for the attributes and models measured here no head we could train, including one given four times the data at matched compute, is ahead of a matrix with no parameters by an interval that excludes zero on any of the sixteen panels, and that for frame-level audio attributes on few-step students the implicit objective of the guidance-head literature is not the right one. Nor do we claim MaJa beats differentiating the decoder everywhere: over sixteen panels it is clearly ahead on two, marginally ahead on two, level on ten, and behind on two, one of them marginally. The recommendation: measure how state-dependent your attribute's decoder Jacobian is, try the precomputed average first, and train a head only if it loses.

Reproducibility Statement

Every number in this paper is produced by a script in the released code from stored artefacts, and every panel is re-scored from its saved audio by an independent implementation of the attributes (Appendix B.1). The release records, for each fitted head, each precomputed Jacobian, each panel and each diagnostic, the command that regenerates it, and states which random seeds were fixed in advance (head initialisations, direction draws, bootstrap resampling, state subsampling for the operators). The protocols for the probe, the targets, the scoring and the calibration were written before the corresponding runs; the one exception, the scoring rule for the spectral centroid, is documented in Appendix B.1 along with when it was changed and what the superseded numbers were.

REFERENCES

Hossein Askari, Yadan Luo, Hongfu Sun, and Fred Roosta. Latent refinement via flow matching for training-free linear inverse problem solving. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. arXiv preprint arXiv:2302.07121, 2023. URL https://arxiv.org/abs/2302.07121.

Hyungjin Chung, Jeongsol Kim, Michael T. Mccann, Marc L. Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *International Conference on Learning Representations (ICLR)*, 2023.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Wojciech Marian Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Świrszcz, and Razvan Pascanu. Sobolev training for neural networks. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

Sanjit Dandapanthula and Nicholas M. Boffi. Are we really tilting? The mechanics of reward guidance in flow and diffusion models. arXiv preprint arXiv:2606.02884, 2026. URL https://arxiv.org/abs/2606.02884.

Nils Demerlé, Philippe Esling, Guillaume Doras, and David Genova. Combining audio control and style transfer using latent diffusion. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, 2024.

Ninon Devis, Nils Demerlé, Sarah Nabi, David Genova, and Philippe Esling. Continuous descriptor-based control for deep audio synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. CLAP: Learning audio concepts from natural language supervision. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

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

Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12652–12665. PMLR, 2024.

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

Rohit Gandikota and David Bau. Distilling diversity and control in diffusion models. arXiv preprint arXiv:2503.10637, 2025. URL https://arxiv.org/abs/2503.10637.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23)*, pp. 3590–3598. Association for Computing Machinery, 2023. doi: 10.1145/3581783.3612348.

Junmin Gong, Yulin Song, Wenxiao Zhao, Sen Wang, Shengyuan Xu, Jing Guo, and Xuerui Yang. ACE-Step 1.5: Pushing the boundaries of open-source music generation. arXiv preprint arXiv:2602.00744, 2026. URL https://arxiv.org/abs/2602.00744. Technical report.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *International Conference on Learning Representations (ICLR)*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598, 2022. URL https://arxiv.org/abs/2207.12598.

Yeobin Hong, Suhyeon Lee, Hyungjin Chung, and Jong Chul Ye. InverseCrafter: Efficient video ReCapture as a latent domain inverse problem. In *European Conference on Computer Vision (ECCV)*, 2026.

Yi-Ting Hsiao, Siavash Khodadadeh, Kevin Duarte, Wei-An Lin, Hui Qu, Mingi Kwon, and Ratheesh Kalarot. Plug-and-play diffusion distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Huayang Huang, Ruoyu Wang, Jinhui Zhao, Wei Deng, Daiguo Zhou, Jian Luan, Yu Wu, and Ye Zhu. Restoring initial noise sensitivity in text-to-image distillation via geometric alignment. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

Bahjat Kawar, Roy Ganz, and Michael Elad. Enhancing diffusion-based image synthesis with robust classifier guidance. *Transactions on Machine Learning Research*, 2023.

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

Junghyun Koo, Gordon Wichern, François G. Germain, Sameer Khurana, and Jonathan Le Roux. SMITIN: Self-monitored inference-time INtervention for generative music transformers. *IEEE Open Journal of Signal Processing*, 6:266–275, 2025. doi: 10.1109/OJSP.2025.3534686.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

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. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 2023a.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023b.

Jiajun Ma, Tianyang Hu, Wenjia Wang, and Jiacheng Sun. Elucidating the design space of classifier-guided diffusion generation. In *International Conference on Learning Representations (ICLR)*, 2024.

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. DITTO: Diffusion inference-time T-optimization for music generation. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38426–38447, 2024a.

Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. DITTO-2: Distilled diffusion inference-time T-optimization for music generation. In *Proceedings of the 25th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 874–881, 2024b.

Zachary Novack, Zach Evans, Zack Zukowski, Josiah Taylor, CJ Carr, Julian Parker, Adnan Al-Sinan, Gian Marco Iodice, Julian McAuley, Taylor Berg-Kirkpatrick, and Jordi Pons. Fast text-to-audio generation with adversarial post-training. arXiv preprint arXiv:2505.08175, 2025a. URL https://arxiv.org/abs/2505.08175.

Zachary Novack, Ge Zhu, Jonah Casebeer, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. Presto! distilling steps and layers for accelerating music generation. In *International Conference on Learning Representations (ICLR)*, 2025b. URL https://openreview.net/forum?id=Gj5JTAwdoy.

Zachary Novack, Zack Zukowski, CJ Carr, Julian Parker, Zach Evans, Josiah Taylor, Taylor Berg-Kirkpatrick, Julian McAuley, and Jordi Pons. Low-resource guidance for controllable latent audio diffusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Kushagra Pandey, Farrin Marouf Sofian, Jan Niklas Groeneveld, Felix Draxler, and Stephan Mandt. Hierarchical variational policies for reward-guided diffusion. arXiv preprint arXiv:2605.21661, 2026. URL https://arxiv.org/abs/2605.21661.

Litu Rout, Negin Raoof, Giannis Daras, Constantine Caramanis, Alexandros G. Dimakis, and Sanjay Shakkottai. Solving linear inverse problems provably via posterior sampling with latent diffusion models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023.

Koichi Saito, Dongjun Kim, Takashi Shibuya, Chieh-Hsin Lai, Zhi Zhong, Yuhta Takida, and Yuki Mitsufuji. SoundCTM: Unifying score-based and consistency models for full-band text-to-sound generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision (ECCV)*, 2024.

Abdullah Al Shafi, Kazi Saeed Alam, Sk Imran Hossain, and Engelbert Mephu Nguifo. DASH: Dual-branch score distillation for guidance-calibrated compact diffusion models. arXiv preprint arXiv:2606.00798, 2026. URL https://arxiv.org/abs/2606.00798.

Yifei Shen, Xinyang Jiang, Yezhen Wang, Yifan Yang, Dongqi Han, and Dongsheng Li. Understanding and improving training-free loss-based diffusion guidance. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Rahul Shenoy, Zhihong Pan, Kaushik Balakrishnan, Qisen Cheng, Yongmoon Jeon, Heejune Yang, and Jaewon Kim. Gradient-free classifier guidance for diffusion model sampling. arXiv preprint arXiv:2411.15393, 2024. URL https://arxiv.org/abs/2411.15393.

Liang Shi and Yun Fu. ExpertGen: Training-free expert guidance for controllable text-to-face generation. arXiv preprint arXiv:2505.17256, 2025. URL https://arxiv.org/abs/2505.17256.

Bowen Song, Soo Min Kwon, Zecheng Zhang, Xinyu Hu, Qing Qu, and Liyue Shen. Solving inverse problems with latent diffusion models via hard data consistency. In *International Conference on Learning Representations (ICLR)*, 2024.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2021.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

Suraj Srinivas and Francois Fleuret. Knowledge transfer with Jacobian matching. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. AudioX: A unified framework for anything-to-audio generation. arXiv preprint arXiv:2503.10522, 2025. URL https://arxiv.org/abs/2503.10522.

Zeyue Tian, Lei Ke, Zhaoyang Liu, Ruibin Yuan, Liumeng Xue, Yujiu Yang, Weijia Chen, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. AudioX-Turbo: A unified framework for efficient anything-to-audio generation. arXiv preprint arXiv:2606.12555, 2026. URL https://arxiv.org/abs/2606.12555.

Junnuo Wang. Audio palette: A diffusion transformer with multi-signal conditioning for controllable foley synthesis. arXiv preprint arXiv:2510.12175, 2025. URL https://arxiv.org/abs/2510.12175.

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan. Music ControlNet: Multiple time-varying controls for music generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2692–2703, 2024. doi: 10.1109/TASLP.2024.3399026.

Jie Xiao, Kai Zhu, Han Zhang, Zhiheng Liu, Yujun Shen, Yu Liu, Xueyang Fu, and Zheng-Jun Zha. CCM: Real-time controllable visual content creation using text-to-image consistency models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Zou, and Stefano Ermon. TFG: Unified training-free guidance for diffusion models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T. Freeman. Improved distribution matching distillation for fast image synthesis. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T. Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024b.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. FreeDoM: Training-free energy-guided conditional diffusion model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and Models*. Springer, 3 edition, 2007.

A Extended Related Work

This section is the full version of Section 2.

Latent audio generators and their decoders. Text-to-audio and text-to-music systems place a diffusion or flow model in the latent space of a neural audio autoencoder, so that every waveform-level quantity is separated from the sampler by a decoder (Liu et al., 2023a; Ghosal et al., 2023; Evans et al., 2024; 2025; Tian et al., 2025); a parallel line generates discrete codec tokens instead (Copet et al., 2023). The autoencoder is usually a residual-vector-quantised convolutional codec or its continuous variant (Kumar et al., 2023; Evans et al., 2025), which is what makes the decoder both differentiable and expensive: the Oobleck autoencoder used by ACE-Step 1.5 (Gong et al., 2026) is the Descript codec's encoder and decoder without the quantiser. Rectified flow and flow matching (Liu et al., 2023b; Lipman et al., 2023) supply the sampler both of our families use, and distillation compresses it to a few steps: consistency models (Song et al., 2023), distribution matching (Yin et al., 2024b;a), adversarial objectives (Sauer et al., 2024; Novack et al., 2025a) and their audio-specific forms (Saito et al., 2025; Novack et al., 2025b). Few-step samplers are the regime our result is about, and they are now the deployed regime.

Controlling a pretrained audio generator. Control is added either by training conditional branches and adapters (Wu et al., 2024; Devis et al., 2023; Demerlé et al., 2024; Wang, 2025) or by optimising at inference time against a differentiable objective on the decoded audio (Novack et al., 2024a;b), which is where differentiable signal processing supplies the objectives (Engel et al., 2020). SMITIN fits logistic probes to attention activations and steers with them while monitoring the effect (Koo et al., 2025); SoundCTM shows training-free intensity control inside a few-step audio model (Saito et al., 2025). DITTO optimises the initial noise through the full sampler and DITTO-2 distils the sampler to make that affordable (Novack et al., 2024a;b), which is the same pressure that motivates a latent head: the decoded objective is what one wants and the decoder is what one cannot afford.

Training-free guidance. Guidance steers a frozen sampler by adding the gradient of an external score to the update (Dhariwal & Nichol, 2021; Song et al., 2021; Ho & Salimans, 2022; Bansal et al., 2023; Yu et al., 2023; Chung et al., 2023), and the loss-based family has been unified and analysed (Shen et al., 2024; Ye et al., 2024; Ma et al., 2024). The bias introduced by plugging an external estimate into a frozen denoiser has been studied on its own (Pandey et al., 2026; Dandapanthula & Boffi, 2026); the error we isolate is different, being a direction error in the direction source rather than a bias in the reward model. Closest to our setting, LatCH trains latent heads for audio attributes and guides with their input gradients (Novack et al., 2026). None of this work reports a state-independent baseline, which our results suggest is the first thing to try.

Is the decoder gradient the right target? A guidance head is trained, explicitly or implicitly, to reproduce it, so the question matters. Kawar et al. (2023) found that adversarially robust classifiers give better guidance gradients than accurate ones, and gradient-free schemes reach comparable control without differentiating the scorer at all (Shenoy et al., 2024). We arrive at the same conclusion from the opposite direction: we make a head match the decoder gradient almost exactly, at a cosine of 0.89 to 0.99, and find that it does not control better than a fixed matrix at a cosine of 0.60 to 0.91.

Linearising the decoder. Replacing the decoder by a linear map is familiar from latent inverse problems, where the measurement is defined in signal space while sampling happens in latent space (Rout et al., 2023; Song et al., 2024). MPGD shortcuts the autoencoder's gradient with a local approximation (He et al., 2024), recent solvers rewrite the decoder-induced measurement so the problem becomes linear in the latent and no decoder backward pass is needed (Hong et al., 2026), and others derive a posterior covariance for latent guidance from a local linearisation of the same kind (Askari et al., 2025). Two things separate our result from that line. Their Jacobian is taken *at the current state*, so it is state-dependent and must be recomputed as sampling proceeds; ours is one matrix, averaged over a few dozen training states and frozen before any of the panels are run. And their purpose is to save computation while conceding accuracy, with the decoder gradient as the reference that linearisation approximates. We find the opposite empirically: the cruder object controls at least as well, which nothing in the efficiency literature implies.

Supervising a network's Jacobian. Sobolev training and Jacobian matching fit a student's derivatives as well as its values (Czarnecki et al., 2017; Srinivas & Fleuret, 2018); we apply them in Section 5.3 to the one function whose derivative is what gets consumed, with labels that are free because the teacher is a differentiable decoder available at training time. GAD matches teacher and student Jacobian-vector products for the generator itself rather than for a direction source (Huang et al., 2026).

What survives distillation. Image ControlNets give only coarse control on a consistency model (Xiao et al., 2024), sliders and LoRAs do transfer (Gandikota & Bau, 2025), guide modules can be distilled along with the sampler (Hsiao et al., 2024), classifier-free guidance does not survive naive distillation (Shafi et al., 2026), and ExpertGen exploits the in-distribution intermediate predictions of a consistency model (Shi & Fu, 2025). Our Section 6.2 adds one item to that list and Section 4.1 removes it again. A fixed operator has no training domain, so it cannot suffer the failure that motivated this line of work: a head fitted on the teacher's trajectories multiplies its error three- to thirteen-fold on the student's states, while an operator built from the teacher's own states guides the student as well as one built from the student's (1.78 and 1.62 dB against 1.64), provided the states it is built from have the attribute well defined. That proviso is the one way we saw the construction fail: the loudness feature floors frame power, so frames of the source state at the floor give exactly zero Jacobian rows, and two of the four teacher states we tried kept only 105 and 72 of 212 rows, against 212 for student states. Counting non-zero rows is a one-line check on a candidate state. What distillation breaks is therefore not guidance and not the attribute, both of which are properties of the shared decoder; it is the Jacobian that a regression head happens to fit on the trajectories it was shown, which is tied to those states and does not survive their displacement. Training the head on the decoder's VJPs repairs it (Section 5.3) because those labels are a property of the decoder rather than of the generator, and taking the operator out of the head removes the question entirely.

B EXPERIMENTAL DETAILS

B.1 PROTOCOL

Models and runtime. Teacher and student checkpoints, the VAE and the T5 text encoder are the public AudioX/AudioX-Turbo release (Tian et al., 2026). We verify that the separately released VAE weights are loaded into both models (365 tensors, value-equal) and record the number of matched backbone tensors (1384) for every model load. All generation uses the upstream sampling code without modification; guidance is injected through a forward hook on the backbone that replaces the returned velocity at the intervention step. Every run asserts that the number of backbone calls equals the schedule length (4 or 32), that a zero-strength hook reproduces the unguided latent and waveform bit-exactly, that a repeated guided run reproduces the guided output bit-exactly, and that the CUDA random state after a guided run equals that after the unguided run, so that the student's stochastic re-noising is identical across methods. Float32 matmul precision is set to highest; TF32 and cuDNN benchmarking are disabled; the FP16 reduced-precision reduction flag is fixed to the value used by the native sampler of each model.

Trajectory collection. For each of 128 prompts (eight instruments $\times$ sixteen performance descriptions) and four generation seeds we store, for both generators, the noisy latent $\mathbf{z}_t$ and the clean estimate $\hat{\mathbf{x}}_t = \mathbf{z}_t - t\, v_\theta(\mathbf{z}_t, t)$ at $t \in \{1, 0.75, 0.5, 0.25\}$, the terminal latent and its decoded waveform. Direct labels are the attribute measured on the decoded clean estimate itself. Six instruments (piano, acoustic guitar, violin, drum kit, flute, saxophone) form the training split and two (synthesizer, cello) the validation split; all seeds of a prompt stay in one split. The equal-cost baseline adds twelve further generation seeds for every training prompt of the student (1152 extra trajectories, same capture).

Heads. A head takes the 64×236 latent estimate, normalises channels with source-training statistics, prepends a Fourier time token and applies an 8-layer, width-256 bidirectional transformer with rotary position embeddings and SwiGLU feed-forward (multiplier 3), followed by a linear projection and a length-4 average pool to 212 frames; 6,921,216 parameters. Regression training: 10,000 AdamW updates, batch 32, learning rate 10^{-4} with 500 warm-up steps, weight decay 10^{-3}, gradient-norm clip 1, bf16 autocast, MSE on the attribute divided by 20. For band contrast the loss and all metrics are restricted to frames that are valid in the label (frame power $\geq 10^{-6}$ and both band shares $\geq 10^{-4}$). No early stopping or checkpoint selection; the final checkpoint is used. Initialisation seeds 29101, 29103 and 29107 were fixed in advance. Terminal-only heads use the same recipe on terminal latents with the time input fixed to zero.

VJP labels and VJP-matched training. Directions: $u_0 = F^{-1/2}\mathbf{1}$ and $K{=}2$ random unit vectors drawn once (seed 29301) and shared by all heads and both generators. For every stored training state at $t \in \{0.75, 0.5, 0.25\}$ (states at $t{=}1$ are not used for guidance) we compute $\mathbf{v}_k(\mathbf{x}) = \partial(u_k \cdot \phi_{\mathrm{dec}}(\mathbf{x}))/\partial\mathbf{x}$ by one backward pass through the decoder per direction (for band contrast, masked by the label's validity mask) and store it in fp16 ($3 \times 64 \times 236$ per state). Training uses the same optimiser, schedule, clipping and 10,000-update budget as regression, but batch 16 in fp32 without autocast, because the loss needs the gradient of the head's own VJP (double backward), and the attention layers are switched to the math kernel of scaled dot-product attention, whose backward is twice differentiable. Each step computes the prediction, $K{+}1$ input gradients with a retained graph, the cosine of each to its stored label, and back-propagates $\mathcal{L}$. $\lambda \in \{0.1, 1, 10\}$ and $K \in \{0, 2\}$ were the only variants trained; all are reported. Training does not diverge or stall in any run; the batch cosine starts near zero at initialisation (-0.01 on the first batch), passes 0.78 by the hundredth update and 0.90 by the two hundredth, and ends at 0.99.

Attributes. The five attributes are loose engineering correlates of perceptual quantities such as loudness and sharpness (Zwicker & Fastl, 2007); we treat them as signal measurements and make no perceptual claim. RMS envelope: stereo power averaged over channels, frame 8192, hop 2048, floor -80 dB. Band contrast: periodic Hann window of the same frame and hop, one-sided power summed over $[50, 1000)$ Hz and $[1000, 8000)$ Hz, $10 \log_{10}$ of the high/low ratio with a 10^{-6} relative floor. Both implementations were cross-checked against independent NumPy/SciPy code (maximum difference 10^{-5} dB) and against analytic values on synthetic two-tone signals.

Targets. Gain: the RMS envelope of the unguided clip after an exact scalar gain of ± 6 dB. Fade: after a linear ramp from ∓ 6 to ± 6 dB over the clip, applied per frame by interpolating the gain between frame centres. Envelope transfer: the envelope of one of two stored student clips of other training prompts (an acoustic-guitar and a flute crescendo), imposed on the unguided clip by a per-frame gain, clipped to ± 12 dB. Brightness: the band contrast of the unguided clip after a zero-phase shelf equaliser with ± 6 dB high-frequency gain ($s(f) = \frac{1}{2}(1 + \tanh(\log_2(f/1000)/0.5))$) followed by a scalar that restores the clip's overall RMS. All targets are attainable by signal processing (the "signal-processing reference" has zero error by construction). A case is eligible for brightness scoring if at least 106 of its 212 frames are valid; ineligible cases are reported, never replaced.

Precomputed Jacobians. The mean Jacobian is $\bar{J} = \frac{1}{N}\sum_n J_\phi(\mathbf{x}_n)^\top$ with J_ϕ the Jacobian of the decoded attribute: for AudioX, $F{=}212$ backward passes through the VAE per state, over $N{=}80$ student training states, 32 at each guidance time $t \in \{0.75, 0.5, 0.25\}$, stored as a $212 \times 64 \times 236$ fp32 array (12.8 MB); for ACE, $F{=}497$ backward passes per state over $N{=}48$ states, eight at each of six of the seven guidance times. States are drawn from the head-training split with a fixed seed, so the baseline sees exactly the data the heads see. The single-state ablation stores $J_\phi(\mathbf{x})^\top$ at one such state (seed 29801). Guidance applies $\bar{J}\, r$ to the residual r and then follows the same normalisation and step as every other direction source. Two ways of obtaining r are reported: from the exact decoder (an upper bound on any state-independent scheme, and the version used in the main comparison) and from a head's prediction, which removes the decoder from inference entirely.

Jacobian state-dependence screen. For each attribute and each of the three fixed directions u_k we compute the decoder VJP at 48 student training states at $t{=}0.75$ (three backward passes per state) and report the mean pairwise cosine across states and the mean cosine between a held-out state's VJP and the mean VJP over the other half. The five attributes share one frame grid (frame 8192, hop 2048, 212 frames): loudness, band contrast, spectral centroid (power-weighted mean frequency over $[50, 16000)$ Hz, reported as $20\log_{10}$ Hz so the decibel-like scale matches the shared loss normalisation), spectral flux (mean positive frame-to-frame increase of the log power spectrum, an onset strength) and high-band loudness ($[4, 16)$ kHz). The same statistics computed from the stored VJP labels give the cross-family comparison.

Local versus final diagnostic. Each panel records the decoder-measured target loss immediately before and after every intervention. For the decoder gradient and for the state-independent baselines these are directly comparable, because both read the loss from the decoder and both move the estimate by the same fraction of its RMS; Table 39 reports the mean fraction of the local loss removed at the first step and over all steps, next to the target error of the finished clip.

Intervention and scoring. At an intervention step the clean estimate is detached, the loss $\frac{1}{|M|}\sum_{i\in M}((\phi(\hat{\mathbf{x}})_i - y_i)/20)^2$ is differentiated with respect to $\hat{\mathbf{x}}$ (through the VAE for ϕ_{dec}, through the head otherwise), the gradient is scaled to unit RMS, and $\hat{\mathbf{x}}$ is moved by η of its RMS against it; the sampler receives $v + (\hat{\mathbf{x}} - \hat{\mathbf{x}}')/t$. The state-independent baselines use the exact decoder value of the current estimate for the residual (they are therefore upper bounds on any state-independent scheme): the fixed direction is the mean of the stored $\mathbf{v}_0$ over student training states at $t{=}0.75$, signed by the mean residual; the mean Jacobian is the mean of the full decoder Jacobian transpose (212 backward passes per state) over 80 student training states, 32 at each guidance time, applied to the residual. Target error is the RMSE in dB between the attribute of the final decoded clip and the target over the 212 frames (brightness: over the baseline-valid frames, with a 40 dB penalty for frames whose RMS falls below -60 dB after guidance). Aggregation averages the target directions and generation seeds within a prompt, then averages prompts with equal weight. Bootstrap intervals resample instrument clusters (10,000 draws, seed 29201) and are descriptive for the fixed panels. Every panel is re-scored from the saved waveforms by an independent script (separate NumPy/SciPy feature code reading the audio back from disk); the largest discrepancy across all panels is below 10^{-3} dB.

A protocol correction, and when it was made. The band-contrast score penalises frames whose level falls below -60 dB after guidance by 40 dB, because the high/low ratio is undefined on silence. When the spectral centroid was added as a third attribute it inherited that rule and the shelf solver's ± 48 dB search range. Both were wrong for it: 17 of 32 downward targets saturated at -48 dB, which strips the treble, and the penalty then fired on frames of the exact signal-processing target itself,

giving the reference a score of 1.82 dB where it should be 0 by construction, and reversing orderings among the direction sources. We noticed this when reading the first centroid panel, that is, after seeing results. The rule for this attribute was then fixed and declared before re-running: the shelf search is capped at ± 18 dB with the achieved shift recorded per case, and the centroid, which is defined wherever a frame carries power, is scored without the activity penalty and with the same validity mask. The superseded panels are archived rather than deleted and their numbers are in the released record; under the penalised rule they read unguided 3.47, decoder gradient 3.55, mean Jacobian 2.97, best regression head 2.52, reference 1.82, and the conclusion that the decoder gradient is not the best direction is the same on both rules and on the unpenalised metric of the first run. Nothing else in the paper uses a rule set after seeing results.

Calibrated multi-step probe. The correction is applied at every non-terminal step of the student ($t \in \{0.75, 0.5, 0.25\}$ for AudioX, $\{0.875, \ldots, 0.125\}$ for ACE) with one per-step strength η chosen from $\{1, 2, 3, 5\}\,\%$ as the value that minimises the *decoder gradient's* mean target error on the development prompts for the gain (loudness) and shelf (brightness) targets, then frozen before any head is run (AudioX loudness 3 %, brightness 2 %; ACE 3 %). The fade and transfer panels reuse the loudness strength without re-calibration. No head, prompt, seed, target or direction was selected after reading results.

Second family. ACE-Step 1.5 SFT (teacher) and Turbo (student) are the public releases; both are loaded from the released safetensors, share the Oobleck codec (48 kHz stereo, 64-channel latent at 25 Hz; the Descript audio codec's convolutional encoder and decoder without the quantiser) and the Qwen3 text encoder, and are conditioned through the upstream text-to-music path with instrumental lyrics and a 20-s duration. The teacher runs 48 linear Euler steps with APG guidance 7 (the release default is 50; 48 makes the student's seven non-terminal times $\{0.875, \ldots, 0.125\}$ exact grid points), the student its native 8-step linear schedule. States are captured by a forward hook on the Turbo decoder and, for the teacher, by wrapping the module's guided-velocity function so that the clean estimate uses the velocity the Euler update consumes. The loudness feature uses frame 7680 and hop 1920 (25 Hz, aligned with the latent) and the head's 497 output frames; VJP labels use the same three directions in the 497-frame space; everything else is as for AudioX. Prompts: eight ensemble classes $\times$ eight performance descriptions (six classes for training, two for validation) $\times$ seeds 28551/28557; twelve new prompts $\times$ seeds 28561/28563; eight prompts over unseen classes with seed 28571; the development captions are the eight platform captions.

Distribution guard. FAD is the Fréchet distance between Gaussian fits of CLAP audio embeddings (unit-normalised, 48 kHz mono) of a method's guided clips and of the unguided clips of the same panel, both directions pooled; CLAP text-audio cosine uses the panel prompt.

Cost measurement. Table 45 times one loss-to-gradient evaluation (five repetitions after one warm-up, CUDA-synchronised) at batch 1 in fp32 on one RTX A6000 with peak allocated memory reset before timing; ACE latents longer than 20 s are tiled copies of a stored 20-s state, so the decoder's cost is representative and the head's is exact for its architecture.

The affine predictor implied by the mean Jacobian. If the decoder's Jacobian were state-independent, the attribute would be an affine function of the latent, $\phi(\mathbf{x}) \approx b + \bar{J}\mathbf{x}$ with b estimated as the mean of $\phi(\mathbf{x}_n) - \bar{J}\mathbf{x}_n$ over the same states that build $\bar{J}$. That predictor has no learned parameters and is far worse than a head. On student states, familiar and then unseen instruments: loudness 4.51 and 6.77 dB against the student head's 0.55 and 2.89 and the VJP head's 0.60 and 0.95; band contrast 5.40 and 4.30 against 1.16 and 3.14; spectral centroid 3.75 and 4.19 against 0.54 and 1.74. On teacher states it is worse again (8.33 dB for loudness). So the operator captures the local direction of the decoder's response without capturing its value, which is consistent with the cosines of Table 19 and is why the decoder-free rule needs an accurate head for the residual.

B.2 Attributes and Targets

Figure 2 sketches the seven target families defined below. Every attribute is a function of the decoded waveform on one common frame grid, so that one intervention, one loss and one step size serve all of them. This section gives the definitions in full. All five are loose engineering correlates of

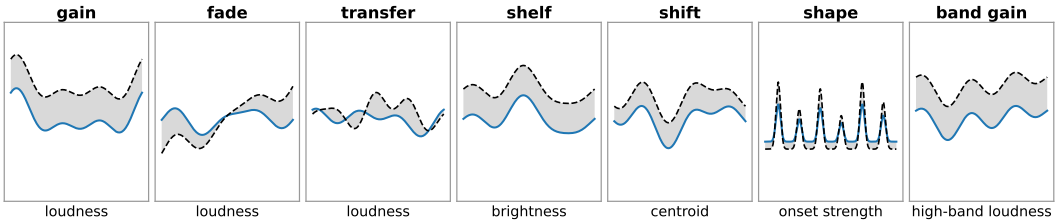

Figure 2: Target families, schematic. Each panel shows an attribute of the unguided clip over time (solid) and the target derived from it by a signal-processing operation (dashed); the shaded area is what guidance has to close. Only one sign of each target is shown.

perceptual quantities such as loudness and sharpness (Zwicker & Fastl, 2007); we treat them as signal measurements and make no perceptual claim.

The frame grid. A decoded clip $x \in \mathbb{R}^{2\times T}$ is cut into frames of $N_{\mathrm{f}} = 8192$ samples with hop 2048, giving $F = 212$ frames for a ten-second AudioX clip at 44.1 kHz and $F = 497$ for a twenty-second ACE clip at 48 kHz. Write $X[f, k]$ for the one-sided power of frame f at bin k under a periodic Hann window w, averaged over the two channels and normalised by $2/(N_{\mathrm{f}} \sum_i w_i^2)$, and write $P[f]$ for the mean square of the raw samples of frame f. A frame is *valid* when $P[f] \geq 10^{-6}$ and its band powers are positive; invalid frames are excluded from scoring and reported, never replaced.

Loudness (RMS in dB). $\phi_{\mathrm{rms}}[f] = 10 \log_{10} \max(P[f], 10^{-8})$, that is stereo power averaged over channels with a floor at -80 dB. Every frame is valid by construction. The floor matters for the mean Jacobian: a state whose frames sit at the floor contributes exactly zero Jacobian rows, which is the failure mode described in Appendix F.

Brightness (band contrast, dB). With $B_{\mathrm{lo}} = [50, 1000)$ Hz and $B_{\mathrm{hi}} = [1000, 8000)$ Hz,

$$\phi_{\mathrm{bright}}[f] = 10 \log_{10} \frac{\sum_{k \in B_{\mathrm{hi}}} X[f, k]}{\sum_{k \in B_{\mathrm{lo}}} X[f, k]}, \tag{4}$$

where a relative floor of 10^{-6} times the total of the two bands is added to both sums before the logarithm. A frame is valid when its raw power clears 10^{-6} and each band carries at least 10^{-4} of the two bands' total; a case is scored only if at least half of its 212 frames are valid. Frames that fall below -60 dB after guidance are penalised by 40 dB rather than scored, because band contrast is undefined on silence; Appendix B.1 records when that rule was fixed.

Spectral centroid ($20 \log_{10}$Hz). Over $[50, 16000)$ Hz,

$$c[f] = \frac{\sum_{k \in B} X[f, k]\, \nu_k}{\sum_{k \in B} X[f, k]}, \qquad \phi_{\mathrm{cent}}[f] = 20 \log_{10} \max(c[f], 20), \tag{5}$$

with ν_k the bin frequency in hertz. The logarithm is what makes the scale decibel-like, so the shared loss normalisation and the shared step sizes apply without change; it also means a target expressed as a shift in $20 \log_{10}$Hz is a ratio of frequencies rather than a difference.

Onset strength (spectral flux, dB). With $D[f, k] = 10 \log_{10} \max(X[f, k], 10^{-20})$,

$$\phi_{\mathrm{flux}}[f] = \frac{1}{K} \sum_{k} \max\big(D[f, k] - D[f{-}1, k],\, 0\big), \tag{6}$$

and $\phi_{\mathrm{flux}}[0] = 0$. Only increases count, which is what makes it an onset measure. This is the most state-dependent attribute in the screen and therefore the test of whether a state-independent matrix finally fails.

High-band loudness (dB). $\phi_{\mathrm{high}}[f] = 10 \log_{10} \max\big(\sum_{k \in [4,16)\mathrm{kHz}} X[f, k], 10^{-20}\big)$. It sits between loudness and brightness in the screen, and a band-limited gain attains any target for it exactly, which is why we added it as high-band loudness.

B.2.1 TARGETS

A target is always the attribute of a waveform that signal processing actually produced, so the "signal-processing reference" row of every panel has zero error by construction and the floor of every panel is zero rather than an estimate.

Gain. ϕ_{rms} of the unguided clip after an exact scalar gain of ± 6 dB, and ± 3 or ± 12 dB in the magnitude sweep.

Fade. After a per-frame gain interpolated linearly between ∓ 6 and ± 6 dB across the clip.

Envelope transfer. The loudness envelope of one of two stored student clips of other training prompts, an acoustic-guitar crescendo and a flute crescendo, imposed on the unguided clip by a per-frame gain clipped to ± 12 dB. The initial error is 9.64 dB, the largest in the paper.

Tremolo, step and random envelope. Three further per-frame gains, all on loudness: $6 \sin(8\pi f/F)$ dB, which changes sign eight times; -6 dB on the first half of the clip and $+6$ on the second; and two band-limited random envelopes, each a fixed cosine series of six terms rescaled to ± 6 dB with the seed written in the code.

Shelf (brightness). A zero-phase shelf equaliser with response $s(\nu) = \frac{1}{2}\big(1 + \tanh(\log_2(\nu/1000)/0.5)\big)$ scaled to ± 6 dB of high-frequency gain, followed by a scalar that restores the clip's overall loudness so that the target changes brightness and not level.

Centroid shift. The same shelf family, with the shelf amount solved by bisection so that the mean centroid of the valid frames moves by exactly the requested amount in $20 \log_{10}$Hz, capped at ± 18 dB of shelf. Solving rather than fixing the shelf is what keeps the target attainable: the map from shelf gain to centroid shift depends on the clip.

Transient shaping (onset strength). The log magnitude spectrogram is split into a part smoothed over five frames along time and the residual; the residual is scaled by $1 + a$ and the signal is resynthesised with its original phase. $a > 0$ sharpens onsets and raises onset strength, $a < 0$ softens them, and we use $a = 2.5$ and $a = -0.95$. The attribute is measured on the resynthesised waveform, so the target remains exactly attainable even though the operation is not a per-frame gain.

Band gain (high-band loudness). A zero-phase gain of ± 6 dB on $[4, 16)$ kHz, which changes the high-band loudness by exactly that amount and leaves the rest of the spectrum alone.

B.3 INDEPENDENT RE-SCORING

Every panel is re-scored from its saved waveforms by a second implementation of the attributes written against NumPy and SciPy rather than PyTorch, reading the audio back from disk. The auditor recomputes the target, the attribute and the error for every row and compares them with what the panel recorded. Tolerances are per attribute and are set from the measured conditioning of the logarithms involved: 10^{-3} dB for loudness, brightness and high-band loudness, 10^{-2} for onset strength, where the clamped logarithm of a small power is worst conditioned. Observed disagreements are far below those tolerances, at 3×10^{-6} dB on the panels re-audited most recently. The auditor refuses to score an attribute it does not implement, which it did not always do; Appendix F records what that cost us.

B.4 HEAD ARCHITECTURE AND TRAINING

Every head in this paper is the same network with the same optimiser; only the training labels, the training states and the budget change. This section states the configuration so that the head route can be reproduced or improved on, and so that a reader can judge whether we crippled it. The inventory of fitted heads with their budgets and prediction errors is Table 71.

Architecture. A head takes the latent clean estimate as a $64 \times L$ array, normalises each channel with the mean and standard deviation of the training states of its own source (standard deviations floored at 0.1), and transposes it to a sequence of L tokens of width 64. A time token is prepended: the scalar t together with $\sin$ and $\cos$ of $t \cdot 2^j \pi$ for $j = 0, \ldots, 7$, so seventeen numbers, passed through a two-layer perceptron with SiLU. The trunk is a bidirectional continuous transformer with width 256, depth 8, head width 64, rotary positional embeddings and a feed-forward multiplier of 3, projecting to one scalar per token. The output is smoothed by an average pool of width 4 along time and mapped to decibels by an affine $20v - 30$, which puts the initialisation in the right range for a loudness envelope. That is $6\,921\,216$ parameters. The capacity ablation widens the trunk to 512 and deepens it to 8 or 12 layers, giving 27.6M and 41.3M.

Optimisation. AdamW, learning rate 10^{-4} with a linear warm-up over 500 steps and no decay afterwards, weight decay 10^{-3}, batch 32 states drawn with replacement, gradient norm clipped to 1, bf16 autocast with the loss accumulated in fp32, $10\,000$ updates. The loss is the mean squared error of the attribute divided by 20, so it is the same normalisation the guidance loss uses. Training asserts that the loss and the gradient norm are finite at every step and records the batch error every hundred steps. Seeds 29101, 29103 and 29107 give the three initialisations of the ablation.

VJP-matched training. The Sobolev term needs the head's own vector-Jacobian product, so the forward pass is run with a retained graph and the cosine between the head's VJP and the stored decoder VJP is differentiated again through the head's parameters. Two details matter in practice. The attention implementation must be the math kernel rather than a fused one, because the fused kernels do not support double backward. And the whole fit runs in fp32 with autocast disabled, because the double backward is not reliable under bf16 autocast. Those two changes are why a VJP-matched fit takes 107 minutes where the regression fit takes eight, on the same data and the same number of updates; the labels themselves cost one decoder backward pass per state and per direction, about an hour once, and are reused by every fit.

Training states. 512 trajectories per generator: 128 prompts covering eight instrument classes, six of which are training classes, with four generation seeds each. For every trajectory the noisy state $\mathbf{z}_t$ and the clean estimate $\hat{\mathbf{x}}_t$ are stored at the four common times, and the attribute is measured on the decoded clean estimate, so labels are *direct* rather than terminal. The equal-cost fits add twelve further generation seeds per prompt, that is 1536 extra trajectories, and are trained in fp32 with batch 16 up to a wall-clock-matched number of updates.

What we did not try. We did not search the architecture, the learning rate or the schedule; we did not try convolutional or state-space trunks; we did not try per-attribute architectures; and we did not try training a single head on several attributes at once. A reader who believes a better head exists is probably right, and the claim of the paper survives that: the comparison is against a matrix with no parameters that costs a quarter of an hour to build, so the interesting question is not whether a head can be improved but whether it needs to exist.

B.5 Constructing the Mean Jacobian

The recipe in Section 4.1 says to compute the decoder Jacobian at about eight states and average. This section is the evidence for the three numbers in that sentence, and for what a practitioner can throw away afterwards.

Number of states. Table 2 pools single-state Jacobians in increasing numbers. Eight states are indistinguishable from the 80-state operator in a paired test, and two already are. The reason to prefer a handful over one is not the mean but the tail: Table 3 reports all eight single-state operators individually, and one of the eight fails outright while the other seven match the average. Averaging that failing state with one other repairs it, which is what makes a small pool the right recommendation. Table 4 builds two operators from non-overlapping draws of training states; they agree with each other and with the mean Jacobian used throughout the paper, so the construction does not depend on which states happen to be drawn.

Rank and precision. Table 5 truncates the $212 \times 15{,}104$ matrix by singular value decomposition and stores it at half precision. Half precision is lossless to three decimal places, so the shipped artefact is 6.4 MB rather than 12.8. Rank 128 is identical to the full matrix, rank 32 is still ahead of the decoder gradient at 1.96 MB, and rank 8 falls behind it. The rank the mean Jacobian needs is larger than its singular-value spectrum suggests: rank 8 already carries a substantial fraction of the Frobenius energy and still loses 0.65 dB to the full matrix, so the directions that matter for guidance are not the directions that carry the energy.

Source states. Table 6 is the full provenance panel in one place: operators built at training states, at unseen-instrument states, at teacher states and at latents encoded from real recordings, plus the random and wrong-attribute controls and the linear head that has $\bar{J}$'s functional form. The ordering is the one Section 4.1 reports, and the only construction that is unreliable is the one built from real recorded audio.

Timing of the steps. Table 7 restricts guidance to the two steps closest to the output and should be read against the early-steps-only block of Table 31. Late steps are where the mean Jacobian gains and early steps are where it is level, which is consistent with the survival measurement of Table 38: a perturbation injected late has fewer remaining steps in which to be re-projected, so the difference between two directions survives to the output instead of being repaired.

Table 2: How many states the operator needs (AudioX loudness, gain, new prompts). Pooling eight single-state Jacobians is indistinguishable from the 80-state operator, and two already are; eight states cost about 0.2 GPU-hours.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.03	[1.83, 2.27]	16
MaJa ($\bar{J}$)	1.64	[1.41, 1.90]	16
8-state pool	1.63	[1.42, 1.86]	16
4-state pool	1.64	[1.47, 1.82]	16
2-state pool	1.68	[1.56, 1.81]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.39	[-0.50, -0.27]	3/16
decoder gradient − 8-state pool	-0.40	[-0.49, -0.30]	2/16
decoder gradient − 4-state pool	-0.39	[-0.48, -0.28]	3/16
decoder gradient − 2-state pool	-0.35	[-0.48, -0.23]	5/16
decoder gradient − unguided	+3.55	[+3.16, +3.88]	16/16
MaJa ($\bar{J}$) − unguided	+3.94	[+3.59, +4.22]	16/16
8-state pool − unguided	+3.96	[+3.61, +4.22]	16/16
4-state pool − unguided	+3.94	[+3.62, +4.20]	16/16
2-state pool − unguided	+3.90	[+3.60, +4.15]	16/16

Table 3: Single-state operators on the gain target (AudioX loudness, new prompts): seven of eight match the training-state average and one fails outright, which is the reason the recipe averages a handful of states.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.03	[1.83, 2.27]	16
MaJa ($\bar{J}$)	1.59	[1.44, 1.72]	16
VJP-matched head (student)	2.09	[1.89, 2.32]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.44	[-0.59, -0.30]	5/16
decoder gradient − VJP-matched head (student)	+0.05	[+0.01, +0.10]	9/16
decoder gradient − unguided	+3.55	[+3.16, +3.88]	16/16
MaJa ($\bar{J}$) − unguided	+4.00	[+3.70, +4.25]	16/16
VJP-matched head (student) − unguided	+3.50	[+3.11, +3.80]	16/16

Table 4: Two disjoint pools of states (AudioX loudness, gain, new prompts): operators built from non-overlapping draws of training states agree with each other and with the 80-state operator, so the construction is not sensitive to which states are drawn.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
pool A (disjoint states)	1.61	[1.41, 1.83]	16
pool B (disjoint states)	1.61	[1.39, 1.86]	16
paired differences			
pool A (disjoint states) − unguided	+3.97	[+3.63, +4.23]	16/16
pool B (disjoint states) − unguided	+3.98	[+3.63, +4.25]	16/16

Table 5: Storage and rank of the operator (AudioX loudness, gain, new prompts, calibrated multi-step probe). Half precision is lossless to three decimals; truncating the $212 \times 15{,}104$ matrix by SVD costs nothing at rank 128 and is still ahead of the exact decoder gradient at rank 32, which is a 2 MB artefact.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.03	[1.83, 2.27]	16
MaJa ($\bar{J}$)	1.64	[1.41, 1.90]	16
fp16 storage	1.64	[1.41, 1.90]	16
rank 8	2.29	[2.06, 2.52]	16
rank 32	1.72	[1.47, 1.99]	16
rank 128	1.64	[1.41, 1.90]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.39	[-0.50, -0.27]	3/16
decoder gradient − fp16 storage	-0.39	[-0.50, -0.27]	3/16
decoder gradient − rank 8	+0.25	[+0.07, +0.43]	10/16
decoder gradient − rank 32	-0.32	[-0.44, -0.19]	4/16
decoder gradient − rank 128	-0.39	[-0.50, -0.27]	3/16
decoder gradient − unguided	+3.55	[+3.16, +3.88]	16/16
MaJa ($\bar{J}$) − unguided	+3.94	[+3.59, +4.22]	16/16
fp16 storage − unguided	+3.94	[+3.59, +4.22]	16/16
rank 8 − unguided	+3.30	[+3.00, +3.55]	16/16
rank 32 − unguided	+3.87	[+3.50, +4.16]	16/16
rank 128 − unguided	+3.94	[+3.59, +4.22]	16/16

Table 6: Every fixed matrix we tried, in one panel (AudioX loudness, gain, new prompts). Only the matrix changes.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.03	[1.83, 2.27]	16
MaJa ($\bar{J}$)	1.64	[1.41, 1.90]	16
random matrix, matched row norms	5.63	[5.42, 5.84]	16
built at an unseen instrument	1.68	[1.49, 1.89]	16
built at a real recording	4.50	[4.31, 4.69]	16
single state 0	1.59	[1.44, 1.72]	16
single state 1	1.59	[1.42, 1.76]	16
single state 2	4.38	[4.16, 4.61]	16
single state 3	1.54	[1.36, 1.72]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.39	[-0.50, -0.27]	3/16
decoder gradient − random matrix, matched row norms	+3.60	[+3.21, +3.92]	16/16
decoder gradient − built at an unseen instrument	-0.35	[-0.44, -0.26]	4/16
decoder gradient − built at a real recording	+2.47	[+2.12, +2.76]	16/16
decoder gradient − single state 0	-0.44	[-0.59, -0.30]	5/16
decoder gradient − single state 1	-0.44	[-0.55, -0.33]	3/16
decoder gradient − single state 2	+2.34	[+1.99, +2.66]	16/16
decoder gradient − single state 3	-0.49	[-0.59, -0.38]	3/16
decoder gradient − unguided	+3.55	[+3.16, +3.88]	16/16
MaJa ($\bar{J}$) − unguided	+3.94	[+3.59, +4.22]	16/16
random matrix, matched row norms − unguided	-0.04	[-0.05, -0.03]	1/16
built at an unseen instrument − unguided	+3.91	[+3.59, +4.16]	16/16
built at a real recording − unguided	+1.09	[+1.02, +1.15]	16/16
single state 0 − unguided	+4.00	[+3.70, +4.25]	16/16
single state 1 − unguided	+4.00	[+3.68, +4.25]	16/16
single state 2 − unguided	+1.21	[+1.12, +1.30]	16/16
single state 3 − unguided	+4.05	[+3.73, +4.30]	16/16

Table 7: Guidance restricted to the two late steps (AudioX loudness, gain, new prompts), against the early-steps-only panel of Table 31. The operator's advantage lives in the steps closest to the output.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.10	[1.98, 2.21]	16
MaJa ($\bar{J}$)	1.58	[1.47, 1.68]	16
regression head (teacher and student states)	2.29	[2.16, 2.42]	16
VJP-matched head (student)	2.10	[1.99, 2.19]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.52	[-0.57, -0.47]	1/16
decoder gradient − regression head (teacher and student states)	+0.19	[+0.12, +0.28]	10/16
decoder gradient − VJP-matched head (student)	-0.01	[-0.03, +0.02]	9/16
decoder gradient − unguided	+3.48	[+3.21, +3.75]	16/16
MaJa ($\bar{J}$) − unguided	+4.01	[+3.75, +4.25]	16/16
regression head (teacher and student states) − unguided	+3.29	[+3.02, +3.57]	16/16
VJP-matched head (student) − unguided	+3.49	[+3.23, +3.74]	16/16

B.6 PROMPTS

Four disjoint prompt sets are used, and no set is reused across roles. Head training and every precomputed operator see only the 128 prompts of Table 8, and only the six instrument classes marked as training splits there. The new-prompt panel (Table 9) was written after all heads and operators were fixed, uses new wording and new performance descriptions over all eight instrument classes of the training manifest, two of which are held out from head training, and uses two generation seeds that appear nowhere in training. The development prompts of Table 10 are the only prompts that influence any choice reported in this paper: the guidance strength is selected on eight of them with the decoder gradient alone and then frozen. The unseen-instrument prompts of the same table use the two instrument classes held out of head training and of every operator, and are scored once.

Prompts are short natural-language descriptions of a single instrument and a performance behaviour, written to make the attributes we control well defined: a clip has to have a loudness envelope with something happening in it for a fade target to be meaningful, and it has to have energy in both bands for band contrast to be defined. Nothing about the prompt text is used by any direction source, which reads only the latent estimate or the decoded waveform.

Table 8: The 128 head-training prompts (eight instrument classes $\times$ sixteen performance descriptions). Classes 0 to 5 are the training instruments and 6 to 7 are held out from head training and from every operator. Each prompt is generated with four seeds (29001, 29003, 29009, 29011) per generator, giving 512 trajectories per generator.

id	split	instrument	style	text
enriched_i00_s00	train	0	0	A solo piano plays slow, gentle phrases with long pauses.
enriched_i00_s01	train	0	1	A solo piano plays a steady, evenly paced pattern.
enriched_i00_s02	train	0	2	A solo piano plays short, detached sounds with clear gaps.
enriched_i00_s03	train	0	3	A solo piano plays a flowing phrase that becomes gradually louder.
enriched_i00_s04	train	0	4	A solo piano plays a repeating phrase that becomes gradually quieter.
enriched_i00_s05	train	0	5	A solo piano plays a quiet passage followed by a strong accent.
enriched_i00_s06	train	0	6	A solo piano plays strong accents separated by softer sounds.
enriched_i00_s07	train	0	7	A solo piano plays an energetic pattern with frequent changes.
enriched_i00_s08	train	0	8	A solo piano plays a sparse pattern with irregular pauses.
enriched_i00_s09	train	0	9	A solo piano plays a smooth, continuous passage at moderate volume.
enriched_i00_s10	train	0	10	A solo piano plays two short phrases separated by a long pause.
enriched_i00_s11	train	0	11	A solo piano plays a fast, regular sequence with sharp attacks.
enriched_i00_s12	train	0	12	A solo piano plays a relaxed, spacious performance.
enriched_i00_s13	train	0	13	A solo piano plays a loud opening followed by a soft ending.
enriched_i00_s14	train	0	14	A solo piano plays a soft opening followed by a loud ending.
enriched_i00_s15	train	0	15	A solo piano plays alternating quiet and loud phrases.
enriched_i01_s00	train	1	0	A solo acoustic guitar plays slow, gentle phrases with long pauses.
enriched_i01_s01	train	1	1	A solo acoustic guitar plays a steady, evenly paced pattern.
enriched_i01_s02	train	1	2	A solo acoustic guitar plays short, detached sounds with clear gaps.
enriched_i01_s03	train	1	3	A solo acoustic guitar plays a flowing phrase that becomes gradually louder.
enriched_i01_s04	train	1	4	A solo acoustic guitar plays a repeating phrase that becomes gradually quieter.
enriched_i01_s05	train	1	5	A solo acoustic guitar plays a quiet passage followed by a strong accent.
enriched_i01_s06	train	1	6	A solo acoustic guitar plays strong accents separated by softer sounds.
enriched_i01_s07	train	1	7	A solo acoustic guitar plays an energetic pattern with frequent changes.

continued

id	split	instrument	style	text
enriched_i01_s08	train	1	8	A solo acoustic guitar plays a sparse pattern with irregular pauses.
enriched_i01_s09	train	1	9	A solo acoustic guitar plays a smooth, continuous passage at moderate volume.
enriched_i01_s10	train	1	10	A solo acoustic guitar plays two short phrases separated by a long pause.
enriched_i01_s11	train	1	11	A solo acoustic guitar plays a fast, regular sequence with sharp attacks.
enriched_i01_s12	train	1	12	A solo acoustic guitar plays a relaxed, spacious performance.
enriched_i01_s13	train	1	13	A solo acoustic guitar plays a loud opening followed by a soft ending.
enriched_i01_s14	train	1	14	A solo acoustic guitar plays a soft opening followed by a loud ending.
enriched_i01_s15	train	1	15	A solo acoustic guitar plays alternating quiet and loud phrases.
enriched_i02_s00	train	2	0	A solo violin plays slow, gentle phrases with long pauses.
enriched_i02_s01	train	2	1	A solo violin plays a steady, evenly paced pattern.
enriched_i02_s02	train	2	2	A solo violin plays short, detached sounds with clear gaps.
enriched_i02_s03	train	2	3	A solo violin plays a flowing phrase that becomes gradually louder.
enriched_i02_s04	train	2	4	A solo violin plays a repeating phrase that becomes gradually quieter.
enriched_i02_s05	train	2	5	A solo violin plays a quiet passage followed by a strong accent.
enriched_i02_s06	train	2	6	A solo violin plays strong accents separated by softer sounds.
enriched_i02_s07	train	2	7	A solo violin plays an energetic pattern with frequent changes.
enriched_i02_s08	train	2	8	A solo violin plays a sparse pattern with irregular pauses.
enriched_i02_s09	train	2	9	A solo violin plays a smooth, continuous passage at moderate volume.
enriched_i02_s10	train	2	10	A solo violin plays two short phrases separated by a long pause.
enriched_i02_s11	train	2	11	A solo violin plays a fast, regular sequence with sharp attacks.
enriched_i02_s12	train	2	12	A solo violin plays a relaxed, spacious performance.
enriched_i02_s13	train	2	13	A solo violin plays a loud opening followed by a soft ending.
enriched_i02_s14	train	2	14	A solo violin plays a soft opening followed by a loud ending.
enriched_i02_s15	train	2	15	A solo violin plays alternating quiet and loud phrases.
enriched_i03_s00	train	3	0	A solo drum kit plays slow, gentle phrases with long pauses.
enriched_i03_s01	train	3	1	A solo drum kit plays a steady, evenly paced pattern.
enriched_i03_s02	train	3	2	A solo drum kit plays short, detached sounds with clear gaps.
enriched_i03_s03	train	3	3	A solo drum kit plays a flowing phrase that becomes gradually louder.
enriched_i03_s04	train	3	4	A solo drum kit plays a repeating phrase that becomes gradually quieter.
enriched_i03_s05	train	3	5	A solo drum kit plays a quiet passage followed by a strong accent.
enriched_i03_s06	train	3	6	A solo drum kit plays strong accents separated by softer sounds.
enriched_i03_s07	train	3	7	A solo drum kit plays an energetic pattern with frequent changes.
enriched_i03_s08	train	3	8	A solo drum kit plays a sparse pattern with irregular pauses.
enriched_i03_s09	train	3	9	A solo drum kit plays a smooth, continuous passage at moderate volume.
enriched_i03_s10	train	3	10	A solo drum kit plays two short phrases separated by a long pause.
enriched_i03_s11	train	3	11	A solo drum kit plays a fast, regular sequence with sharp attacks.
enriched_i03_s12	train	3	12	A solo drum kit plays a relaxed, spacious performance.
enriched_i03_s13	train	3	13	A solo drum kit plays a loud opening followed by a soft ending.
enriched_i03_s14	train	3	14	A solo drum kit plays a soft opening followed by a loud ending.
enriched_i03_s15	train	3	15	A solo drum kit plays alternating quiet and loud phrases.
enriched_i04_s00	train	4	0	A solo flute plays slow, gentle phrases with long pauses.
enriched_i04_s01	train	4	1	A solo flute plays a steady, evenly paced pattern.
enriched_i04_s02	train	4	2	A solo flute plays short, detached sounds with clear gaps.
enriched_i04_s03	train	4	3	A solo flute plays a flowing phrase that becomes gradually louder.
enriched_i04_s04	train	4	4	A solo flute plays a repeating phrase that becomes gradually quieter.
enriched_i04_s05	train	4	5	A solo flute plays a quiet passage followed by a strong accent.
enriched_i04_s06	train	4	6	A solo flute plays strong accents separated by softer sounds.
enriched_i04_s07	train	4	7	A solo flute plays an energetic pattern with frequent changes.
enriched_i04_s08	train	4	8	A solo flute plays a sparse pattern with irregular pauses.
enriched_i04_s09	train	4	9	A solo flute plays a smooth, continuous passage at moderate volume.
enriched_i04_s10	train	4	10	A solo flute plays two short phrases separated by a long pause.
enriched_i04_s11	train	4	11	A solo flute plays a fast, regular sequence with sharp attacks.
enriched_i04_s12	train	4	12	A solo flute plays a relaxed, spacious performance.
enriched_i04_s13	train	4	13	A solo flute plays a loud opening followed by a soft ending.
enriched_i04_s14	train	4	14	A solo flute plays a soft opening followed by a loud ending.
enriched_i04_s15	train	4	15	A solo flute plays alternating quiet and loud phrases.
enriched_i05_s00	train	5	0	A solo saxophone plays slow, gentle phrases with long pauses.
enriched_i05_s01	train	5	1	A solo saxophone plays a steady, evenly paced pattern.
enriched_i05_s02	train	5	2	A solo saxophone plays short, detached sounds with clear gaps.
enriched_i05_s03	train	5	3	A solo saxophone plays a flowing phrase that becomes gradually louder.
enriched_i05_s04	train	5	4	A solo saxophone plays a repeating phrase that becomes gradually quieter.
enriched_i05_s05	train	5	5	A solo saxophone plays a quiet passage followed by a strong accent.
enriched_i05_s06	train	5	6	A solo saxophone plays strong accents separated by softer sounds.
enriched_i05_s07	train	5	7	A solo saxophone plays an energetic pattern with frequent changes.
enriched_i05_s08	train	5	8	A solo saxophone plays a sparse pattern with irregular pauses.
enriched_i05_s09	train	5	9	A solo saxophone plays a smooth, continuous passage at moderate volume.
enriched_i05_s10	train	5	10	A solo saxophone plays two short phrases separated by a long pause.
enriched_i05_s11	train	5	11	A solo saxophone plays a fast, regular sequence with sharp attacks.
enriched_i05_s12	train	5	12	A solo saxophone plays a relaxed, spacious performance.
enriched_i05_s13	train	5	13	A solo saxophone plays a loud opening followed by a soft ending.
enriched_i05_s14	train	5	14	A solo saxophone plays a soft opening followed by a loud ending.
enriched_i05_s15	train	5	15	A solo saxophone plays alternating quiet and loud phrases.
enriched_i06_s00	validation	6	0	A solo synthesizer plays slow, gentle phrases with long pauses.
enriched_i06_s01	validation	6	1	A solo synthesizer plays a steady, evenly paced pattern.
enriched_i06_s02	validation	6	2	A solo synthesizer plays short, detached sounds with clear gaps.
enriched_i06_s03	validation	6	3	A solo synthesizer plays a flowing phrase that becomes gradually louder.
enriched_i06_s04	validation	6	4	A solo synthesizer plays a repeating phrase that becomes gradually quieter.

continued

id	split	instrument	style	text
enriched_i06_s05	validation	6	5	A solo synthesizer plays a quiet passage followed by a strong accent.
enriched_i06_s06	validation	6	6	A solo synthesizer plays strong accents separated by softer sounds.
enriched_i06_s07	validation	6	7	A solo synthesizer plays an energetic pattern with frequent changes.
enriched_i06_s08	validation	6	8	A solo synthesizer plays a sparse pattern with irregular pauses.
enriched_i06_s09	validation	6	9	A solo synthesizer plays a smooth, continuous passage at moderate volume.
enriched_i06_s10	validation	6	10	A solo synthesizer plays two short phrases separated by a long pause.
enriched_i06_s11	validation	6	11	A solo synthesizer plays a fast, regular sequence with sharp attacks.
enriched_i06_s12	validation	6	12	A solo synthesizer plays a relaxed, spacious performance.
enriched_i06_s13	validation	6	13	A solo synthesizer plays a loud opening followed by a soft ending.
enriched_i06_s14	validation	6	14	A solo synthesizer plays a soft opening followed by a loud ending.
enriched_i06_s15	validation	6	15	A solo synthesizer plays alternating quiet and loud phrases.
enriched_i07_s00	validation	7	0	A solo cello plays slow, gentle phrases with long pauses.
enriched_i07_s01	validation	7	1	A solo cello plays a steady, evenly paced pattern.
enriched_i07_s02	validation	7	2	A solo cello plays short, detached sounds with clear gaps.
enriched_i07_s03	validation	7	3	A solo cello plays a flowing phrase that becomes gradually louder.
enriched_i07_s04	validation	7	4	A solo cello plays a repeating phrase that becomes gradually quieter.
enriched_i07_s05	validation	7	5	A solo cello plays a quiet passage followed by a strong accent.
enriched_i07_s06	validation	7	6	A solo cello plays strong accents separated by softer sounds.
enriched_i07_s07	validation	7	7	A solo cello plays an energetic pattern with frequent changes.
enriched_i07_s08	validation	7	8	A solo cello plays a sparse pattern with irregular pauses.
enriched_i07_s09	validation	7	9	A solo cello plays a smooth, continuous passage at moderate volume.
enriched_i07_s10	validation	7	10	A solo cello plays two short phrases separated by a long pause.
enriched_i07_s11	validation	7	11	A solo cello plays a fast, regular sequence with sharp attacks.
enriched_i07_s12	validation	7	12	A solo cello plays a relaxed, spacious performance.
enriched_i07_s13	validation	7	13	A solo cello plays a loud opening followed by a soft ending.
enriched_i07_s14	validation	7	14	A solo cello plays a soft opening followed by a loud ending.
enriched_i07_s15	validation	7	15	A solo cello plays alternating quiet and loud phrases.

Table 9: The sixteen prompts of the new-prompt panel, written after all heads and operators were fixed. Two new generation seeds (32801, 32803) each, so this panel shares no prompt and no seed with head training. The prompts cover all eight instrument classes of the training manifest, two per class, so four of the sixteen are on the classes held out from head training; every prompt has new wording and a new performance description.

id	instrument	style	text
rep_i00_s00	piano	0	A close recording of a solo piano playing a short phrase, pausing, and repeating the phrase with a softer ending.
rep_i00_s01	piano	1	A solo piano performance that begins with sparse, quiet notes and builds to a louder, busier phrase.
rep_i01_s00	acoustic guitar	0	A close recording of a solo acoustic guitar playing a short phrase, pausing, and repeating the phrase with a softer ending.
rep_i01_s01	acoustic guitar	1	A solo acoustic guitar performance that begins with sparse, quiet notes and builds to a louder, busier phrase.
rep_i02_s00	violin	0	A close recording of a solo violin playing a short phrase, pausing, and repeating the phrase with a softer ending.
rep_i02_s01	violin	1	A solo violin performance that begins with sparse, quiet notes and builds to a louder, busier phrase.
rep_i03_s00	drum kit	0	A close recording of a solo drum kit playing a short phrase, pausing, and repeating the phrase with a softer ending.
rep_i03_s01	drum kit	1	A solo drum kit performance that begins with sparse, quiet notes and builds to a louder, busier phrase.
rep_i04_s00	flute	0	A close recording of a solo flute playing a short phrase, pausing, and repeating the phrase with a softer ending.
rep_i04_s01	flute	1	A solo flute performance that begins with sparse, quiet notes and builds to a louder, busier phrase.
rep_i05_s00	saxophone	0	A close recording of a solo saxophone playing a short phrase, pausing, and repeating the phrase with a softer ending.
rep_i05_s01	saxophone	1	A solo saxophone performance that begins with sparse, quiet notes and builds to a louder, busier phrase.
rep_i06_s00	synthesizer	0	A close recording of a solo synthesizer playing a short phrase, pausing, and repeating the phrase with a softer ending.
rep_i06_s01	synthesizer	1	A solo synthesizer performance that begins with sparse, quiet notes and builds to a louder, busier phrase.
rep_i07_s00	cello	0	A close recording of a solo cello playing a short phrase, pausing, and repeating the phrase with a softer ending.
rep_i07_s01	cello	1	A solo cello performance that begins with sparse, quiet notes and builds to a louder, busier phrase.

Table 10: The development and unseen-instrument prompts. The development prompts are the only ones used to choose the guidance strength (Appendix B.1); the unseen-instrument prompts use eight instruments that appear nowhere in the training manifest. Each panel generates every prompt with one seed (202); the pilot file lists three seeds of which the panels use one.

id	split	seeds	text
dev_00	dev	202	A solo piano plays a slow and gentle melody.
dev_01	dev	202	An acoustic guitar strums a steady chord progression.
dev_02	dev	202	A violin plays sustained expressive notes.
dev_03	dev	202	A drum kit plays a steady rock beat.
dev_04	dev	202	A flute plays a bright flowing melody.
dev_05	dev	202	A saxophone plays a relaxed jazz melody.
dev_06	dev	202	A synthesizer plays a pulsing electronic sequence.
dev_07	dev	202	A cello plays a low sustained melody.
heldout_08	heldout	202	A trumpet plays a short brass fanfare.
heldout_09	heldout	202	A marimba plays a rhythmic melodic pattern.
heldout_10	heldout	202	An electric bass plays a repeating funk groove.
heldout_11	heldout	202	A harp plays soft ascending arpeggios.
heldout_12	heldout	202	A clarinet plays a smooth lyrical melody.
heldout_13	heldout	202	A banjo plays a quick rolling pattern.
heldout_14	heldout	202	An organ plays sustained rich chords.
heldout_15	heldout	202	A hand drum plays a syncopated rhythm.

C Analysis of the Mean Jacobian

This section states two conditions under which the mean Jacobian is a sensible guidance direction, and three simple consequences. None of it shows that $\bar{J}$ is optimal. The first result shows that, for the local loss, $\bar{J}$ is slower than the decoder gradient by a known factor; the second gives a condition under which it can nevertheless do better at the output. Section 6.5 and Appendix E.1 show that this condition is only part of the explanation.

Notation. Flatten the clean estimate to $\hat{x} \in \mathbb{R}^D$ with $D = CL$. Let $J(x) = \partial\phi/\partial x \in \mathbb{R}^{F\times D}$, $r = \phi(\hat{x}) - y$, and $\ell = \frac{1}{2c}\|r\|^2$ with $c = 200F$, which is the $1/(20^2F)$ normalisation of Equation (1), so that the decoder gradient is $\nabla\ell = \frac{1}{c}J(\hat{x})^\top r$. The mean Jacobian is $\bar{J} = \frac{1}{N}\sum_n J(x_n)^\top \in \mathbb{R}^{D\times F}$ (Equation (2)). Write

$$\cos\theta(\hat{x}) = \frac{\langle J(\hat{x})^\top r,\ \bar{J}r\rangle}{\|J(\hat{x})^\top r\|\,\|\bar{J}r\|}$$

for the cosine between the decoder gradient and the mean-Jacobian direction at $\hat{x}$.

Proposition 1 (descent condition). The direction $d = \bar{J}r$ used in Equation (1) is a local descent direction for ℓ at $\hat{x}$ if and only if $\cos\theta(\hat{x}) > 0$. A sufficient condition is

$$\|(\bar{J} - J(\hat{x})^\top)\,r\| < \|J(\hat{x})^\top r\|,$$

and, if $J(\hat{x})$ has full row rank, $\|\bar{J} - J(\hat{x})^\top\|_{\mathrm{op}} < \sigma_{\min}(J(\hat{x}))$ suffices for every $r \neq 0$. Under the normalised step of Equation (1), the first-order rate of decrease of ℓ along d is $\|\nabla\ell\|\cos\theta(\hat{x})$, a factor $\cos\theta$ below the rate along the decoder gradient.

Proof. The directional derivative of ℓ along $-d/\|d\|$ is $-\frac{1}{c}\langle J^\top r, \bar{J}r\rangle/\|\bar{J}r\|$, which is negative if and only if $\cos\theta > 0$, and equals $-\|\nabla\ell\|\cos\theta$. With $e = (\bar{J} - J^\top)r$, $\langle J^\top r, \bar{J}r\rangle = \|J^\top r\|^2 + \langle J^\top r, e\rangle \geq \|J^\top r\|\,(\|J^\top r\| - \|e\|)$, which is positive when $\|e\| < \|J^\top r\|$. If J has full row rank, $\|J^\top r\| \geq \sigma_{\min}(J)\|r\|$ and $\|e\| \leq \|\bar{J} - J^\top\|_{\mathrm{op}}\|r\|$.

The screen of Section 4.3 estimates $\cos\theta$ on fixed attribute directions u_k in place of r; Proposition 1 is what that number means. It also makes explicit that the local theory predicts $\bar{J}$ to be *worse* than the decoder gradient, by the factor $\cos\theta$, which is 0.60 to 0.91 in our panels. Any advantage of $\bar{J}$ must therefore come from what happens after the step.

Proposition 2 (normalised steps and discarded components). Let Φ map the clean estimate at the guided step to the final latent through the remaining sampler steps, and assume:

(A1) $\Phi(\hat{x} + \delta) \approx \Phi(\hat{x}) + A\delta$ for the displacements used, so that the change in the final attribute is approximately $J_0 A\delta$ with $J_0 = J(\hat{x}_0)$;

(A2) writing the decoder gradient as $g = \alpha\,\bar{J}r + \varepsilon$ with $\varepsilon \perp \bar{J}r$ and $\alpha = \langle g, \bar{J}r\rangle/\|\bar{J}r\|^2 > 0$, the remaining steps discard the orthogonal part: $A\varepsilon = 0$.

Let $E(d) = \langle J_0 A d, r\rangle/\|d\|$ be the first-order reduction of the final residual per unit step along d, and assume $\langle J_0 A\bar{J}r, r\rangle > 0$. Then

$$\frac{E(\bar{J}r)}{E(g)} = \frac{1}{\cos\theta}.$$

Proof. By (A2), $Ag = \alpha A\bar{J}r$, so $E(g) = \alpha\langle J_0 A\bar{J}r, r\rangle/\|g\|$ and $E(\bar{J}r) = \langle J_0 A\bar{J}r, r\rangle/\|\bar{J}r\|$. Their ratio is $\|g\|/(\alpha\|\bar{J}r\|)$, and $\alpha\|\bar{J}r\|/\|g\| = \langle g, \bar{J}r\rangle/(\|g\|\,\|\bar{J}r\|) = \cos\theta$.

Under (A1) and (A2) the mean Jacobian delivers $1/\cos\theta$ times the decoder gradient's effect at the output, 1.10 to 1.67 for the cosines in our panels, because a normalised step spends none of its length on components the sampler removes. This is consistent in direction with the survival measurements of Section 6.5. The analysis is incomplete in a way the experiments expose: under (A1) alone, the best first-order direction for the final loss is $A^\top J_0^\top r$, the end-to-end gradient, and Appendix E.1 finds that direction worse than both $\bar{J}r$ and the decoder gradient at the step sizes we use. The remaining steps are therefore not well described by a single linear map at those step sizes, and Proposition 2 is a condition under which $\bar{J}$ can win, not an account of why it does.

Remark A (why the mean). $\bar{J}$ is the least-squares state-independent approximation, $\bar{J} = \arg\min_M \frac{1}{N}\sum_n \|J(x_n)^\top - M\|_F^2$, and $\frac{1}{N}\sum_n \|(J(x_n)^\top - \bar{J})\,r\|^2$ is the average squared size of the error term e in Proposition 1.

Remark B (decoder-free error). When a head supplies the value, $\hat{r} = h(\hat{x}, t) - y$ replaces r, and $\|\bar{J}\hat{r} - \bar{J}r\| \le \|\bar{J}\|_{\mathrm{op}}\,\|h(\hat{x}, t) - \phi(\hat{x})\|$. The head's prediction error enters the direction linearly, which is why the decoder-free rule loses ground where the head's value is least accurate (Section 6.7).

Remark C (composition). For $\ell = \ell_1 + \ell_2$ the gradient is the sum of the two gradients, so $\bar{J}_1 r_1 + \bar{J}_2 r_2$ applies the same approximation to each term and introduces no error beyond it. This is why two precomputed matrices can simply be added (Section 6.7).

D Additional Results

Each subsection supplements one part of Section 6 and is referenced from it.

D.1 Main Results and Paired Statistics

This section supplements Section 6.1. Table 11 lists every comparison the paper asserts, with the mean paired difference, the number of prompts in favour and a 95 % cluster bootstrap interval over instrument clusters (10,000 draws, seed 29201, fixed before the panels were run). Table 12 does the same across panels that share prompts, seeds, strength and unguided baseline, which is how the operator is compared with the equal-cost regression head; Table 13 does the same for the VJP-matched head and the fixed direction on the panels where those ran as separate fill panels. Two cautions apply to all three. The new-prompt panel has sixteen prompts in eight instrument clusters (six for ACE) and the unseen-instrument panel has eight prompts that are each their own cluster, so the intervals on the unseen-instrument panels rest on eight single-prompt clusters, are wide for a structural reason, and a tie there is weaker evidence than a tie on the new-prompt panel. And we report no correction for multiple comparisons: the intervals are descriptive for the panels as run, and the claims the paper makes are the ones where the interval excludes zero on the new-prompt panel and the direction is consistent elsewhere.

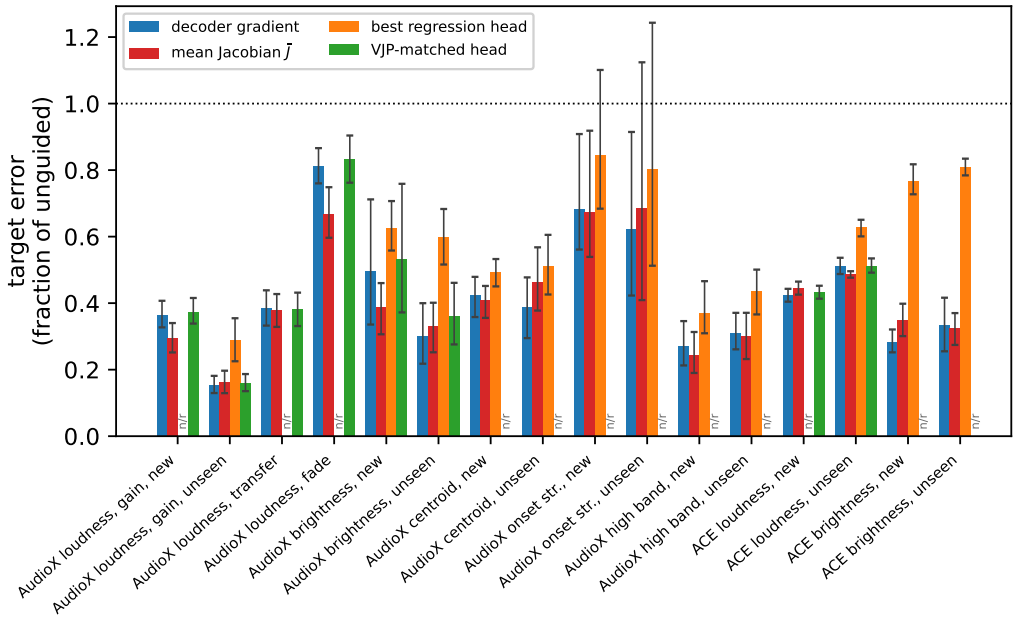

Figure 3: Target error on the sixteen calibrated panels, lower is better. Each panel's errors are divided by its own unguided error, so panels with different units and different difficulty share one axis. Whiskers are 95 % cluster bootstrap intervals on the panel mean; `n/r` marks a direction source that was not run on that panel. The mean Jacobian is below the decoder gradient on four panels, level on ten, and above it on the two ACE new-prompt panels, one of them marginally, and no learned head is ahead of it by an interval that excludes zero on any panel.

Table 11: Every comparison this paper asserts, paired by prompt, *including the sixteen main panels of Table 1* so that this table is self-contained. All direction sources run on the same prompts, seeds and noise, so the difference is taken per prompt; the interval is a 95 % cluster bootstrap over instrument clusters (10,000 draws, seed fixed before the panels were run). A positive benefit favours the first direction source. Intervals that contain zero are marked and are described as ties in the text. No correction is applied for multiple comparisons; see Appendix G.

Panel	Comparison	benefit (dB)	prompts	95% CI
AudioX loudness, gain, new prompts	MaJa ($\bar{J}$) vs. decoder gradient	+0.39	13/16	[+0.27, +0.50]
AudioX loudness, gain, unseen instruments	MaJa ($\bar{J}$) vs. decoder gradient	-0.05	3/8	[-0.14, +0.04]†
	MaJa ($\bar{J}$) vs. regression (teacher and student states)	+0.76	8/8	[+0.54, +0.98]
AudioX loudness, envelope transfer	MaJa ($\bar{J}$) vs. decoder gradient	+0.07	9/16	[+0.00, +0.15]
AudioX loudness, fade	MaJa ($\bar{J}$) vs. decoder gradient	+0.47	16/16	[+0.38, +0.56]
AudioX brightness, new prompts	MaJa ($\bar{J}$) vs. decoder gradient	+0.51	6/9	[-0.04, +1.32]†
	MaJa ($\bar{J}$) vs. VJP head	+0.69	7/9	[+0.14, +1.55]
	VJP head vs. regression (teacher and student states)	+0.45	8/9	[-0.27, +0.97]†
AudioX brightness, unseen instruments	MaJa ($\bar{J}$) vs. decoder gradient	-0.15	2/7	[-0.44, +0.24]†
	MaJa ($\bar{J}$) vs. regression (teacher and student states)	+1.25	7/7	[+0.85, +1.65]
AudioX centroid, new prompts	MaJa ($\bar{J}$) vs. decoder gradient	+0.05	5/9	[-0.05, +0.14]†
	MaJa ($\bar{J}$) vs. regression (teacher and student states)	+0.33	9/9	[+0.23, +0.44]
AudioX centroid, unseen instruments	MaJa ($\bar{J}$) vs. decoder gradient	-0.22	2/8	[-0.46, +0.06]†
	MaJa ($\bar{J}$) vs. regression (teacher and student states)	+0.20	7/8	[+0.07, +0.34]
ACE loudness, new prompts	MaJa ($\bar{J}$) vs. decoder gradient	-0.14	4/12	[-0.25, -0.00]
	VJP head vs. MaJa ($\bar{J}$)	+0.08	8/12	[-0.05, +0.19]†
ACE loudness, unseen classes	MaJa ($\bar{J}$) vs. decoder gradient	+0.15	4/8	[-0.03, +0.33]†
	MaJa ($\bar{J}$) vs. regression (teacher and student states)	+0.85	8/8	[+0.73, +0.95]
ACE brightness, new prompts	MaJa ($\bar{J}$) vs. decoder gradient	-0.34	2/12	[-0.57, -0.12]
	MaJa ($\bar{J}$) vs. regression (teacher and student states)	+2.17	12/12	[+2.01, +2.32]
ACE brightness, unseen classes	MaJa ($\bar{J}$) vs. decoder gradient	+0.04	3/8	[-0.32, +0.38]†
	MaJa ($\bar{J}$) vs. regression (teacher and student states)	+2.77	8/8	[+2.59, +2.95]
AudioX onset strength, new prompts	MaJa ($\bar{J}$) vs. decoder gradient	+0.01	5/9	[-0.02, +0.05]†
	MaJa ($\bar{J}$) vs. regression (teacher and student states)	+0.22	9/9	[+0.18, +0.27]
AudioX onset strength, unseen instruments	MaJa ($\bar{J}$) vs. decoder gradient	-0.10	5/8	[-0.35, +0.03]†
	MaJa ($\bar{J}$) vs. regression (teacher and student states)	+0.19	8/8	[+0.15, +0.23]
AudioX loudness, inside the 32-step teacher	MaJa ($\bar{J}$) vs. decoder gradient	-0.02	3/16	[-0.05, -0.01]
	decoder gradient vs. unguided	+0.69	16/16	[+0.55, +0.85]
AudioX loudness, cheaper decodes and a per-time operator	MaJa ($\bar{J}$) vs. decoder gradient	+0.39	13/16	[+0.27, +0.50]
	MaJa ($\bar{J}$) vs. bf16 decode	+0.38	12/16	[+0.27, +0.49]
	MaJa ($\bar{J}$) vs. half-length decode	+3.39	16/16	[+3.21, +3.56]
	per-step $\bar{J}$ vs. MaJa ($\bar{J}$)	+0.01	11/16	[+0.00, +0.02]
	bf16 decode vs. decoder gradient	+0.01	13/16	[-0.01, +0.02]†
AudioX loudness, where the fixed matrix comes from	MaJa ($\bar{J}$) vs. linear regression head	+0.78	12/16	[+0.30, +1.19]
	MaJa ($\bar{J}$) vs. brightness $\bar{J}$ on a loudness target	+4.47	16/16	[+4.04, +4.86]
	decoder gradient vs. linear regression head	+0.39	8/16	[-0.09, +0.78]†
AudioX loudness, a 27.6M head	MaJa ($\bar{J}$) vs. regr. 27.6M	+1.06	16/16	[+0.78, +1.34]
	regression (teacher and student states) vs. regr. 27.6M	+0.36	16/16	[+0.20, +0.51]
AudioX loudness, a 41.3M head	MaJa ($\bar{J}$) vs. regr. 41.3M	+1.05	16/16	[+0.74, +1.36]
	regression (teacher and student states) vs. regr. 41.3M	+0.35	14/16	[+0.16, +0.53]
AudioX loudness, a +-3 dB target	MaJa ($\bar{J}$) vs. decoder gradient	+0.36	15/16	[+0.29, +0.42]
AudioX loudness, a +-12 dB target	MaJa ($\bar{J}$) vs. decoder gradient	+0.06	9/16	[-0.03, +0.14]†
AudioX loudness, one guidance step	MaJa ($\bar{J}$) vs. decoder gradient	-0.03	6/16	[-0.06, -0.01]
AudioX loudness, two guidance steps	MaJa ($\bar{J}$) vs. decoder gradient	-0.06	8/16	[-0.13, +0.00]†
AudioX loudness, operator precision and rank	MaJa ($\bar{J}$) vs. fp16 $\bar{J}$	-0.00	2/16	[-0.00, -0.00]
	MaJa ($\bar{J}$) vs. rank-32 $\bar{J}$	+0.08	16/16	[+0.05, +0.11]
	MaJa ($\bar{J}$) vs. rank-8 $\bar{J}$	+0.64	16/16	[+0.55, +0.74]
	rank-32 $\bar{J}$ vs. decoder gradient	+0.32	12/16	[+0.19, +0.44]
AudioX brightness at 3 per cent	MaJa ($\bar{J}$) vs. decoder gradient	+1.61	9/9	[+0.86, +2.35]
AudioX centroid at 5 per cent	MaJa ($\bar{J}$) vs. decoder gradient	+0.44	8/9	[+0.24, +0.67]
AudioX loudness, early steps only	MaJa ($\bar{J}$) vs. decoder gradient	-0.06	8/16	[-0.13, +0.00]†
AudioX loudness, late steps only	MaJa ($\bar{J}$) vs. decoder gradient	+0.52	15/16	[+0.47, +0.57]
AudioX loudness, VJP vs regression	VJP head vs. regression (teacher and student states)	+0.26	12/16	[+0.13, +0.40]
AudioX loudness, how many states the operator needs	operator from 8 states vs. MaJa ($\bar{J}$)	+0.01	9/16	[-0.03, +0.06]†
	operator from 2 states vs. MaJa ($\bar{J}$)	-0.04	6/16	[-0.18, +0.12]†
AudioX loudness, preconditioning the exact gradient	MaJa ($\bar{J}$) vs. decoder gradient	+0.39	13/16	[+0.27, +0.50]
	MaJa ($\bar{J}$) vs. decoder_lowpass	+0.42	16/16	[+0.33, +0.51]
	decoder_clip vs. decoder gradient	+0.02	8/16	[-0.04, +0.07]†
	decoder_lowpass vs. decoder gradient	-0.03	7/16	[-0.10, +0.02]†
AudioX loudness, operator controls	MaJa ($\bar{J}$) vs. random matrix, matched row norms	+3.99	16/16	[+3.64, +4.26]
	MaJa ($\bar{J}$) vs. $\bar{J}$ built at a real recording	+2.86	16/16	[+2.56, +3.09]
AudioX loudness, head value + fixed direction	head value + $\bar{J}$ direction vs. VJP head	+0.43	14/16	[+0.34, +0.53]
AudioX loudness, tremolo target	MaJa ($\bar{J}$) vs. decoder gradient	+0.60	16/16	[+0.53, +0.68]
	MaJa ($\bar{J}$) vs. fixed direction	+3.20	16/16	[+2.92, +3.48]
	MaJa ($\bar{J}$) vs. regression (student)	+1.15	15/16	[+0.94, +1.35]
AudioX loudness, mid-clip step target	MaJa ($\bar{J}$) vs. decoder gradient	+0.53	15/16	[+0.45, +0.61]
	MaJa ($\bar{J}$) vs. fixed direction	+4.34	16/16	[+3.95, +4.72]
	MaJa ($\bar{J}$) vs. regression (student)	+1.27	15/16	[+1.11, +1.41]
AudioX loudness, random envelope targets	MaJa ($\bar{J}$) vs. decoder gradient	+0.33	16/16	[+0.28, +0.39]
	MaJa ($\bar{J}$) vs. fixed direction	+1.25	15/16	[+0.93, +1.54]
	MaJa ($\bar{J}$) vs. regression (student)	+0.23	11/16	[-0.05, +0.44]†
AudioX loudness, 80 against 128 states	$\bar{J}$ from 128 states vs. MaJa ($\bar{J}$)	+0.00	12/16	[-0.00, +0.01]†
	MaJa ($\bar{J}$) vs. decoder gradient	+0.39	13/16	[+0.27, +0.50]
AudioX loudness, smoothing the gradient in state space	MaJa ($\bar{J}$) vs. decoder gradient	+0.39	13/16	[+0.27, +0.50]
	MaJa ($\bar{J}$) vs. decoder_jitter	+0.41	14/16	[+0.30, +0.52]
	decoder_jitter vs. decoder gradient	-0.02	5/16	[-0.03, -0.01]
	decoder_jitter8 vs. decoder gradient	-0.02	3/16	[-0.04, -0.01]
AudioX loudness, operators from other attributes	MaJa ($\bar{J}$) vs. fixedjac_attrspectral	+4.47	16/16	[+4.04, +4.86]
	MaJa ($\bar{J}$) vs. fixedjac_attrcentroid	+5.62	16/16	[+5.17, +6.06]
	MaJa ($\bar{J}$) vs. fixedjac_attrflux	+3.36	16/16	[+2.98, +3.64]
	unguided vs. fixedjac_attrspectral	+0.52	13/16	[+0.24, +0.80]
AudioX brightness, operators from other attributes	MaJa ($\bar{J}$) vs. fixedjac_attrrms	+5.47	9/9	[+4.45, +6.43]
	MaJa ($\bar{J}$) vs. fixedjac_attrcentroid	+0.64	8/9	[+0.20, +1.33]
	MaJa ($\bar{J}$) vs. fixedjac_attrflux	+3.47	9/9	[+3.04, +3.94]
AudioX centroid, operators from other attributes	MaJa ($\bar{J}$) vs. fixedjac_attrrms	+2.31	9/9	[+1.88, +2.77]
	MaJa ($\bar{J}$) vs. fixedjac_attrspectral	-0.10	0/9	[-0.16, -0.04]

Table 12: The zero-parameter operator against the equal-cost regression head (four times the student trajectories, the VJP fit's precision and batch, a wall-clock-matched update budget), paired by prompt across the two panels, which share prompts, seeds, strength and unguided baseline. A positive benefit favours the operator.

Panel	benefit (dB)	prompts	95% CI
AudioX loudness, gain, new prompts	+0.38	15/16	[+0.32, +0.44]
AudioX loudness, gain, unseen instruments	+0.08	6/8	[-0.02, +0.16]†
AudioX brightness, new prompts	+0.53	8/9	[+0.20, +1.05]
AudioX brightness, unseen instruments	+0.25	5/7	[+0.00, +0.55]
AudioX centroid, new prompts	+0.05	7/9	[-0.02, +0.13]†
AudioX centroid, unseen instruments	-0.07	2/8	[-0.27, +0.16]†
AudioX onset strength, new prompts	+0.10	8/9	[+0.04, +0.16]
AudioX onset strength, unseen instruments	+0.03	7/8	[-0.13, +0.14]†
AudioX high-band loudness, new prompts	+0.62	9/9	[+0.45, +0.85]
AudioX high-band loudness, unseen instruments	+0.68	8/8	[+0.52, +0.86]
ACE loudness, new prompts	+0.76	12/12	[+0.64, +0.88]
ACE loudness, unseen classes	+0.79	8/8	[+0.64, +0.93]
ACE brightness, new prompts	+1.78	12/12	[+1.63, +1.96]
ACE brightness, unseen classes	+2.03	8/8	[+1.66, +2.32]
AudioX CLAP contrast, new prompts	-0.19	4/16	[-0.36, -0.04]
AudioX CLAP contrast, unseen instruments	-0.02	3/8	[-0.16, +0.15]†

†interval contains zero.

Table 13: MaJa against the VJP-matched head and against the fixed direction on the panels where those direction sources ran as separate fill panels (same prompts, seeds, step size and unguided baseline), paired by prompt. A positive benefit favours MaJa. The loudness panels, where all direction sources ran in one panel, are in Table 11.

Panel	against	benefit (dB)	prompts	95% CI
AudioX onset strength, new prompts	VJP-matched head	+0.12	8/9	[+0.06, +0.18]
AudioX onset strength, unseen instruments	VJP-matched head	+0.04	7/8	[-0.18, +0.16]†
AudioX high-band loudness, new prompts	VJP-matched head	+0.60	9/9	[+0.45, +0.79]
AudioX high-band loudness, unseen instruments	VJP-matched head	+0.67	8/8	[+0.47, +0.87]
ACE brightness, new prompts	VJP-matched head	+0.23	7/12	[-0.04, +0.48]†
ACE brightness, unseen classes	VJP-matched head	+0.63	7/8	[+0.36, +0.88]
AudioX CLAP contrast, new prompts	VJP-matched head	-0.28	3/16	[-0.49, -0.08]
AudioX CLAP contrast, unseen instruments	VJP-matched head	-0.07	5/8	[-0.58, +0.31]†
AudioX brightness, new prompts	fixed direction	+0.08	4/9	[-0.09, +0.31]†
AudioX brightness, unseen instruments	fixed direction	+0.04	4/7	[-0.03, +0.13]†
AudioX centroid, new prompts	fixed direction	+0.23	8/9	[+0.12, +0.31]
AudioX centroid, unseen instruments	fixed direction	+0.19	8/8	[+0.08, +0.32]
AudioX onset strength, new prompts	fixed direction	+0.41	9/9	[+0.30, +0.56]
AudioX onset strength, unseen instruments	fixed direction	+0.49	8/8	[+0.33, +0.65]
AudioX high-band loudness, new prompts	fixed direction	+0.62	8/9	[+0.33, +0.94]
AudioX high-band loudness, unseen instruments	fixed direction	+0.80	8/8	[+0.65, +0.93]
ACE loudness, new prompts	fixed direction	+0.62	12/12	[+0.41, +0.82]
ACE loudness, unseen classes	fixed direction	+1.08	8/8	[+0.82, +1.35]
ACE brightness, new prompts	fixed direction	+0.68	11/12	[+0.37, +0.96]
ACE brightness, unseen classes	fixed direction	+1.05	8/8	[+0.70, +1.37]
AudioX CLAP contrast, new prompts	fixed direction	-0.07	2/16	[-0.11, -0.04]
AudioX CLAP contrast, unseen instruments	fixed direction	-0.00	5/8	[-0.04, +0.03]†

†interval contains zero.

D.2 Learned Heads

This section supplements Section 6.2. Three tables measure how well a direction source points, separately from how well it controls. Table 14 is measured on stored states, so it is the number a head-training pipeline can see without running the sampler; Table 15 is measured at the states the sampler actually visits during a panel, which is the number that matters and which no training pipeline observes. The two agree on the ordering and disagree on the level, regression heads losing about 0.1 in cosine between the two measurements. Table 16 asks whether a VJP-matched head has learned the Jacobian or only the three directions it was trained on: on unseen-instrument states it reaches 0.95 to 0.96 for a new random envelope, for low-order cosine modes and for a fade, and 0.87 for five-frame local impulses, which are the hardest cotangents because they ask about one frame at a time. Table 17 is the terminal-only control: a head trained on the endpoint latents alone, which is the cheapest data one could collect, and which does not work at the guidance step.

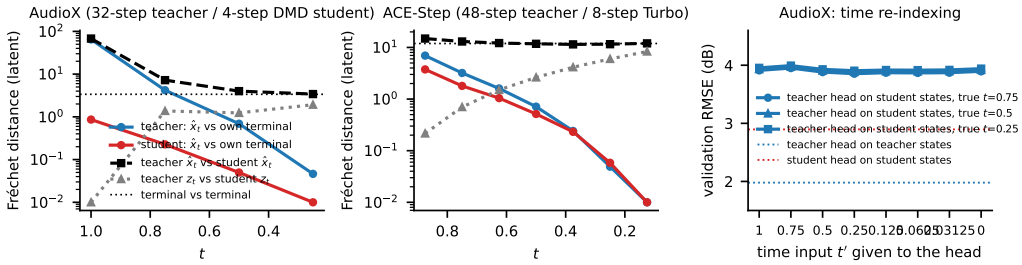

Figure 4: Where distillation moves the states. Fréchet distance in latent space between clean estimates $\hat{\mathbf{x}}_t$ and terminal latents per time. **Left:** AudioX. **Middle:** ACE, where the teacher-to-student distance is dominated by the two released models generating different music. **Right:** giving the AudioX teacher head a different time input on student states does not help.

Table 14: Direction accuracy on stored states: mean cosine between a head's input gradient and the decoder gradient at t=0.75 states (higher is better), for loudness and brightness, on teacher and student states, familiar (training prompts) and unseen instruments. Regression heads point 40 to 60° away from the decoder gradient even on their own states; VJP-matched heads reach 0.98 to 0.99 on familiar instruments whichever generator's states they were trained on. The last row is the state-independent floor: the cosine between the decoder gradient at a held-out state and the mean decoder gradient over training states.
textttn/r: the floor was measured on the student
textquotesingle s held-out loudness states only;
textttn/a: not applicable to brightness, which has its own operator.

	loudness				brightness			
	familiar		held-out		familiar		held-out	
Head	teacher	student	teacher	student	teacher	student	teacher	student
regression (teacher)	0.63	0.63	0.47	0.47	0.47	0.48	0.46	0.52
regression (student)	0.54	0.78	0.40	0.55	0.39	0.58	0.34	0.55
regression (teacher and student states)	0.74	0.83	0.55	0.58	0.60	0.65	0.55	0.65
regression (terminal)	0.55	0.71	0.40	0.51	n/r	n/r	n/r	n/r
VJP-matched (teacher)	0.99	0.98	0.86	0.75	0.94	0.92	0.89	0.82
VJP-matched (student)	0.97	0.99	0.64	0.75	0.89	0.97	0.90	0.90
VJP-matched (teacher and student states)	0.98	0.98	0.83	0.74	0.93	0.97	0.91	0.91
fixed vector (mean training gradient)	n/r	n/r	n/r	0.88	n/a	n/a	n/a	n/a

Table 15: Guidance-time alignment: mean cosine between each direction source's direction and the decoder gradient at the states the sampler actually visits, on prompts and seeds no head saw. The fixed operator points less like the decoder gradient than a VJP-matched head does and controls at least as well. `n/r`: the fixed direction and the equal-cost head on these panels ran in separate fill panels without the decoder gradient in the same pass, so their guidance-time cosine was not recorded.

Setting	fixed dir.	MaJa ($\bar{J}$)	regr.	regr. eq.-cost	VJP
AudioX loudness, gain, new	0.79	0.91	0.74	0.95	0.98
AudioX loudness, gain, unseen	0.91	0.95	0.84	0.97	0.99
AudioX loudness, transfer	0.58	0.88	0.74	0.95	0.97
AudioX brightness, new	n/r	0.79	0.51	0.83	0.91
AudioX brightness, unseen	n/r	0.81	0.53	0.85	0.92
AudioX centroid, new	n/r	0.68	0.62	0.80	0.92
AudioX centroid, unseen	n/r	0.60	0.56	0.76	0.89
AudioX onset strength, new	n/r	0.54	0.37	n/r	n/r
AudioX onset strength, unseen	n/r	0.51	0.35	n/r	n/r
ACE loudness, new	n/r	0.83	0.62	n/r	0.98
ACE brightness, new	n/r	0.68	0.23	n/r	n/r
ACE brightness, unseen	n/r	0.61	0.12	n/r	n/r

Table 16: Jacobian generalisation to cotangents never used in training. Mean cosine between a head's VJP and the decoder's for a given direction u in loudness space, on 64 held-out-instrument $t{=}0.75$ states per generator. Training used the constant direction and two fixed random envelopes; new random, DCT modes 1 to 4, a linear fade and five-frame local impulses were never seen. The fixed-vector row is the cosine between the decoder VJP at a held-out state and the mean decoder VJP over training states for the same direction (the state-independent floor; only available for the trained directions).

	student states					teacher states				
Head	constant	new random	DCT 1 to 4	fade	local impulses	constant	new random	DCT 1 to 4	fade	local impulses
regression (teacher)	0.65	0.64	0.69	0.69	0.65	0.51	0.48	0.55	0.56	0.49
regression (student)	0.74	0.61	0.69	0.70	0.63	0.55	0.42	0.52	0.52	0.45
regression (teacher and student states)	0.78	0.80	0.84	0.83	0.78	0.59	0.62	0.68	0.69	0.58
VJP-matched (teacher)	0.96	0.94	0.96	0.96	0.87	0.86	0.83	0.86	0.87	0.66
VJP-matched (student)	0.96	0.95	0.96	0.96	0.87	0.81	0.79	0.82	0.82	0.66
VJP-matched (teacher and student states)	0.96	0.95	0.96	0.96	0.88	0.86	0.83	0.86	0.87	0.67
VJP-matched (student), $K{=}0$	0.96	0.93	0.96	0.96	0.87	0.82	0.76	0.81	0.82	0.65
fixed vector (mean training VJP)	0.88	0.85*	n/r	n/r	n/r	n/r	n/r	n/r	n/a	n/a

*for the two random directions used in training, not for the new one.

Table 17: Heads that never see intermediate states (terminal latents only, no time input) versus trajectory heads, AudioX loudness: prediction error on student $t{=}0.75$ states of the training prompts, and control error on the new-prompt panel (single step / calibrated multi-step).

Head	pred. error $t{=}0.75$	single	multi
trajectory, teacher	3.95	4.17	3.07
trajectory, student	2.86	3.74	2.62
trajectory, both	0.91	3.59	2.35
terminal, teacher	3.48	4.42	3.51
terminal, student	2.29	4.00	2.89
terminal, both	2.07	4.00	2.97

D.3 Ablations on the Mean Jacobian

This section supplements Section 6.3. Four tables isolate what makes the mean Jacobian work. Table 18 varies the matrix's provenance while holding the rule, the step size and the residual fixed: a random Gaussian matrix with matched row norms does nothing, the brightness operator applied to a loudness target is worse than not guiding, and a linear head with $\bar{J}$'s exact functional form but trained by regression on the attribute reaches only 2.42 dB. State-independence is not the active ingredient. Table 19 varies the state the matrix is built at, and Table 20 reports every single-state operator individually, which is where the one failure is visible. Table 21 is the negative control described in Appendix F: four ways of flattening the decoder gradient, none of which reproduces the mean Jacobian's advantage.

Structure of the matrix. Section 4.1 shows that the label has to belong to the attribute and that a random matrix of the same shape does nothing. This section asks the complementary question: which parts of $\bar{J}$'s structure are load bearing. Every row of Tables 23 and 24 is the same $\bar{J}$ modified in one stated way, applied through the same rule with the same step size and the same residual, so the only thing that changes is the matrix.

Table 23 removes information, and two rows in it are worth more than the rest. Truncating the singular value decomposition destroys the matrix: rank 1, 2 and 4 give 5.31, 4.96 and 3.56 dB against 5.59 unguided, so the leading singular directions are close to useless for guidance even though they carry a large share of the Frobenius energy. Keeping only the latent columns temporally aligned with each attribute frame is almost as bad (4.78), so the long-range structure along the latent time axis is doing real work rather than decorating a diagonal. Against that, the *separable* matrix, which places one averaged channel direction at each frame-aligned column and scales it by that row's norm, reaches 1.80 dB with 276 numbers instead of three million, ahead of the decoder gradient's 2.03. That result does not survive a change of target or attribute, as Table 22 and Table 25 show below. And 8-bit quantisation is free (1.64, the full matrix's value to two decimals), so the shipped artefact could be 3.2 MB rather than 12.8.

Taken together these say what the matrix is not: it is not low rank and it is not local in time. On a uniform gain target it is close to one channel pattern replicated along the clip, but that description fails as soon as the target varies, so it is a property of the easiest target rather than of the operator.

Table 24 destroys alignment while preserving statistics, and reading it together with Table 22 taught us that one of our controls was vacuous on the target we first ran it on.

On the uniform gain target, permuting the latent channels costs 3.68 dB and lands the matrix at the unguided error, so the channel pattern is load bearing; reversing the latent time axis costs 0.82; equalising the row norms costs nothing at all, an interval of $[-0.01, +0.00]$; and permuting the *rows* costs only 0.43 dB, leaving the shuffled matrix still ahead of the decoder gradient. The last of those is not a control failing to fire, it is a control that cannot fire: a uniform gain target puts the same residual in every frame, so permuting rows permutes identical numbers.

Table 22 repeats the permutations on the four-cycle tremolo target, where the residual changes sign eight times along the clip, and there they behave as controls should. Against the full matrix's 1.49 dB and an unguided error of 3.93, permuting rows gives 4.75, reversing latent time gives 4.49 and permuting channels gives 3.75; the first two are worse than not guiding at all. Equalising the row norms again costs nothing (1.484 against 1.486), on a target where the per-frame scale could have mattered, so that result is now established rather than an artefact of a flat residual.

The same table bounds the separable matrix. It reached 1.80 dB on the gain target, close to the full matrix and ahead of the decoder gradient, but on the tremolo target it gives 4.63, worse than not guiding, and on brightness and the centroid it gives 2.44 and 1.97 against 1.87 and 1.25. The 276-number form is therefore a curiosity specific to a uniform loudness target, not a compressed version of the mean Jacobian, and we report it as such.

Table 18: Where the fixed matrix has to come from (AudioX loudness, gain, new prompts). Each row is one fixed matrix of the same shape used through the same rule. A linear head has the operator's form and the regression heads' data and does not reach it; the brightness operator on loudness is worse than no guidance. What matters is the label, and it has to belong to the attribute being controlled.

Direction	error (dB)	cos to exact
no guidance	5.59	n/r
decoder gradient	2.03	1.00
a random matrix, matched row norms	5.63	-0.00
the brightness operator on loudness	6.11	-0.09
a linear head, trained by regression	2.42	0.34
the loudness mean decoder Jacobian	1.64	0.91

Table 19: Which fixed matrices guide (AudioX loudness, gain, new prompts). Only the matrix changes; the rule, the step size and the residual are identical. A random matrix does nothing, operators taken at an unseen instrument's state or at a teacher state work as well as the training-state average; operators taken at latents encoded from real recordings are less reliable, with a median of 1.93 dB over eight recordings but two above 2.3 and one failure at 4.50; single generated states mostly match the average but one of eight fails outright.

Fixed matrix	error (dB)	cos to exact
unguided	5.59	n/r
decoder gradient (reference)	2.03	1.00
mean over 80 training states	1.64	0.91
random matrix, matched row norms	5.63	-0.00
one state of an unseen instrument	1.68	0.87
one latent encoded from real audio	4.50	0.35
mean over 8 states	1.63	0.89
mean over 4 states	1.64	0.87
mean over 2 states (one of them the failing one)	1.68	0.83
real audio, recording 2	2.40	n/r
real audio, recording 3	1.68	n/r
real audio, recording 4	1.75	n/r
real audio, recording 5	1.82	n/r
real audio, recording 6	1.67	n/r
real audio, recording 7	2.04	n/r
real audio, recording 8	2.18	n/r
one teacher state	1.78	n/r
teacher state 2	1.62	n/r
unseen instrument 2	1.58	n/r
unseen instrument 3	1.65	n/r
unseen instrument 4	1.60	n/r
single generated state 1	1.54	0.85
single generated state 2	1.57	n/r
single generated state 3	1.58	n/r
single generated state 4	1.59	n/r
single generated state 5	1.59	0.82
single generated state 6	1.59	0.84
single generated state 7	1.63	n/r
single generated state 8	4.38	0.53

Table 20: Single-step probe (one 5 % correction at t=0.75, no calibration): mean target error in dB on the new-prompt panels. With one small step the gap between direction sources is compressed; the calibrated multi-step probe of Table 1 is the informative one.

Feature function	AudioX loudness	AudioX brightness	AudioX transfer	AudioX fade
unguided	5.59	4.82	9.64	3.26
decoder gradient	3.53	3.50	7.88	2.49
regression (teacher and student states)	3.59	3.76	7.91	2.46
regression (teacher)	4.17	3.77	8.24	2.83
regression (student)	3.74	3.97	7.97	2.63
VJP-matched (teacher)	3.55	3.37	7.92	2.46
VJP-matched (student)	3.53	3.38	7.88	2.46
VJP-matched (teacher and student states)	3.53	3.42	7.88	2.46

Table 21: Flattening the decoder gradient does not reproduce the operator's advantage (AudioX loudness, gain, new prompts). Each row applies one fixed preconditioner to the decoder gradient and then follows the same rule and step size.

Direction	error (dB)	cos to exact
unguided	5.59	n/r
decoder gradient	2.03	1.00
clipped at the 99th percentile	2.02	0.92
low-passed along the latent time axis	2.07	0.95
per-channel whitened	2.11	0.47
sign only	2.96	0.28
mean Jacobian (no parameters)	1.64	0.91

Table 22: The permutation controls on a target that varies along the clip (AudioX loudness, four-cycle tremolo, new prompts). A uniform gain target makes the row order of the matrix nearly irrelevant, because the residual is the same in every frame; a tremolo target does not.

direction source	error (dB)	95% interval	prompts
unguided	3.93	[3.78, 4.07]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.09	[1.94, 2.22]	16
MaJa ($\bar{J}$)	1.49	[1.36, 1.62]	16
rows permuted	4.74	[4.52, 4.99]	16
channels permuted	3.75	[3.59, 3.90]	16
latent time reversed	4.49	[4.17, 4.86]	16
row norms equalised	1.48	[1.36, 1.62]	16
separable (one channel direction)	4.62	[4.49, 4.73]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.60	[-0.68, -0.53]	0/16
decoder gradient − rows permuted	+2.66	[+2.38, +2.93]	16/16
decoder gradient − channels permuted	+1.66	[+1.39, +1.90]	16/16
decoder gradient − latent time reversed	+2.41	[+2.03, +2.79]	16/16
decoder gradient − row norms equalised	-0.60	[-0.68, -0.53]	0/16
decoder gradient − separable (one channel direction)	+2.54	[+2.33, +2.74]	16/16
decoder gradient − unguided	+1.84	[+1.58, +2.07]	16/16
MaJa ($\bar{J}$) − unguided	+2.44	[+2.21, +2.63]	16/16
rows permuted − unguided	-0.82	[-0.98, -0.67]	0/16
channels permuted − unguided	+0.18	[+0.13, +0.23]	16/16
latent time reversed − unguided	-0.57	[-0.82, -0.34]	1/16
row norms equalised − unguided	+2.44	[+2.21, +2.63]	16/16
separable (one channel direction) − unguided	-0.70	[-0.85, -0.56]	1/16

Tables 25 and 26 repeat the variants on brightness and on the spectral centroid, and the answer is that the separable form is a property of loudness rather than of the construction. On brightness it reaches 2.44 dB against the full matrix's 1.87 and the decoder gradient's 2.38, so it loses most of the advantage and is no better than differentiating the decoder. Loudness is close to a per-frame function of local latent energy, which is what a separable matrix can express; brightness is a ratio between two bands, which it cannot. On the spectral centroid the separable matrix reaches 1.97 against the full matrix's 1.25, the same picture. Eight-bit quantisation remains free on both, reproducing the full matrix to three decimals, and the low-rank and frame-local truncations fail on both.

Table 23: Structural variants of the mean Jacobian, part one (AudioX loudness, gain, new prompts, calibrated multi-step probe). Only the matrix changes; the rule, the step size and the residual are identical. Low-rank truncations, keeping only the frame-aligned band of columns, a separable hand-built matrix with 276 numbers, and 8-bit quantisation.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.03	[1.83, 2.27]	16
MaJa ($\bar{J}$)	1.64	[1.41, 1.90]	16
rank 1	5.31	[5.16, 5.47]	16
rank 2	4.96	[4.78, 5.16]	16
rank 4	3.56	[3.31, 3.83]	16
frame-local band only	4.78	[4.64, 4.95]	16
separable (one channel direction)	1.80	[1.68, 1.92]	16
8-bit quantised	1.64	[1.40, 1.90]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.39	[-0.50, -0.27]	3/16
decoder gradient − rank 1	+3.28	[+2.98, +3.55]	16/16
decoder gradient − rank 2	+2.92	[+2.64, +3.16]	16/16
decoder gradient − rank 4	+1.53	[+1.24, +1.81]	15/16
decoder gradient − frame-local band only	+2.75	[+2.48, +2.95]	16/16
decoder gradient − separable (one channel direction)	-0.24	[-0.39, -0.11]	6/16
decoder gradient − 8-bit quantised	-0.39	[-0.50, -0.28]	3/16
decoder gradient − unguided	+3.55	[+3.16, +3.88]	16/16
MaJa ($\bar{J}$) − unguided	+3.94	[+3.59, +4.22]	16/16
rank 1 − unguided	+0.27	[+0.18, +0.33]	15/16
rank 2 − unguided	+0.63	[+0.49, +0.75]	16/16
rank 4 − unguided	+2.03	[+1.82, +2.24]	16/16
frame-local band only − unguided	+0.80	[+0.66, +0.95]	16/16
separable (one channel direction) − unguided	+3.79	[+3.53, +4.02]	16/16
8-bit quantised − unguided	+3.95	[+3.60, +4.22]	16/16

Table 24: Structural variants of the mean Jacobian, part two (AudioX loudness, gain, new prompts). Equalising row norms, keeping only signs, permuting rows, permuting latent channels and reversing the latent time axis. The three permutations are controls: they preserve the matrix statistics and destroy its alignment.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.03	[1.83, 2.27]	16
MaJa ($\bar{J}$)	1.64	[1.41, 1.90]	16
row norms equalised	1.64	[1.41, 1.89]	16
signs only	2.76	[2.49, 3.02]	16
rows permuted	2.07	[1.76, 2.40]	16
channels permuted	5.32	[5.09, 5.55]	16
latent time reversed	2.46	[2.26, 2.68]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.39	[-0.50, -0.27]	3/16
decoder gradient − row norms equalised	-0.39	[-0.50, -0.28]	3/16
decoder gradient − signs only	+0.73	[+0.21, +1.17]	9/16
decoder gradient − rows permuted	+0.04	[-0.24, +0.31]	10/16
decoder gradient − channels permuted	+3.28	[+2.86, +3.66]	16/16
decoder gradient − latent time reversed	+0.42	[+0.21, +0.61]	11/16
decoder gradient − unguided	+3.55	[+3.16, +3.88]	16/16
MaJa ($\bar{J}$) − unguided	+3.94	[+3.59, +4.22]	16/16
row norms equalised − unguided	+3.95	[+3.60, +4.22]	16/16
signs only − unguided	+2.82	[+2.59, +3.11]	16/16
rows permuted − unguided	+3.51	[+3.23, +3.77]	16/16
channels permuted − unguided	+0.27	[+0.20, +0.33]	16/16
latent time reversed − unguided	+3.13	[+2.91, +3.33]	16/16

Table 25: Structural variants on a brightness target (AudioX, shelf, new prompts): does the separable 276-number matrix generalise beyond loudness?

direction source	error (dB)	95% interval	prompts
unguided	4.82	[4.53, 5.13]	9
signal-processing reference	1.17	[0.25, 1.96]	9
decoder gradient	2.38	[1.62, 3.43]	9
MaJa ($\bar{J}$)	1.87	[1.48, 2.22]	9
separable (one channel direction)	2.44	[2.03, 3.00]	9
8-bit quantised	1.87	[1.48, 2.22]	9
rank 4	3.75	[3.55, 4.02]	9
frame-local band only	4.10	[3.73, 4.51]	9
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.51	[-1.32, +0.04]	3/9
decoder gradient − separable (one channel direction)	+0.05	[-0.48, +0.49]	6/9
decoder gradient − 8-bit quantised	-0.51	[-1.32, +0.04]	3/9
decoder gradient − rank 4	+1.37	[+0.52, +2.07]	8/9
decoder gradient − frame-local band only	+1.71	[+0.98, +2.47]	8/9
decoder gradient − unguided	+2.44	[+1.59, +3.23]	8/9
MaJa ($\bar{J}$) − unguided	+2.95	[+2.53, +3.53]	9/9
separable (one channel direction) − unguided	+2.39	[+1.97, +2.84]	9/9
8-bit quantised − unguided	+2.95	[+2.53, +3.53]	9/9
rank 4 − unguided	+1.07	[+0.85, +1.30]	9/9
frame-local band only − unguided	+0.73	[+0.58, +0.83]	9/9

Table 26: Structural variants on a spectral-centroid target (AudioX, new prompts).

direction source	error (dB)	95% interval	prompts
unguided	3.08	[2.78, 3.39]	9
signal-processing reference	0.00	[0.00, 0.00]	9
decoder gradient	1.31	[1.10, 1.47]	9
MaJa ($\bar{J}$)	1.25	[1.09, 1.39]	9
separable (one channel direction)	1.97	[1.78, 2.12]	9
8-bit quantised	1.25	[1.10, 1.39]	9
rank 4	2.64	[2.39, 2.91]	9
frame-local band only	2.68	[2.41, 2.95]	9
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.05	[-0.14, +0.05]	4/9
decoder gradient − separable (one channel direction)	+0.66	[+0.59, +0.76]	9/9
decoder gradient − 8-bit quantised	-0.05	[-0.14, +0.05]	4/9
decoder gradient − rank 4	+1.34	[+1.09, +1.63]	9/9

continued

direction source	error (dB)	95% interval	prompts
decoder gradient − frame-local band only	+1.38	[+1.12, +1.67]	9/9
decoder gradient − unguided	+1.77	[+1.46, +2.12]	9/9
MaJa ($\bar{J}$) − unguided	+1.83	[+1.52, +2.17]	9/9
separable (one channel direction) − unguided	+1.11	[+0.85, +1.39]	9/9
8-bit quantised − unguided	+1.82	[+1.52, +2.17]	9/9
rank 4 − unguided	+0.43	[+0.36, +0.51]	9/9
frame-local band only − unguided	+0.39	[+0.34, +0.46]	9/9

Operators from other attributes. Section 4.1 reports one wrong-attribute control, the brightness matrix applied to a loudness target. This section runs the full grid on four attributes. For each target attribute, guidance is performed with that attribute's own mean Jacobian and with each of the other three, through the same rule and the same residual.

The result is not a clean diagonal, and the exceptions are informative. On a loudness target the three foreign matrices give 6.11, 7.27 and 5.00 dB against 5.59 unguided, so they are at best useless and at worst worse than doing nothing. On an onset-strength target all three give 1.27 to 1.31 against 1.29 unguided, that is exactly nothing. On a brightness target the centroid matrix reaches 2.51 against 4.82 unguided, and on a centroid target the brightness matrix reaches 1.16, which is better than the centroid's own matrix (1.25) and better than differentiating the decoder (1.31).

That last row is the honest qualification of the wrong-attribute control. Brightness and the spectral centroid are two ways of measuring the same spectral tilt, so their decoder Jacobians point the same way and their matrices are interchangeable. What the control establishes is therefore narrower than "the matrix must belong to the attribute" and more useful: the matrix must belong to the *direction in signal space* the attribute measures. Loudness, onset strength and spectral tilt are three such directions here, and a matrix built for one of them does nothing for the others.

Table 27: Operators from other attributes on a loudness target (AudioX, gain, new prompts): the brightness, spectral-centroid and onset-strength mean Jacobians applied to the loudness residual through the same rule.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.03	[1.83, 2.27]	16
MaJa ($\bar{J}$)	1.64	[1.41, 1.90]	16
brightness operator	6.11	[5.76, 6.52]	16
centroid operator	7.27	[6.84, 7.77]	16
onset-strength operator	5.00	[4.79, 5.19]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.39	[-0.50, -0.27]	3/16
decoder gradient − brightness operator	+4.07	[+3.61, +4.51]	16/16
decoder gradient − centroid operator	+5.23	[+4.74, +5.72]	16/16
decoder gradient − onset-strength operator	+2.97	[+2.56, +3.29]	16/16
decoder gradient − unguided	+3.55	[+3.16, +3.88]	16/16
MaJa ($\bar{J}$) − unguided	+3.94	[+3.59, +4.22]	16/16
brightness operator − unguided	-0.52	[-0.80, -0.24]	3/16
centroid operator − unguided	-1.68	[-2.06, -1.34]	0/16
onset-strength operator − unguided	+0.59	[+0.53, +0.64]	16/16

Table 28: Operators from other attributes on a brightness target (AudioX, shelf, new prompts).

direction source	error (dB)	95% interval	prompts
unguided	4.82	[4.53, 5.13]	9
signal-processing reference	1.17	[0.25, 1.96]	9
decoder gradient	2.38	[1.62, 3.43]	9
MaJa ($\bar{J}$)	1.87	[1.48, 2.22]	9
loudness operator	7.34	[6.10, 8.52]	9
centroid operator	2.51	[1.86, 3.45]	9
onset-strength operator	5.34	[4.86, 5.94]	9
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.51	[-1.32, +0.04]	3/9
decoder gradient − loudness operator	+4.95	[+4.20, +5.52]	9/9
decoder gradient − centroid operator	+0.13	[-0.24, +0.47]	6/9
decoder gradient − onset-strength operator	+2.96	[+2.44, +3.47]	9/9

continued

direction source	error (dB)	95% interval	prompts
decoder gradient − unguided	+2.44	[+1.59, +3.23]	8/9
MaJa ($\bar{J}$) − unguided	+2.95	[+2.53, +3.53]	9/9
loudness operator − unguided	-2.52	[-3.59, -1.19]	1/9
centroid operator − unguided	+2.31	[+1.58, +2.99]	9/9
onset-strength operator − unguided	-0.52	[-0.91, -0.16]	2/9

Table 29: Operators from other attributes on a spectral-centroid target (AudioX, new prompts).

direction source	error (dB)	95% interval	prompts
unguided	3.08	[2.78, 3.39]	9
signal-processing reference	0.00	[0.00, 0.00]	9
decoder gradient	1.31	[1.10, 1.47]	9
MaJa ($\bar{J}$)	1.25	[1.09, 1.39]	9
loudness operator	3.57	[3.09, 4.00]	9
brightness operator	1.16	[1.00, 1.31]	9
onset-strength operator	2.95	[2.66, 3.23]	9
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.05	[-0.14, +0.05]	4/9
decoder gradient − loudness operator	+2.26	[+1.85, +2.71]	9/9
decoder gradient − brightness operator	-0.15	[-0.23, -0.07]	2/9
decoder gradient − onset-strength operator	+1.64	[+1.36, +1.95]	9/9
decoder gradient − unguided	+1.77	[+1.46, +2.12]	9/9
MaJa ($\bar{J}$) − unguided	+1.83	[+1.52, +2.17]	9/9
loudness operator − unguided	-0.49	[-0.67, -0.26]	1/9
brightness operator − unguided	+1.92	[+1.59, +2.30]	9/9
onset-strength operator − unguided	+0.13	[+0.09, +0.19]	9/9

Table 30: Operators from other attributes on an onset-strength target (AudioX, new prompts).

direction source	error (dB)	95% interval	prompts
unguided	1.29	[1.01, 1.74]	9
signal-processing reference	0.00	[0.00, 0.00]	9
decoder gradient	0.88	[0.73, 1.18]	9
MaJa ($\bar{J}$)	0.87	[0.70, 1.19]	9
loudness operator	1.31	[1.05, 1.72]	9
brightness operator	1.27	[0.98, 1.73]	9
centroid operator	1.27	[0.99, 1.72]	9
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.01	[-0.05, +0.02]	4/9
decoder gradient − loudness operator	+0.43	[+0.31, +0.58]	9/9
decoder gradient − brightness operator	+0.39	[+0.24, +0.58]	9/9
decoder gradient − centroid operator	+0.39	[+0.25, +0.57]	9/9
decoder gradient − unguided	+0.41	[+0.27, +0.59]	9/9
MaJa ($\bar{J}$) − unguided	+0.42	[+0.31, +0.57]	9/9
loudness operator − unguided	-0.02	[-0.07, +0.03]	4/9
brightness operator − unguided	+0.02	[+0.00, +0.03]	6/9
centroid operator − unguided	+0.02	[+0.00, +0.04]	6/9

D.4 Conditions for the Advantage

This section supplements Section 6.4. Table 31 varies one thing at a time around the protocol values: the number of sampler steps, how many of them are guided and which, the magnitude of the target, the size of the head, and the fidelity of the decode. The advantage needs a few-step sampler, needs the direction to be repeated and repeated late, decays as the target moves away from the clip, and is not closed by a larger head or a cheaper decoder. Each block is a separate panel and is listed in Table 72.

Further target families. Table 1 uses four target families on loudness: a uniform gain, a linear fade, an envelope transferred from another clip, and a shelf on brightness. A natural question is whether the fade and the transfer are the only non-constant targets where a matrix beats a gradient, so this section adds three more, all on loudness and all with the same calibrated step size.

The tremolo target of Table 32 asks for four cycles of a ± 6 dB sinusoidal envelope, which changes sign eight times along a ten-second clip and is the fastest-varying target in the paper. The step

Table 31: Sweeps. Target error in dB; each block varies one thing and leaves the rest at the protocol values. The operator's advantage needs a few-step sampler, survives a threefold change of target magnitude, and is not closed by making the head four or six times larger.

Guidance steps (gain, new prompts)					
setting	exact	MaJa ($\bar{J}$, no parameters)	regression (teacher and student states)	VJP	
one step, at t=0.75	4.32	4.35	4.35	4.31	
two steps	2.34	2.41	2.71	2.35	
three steps	2.03	1.64	2.35	2.09	
Target magnitude (gain, new prompts)					
setting	exact	MaJa ($\bar{J}$, no parameters)	regression (teacher and student states)	VJP	
$\pm$3 dB	3.09	2.74	2.79	3.11	
$\pm$6 dB	2.03	1.64	2.35	2.09	
$\pm$12 dB	4.23	4.18	5.07	4.22	
Head capacity (gain, new prompts)					
setting	exact	MaJa ($\bar{J}$, no parameters)	regression (teacher and student states)	regr. 27.6M	regr. 41.3M
27.6M head	2.03	1.64	2.35	2.70	n/r
41.3M head	2.03	1.64	2.35	n/r	2.69
Sampler length (gain, new prompts)					
setting	exact	MaJa ($\bar{J}$, no parameters)	regression (teacher and student states)	VJP	
four-step student	2.03	1.64	2.35	2.09	
32-step teacher	4.21	4.24	4.33	4.22	

target of Table 33 asks for a discontinuity in the middle of the clip, which no smooth operator can express exactly and which therefore measures how gracefully each direction source fails. The random-envelope targets of Table 34 are two band-limited random envelopes with seeds fixed in the code, included because they are not targets we designed to suit any method.

Two of the three results are clean wins for the mean Jacobian and the third needs a caveat. On the tremolo target it reaches 1.86 dB against the decoder gradient's 2.91, and on the step target 1.71 against 2.24, with the fixed direction failing outright on both (3.32 and 6.05) because neither target can be expressed by one direction. On the random envelopes the ordering is the same but the setting is degenerate: a band-limited random envelope with zero mean is close to the unguided clip to begin with, so the unguided error is only 2.51 dB and there is little headroom. In that narrow regime the mean Jacobian is the only direction source that improves on not guiding at all (2.38), while the decoder gradient (2.72), the VJP-matched head (2.80) and the regression head (2.62) all make the clip worse than leaving it alone. We report this because it is the sharpest version of the paper's claim that we found, and we flag it as a small-headroom setting rather than leaning on it.

Each target is also run on the unseen-instrument panel (Tables 35 to 37), and this is where the three new targets say something the four original ones do not. On the gain and shelf targets the mean Jacobian's advantage turns slightly negative on unseen instrument classes. On all three time-varying targets it stays positive there: 0.85 against the decoder gradient's 0.93 dB on tremolo, 1.49 against 1.54 on the step, and 1.54 against 1.71 on the random envelopes, with the fixed direction failing on all three. Whether the advantage survives a change of instrument therefore depends on the target, and it survives on the targets a fixed direction cannot express at all.

Table 32: A four-cycle $\pm$6 dB tremolo target (AudioX loudness, new prompts). The target changes sign eight times along the clip, so a single fixed direction cannot express it.

direction source	error (dB)	95% interval	prompts
unguided	3.93	[3.78, 4.07]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.09	[1.94, 2.22]	16
fixed direction	4.69	[4.43, 4.97]	16
MaJa ($\bar{J}$)	1.49	[1.36, 1.62]	16
regression head (student)	2.64	[2.47, 2.78]	16
VJP-matched head (student)	2.06	[1.91, 2.22]	16
paired differences			
decoder gradient − fixed direction	+2.60	[+2.27, +2.93]	16/16
decoder gradient − MaJa ($\bar{J}$)	-0.60	[-0.68, -0.53]	0/16
decoder gradient − regression head (student)	+0.55	[+0.32, +0.75]	14/16
decoder gradient − VJP-matched head (student)	-0.03	[-0.12, +0.05]	8/16
decoder gradient − unguided	+1.84	[+1.58, +2.07]	16/16

continued

direction source	error (dB)	95% interval	prompts
fixed direction − unguided	-0.76	[-0.92, -0.63]	0/16
MaJa ($\bar{J}$) − unguided	+2.44	[+2.21, +2.63]	16/16
regression head (student) − unguided	+1.29	[+1.16, +1.44]	16/16
VJP-matched head (student) − unguided	+1.87	[+1.60, +2.07]	16/16

Table 33: A mid-clip step target, -6 dB on the first half and $+6$ dB on the second (AudioX loudness, new prompts).

direction source	error (dB)	95% interval	prompts
unguided	5.57	[5.36, 5.77]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.24	[2.08, 2.40]	16
fixed direction	6.05	[5.70, 6.38]	16
MaJa ($\bar{J}$)	1.71	[1.55, 1.84]	16
regression head (student)	2.98	[2.78, 3.16]	16
VJP-matched head (student)	2.22	[2.09, 2.34]	16
paired differences			
decoder gradient − fixed direction	+3.81	[+3.38, +4.23]	16/16
decoder gradient − MaJa ($\bar{J}$)	-0.53	[-0.61, -0.45]	1/16
decoder gradient − regression head (student)	+0.74	[+0.56, +0.91]	14/16
decoder gradient − VJP-matched head (student)	-0.02	[-0.11, +0.06]	10/16
decoder gradient − unguided	+3.33	[+3.02, +3.63]	16/16
fixed direction − unguided	-0.48	[-0.63, -0.34]	1/16
MaJa ($\bar{J}$) − unguided	+3.86	[+3.59, +4.12]	16/16
regression head (student) − unguided	+2.59	[+2.35, +2.84]	16/16
VJP-matched head (student) − unguided	+3.35	[+3.06, +3.61]	16/16

Table 34: Two band-limited random envelope targets with fixed seeds (AudioX loudness, new prompts).

direction source	error (dB)	95% interval	prompts
unguided	2.51	[2.37, 2.66]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.72	[2.58, 2.87]	16
fixed direction	3.64	[3.37, 3.94]	16
MaJa ($\bar{J}$)	2.38	[2.21, 2.57]	16
regression head (student)	2.62	[2.41, 2.84]	16
VJP-matched head (student)	2.80	[2.58, 3.04]	16
paired differences			
decoder gradient − fixed direction	+0.92	[+0.62, +1.19]	13/16
decoder gradient − MaJa ($\bar{J}$)	-0.33	[-0.39, -0.28]	0/16
decoder gradient − regression head (student)	-0.10	[-0.35, +0.08]	8/16
decoder gradient − VJP-matched head (student)	+0.08	[-0.01, +0.18]	13/16
decoder gradient − unguided	-0.21	[-0.44, -0.01]	7/16
fixed direction − unguided	-1.12	[-1.32, -0.97]	0/16
MaJa ($\bar{J}$) − unguided	+0.13	[-0.13, +0.34]	7/16
regression head (student) − unguided	-0.11	[-0.25, +0.03]	8/16
VJP-matched head (student) − unguided	-0.29	[-0.61, -0.04]	7/16

Table 35: The tremolo target on unseen instrument classes (AudioX loudness).

direction source	error (dB)	95% interval	prompts
unguided	4.20	[4.19, 4.22]	8
signal-processing reference	0.00	[0.00, 0.00]	8
decoder gradient	0.93	[0.76, 1.09]	8
fixed direction	4.71	[4.60, 4.82]	8
MaJa ($\bar{J}$)	0.85	[0.69, 1.01]	8
regression head (student)	2.42	[1.96, 2.92]	8
VJP-matched head (student)	1.07	[0.89, 1.24]	8
paired differences			
decoder gradient − fixed direction	+3.78	[+3.68, +3.88]	8/8
decoder gradient − MaJa ($\bar{J}$)	-0.08	[-0.17, +0.03]	2/8
decoder gradient − regression head (student)	+1.49	[+1.15, +1.88]	8/8
decoder gradient − VJP-matched head (student)	+0.14	[+0.07, +0.21]	8/8
decoder gradient − unguided	+3.27	[+3.11, +3.44]	8/8
fixed direction − unguided	-0.50	[-0.62, -0.40]	0/8
MaJa ($\bar{J}$) − unguided	+3.35	[+3.19, +3.51]	8/8

continued

direction source	error (dB)	95% interval	prompts
regression head (student) − unguided	+1.79	[+1.28, +2.24]	8/8
VJP-matched head (student) − unguided	+3.13	[+2.96, +3.31]	8/8

Table 36: The mid-clip step target on unseen instrument classes (AudioX loudness).

direction source	error (dB)	95% interval	prompts
unguided	5.96	[5.96, 5.97]	8
signal-processing reference	0.00	[0.00, 0.00]	8
decoder gradient	1.54	[1.28, 1.80]	8
fixed direction	6.33	[6.25, 6.41]	8
MaJa ($\bar{J}$)	1.49	[1.16, 1.84]	8
regression head (student)	3.01	[2.66, 3.40]	8
VJP-matched head (student)	1.62	[1.36, 1.89]	8
paired differences			
decoder gradient − fixed direction	+4.79	[+4.47, +5.09]	8/8
decoder gradient − MaJa ($\bar{J}$)	-0.05	[-0.16, +0.07]	4/8
decoder gradient − regression head (student)	+1.47	[+1.17, +1.79]	8/8
decoder gradient − VJP-matched head (student)	+0.08	[+0.03, +0.13]	7/8
decoder gradient − unguided	+4.42	[+4.16, +4.69]	8/8
fixed direction − unguided	-0.36	[-0.45, -0.29]	0/8
MaJa ($\bar{J}$) − unguided	+4.48	[+4.12, +4.80]	8/8
regression head (student) − unguided	+2.96	[+2.57, +3.30]	8/8
VJP-matched head (student) − unguided	+4.35	[+4.07, +4.61]	8/8

Table 37: The random envelope targets on unseen instrument classes (AudioX loudness).

direction source	error (dB)	95% interval	prompts
unguided	2.80	[2.80, 2.80]	8
signal-processing reference	0.00	[0.00, 0.00]	8
decoder gradient	1.70	[1.22, 2.20]	8
fixed direction	3.59	[3.41, 3.80]	8
MaJa ($\bar{J}$)	1.54	[1.10, 2.00]	8
regression head (student)	2.08	[1.42, 2.86]	8
VJP-matched head (student)	1.81	[1.39, 2.26]	8
paired differences			
decoder gradient − fixed direction	+1.89	[+1.58, +2.22]	8/8
decoder gradient − MaJa ($\bar{J}$)	-0.17	[-0.29, -0.07]	0/8
decoder gradient − regression head (student)	+0.37	[-0.15, +0.96]	4/8
decoder gradient − VJP-matched head (student)	+0.11	[-0.03, +0.23]	6/8
decoder gradient − unguided	+1.10	[+0.60, +1.58]	8/8
fixed direction − unguided	-0.79	[-1.00, -0.61]	0/8
MaJa ($\bar{J}$) − unguided	+1.26	[+0.80, +1.69]	8/8
regression head (student) − unguided	+0.72	[-0.06, +1.38]	7/8
VJP-matched head (student) − unguided	+0.99	[+0.54, +1.41]	7/8

D.5 MECHANISM

This section supplements Section 6.5. Figure 5 sketches where the three reference directions are taken and how survival is measured. Figure 6 and Table 38 measure what reaches the output. The intervention displaces the clean estimate by a known vector d; the guided run's final latent differs from the unguided one by D; survival is $\langle D, d\rangle / \|d\|^2$. More of the mean Jacobian's displacement survives the remaining steps than of the decoder gradient's, on every panel. Table 39 pairs that with the local picture: the decoder-measured target loss immediately before and after each intervention, which shows that the decoder gradient reduces the local loss more and the final error less. Together they say the difference is not in how good the step is where it is taken but in how much of it the sampler keeps.

Two further controls. Table 40 is the state-space smoothing control. If the mean Jacobian wins because averaging discards the state-specific part of the decoder gradient, then averaging the decoder gradient over small random perturbations of the state should recover part of the gap; it is the same kind of averaging, performed at guidance time instead of in advance. Table 41 is the saturation control: an operator built from the 80 states we used against one built from those 80 plus 48 states

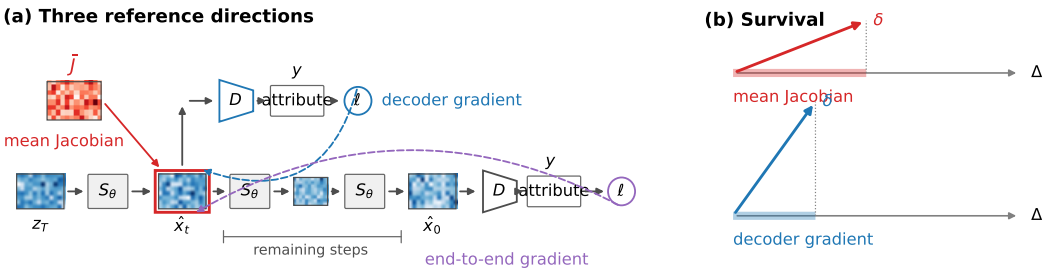

Figure 5: (a) Where each direction is taken at a guided state $\hat{x}_t$. The decoder gradient backpropagates the attribute loss through the decoder at $\hat{x}_t$ (blue). The end-to-end gradient backpropagates the loss of the final output through the remaining sampler steps and the decoder (purple). The mean Jacobian is precomputed offline and does not depend on $\hat{x}_t$ (red). (b) Survival: the injected displacement δ is projected onto the difference Δ between the guided and unguided final latents; the shaded segment is the part of the intervention that reaches the output. Schematic; measured values are in Table 38.

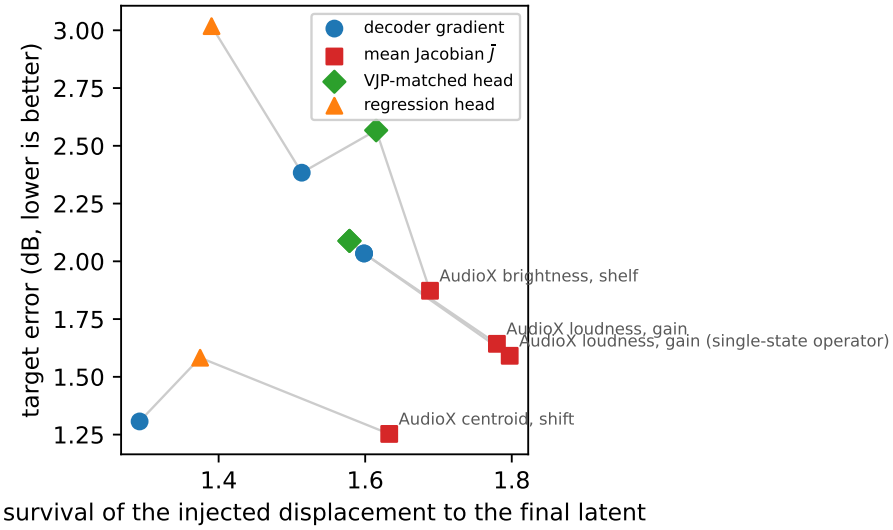

Figure 6: Survival against target error, one point per panel and direction source, with the points of one panel joined. Survival is $\langle D, d\rangle / \|d\|^2$ for d the injected displacement of the clean estimate and D the difference between the guided and unguided final latents, so it measures how much of the intervention reaches the output. In all four panels the mean Jacobian has both the highest survival and the lowest error. Survival does not order the rest: on brightness and on the centroid a head has higher survival than the decoder gradient and larger error, so survival is necessary for the advantage and not sufficient for good control.

drawn disjointly afterwards. It was already enough. The larger operator gives 1.639 dB against 1.643, a difference of four thousandths of a decibel, which together with the eight-state result of Table 2 brackets the useful range of N between about eight and eighty and says nothing is gained above it.

Table 38: What reaches the output. The intervention displaces the clean estimate by d; the guided run's final latent differs from the unguided one by D. Survival is $\langle D, d\rangle / \|d\|^2$, so 1 would mean the displacement passed through untouched and more than 1 that later steps compounded it. Participation ratio and kurtosis describe how concentrated the direction is across latent coordinates; a small participation ratio means a few coordinates carry it.

Panel	Direction	survival	$\cos(D, d)$	partic.	kurtosis
loudness, gain	decoder gradient	1.60	0.66	0.0112	112
	MaJa ($\bar{J}$, no parameters)	1.78	0.73	0.0132	85
	VJP-matched (student)	1.58	0.65	0.0110	111
brightness, shelf	decoder gradient	1.51	0.60	0.0248	50
	MaJa ($\bar{J}$, no parameters)	1.69	0.65	0.0268	39
	VJP-matched (student)	1.61	0.63	0.0239	51
	regression (teacher and student states)	1.39	0.61	0.0787	14
loudness, gain (single-state operator)	decoder gradient	1.60	0.66	0.0112	112
	MaJa ($\bar{J}$, no parameters)	1.80	0.73	0.0149	75
	VJP-matched (student)	1.58	0.65	0.0110	111
centroid, shift	decoder gradient	1.29	0.56	0.0092	169
	MaJa ($\bar{J}$, no parameters)	1.63	0.67	0.0209	58
	regression (teacher and student states)	1.37	0.61	0.0242	80
	regression (student)	1.30	0.58	0.0185	103
	regression (teacher)	1.33	0.62	0.0927	20

Table 39: The decoder gradient is the better local step and the worse guidance direction. Every direction source moves the clean estimate by the same fraction of its RMS on a given panel, 3 % for loudness and 2 % for brightness, so only the direction differs, and the local losses below are measured with the decoder for both rows. The local columns give the fraction of the decoder-measured target loss removed by the interventions themselves; the last column is the target error of the finished clip.

Panel	Direction	local, step 1	local, all steps	final (dB)
loudness, gain	decoder gradient	+0.722	+0.863	2.03
	MaJa ($\bar{J}$, no parameters)	+0.733	+0.913	1.64
loudness, envelope transfer	decoder gradient	+0.556	+0.850	3.71
	MaJa ($\bar{J}$, no parameters)	+0.548	+0.859	3.64
brightness, shelf	decoder gradient	+0.477	+0.938	2.38
	MaJa ($\bar{J}$, no parameters)	+0.436	+0.930	1.87
loudness, gain (single-state Jacobian)	decoder gradient	+0.722	+0.863	2.03
	MaJa ($\bar{J}$, no parameters)	+0.682	+0.922	1.59

Table 40: Smoothing the decoder gradient in state space rather than in coordinate space (AudioX loudness, gain, new prompts): the decoder gradient averaged over small random perturbations of the clean estimate, with two and eight probes. If the mean Jacobian won by discarding the state-specific part of the decoder gradient, this should recover part of the gap.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.03	[1.83, 2.27]	16
decoder gradient, state-space smoothed (2 probes)	2.05	[1.85, 2.29]	16
decoder gradient, state-space smoothed (8 probes)	2.05	[1.85, 2.30]	16
MaJa ($\bar{J}$)	1.64	[1.41, 1.90]	16
paired differences			
decoder gradient − decoder gradient, state-space smoothed (2 probes)	+0.02	[+0.01, +0.03]	11/16
decoder gradient − decoder gradient, state-space smoothed (8 probes)	+0.02	[+0.01, +0.04]	13/16
decoder gradient − MaJa ($\bar{J}$)	-0.39	[-0.50, -0.27]	3/16
decoder gradient − unguided	+3.55	[+3.16, +3.88]	16/16
decoder gradient, state-space smoothed (2 probes) − unguided	+3.53	[+3.14, +3.86]	16/16
decoder gradient, state-space smoothed (8 probes) − unguided	+3.53	[+3.13, +3.85]	16/16
MaJa ($\bar{J}$) − unguided	+3.94	[+3.59, +4.22]	16/16

Table 41: Saturation in the number of states (AudioX loudness, gain, new prompts): the mean Jacobian in use, built from 80 states, against one built from those 80 plus 48 disjoint further states.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16

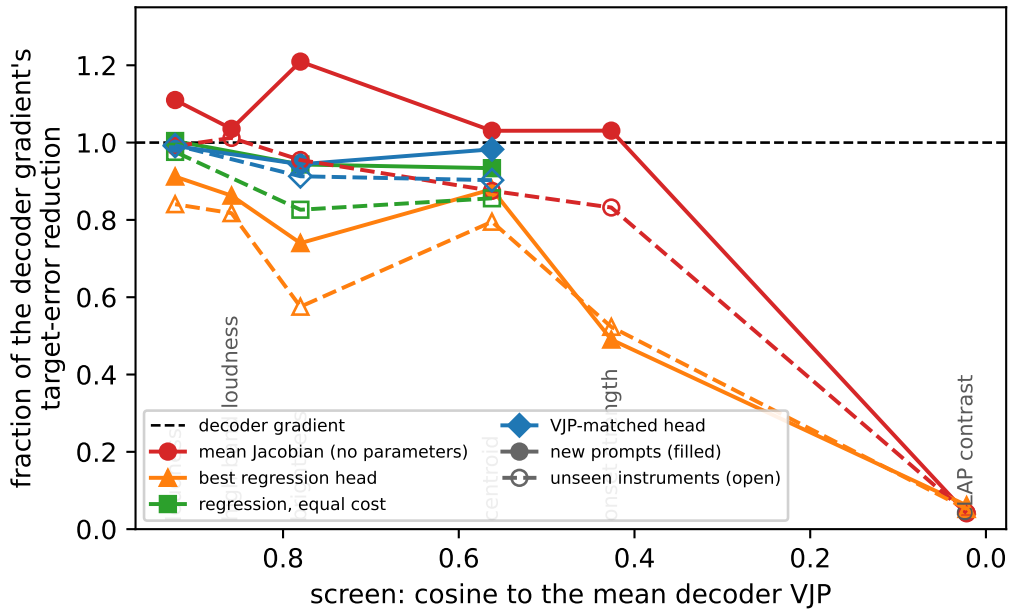

Figure 7: What a learned head buys, as a function of how state-dependent the attribute's decoder Jacobian is. The horizontal axis is the *fixed* column of the lower block of Table 42, the cosine between a held-out state's decoder vector-Jacobian product and the mean over the other states, measured on the same 48 AudioX student states at t=0.75 for all four attributes; it decreases to the right. The vertical axis is the fraction of the unguided target error that a direction source removes, divided by what the decoder gradient removes on the same panel, so 1.0 means as good as differentiating the decoder. Filled markers and solid lines are the new-prompt panels; open markers and dashed lines are the unseen-instrument panels, which are the panels where the mean Jacobian's advantage turns slightly negative. The mean Jacobian is above 1.0 on all four new-prompt panels and within 0.12 of it on all four unseen-instrument panels; every learned head is below it everywhere.

Table 42: The screen to run before training a guidance head: how much the decoder's Jacobian varies from state to state, measured with three backward passes per state. Pairwise: mean cosine between the decoder VJP at two different states. Fixed: mean cosine between a held-out state's VJP and the mean VJP over other states, i.e. the direction accuracy a zero-parameter rule can reach. The two blocks are different measurements and are not comparable with each other: the upper block uses the stored VJP labels, which cover three directions at three guidance times and both generators, while the lower block uses 48 AudioX student states at t=0.75 with the same three directions, which is the measurement the ordering in Figure 7 uses. Within one generator a higher value goes with less for a learned head to add; across generators the two columns do not order the settings correctly, as Section 4.1 reports.

Family	Attribute	pairwise	fixed
AudioX	loudness	0.92	0.95
AudioX	brightness	0.80	0.88
ACE-Step	loudness	0.72	0.81
further frame-level attributes, AudioX student states			
	loudness (RMS dB)	0.84	0.92
	band contrast	0.59	0.78
	spectral centroid	0.35	0.56
	spectral flux	0.20	0.43
	high-band loudness	0.71	0.86
	CLAP contrast	0.01	0.02

Table 43: The state-dependence screen in full: every attribute and every fixed direction in attribute space, on the same 48 AudioX student states at t=0.75. Pairwise is the mean cosine between the decoder vector-Jacobian product at two different states; the last column is the mean cosine between a held-out state's product and the mean over the others, which is the direction accuracy a zero-parameter rule can reach. u_0 is the constant direction and u_1, u_2 are the two fixed random directions. Table 42 in the main appendix reports the per-attribute means of these rows.

attribute	direction	states	pairwise cosine	cosine to mean
loudness (RMS dB)	u_0	48	0.876	0.942
loudness (RMS dB)	u_1	48	0.812	0.910
loudness (RMS dB)	u_2	48	0.825	0.916
loudness (RMS dB)	mean	48	**0.838**	**0.923**
band contrast	u_0	48	0.633	0.811
band contrast	u_1	48	0.574	0.769
band contrast	u_2	48	0.557	0.761
band contrast	mean	48	**0.588**	**0.780**
spectral centroid	u_0	48	0.377	0.588
spectral centroid	u_1	48	0.332	0.543
spectral centroid	u_2	48	0.334	0.556
spectral centroid	mean	48	**0.348**	**0.562**
spectral flux	u_0	48	0.056	0.195
spectral flux	u_1	48	0.273	0.545
spectral flux	u_2	48	0.268	0.539
spectral flux	mean	48	**0.199**	**0.426**
high-band loudness	u_0	48	0.746	0.883
high-band loudness	u_1	48	0.690	0.836
high-band loudness	u_2	48	0.688	0.857
high-band loudness	mean	48	**0.708**	**0.859**
CLAP contrast	u_0	48	0.005	0.023
CLAP contrast	u_1	48	0.006	0.027
CLAP contrast	u_2	48	0.005	0.016
CLAP contrast	mean	48	**0.005**	**0.022**

D.7 Decoder-Free Guidance, Cost, and Quality Guards

This section supplements Section 6.7. Table 44 is the decoder-free rule: a head supplies the attribute's value and the precomputed matrix supplies the direction. Table 45 times one loss-to-gradient evaluation and records peak memory for both paths at several clip lengths, which is where the decoder path stops fitting a 48 GB GPU while the head path does not. Table 46 and Table 47 are the head-side controls: the equal-cost regression head, and the sensitivity of a VJP-matched head to its initialisation, to λ and to the number of cotangent directions.

Two attributes at once. Table 48 requests a gain and a shelf together and scores both attributes on the finished clip, so a method that buys one attribute at the other's expense is visible. Summing two precomputed operators works; summing two heads' gradients does not.

Table 44: Value and direction are separate jobs (AudioX loudness, gain target, calibrated multi-step). Each row pairs the value of one head with the direction of the precomputed Jacobian; neither calls the decoder at inference. The head alone and the decoder gradient are shown for reference, as is the mean Jacobian reading its residual from the decoder.

Panel	Value from	head alone	+ mean Jacobian
new prompts	VJP-matched head	2.09	1.65
	regression head, equal cost	2.02	1.71
	regression head (both)	2.35	2.11
unseen instr.	VJP-matched head	0.96	1.01
	regression head, equal cost	1.05	1.08
	regression head (both)	1.73	1.52
new prompts	exact decoder	2.03	1.64
unseen instr.	exact decoder	0.93	0.97

Table 45: Cost of one guidance gradient (loss to gradient w.r.t. the latent clean estimate), batch 1, fp32, one RTX A6000 (48 GB): wall time in ms and peak memory in GB. ACE latents beyond 20 s are tiled copies of a 20 s state.

	exact decoder		VJP-matched head	
Clip	ms	GB	ms	GB
AudioX, 10 s	465	11.1	35	0.9
ACE, 20 s	465	26.8	32	6.6
ACE, 40 s	out of memory		34	6.6
ACE, 60 s	out of memory		34	6.6
ACE, 120 s	out of memory		59	6.7

Table 46: Equal-cost comparison (AudioX loudness). The four middle rows are the equal-cost baselines: 12 extra generation seeds per training prompt ($4\times$ the student data), the VJP fits' precision and batch (fp32, 16), and up to a wall-clock-matched number of updates, with no Jacobian term. The two top rows are the standard regression heads as trained throughout the paper (bf16 autocast, batch 32, 10k updates) and are shown for reference, not as equal-cost baselines. Budget: training states, updates, training wall time on one A6000 (label time for the VJP heads: about one hour of decoder backward passes, once). Outcome: cosine to the decoder gradient on student states (familiar / unseen instruments) and calibrated multi-step control error on the new-prompt and held-out panels.

Head	states	updates	prec./batch	train s	cos fam.	cos h.-o.	new	h.-o.
regression (student)	1536	10k	bf16/32	481	0.78	0.55	2.62	2.08
regression (both)	3072	10k	bf16/32	1296	0.83	0.58	2.35	1.73
regression (both, $4\times$ student data, 10k upd.)	7680	10k	fp32/16	376	0.89	0.65	2.12	1.38
regression (both, $4\times$ student data, 175k upd.)	7680	175k	fp32/16	11716	0.93	0.72	2.06	1.18
regression (student, $4\times$ student data, 10k upd.)	6144	10k	fp32/16	370	0.95	0.74	2.02	1.23
regression (student, $4\times$ student data, 175k upd.)	6144	175k	fp32/16	8139	0.96	0.77	2.02	1.05
VJP-matched (student)	1152	10k	fp32/16	6436	0.99	0.75	2.09	0.96
VJP-matched (both)	2304	10k	fp32/16	6428	0.98	0.74	2.06	1.04

Table 47: Ablations of VJP matching (loudness, AudioX student). Cosine to the decoder gradient on student states (familiar / unseen instruments) and calibrated multi-step control error on the new-prompt and held-out panels; the decoder gradient gives 2.03 / 0.93 dB and the regression heads 2.3 to 3.1 / 1.7 to 2.4 dB.

Variant	cos fam.	cos h.-o.	new	h.-o.
reference (student, 29101)	0.988	0.749	2.09	0.96
student, 29103	0.988	0.750	2.06	1.01
student, 29107	0.987	0.801	2.12	0.99
teacher, 29103	0.978	0.760	2.13	1.07
teacher, 29107	0.979	0.788	2.06	1.04
$\lambda = 0.1$	0.982	0.767	2.15	1.03
$\lambda = 10$	0.993	0.714	2.04	0.99
constant dir. only	0.990	0.724	2.09	1.04

Table 48: Controlling two attributes at once (AudioX, a ± 6 dB gain and a ± 6 dB shelf requested together, calibrated multi-step probe). Both errors are measured on the finished clip, so a method that buys one attribute with the other is visible. Summing the two precomputed operators works; summing the two heads' gradients does not.

	new prompts		unseen instruments	
Feature function	loudness	brightness	loudness	brightness
no guidance	5.97	4.85	6.02	4.68
signal-processing reference	0.00	0.00	0.00	0.00
decoder gradient, both terms	1.68	1.56	1.54	1.48
the two operators, summed	1.48	1.67	1.41	1.75
the two regression heads, summed	2.52	2.86	2.56	2.84

Quality guards. Table 49 is the distribution guard: the Fréchet distance between CLAP audio embeddings of the guided and unguided sets of the same panel, and the change in CLAP text-audio cosine. The purpose of the guard is to rule out the explanation that a direction source wins by degrading the audio more, and it does rule that out: the signal-processing reference, which attains the target exactly and is by construction the intended change, moves these numbers by as much as the guidance methods do. Neither number is a perceptual measure and we make no perceptual claim; a listening study is the obvious missing piece and is named in the limitations.

Table 49: Distribution and semantic guards per panel and direction source. FAD: Fréchet distance between CLAP audio embeddings of the guided clips and of the unguided clips of the same panel (lower is closer). ΔCLAP: mean change in CLAP text-audio cosine (guided minus unguided). corr.: mean gain-invariant waveform correlation to the unguided clip. The signal-processing reference row of each panel attains the target exactly and is therefore the size of the intended change, not an artefact. None of the three is a perceptual quality measure and we make no perceptual claim.

Panel	Feature function	FAD	ΔCLAP	corr.
AudioX loudness, gain, new	signal-processing reference	0.015	n/r	n/r
	decoder gradient	0.027	-0.004	0.972
	regression (teacher)	0.017	+0.003	0.982
	regression (student)	0.020	-0.000	0.979
	regression (teacher and student states)	0.021	-0.002	0.973
	VJP-matched (student)	0.028	-0.004	0.971
	VJP-matched (teacher and student states)	0.030	-0.003	0.972
	VJP-matched (teacher)	0.028	-0.004	0.971
AudioX loudness, gain, new ($\bar{J}$ run)	signal-processing reference	0.015	n/r	n/r
	decoder gradient	0.027	-0.004	0.972
	MaJa ($\bar{J}$, no parameters)	0.033	-0.006	0.977
	VJP-matched (student)	0.028	-0.004	0.971
AudioX loudness, gain, unseen instruments	signal-processing reference	0.027	n/r	n/r
	decoder gradient	0.030	n/r	0.994
	regression (teacher)	0.019	n/r	0.992
	regression (student)	0.024	n/r	0.993
	regression (teacher and student states)	0.024	n/r	0.993
	VJP-matched (student)	0.029	n/r	0.994
	VJP-matched (teacher and student states)	0.029	n/r	0.993
	VJP-matched (teacher)	0.029	n/r	0.993
AudioX loudness, gain, unseen instruments ($\bar{J}$ run)	signal-processing reference	0.027	n/r	n/r
	decoder gradient	0.030	n/r	0.994
	fixed direction (no parameters)	0.036	n/r	0.992
	MaJa ($\bar{J}$, no parameters)	0.035	n/r	0.994
	VJP-matched (student)	0.029	n/r	0.994
	regression (teacher and student states)	0.024	n/r	0.993
AudioX loudness, transfer, new	signal-processing reference	0.069	n/r	n/r
	decoder gradient	0.047	+0.005	0.960
	regression (teacher)	0.024	+0.005	0.966
	regression (student)	0.032	-0.002	0.970
	regression (teacher and student states)	0.037	+0.004	0.960
	VJP-matched (student)	0.046	+0.004	0.960
	VJP-matched (teacher and student states)	0.047	+0.005	0.959
	VJP-matched (teacher)	0.045	+0.006	0.960
	VJP-matched (student), $K=0$	0.048	+0.004	0.960
	decoder gradient	0.044	-0.001	0.960
	regression (teacher)	0.025	-0.001	0.966
	regression (student)	0.031	-0.008	0.970
	regression (teacher and student states)	0.036	-0.003	0.960
	VJP-matched (student)	0.044	-0.002	0.960
	VJP-matched (teacher and student states)	0.044	-0.002	0.959
	VJP-matched (teacher)	0.042	-0.001	0.960
	VJP-matched (student), $K=0$	0.045	-0.003	0.960
AudioX loudness, transfer, new ($\bar{J}$ run)	signal-processing reference	0.069	n/r	n/r
	decoder gradient	0.047	+0.005	0.960
	MaJa ($\bar{J}$, no parameters)	0.045	+0.001	0.963
	VJP-matched (student)	0.046	+0.004	0.960
	decoder gradient	0.044	-0.001	0.960
	MaJa ($\bar{J}$, no parameters)	0.043	-0.006	0.963
	VJP-matched (student)	0.044	-0.002	0.960
AudioX loudness, fade, new	signal-processing reference	0.006	n/r	n/r
	decoder gradient	0.021	-0.002	0.924
	regression (teacher)	0.016	+0.004	0.969
	regression (student)	0.019	+0.003	0.959
	regression (teacher and student states)	0.023	+0.000	0.937
	VJP-matched (student)	0.019	-0.001	0.925
	VJP-matched (teacher and student states)	0.022	+0.000	0.923
	VJP-matched (teacher)	0.021	+0.001	0.923
	VJP-matched (student), $K=0$	0.020	+0.001	0.927

Panel	Feature function	FAD	ΔCLAP	corr.
	decoder gradient	0.026	-0.002	0.924
	regression (teacher)	0.019	+0.005	0.969
	regression (student)	0.023	+0.000	0.959
	regression (teacher and student states)	0.025	+0.005	0.937
	VJP-matched (student)	0.028	-0.001	0.925
	VJP-matched (teacher and student states)	0.030	-0.000	0.923
	VJP-matched (teacher)	0.030	+0.000	0.923
	VJP-matched (student), $K=0$	0.028	-0.000	0.927
AudioX loudness, fade, new (fixed-direction run)	signal-processing reference	0.006	n/r	n/r
	decoder gradient	0.021	-0.002	0.924
	fixed direction (no parameters)	0.007	-0.003	0.997
	VJP-matched (student)	0.019	-0.001	0.925
	regression (teacher and student states)	0.023	+0.000	0.937
	decoder gradient	0.026	-0.002	0.924
	fixed direction (no parameters)	0.009	-0.002	0.997
	VJP-matched (student)	0.028	-0.001	0.925
	regression (teacher and student states)	0.025	+0.005	0.937
AudioX loudness, gain, new (fixed-direction run)	signal-processing reference	0.015	n/r	n/r
	decoder gradient	0.027	-0.004	0.972
	fixed direction (no parameters)	0.033	-0.006	0.980
	VJP-matched (student)	0.028	-0.004	0.971
	regression (teacher and student states)	0.021	-0.002	0.973
AudioX loudness, transfer, new (fixed-direction run)	signal-processing reference	0.069	n/r	n/r
	decoder gradient	0.047	+0.005	0.960
	fixed direction (no parameters)	0.040	-0.006	0.983
	VJP-matched (student)	0.046	+0.004	0.960
	regression (teacher and student states)	0.037	+0.004	0.960
	decoder gradient	0.044	-0.001	0.960
	fixed direction (no parameters)	0.040	-0.010	0.983
	VJP-matched (student)	0.044	-0.002	0.960
	regression (teacher and student states)	0.036	-0.003	0.960
AudioX brightness, shelf, new	signal-processing reference	0.020	n/r	n/r
	decoder gradient	0.345	-0.003	0.995
	regression (teacher)	0.348	-0.002	0.998
	regression (student)	0.348	-0.001	0.998
	regression (teacher and student states)	0.348	-0.002	0.997
	VJP-matched (student)	0.345	-0.002	0.995
	VJP-matched (teacher and student states)	0.345	-0.002	0.995
	VJP-matched (teacher)	0.346	-0.001	0.995
AudioX brightness, shelf, new ($\bar{J}$ run)	signal-processing reference	0.020	n/r	n/r
	decoder gradient	0.345	n/r	0.995
	MaJa ($\bar{J}$, no parameters)	0.345	n/r	0.995
	VJP-matched (student)	0.345	n/r	0.995
	regression (teacher and student states)	0.348	n/r	0.997
AudioX brightness, shelf, unseen instruments ($\bar{J}$ run)	signal-processing reference	0.018	n/r	n/r
	decoder gradient	0.123	n/r	0.993
	MaJa ($\bar{J}$, no parameters)	0.126	n/r	0.994
	VJP-matched (student)	0.123	n/r	0.993
	regression (teacher and student states)	0.121	n/r	0.997
AudioX centroid, new	signal-processing reference	0.073	n/r	n/r
	decoder gradient	0.344	-0.001	0.987
	MaJa ($\bar{J}$, no parameters)	0.349	-0.004	0.988
	regression (teacher and student states)	0.347	-0.003	0.991
	regression (student)	0.348	-0.003	0.991
	regression (teacher)	0.352	-0.003	0.995
AudioX centroid, new (VJP-head run)	signal-processing reference	0.073	n/r	n/r
	decoder gradient	0.344	-0.001	0.987
	VJP-matched (student)	0.344	-0.002	0.987
	VJP-matched (teacher and student states)	0.344	-0.001	0.986
	VJP-matched (teacher)	0.346	-0.003	0.987
ACE loudness, gain, new	signal-processing reference	0.079	n/r	n/r
	decoder gradient	0.063	-0.015	0.977
	regression (teacher)	0.074	-0.018	0.983
	regression (student)	0.058	-0.018	0.985
	regression (teacher and student states)	0.064	-0.016	0.983
	VJP-matched (student)	0.064	-0.013	0.975
	VJP-matched (teacher and student states)	0.068	-0.001	0.976
	VJP-matched (teacher)	0.073	-0.026	0.975
ACE loudness, gain, new ($\bar{J}$ run)	signal-processing reference	0.075	n/r	n/r
	decoder gradient	0.067	n/r	0.977
	MaJa ($\bar{J}$, no parameters)	0.079	n/r	0.975
	VJP-matched (student)	0.067	n/r	0.975
ACE loudness, gain, unseen classes	signal-processing reference	0.105	n/r	n/r
	decoder gradient	0.131	+0.004	0.978
	regression (teacher)	0.108	-0.010	0.987

Panel	Feature function	FAD	ΔCLAP	corr.
	regression (student)	0.068	+0.004	0.990
	regression (teacher and student states)	0.124	-0.000	0.987
ACE loudness, gain, unseen classes ($\bar{J}$ run)	signal-processing reference	0.117	n/r	n/r
	decoder gradient	0.120	n/r	0.978
	MaJa ($\bar{J}$, no parameters)	0.122	n/r	0.979
	VJP-matched (student)	0.124	n/r	0.978
	regression (teacher and student states)	0.117	n/r	0.987
AudioX loudness, gain, new (equal-cost run)	signal-processing reference	0.015	n/r	n/r
	decoder gradient	0.027	-0.004	0.972
	regression (teacher and student states, $4\times$ data)	0.024	-0.005	0.973
	regression (teacher and student states, $4\times$ data, equal cost)	0.027	-0.004	0.972
	regression (student, $4\times$ data)	0.026	-0.005	0.973
	regression (student, $4\times$ data, equal cost)	0.028	-0.004	0.973
	VJP-matched (student)	0.028	-0.004	0.971
	regression (teacher and student states)	0.021	-0.002	0.973
AudioX loudness, gain, unseen instruments (equal-cost run)	signal-processing reference	0.027	n/r	n/r
	decoder gradient	0.030	-0.007	0.994
	regression (teacher and student states, $4\times$ data)	0.026	-0.004	0.993
	regression (teacher and student states, $4\times$ data, equal cost)	0.028	-0.005	0.993
	regression (student, $4\times$ data)	0.029	-0.006	0.993
	regression (student, $4\times$ data, equal cost)	0.030	-0.007	0.994
	VJP-matched (student)	0.029	-0.008	0.994
	regression (teacher and student states)	0.024	-0.003	0.993

D.8 High-Band Loudness

High-band loudness, the level of the $[4, 16)$ kHz band in dB, was added after the other four for two reasons. It sits between loudness and brightness in the screen, so it fills the gap in Figure 7; and a zero-phase gain restricted to that band changes it by exactly the requested amount while leaving the rest of the spectrum alone, so the target is attainable by a simpler operation than either the shelf or the transient shaping. Tables 50 and 51 are the two panels, built with the same pipeline as the others: labels collected from the same 512 stored trajectories, three regression heads, an eight-state mean Jacobian, a strength calibrated on the development prompts with the decoder gradient alone, and then both panels scored once.

Table 50: High-band loudness, a ± 6 dB gain on $[4, 16)$ kHz, new prompts (AudioX). The fifth attribute, added because a band-limited gain attains its target exactly and because it sits between loudness and brightness in the screen.

direction source	error (dB)	95% interval	prompts
unguided	5.99	[5.98, 6.00]	9
signal-processing reference	0.00	[0.00, 0.00]	9
decoder gradient	1.62	[1.27, 2.07]	9
MaJa ($\bar{J}$)	1.47	[1.14, 1.88]	9
regression head (student)	2.22	[1.85, 2.79]	9
regression head (teacher and student states)	2.39	[1.99, 2.97]	9
regression head (teacher)	3.66	[3.18, 4.35]	9
paired differences			
regression head (student) − regression head (teacher)	+1.44	[+1.21, +1.68]	9/9
regression head (teacher and student states) − regression head (teacher)	+1.27	[+1.06, +1.51]	9/9
decoder gradient − MaJa ($\bar{J}$)	-0.16	[-0.36, -0.00]	3/9
decoder gradient − regression head (student)	+0.60	[+0.48, +0.77]	9/9
decoder gradient − regression head (teacher and student states)	+0.77	[+0.65, +0.93]	9/9
decoder gradient − regression head (teacher)	+2.04	[+1.76, +2.36]	9/9
decoder gradient − unguided	+4.37	[+3.92, +4.72]	9/9
MaJa ($\bar{J}$) − unguided	+4.53	[+4.12, +4.85]	9/9
regression head (student) − unguided	+3.77	[+3.20, +4.14]	9/9
regression head (teacher and student states) − unguided	+3.60	[+3.02, +4.00]	9/9
regression head (teacher) − unguided	+2.33	[+1.64, +2.82]	9/9

Table 51: High-band loudness on unseen instrument classes (AudioX).

direction source	error (dB)	95% interval	prompts
unguided	6.00	[5.99, 6.00]	8
signal-processing reference	0.00	[0.00, 0.00]	8
decoder gradient	1.85	[1.57, 2.22]	8
MaJa ($\bar{J}$)	1.80	[1.39, 2.22]	8
regression head (student)	2.60	[2.20, 3.00]	8
regression head (teacher and student states)	2.86	[2.32, 3.36]	8
regression head (teacher)	3.45	[2.93, 3.84]	8
paired differences			
regression head (student) − regression head (teacher)	+0.85	[+0.41, +1.29]	7/8
regression head (teacher and student states) − regression head (teacher)	+0.60	[+0.01, +1.18]	6/8
decoder gradient − MaJa ($\bar{J}$)	-0.05	[-0.27, +0.23]	2/8
decoder gradient − regression head (student)	+0.76	[+0.40, +1.11]	7/8
decoder gradient − regression head (teacher and student states)	+1.01	[+0.56, +1.50]	8/8
decoder gradient − regression head (teacher)	+1.60	[+0.97, +2.14]	7/8
decoder gradient − unguided	+4.15	[+3.77, +4.43]	8/8
MaJa ($\bar{J}$) − unguided	+4.20	[+3.77, +4.61]	8/8
regression head (student) − unguided	+3.39	[+2.99, +3.80]	8/8
regression head (teacher and student states) − unguided	+3.14	[+2.64, +3.68]	8/8
regression head (teacher) − unguided	+2.55	[+2.15, +3.06]	8/8

D.9 Results on ACE-Step

ACE-Step 1.5 is the second teacher/student pair and the one where the mean Jacobian loses. This section gives the panels and our reading of why.

Differences from AudioX. The generator is a 48 kHz stereo music model rather than a 44.1 kHz general-audio model; its codec is an Oobleck autoencoder over 497 latent frames rather than a 44.1 kHz variational autoencoder over 212; its student takes eight linear Euler steps rather than four stochastic ones, so guidance acts at seven non-terminal times instead of three; and its teacher and student are separate public releases that generate different music from the same prompt, which is why the teacher-to-student latent distance in Figure 4 is dominated by that difference rather than by distillation. The operator is built from 48 states, eight at each of six times.

Results. Tables 52 and 54 are the two new-prompt panels. On both the mean Jacobian is behind the decoder gradient by an interval that excludes zero, by 0.14 dB on loudness and 0.34 on brightness, and on both it is still ahead of every regression head. Tables 53 and 55 are the unseen-class panels, where the two are level. These are the only two losses in the paper and we do not explain them away.

Interpretation. Two differences are candidates. The first is the number of guidance steps: ACE guides at seven times spread over the whole trajectory, and Table 31 shows in the first family that the mean Jacobian's advantage comes from the late steps and vanishes when guidance is spread thin or applied early. The second is state dependence: ACE loudness screens at 0.72 pairwise and 0.81 to the mean, the lowest loudness figure in Table 42, so one fixed matrix has more to average over. Against the first, the per-time operators of Table 56 narrow the gap without closing it: an operator built from the eight states at $t{=}0.25$ and applied at every step reaches 2.58 dB against the pooled operator's 2.67 and the decoder gradient's 2.54, while the operators from $t{=}0.875$ and $t{=}0.125$ are worse than the pooled one. One matrix for seven times is therefore not the whole problem, although the source time matters more here than in the first family, where a per-time operator changes nothing. Against the second, AudioX brightness screens lower on the pairwise column and the mean Jacobian wins there. We therefore report the two losses as a boundary we can locate but not yet explain, which is also why Section 4.3 presents the screen as a measure of how little a head can add rather than as a decision rule.

Controls. Tables 57 and 58 run the wrong-attribute control in ACE-Step: the brightness operator on a loudness target and the loudness operator on a brightness target. Both are worse than not guiding, as in the first family, so whatever ACE changes it does not change the fact that the matrix has to belong to the attribute.

Table 52: ACE-Step loudness, ± 6 dB gain, new prompts, calibrated seven-step probe. This is one of the two panels where the mean Jacobian is behind the decoder gradient.

direction source	error (dB)	95% interval	prompts
unguided	6.00	[6.00, 6.00]	12
signal-processing reference	0.00	[0.00, 0.00]	12
decoder gradient	2.54	[2.43, 2.66]	12
MaJa ($\bar{J}$)	2.67	[2.55, 2.79]	12
VJP-matched head (student)	2.59	[2.48, 2.71]	12
paired differences			
decoder gradient − MaJa ($\bar{J}$)	+0.14	[+0.00, +0.25]	8/12
decoder gradient − VJP-matched head (student)	+0.06	[+0.04, +0.07]	12/12
decoder gradient − unguided	+3.46	[+3.34, +3.57]	12/12
MaJa ($\bar{J}$) − unguided	+3.33	[+3.21, +3.45]	12/12
VJP-matched head (student) − unguided	+3.41	[+3.29, +3.52]	12/12

Table 53: ACE-Step loudness, gain, unseen classes.

direction source	error (dB)	95% interval	prompts
unguided	6.00	[6.00, 6.00]	8
signal-processing reference	0.00	[0.00, 0.00]	8
decoder gradient	3.06	[2.93, 3.22]	8
MaJa ($\bar{J}$)	2.92	[2.86, 2.98]	8
VJP-matched head (student)	3.07	[2.95, 3.21]	8
regression head (teacher and student states)	3.76	[3.60, 3.90]	8
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.15	[-0.33, +0.03]	4/8
decoder gradient − VJP-matched head (student)	+0.01	[-0.03, +0.04]	6/8
decoder gradient − regression head (teacher and student states)	+0.70	[+0.47, +0.92]	8/8
decoder gradient − unguided	+2.94	[+2.78, +3.07]	8/8
MaJa ($\bar{J}$) − unguided	+3.08	[+3.02, +3.14]	8/8
VJP-matched head (student) − unguided	+2.93	[+2.79, +3.05]	8/8
regression head (teacher and student states) − unguided	+2.24	[+2.10, +2.40]	8/8

Table 54: ACE-Step brightness, ± 6 dB shelf, new prompts. The second panel where the mean Jacobian is behind.

direction source	error (dB)	95% interval	prompts
unguided	5.17	[4.99, 5.39]	12
signal-processing reference	0.29	[0.13, 0.49]	12
decoder gradient	1.46	[1.30, 1.66]	12
MaJa ($\bar{J}$)	1.80	[1.56, 2.06]	12
regression head (student)	4.16	[3.95, 4.43]	12
regression head (teacher and student states)	3.97	[3.76, 4.23]	12
regression head (teacher)	4.10	[3.91, 4.30]	12
paired differences			
regression head (student) − regression head (teacher)	-0.06	[-0.14, +0.00]	3/12
regression head (teacher and student states) − regression head (teacher)	+0.13	[+0.02, +0.25]	9/12
decoder gradient − MaJa ($\bar{J}$)	+0.34	[+0.12, +0.57]	10/12
decoder gradient − regression head (student)	+2.70	[+2.39, +3.01]	12/12
decoder gradient − regression head (teacher and student states)	+2.51	[+2.18, +2.81]	12/12
decoder gradient − regression head (teacher)	+2.64	[+2.36, +2.89]	12/12
decoder gradient − unguided	+3.71	[+3.40, +3.97]	12/12
MaJa ($\bar{J}$) − unguided	+3.37	[+3.20, +3.53]	12/12
regression head (student) − unguided	+1.01	[+0.94, +1.07]	12/12
regression head (teacher and student states) − unguided	+1.20	[+1.11, +1.30]	12/12
regression head (teacher) − unguided	+1.07	[+1.01, +1.15]	12/12

Table 55: ACE-Step brightness, shelf, unseen classes.

direction source	error (dB)	95% interval	prompts
unguided	4.98	[4.74, 5.22]	8
signal-processing reference	0.31	[0.00, 0.73]	8
decoder gradient	1.65	[1.27, 2.07]	8
MaJa ($\bar{J}$)	1.61	[1.37, 1.84]	8

continued

direction source	error (dB)	95% interval	prompts
regression head (student)	4.39	[4.07, 4.72]	8
regression head (teacher and student states)	4.38	[4.11, 4.62]	8
regression head (teacher)	4.03	[3.91, 4.16]	8
paired differences			
regression head (student) − regression head (teacher)	-0.36	[-0.57, -0.15]	2/8
regression head (teacher and student states) − regression head (teacher)	-0.35	[-0.52, -0.16]	1/8
decoder gradient − MaJa ($\bar{J}$)	-0.04	[-0.38, +0.32]	5/8
decoder gradient − regression head (student)	+2.74	[+2.08, +3.33]	8/8
decoder gradient − regression head (teacher and student states)	+2.73	[+2.23, +3.21]	8/8
decoder gradient − regression head (teacher)	+2.38	[+1.88, +2.81]	8/8
decoder gradient − unguided	+3.33	[+2.73, +3.86]	8/8
MaJa ($\bar{J}$) − unguided	+3.37	[+3.06, +3.60]	8/8
regression head (student) − unguided	+0.59	[+0.47, +0.69]	8/8
regression head (teacher and student states) − unguided	+0.60	[+0.42, +0.73]	8/8
regression head (teacher) − unguided	+0.95	[+0.83, +1.08]	8/8

Table 56: One operator per guidance time against one operator for all of them (ACE loudness, gain, new prompts). Each per-time operator is built from the eight states stored at that time and then applied at every guidance step, so this measures how much the Jacobian moves along the trajectory in the second family.

direction source	error (dB)	95% interval	prompts
unguided	6.00	[6.00, 6.00]	12
signal-processing reference	0.00	[0.00, 0.00]	12
decoder gradient	2.54	[2.43, 2.66]	12
MaJa ($\bar{J}$)	2.67	[2.55, 2.79]	12
t875	2.84	[2.73, 2.95]	12
t250	2.58	[2.45, 2.71]	12
t125	2.69	[2.57, 2.81]	12
paired differences			
decoder gradient − MaJa ($\bar{J}$)	+0.14	[+0.00, +0.25]	8/12
decoder gradient − t875	+0.31	[+0.16, +0.42]	11/12
decoder gradient − t250	+0.04	[-0.09, +0.16]	8/12
decoder gradient − t125	+0.16	[+0.02, +0.27]	8/12
decoder gradient − unguided	+3.46	[+3.34, +3.57]	12/12
MaJa ($\bar{J}$) − unguided	+3.33	[+3.21, +3.45]	12/12
t875 − unguided	+3.16	[+3.05, +3.27]	12/12
t250 − unguided	+3.42	[+3.29, +3.55]	12/12
t125 − unguided	+3.31	[+3.19, +3.43]	12/12

Table 57: The wrong-attribute control in the second family: the ACE brightness operator applied to an ACE loudness target.

direction source	error (dB)	95% interval	prompts
unguided	6.00	[6.00, 6.00]	12
signal-processing reference	0.00	[0.00, 0.00]	12
decoder gradient	2.54	[2.43, 2.66]	12
MaJa ($\bar{J}$)	2.67	[2.55, 2.79]	12
brightness operator	6.42	[6.26, 6.60]	12
paired differences			
decoder gradient − MaJa ($\bar{J}$)	+0.14	[+0.00, +0.25]	8/12
decoder gradient − brightness operator	+3.88	[+3.63, +4.10]	12/12
decoder gradient − unguided	+3.46	[+3.34, +3.57]	12/12
MaJa ($\bar{J}$) − unguided	+3.33	[+3.21, +3.45]	12/12
brightness operator − unguided	-0.42	[-0.60, -0.26]	0/12

Table 58: The wrong-attribute control in the second family: the ACE loudness operator applied to an ACE brightness target.

direction source	error (dB)	95% interval	prompts
unguided	5.17	[4.99, 5.39]	12
signal-processing reference	0.29	[0.13, 0.49]	12
decoder gradient	1.46	[1.30, 1.66]	12
MaJa ($\bar{J}$)	1.80	[1.56, 2.06]	12

continued

direction source	error (dB)	95% interval	prompts
loudness operator	5.20	[4.89, 5.61]	12
paired differences			
decoder gradient − MaJa ($\bar{J}$)	+0.34	[+0.12, +0.57]	10/12
decoder gradient − loudness operator	+3.74	[+3.44, +4.15]	12/12
decoder gradient − unguided	+3.71	[+3.40, +3.97]	12/12
MaJa ($\bar{J}$) − unguided	+3.37	[+3.20, +3.53]	12/12
loudness operator − unguided	-0.03	[-0.28, +0.17]	7/12

E Robustness Checks

The checks in this section were run after the sixteen main panels were frozen. The two that change the protocol (Appendices E.3 and E.4) are reported as robustness checks and are not folded into Table 1.

E.1 End-to-End Gradient Through the Remaining Steps

Motivation. The main text argues that the decoder gradient is not the right reference for a head to imitate. The reference a head should arguably imitate if the goal is the final output is the gradient of the final clip's loss with respect to the current clean estimate, through the remaining sampler steps (Figure 5a). How does that direction compare?

Setup. At each intervention the remaining steps of the four-step DMD sampler are run differentiably from the current clean estimate, with the same future noise the native sampler will draw (the generator state is saved before and restored after), the final latent is decoded, and the attribute loss is back-propagated to the clean estimate. This direction source, *end-to-end gradient*, then follows the same rule and step size as every other. It costs three additional backbone passes with gradient per intervention.

Result. The end-to-end gradient is the worst direction source on both new-prompt panels (Tables 59 and 61). On the gain target it reaches 2.37 dB against 2.03 for the decoder gradient and 1.64 for the mean Jacobian, and the decoder gradient is ahead of it by 0.34 dB $[0.16, 0.50]$ on 13 of 16 prompts; on the fade target it reaches 2.96 against 2.64 and 2.17, with the decoder gradient ahead by 0.32 $[0.16, 0.54]$ on 14 of 16. On unseen instruments the three are level (1.07, 0.93, 0.97; Table 60). Its cosine to the decoder gradient is 0.50 on new prompts and 0.73 on unseen instruments. Its own development-optimal step size is the protocol's $3\,\%$, so this is not a calibration artefact. The gradient through the remaining steps is the correct first-order object for the final loss, and under a step of this size it is a worse direction than either the decoder gradient or the averaged matrix. We read this as the strongest version of the paper's point: neither the local reference nor the end-to-end reference is what a head should imitate on these samplers.

Table 59: The end-to-end gradient as a direction source (AudioX loudness, gain, new prompts): the gradient of the final output's loss through the remaining sampler steps, with the same future noise as the native run, and the decoder.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.03	[1.83, 2.27]	16
MaJa ($\bar{J}$)	1.64	[1.41, 1.90]	16
end-to-end gradient (through the remaining steps)	2.37	[2.06, 2.73]	16
VJP-matched head (student)	2.09	[1.89, 2.32]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.39	[-0.50, -0.27]	3/16
decoder gradient − end-to-end gradient (through the remaining steps)	+0.34	[+0.16, +0.50]	13/16
decoder gradient − VJP-matched head (student)	+0.05	[+0.01, +0.10]	9/16
decoder gradient − unguided	+3.55	[+3.16, +3.88]	16/16
MaJa ($\bar{J}$) − unguided	+3.94	[+3.59, +4.22]	16/16
end-to-end gradient (through the remaining steps) − unguided	+3.22	[+2.72, +3.59]	16/16
VJP-matched head (student) − unguided	+3.50	[+3.11, +3.80]	16/16

Table 60: The end-to-end gradient on unseen instruments (AudioX loudness, gain).

direction source	error (dB)	95% interval	prompts
unguided	6.00	[6.00, 6.00]	8
signal-processing reference	0.00	[0.00, 0.00]	8
decoder gradient	0.93	[0.78, 1.09]	8
MaJa ($\bar{J}$)	0.97	[0.78, 1.18]	8
end-to-end gradient (through the remaining steps)	1.07	[0.75, 1.44]	8
VJP-matched head (student)	0.96	[0.81, 1.12]	8
paired differences			
decoder gradient − MaJa ($\bar{J}$)	+0.05	[-0.04, +0.14]	5/8
decoder gradient − end-to-end gradient (through the remaining steps)	+0.14	[-0.06, +0.37]	4/8
decoder gradient − VJP-matched head (student)	+0.03	[+0.02, +0.05]	8/8
decoder gradient − unguided	+5.07	[+4.91, +5.22]	8/8
MaJa ($\bar{J}$) − unguided	+5.03	[+4.82, +5.22]	8/8
end-to-end gradient (through the remaining steps) − unguided	+4.93	[+4.56, +5.25]	8/8
VJP-matched head (student) − unguided	+5.04	[+4.88, +5.19]	8/8

Table 61: The end-to-end gradient on the fade target (AudioX loudness, new prompts).

direction source	error (dB)	95% interval	prompts
unguided	3.26	[3.11, 3.39]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	2.64	[2.47, 2.82]	16
MaJa ($\bar{J}$)	2.17	[1.94, 2.44]	16
end-to-end gradient (through the remaining steps)	2.96	[2.70, 3.22]	16
VJP-matched head (student)	2.71	[2.48, 2.94]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.47	[-0.56, -0.38]	0/16
decoder gradient − end-to-end gradient (through the remaining steps)	+0.32	[+0.16, +0.54]	14/16
decoder gradient − VJP-matched head (student)	+0.07	[-0.02, +0.14]	13/16
decoder gradient − unguided	+0.62	[+0.36, +0.83]	11/16
MaJa ($\bar{J}$) − unguided	+1.08	[+0.76, +1.36]	14/16
end-to-end gradient (through the remaining steps) − unguided	+0.29	[-0.03, +0.60]	9/16
VJP-matched head (student) − unguided	+0.55	[+0.25, +0.80]	10/16

E.2 Equivalence Test at 48 Prompts

Motivation. Ten of the sixteen main comparisons are reported as level because the $95\,\%$ interval contains zero, but an interval that contains zero is not evidence of equivalence, least of all on eight-prompt panels. What does a larger panel with an explicit equivalence test say?

Setup. A new-prompt panel three times larger: the sixteen original prompts plus thirty-two written for this purpose, four new performance descriptions on each of the eight instrument classes, with the same two unseen seeds, so 96 cases in eight instrument clusters. Before running it we fixed an equivalence bound of ± 0.2 dB and the rule that "level with the decoder gradient" would be claimed only where the $90\,\%$ cluster bootstrap interval of the paired difference lies inside that bound (two one-sided tests at $\alpha = 0.05$). The step sizes are those of the main panels; nothing was re-tuned.

Result. More prompts did not turn the ties into equivalences; they turned one of them into a win and sharpened the others (Tables 62 and 63). On loudness the mean Jacobian is ahead of the decoder gradient by 0.28 dB, $95\,\%$ interval $[0.20, 0.36]$, on 35 of 48 prompts, and ahead of the VJP-matched and regression heads by 0.37 $[0.27, 0.45]$ and 0.39 $[0.18, 0.58]$; the panel is harder on average than the sixteen-prompt one (every error rises by about 0.2 dB) and the lead narrows from 0.39 to 0.28 without its interval approaching zero. On brightness the sixteen-prompt tie ($+0.51$, $[-0.04, +1.32]$) becomes $+0.42$, $[0.16, 0.71]$, on 24 of 36 prompts with a valid score, and the leads over the VJP-matched and regression heads are 0.73 $[0.37, 1.24]$ and 1.21 $[1.04, 1.37]$. One comparison does meet the pre-registered equivalence bound: the VJP-matched head against the decoder gradient on loudness, -0.09 dB with a $90\,\%$ interval of $[-0.12, -0.06]$, which is the claim that this head reproduces the reference, made quantitative. One of the 96 loudness cases was excluded before scoring because the decoder gradient was exactly zero on it (a clip at the loudness floor at t=0.75), leaving 48 prompts; the exclusion applies to every direction source. The main text's "level" statements for the

eight-prompt unseen-instrument panels remain descriptive, as Appendix G says; we did not enlarge those panels.

Table 62: The 48-prompt new-prompt panel (AudioX loudness, gain, calibrated multi-step probe, 96 cases, eight instrument clusters). Pre-registered equivalence bound ± 0.2 dB on the 90 % cluster bootstrap interval.

direction source	error (dB)	95% interval	prompts
unguided	5.61	[5.41, 5.81]	48
signal-processing reference	0.00	[0.00, 0.00]	48
decoder gradient	2.18	[1.96, 2.47]	48
MaJa ($\bar{J}$)	1.91	[1.70, 2.16]	48
regression head (teacher and student states)	2.30	[2.18, 2.40]	48
VJP-matched head (student)	2.27	[2.04, 2.56]	48
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.28	[-0.36, -0.20]	13/48
decoder gradient − regression head (teacher and student states)	+0.11	[-0.11, +0.29]	31/48
decoder gradient − VJP-matched head (student)	+0.09	[+0.05, +0.12]	36/48
decoder gradient − unguided	+3.43	[+2.96, +3.79]	47/48
MaJa ($\bar{J}$) − unguided	+3.71	[+3.29, +4.03]	48/48
regression head (teacher and student states) − unguided	+3.32	[+3.05, +3.55]	48/48
VJP-matched head (student) − unguided	+3.34	[+2.87, +3.69]	48/48

Table 63: The 48-prompt new-prompt panel for brightness (AudioX, shelf, calibrated multi-step probe).

direction source	error (dB)	95% interval	prompts
unguided	4.89	[4.59, 5.15]	36
signal-processing reference	0.59	[0.13, 1.10]	36
decoder gradient	2.01	[1.48, 2.54]	36
MaJa ($\bar{J}$)	1.59	[1.24, 1.91]	36
regression head (teacher and student states)	2.80	[2.48, 3.12]	36
VJP-matched head (student)	2.32	[1.70, 3.06]	36
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.42	[-0.71, -0.16]	12/36
decoder gradient − regression head (teacher and student states)	+0.80	[+0.52, +1.05]	32/36
decoder gradient − VJP-matched head (student)	+0.31	[+0.16, +0.56]	30/36
decoder gradient − unguided	+2.88	[+2.35, +3.48]	35/36
MaJa ($\bar{J}$) − unguided	+3.30	[+2.84, +3.78]	36/36
regression head (teacher and student states) − unguided	+2.09	[+1.67, +2.50]	36/36
VJP-matched head (student) − unguided	+2.57	[+1.89, +3.25]	33/36

E.3 Per-Method Step Sizes

Motivation. The step size is calibrated on the decoder gradient alone, which Section 6.4 shows disadvantages the mean Jacobian; the same logic says it may disadvantage the heads. Does letting every direction source choose its own step size change the ordering?

Setup. Each direction source chooses its own step size from $\{1, 2, 3, 5\}$ % by its own error on the eight development prompts, frozen, and the new-prompt panels are rerun with those step sizes. For loudness the development runs at all four step sizes already existed for every direction source; for brightness they were run for this purpose.

Result. On loudness every direction source chose 3 %, the protocol's value, so the per-method panel coincides with the main one and is not repeated. On brightness the decoder gradient and the VJP-matched head chose 2 % and the mean Jacobian and the regression head chose 3 %; Table 64 gives the panel at those step sizes. At its own step size the mean Jacobian reaches 1.49 dB against 2.38 for the decoder gradient, 2.57 for the VJP-matched head and 2.76 for the regression head, so its lead over the decoder gradient grows from 0.51 under the shared step size to 0.89. Calibrating on the decoder gradient alone, as the main protocol does, is the conservative choice for the paper's claim, and the heads gain nothing from choosing their own step.

Table 64: Per-method step sizes on brightness (AudioX, shelf, new prompts).

direction source	error (dB)	95% interval	prompts
unguided	4.82	[4.53, 5.13]	9
signal-processing reference	1.17	[0.25, 1.96]	9
decoder gradient	2.38	[1.62, 3.43]	9
MaJa ($\bar{J}$)	1.49	[1.20, 1.82]	9
regression head (teacher and student states)	2.75	[2.38, 3.18]	9
VJP-matched head (student)	2.57	[1.79, 3.66]	9
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-0.89	[-1.74, -0.17]	2/9
decoder gradient − regression head (teacher and student states)	+0.37	[-0.37, +0.98]	4/9
decoder gradient − VJP-matched head (student)	+0.18	[+0.07, +0.30]	6/9
decoder gradient − unguided	+2.44	[+1.59, +3.23]	8/9
MaJa ($\bar{J}$) − unguided	+3.33	[+2.98, +3.72]	9/9
regression head (teacher and student states) − unguided	+2.07	[+1.63, +2.56]	9/9
VJP-matched head (student) − unguided	+2.26	[+1.35, +3.05]	8/9

E.4 DPS-STYLE UPDATE

Motivation. Every result uses one intervention, a fixed relative step along the normalised direction. Does the conclusion depend on that form?

Setup. The update of diffusion posterior sampling, a step proportional to the gradient and inversely proportional to the residual norm, so that its size varies with the state and the step. Direction sources have gradients of very different magnitude under this rule, so each is given a scale set on the development prompts such that its mean relative perturbation equals the $3\,\%$ of the main protocol; the state-to-state variation of the step is then the DPS law's and the comparison remains matched on average.

Result. Under the DPS-style update the mean Jacobian reaches 2.04 dB, the regression head 3.11, the VJP-matched head 4.47 and the decoder gradient 4.97, against 5.59 unguided (Table 66). The ordering is the paper's, with the gaps much larger, and the reason is visible in the step sizes the rule produces. The scales were set so that every source took a $3\,\%$ step on average on the development prompts; on the new prompts the mean Jacobian's step stayed at $3.3\,\%$ (standard deviation $0.2\,\%$) while the decoder gradient's rose to $5.1\,\%$ and the VJP-matched head's to $4.9\,\%$ (standard deviations $1.9\,\%$), because the magnitude of a gradient is state dependent and the DPS law passes that dependence into the step, whereas $\bar{J}r$ scales with the residual and the residual normalisation cancels it. Choosing each source's scale by its own development error instead, over half, one and two times the matched scale, selected the matched scale for all four (Table 65), so this is not a calibration artefact either. The conclusion does not depend on the form of the intervention; the DPS form is simply harsher on gradient-based sources.

Table 65: The DPS-style update with each direction source's scale chosen by its own development error over half, one and two times the mean-step-matched scale (AudioX loudness, gain, new prompts).

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	4.97	[4.51, 5.48]	16
MaJa ($\bar{J}$)	2.04	[1.78, 2.33]	16
regression head (teacher and student states)	3.11	[2.85, 3.39]	16
VJP-matched head (student)	4.47	[4.02, 4.97]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-2.93	[-3.16, -2.72]	0/16
decoder gradient − regression head (teacher and student states)	-1.86	[-2.12, -1.62]	5/16
decoder gradient − VJP-matched head (student)	-0.50	[-0.59, -0.42]	3/16
decoder gradient − unguided	+0.62	[+0.04, +1.10]	8/16
MaJa ($\bar{J}$) − unguided	+3.55	[+3.18, +3.86]	16/16
regression head (teacher and student states) − unguided	+2.47	[+2.13, +2.77]	16/16
VJP-matched head (student) − unguided	+1.12	[+0.57, +1.58]	8/16

Table 66: A DPS-style update instead of the fixed relative step (AudioX loudness, gain, new prompts): the step is the gradient divided by the residual norm, so its size varies with the state; each direction source has a scale set on the development prompts so that its mean relative perturbation matches the 3 % of the main protocol.

direction source	error (dB)	95% interval	prompts
unguided	5.59	[5.37, 5.80]	16
signal-processing reference	0.00	[0.00, 0.00]	16
decoder gradient	4.97	[4.51, 5.48]	16
MaJa ($\bar{J}$)	2.04	[1.78, 2.33]	16
regression head (teacher and student states)	3.11	[2.85, 3.39]	16
VJP-matched head (student)	4.47	[4.02, 4.97]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-2.93	[-3.16, -2.72]	0/16
decoder gradient − regression head (teacher and student states)	-1.86	[-2.12, -1.62]	5/16
decoder gradient − VJP-matched head (student)	-0.50	[-0.59, -0.42]	3/16
decoder gradient − unguided	+0.62	[+0.04, +1.10]	8/16
MaJa ($\bar{J}$) − unguided	+3.55	[+3.18, +3.86]	16/16
regression head (teacher and student states) − unguided	+2.47	[+2.13, +2.77]	16/16
VJP-matched head (student) − unguided	+1.12	[+0.57, +1.58]	8/16

E.5 A Learned, Non-Linear Attribute

Motivation. Every target in the main panels is attainable by signal processing, and all five attributes are frame-level energy or spectral statistics that are close to linear functions of the waveform. Does the comparison change on an attribute defined by a learned model, and does such an attribute supply the end of the screen on which a head should win?

Setup. A sixth attribute defined by a learned model: the CLAP text-audio contrast $100\,(\cos(a(x), t_{\text{bright}}) - \cos(a(x), t_{\text{dark}}))$ for the texts "a bright, sharp, treble-heavy sound" and "a dark, muffled, bass-heavy sound", computed on five equal windows of the clip so that it lives on the same 212-frame grid as the other attributes (piecewise constant over windows). The CLAP model is the local checkpoint already used for the distribution guard; the score is differentiable end to end through the resampling, the HTSAT audio branch and the cosine, so the decoder gradient, the labels, the screen and the mean Jacobian are obtained exactly as for the other attributes. The target is the clip's own score shifted by ± 5 points; there is no signal-processing reference, so the reference row is the unguided clip. Direction sources: decoder gradient, mean Jacobian from eight states, two regression heads, and a random matrix with the mean Jacobian's row norms. The step size was calibrated on the decoder gradient on the development prompts as always, and the calibration itself is the first result: the decoder gradient improves on not guiding only at the smallest step, 2.52 points at $1\,\%$ against 5.00 unguided, and makes things worse at 2, 3 and $5\,\%$ (5.15, 6.10, 8.27). The score is a valid linearisation of this attribute over a much shorter range than for the signal-processing attributes, whose calibrated steps are 2 to $3\,\%$. The panels below therefore use $1\,\%$; the $2\,\%$ panels are kept as a step-sensitivity check. Because the CLAP backward pass is not bit-reproducible on the GPU, the repeat-run checks record the discrepancy for this attribute instead of asserting equality; the repeat gradients agree to cosine 1.000.

Result. This is the regime the screen was built to find, and it finds it. On the same 48 states used for the other attributes the CLAP contrast has a mean pairwise cosine of 0.005 between decoder vector-Jacobian products and a cosine of 0.022 between a held-out state's product and the mean over the others (Table 42), an order of magnitude below onset strength, the lowest signal-processing attribute (0.20/0.43). Control follows the screen exactly (Tables 67 and 68). At the calibrated $1\,\%$ step the decoder gradient moves the score most of the way to the target, from 5.00 to 2.76 points on new prompts and to 1.98 on unseen instruments, while the mean Jacobian (4.91 and 4.87), the fixed direction (4.83 and 4.87), the two regression heads (4.86 to 4.93), the equal-cost regression head (4.72 and 4.86), the VJP-matched head (4.63 and 4.81) and the random matrix (5.04 and 4.87) all leave it essentially where it was. The equal-cost and VJP-matched heads are ahead of the mean Jacobian on new prompts by 0.19 and 0.28 points with intervals that exclude zero (Tables 12 and 13). The mean Jacobian's and the heads' directions have cosine 0.00 to 0.02 to the decoder gradient, the same as the random matrix's. At $2\,\%$ (Tables 69 and 70) the decoder gradient itself overshoots (6.13 on new prompts) and nothing else changes. Two conclusions. On an attribute whose decoder

Jacobian is this state-dependent, neither a precomputed matrix nor a regression head is of any use, and the paper's recommendation to try the matrix first would be wrong here; the screen says so before any panel is run, at three backward passes per state. And the reference that does work is the decoder gradient at a small step, not a head: the heads trained on this attribute are as uncorrelated with the decoder gradient as the matrix is. The screen therefore separates two regimes rather than measuring one quantity: above roughly 0.4 on the second column the matrix matches or beats the decoder gradient on every attribute we measured, and at 0.02 it does nothing.

Table 67: A learned, non-linear attribute (AudioX, new prompts, calibrated step size 1 %): the CLAP bright-minus-dark text contrast on five windows of the clip, in points of cosine. There is no signal-processing reference, so the reference row is the unguided clip and the target is the clip's own score shifted by ± 5 points.

direction source	error (dB)	95% interval	prompts
unguided	5.00	[5.00, 5.00]	16
signal-processing reference	5.00	[5.00, 5.00]	16
decoder gradient	2.76	[2.51, 3.01]	16
MaJa ($\bar{J}$)	4.91	[4.82, 5.00]	16
random matrix, matched row norms	5.04	[4.99, 5.09]	16
regression head (student)	4.86	[4.79, 4.94]	16
regression head (teacher and student states)	4.93	[4.89, 4.97]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	+2.15	[+1.89, +2.40]	16/16
decoder gradient − random matrix, matched row norms	+2.29	[+2.01, +2.56]	16/16
decoder gradient − regression head (student)	+2.10	[+1.81, +2.39]	16/16
decoder gradient − regression head (teacher and student states)	+2.17	[+1.89, +2.44]	16/16
decoder gradient − unguided	+2.24	[+1.99, +2.49]	16/16
MaJa ($\bar{J}$) − unguided	+0.09	[+0.00, +0.18]	14/16
random matrix, matched row norms − unguided	-0.04	[-0.09, +0.01]	6/16
regression head (student) − unguided	+0.14	[+0.06, +0.21]	14/16
regression head (teacher and student states) − unguided	+0.07	[+0.03, +0.11]	13/16

Table 68: The CLAP contrast attribute on unseen instruments (AudioX, 1 %).

direction source	error (dB)	95% interval	prompts
unguided	5.00	[5.00, 5.00]	8
signal-processing reference	5.00	[5.00, 5.00]	8
decoder gradient	1.98	[1.44, 2.60]	8
MaJa ($\bar{J}$)	4.87	[4.73, 5.02]	8
random matrix, matched row norms	4.87	[4.77, 4.98]	8
regression head (student)	4.84	[4.68, 4.99]	8
regression head (teacher and student states)	4.89	[4.81, 4.97]	8
paired differences			
decoder gradient − MaJa ($\bar{J}$)	+2.89	[+2.22, +3.50]	8/8
decoder gradient − random matrix, matched row norms	+2.89	[+2.25, +3.44]	8/8
decoder gradient − regression head (student)	+2.86	[+2.21, +3.43]	8/8
decoder gradient − regression head (teacher and student states)	+2.91	[+2.28, +3.45]	8/8
decoder gradient − unguided	+3.02	[+2.40, +3.56]	8/8
MaJa ($\bar{J}$) − unguided	+0.13	[-0.02, +0.27]	6/8
random matrix, matched row norms − unguided	+0.13	[+0.02, +0.23]	6/8
regression head (student) − unguided	+0.16	[+0.01, +0.32]	6/8
regression head (teacher and student states) − unguided	+0.11	[+0.03, +0.19]	6/8

Table 69: The CLAP contrast attribute at 2 %, twice the calibrated step (AudioX, new prompts): a step-sensitivity check.

direction source	error (dB)	95% interval	prompts
unguided	5.00	[5.00, 5.00]	16
signal-processing reference	5.00	[5.00, 5.00]	16
decoder gradient	6.13	[5.34, 6.83]	16
MaJa ($\bar{J}$)	4.78	[4.63, 4.92]	16
random matrix, matched row norms	5.12	[5.00, 5.21]	16
regression head (student)	4.75	[4.60, 4.88]	16
regression head (teacher and student states)	4.85	[4.77, 4.93]	16
paired differences			
decoder gradient − MaJa ($\bar{J}$)	-1.35	[-2.08, -0.57]	3/16

continued

direction source	error (dB)	95% interval	prompts
decoder gradient − random matrix, matched row norms	-1.01	[-1.77, -0.19]	4/16
decoder gradient − regression head (student)	-1.38	[-2.04, -0.69]	4/16
decoder gradient − regression head (teacher and student states)	-1.28	[-1.99, -0.51]	4/16
decoder gradient − unguided	-1.13	[-1.83, -0.34]	4/16
MaJa ($\bar{J}$) − unguided	+0.22	[+0.08, +0.37]	11/16
random matrix, matched row norms − unguided	-0.12	[-0.21, +0.00]	4/16
regression head (student) − unguided	+0.25	[+0.12, +0.40]	15/16
regression head (teacher and student states) − unguided	+0.15	[+0.07, +0.23]	14/16

Table 70: The CLAP contrast attribute at 2 % on unseen instruments (AudioX).

direction source	error (dB)	95% interval	prompts
unguided	5.00	[5.00, 5.00]	8
signal-processing reference	5.00	[5.00, 5.00]	8
decoder gradient	4.26	[2.94, 5.90]	8
MaJa ($\bar{J}$)	4.79	[4.52, 5.08]	8
random matrix, matched row norms	4.81	[4.62, 4.99]	8
regression head (student)	4.78	[4.51, 5.02]	8
regression head (teacher and student states)	4.82	[4.68, 4.97]	8
paired differences			
decoder gradient − MaJa ($\bar{J}$)	+0.53	[-1.18, +1.97]	5/8
decoder gradient − random matrix, matched row norms	+0.55	[-1.11, +1.87]	6/8
decoder gradient − regression head (student)	+0.51	[-1.14, +1.85]	6/8
decoder gradient − regression head (teacher and student states)	+0.56	[-1.12, +1.91]	6/8
decoder gradient − unguided	+0.74	[-0.90, +2.06]	6/8
MaJa ($\bar{J}$) − unguided	+0.21	[-0.08, +0.48]	6/8
random matrix, matched row norms − unguided	+0.19	[+0.01, +0.38]	6/8
regression head (student) − unguided	+0.22	[-0.02, +0.49]	6/8
regression head (teacher and student states) − unguided	+0.18	[+0.03, +0.32]	6/8

F Negative Results

This section collects interventions that did not improve control and checks that constrain the explanation of the main result.

Matching the decoder gradient. Supervising a head on the decoder's vector-Jacobian products raises its guidance-time alignment to the decoder gradient from 0.38 to 0.84 up to 0.89 to 0.99 (Table 15), out of sample on prompts and seeds no head saw, and makes it indifferent to which generator's states it was trained on. Control does not follow: the VJP-matched head matches the decoder gradient and is beaten by a matrix with no parameters at a cosine of 0.60 to 0.91 to the same gradient. Reproducing the reference is not what makes a direction source control well.

Preconditioning the decoder gradient. Clipping, low-passing, whitening and taking the sign of the decoder gradient, and averaging it over small perturbations of the state, all leave control unchanged or worse; the numbers are in Appendix D.5.

The affine value model implied by $\bar{J}$. If the decoder's Jacobian were state-independent the attribute would be an affine function of the latent, $\phi(\mathbf{x}) \approx b + \bar{J}\mathbf{x}$. That predictor reaches 4.51 dB on loudness, 5.40 on brightness and 3.75 on the centroid, against 0.54 to 1.16 dB for the learned heads. The direction the Jacobian defines transfers between states; its scale does not. This is the reason the decoder-free rule needs a head at all, and it is also the sharpest available argument that $\bar{J}$ is not secretly a linear model of the attribute.

Repairs to teacher-trained heads. Three attempts to rescue a teacher-trained head on student states failed. Affine recalibration of its output on student states recovers about 0.2 dB of a 2 dB gap. Replacing the head's time input with a remapped t', swept over seven values, changes error by less than 0.1 dB (Figure 4, right). Training on terminal latents only, so that the head never sees an intermediate state, gives 2.07 dB at the guidance step against 0.55 to 0.91 for heads trained

on intermediate states (Table 17). The displacement of the states is not a calibration problem, a time-indexing problem or a labelling problem.

Larger heads. Heads of 27.6M and 41.3M parameters are worse than the 6.9M head on every prompt; see Appendix D.2.

Cheaper decoding. Reduced-precision decoding reproduces the decoder gradient at a small saving and decoding part of the clip destroys the direction; see Appendix D.4.

What the screen does not predict. Within one generator the screen tracks a trend, the mean Jacobian's advantage decaying along it, but across generators its two columns do not order the settings: ACE loudness screens at 0.72/0.81 and is one of the two panels where the mean Jacobian loses, while AudioX brightness screens lower on the pairwise column and the mean Jacobian is ahead there. The screen is therefore reported as a measure of how little a head can add on the signal-processing attributes, and Appendix E.5 shows the regime it does identify.

G Scope and Limitations

Relation to the Published Head Design. Our regression heads are the component of LatCH (Novack et al., 2026) that our claim is about: a small network over the latent clean estimate, trained by regression on attributes measured from decoded latents, whose input gradient stands in for the decoder gradient. We did not reimplement that system, and our numbers are not a reproduction of its reported results. The differences a reader should weigh are these. The generators are ours (AudioX and AudioX-Turbo, ACE-Step 1.5 SFT and Turbo), not theirs. The attribute set is ours, and the targets are constructed so that signal processing attains them exactly, which makes the floor of every panel zero. The intervention is a single rule applied identically to every direction source, one RMS-normalised descent step on the detached clean estimate at each non-terminal sampler step, with one strength per attribute chosen on development prompts using the decoder gradient; we do not use a guidance schedule, a selective or adaptive rule, or any per-method tuning. The head architecture follows the published design in spirit, a bidirectional transformer over latent frames with a Fourier time token, but is not matched parameter for parameter.

The consequence is that absolute errors in this paper are not comparable to published numbers, and we never compare them. Every claim is relative and within-panel: the same prompts, the same generation seeds, the same unguided baseline, the same target, the same rule and the same step size, with only the direction source changing. That is what makes a head-against-matrix comparison valid even though the setup is ours. Where we could give the head an advantage we did: three initialisations, three choices of training-state source, capacities of 6.9M, 27.6M and 41.3M, four times the trajectories at a wall-clock-matched update budget, and supervision on the decoder's vector-Jacobian products. None of these makes a head significantly better than the precomputed matrix on any panel.

We did not use the official implementation of the guidance heads, nor a third few-step generator; the differences between our reimplementation and the published design are those listed above: the same generators are not used; the attribute set and targets are ours and are attainable by signal processing; the intervention is a single rule with a step size chosen on the decoder gradient rather than a guidance schedule; and the head follows the published architecture in spirit but is not matched parameter for parameter. Every comparison is within-panel, so these differences change the absolute numbers, not the head-against-matrix comparison, and we gave the head every advantage we could (initialisations, training sources, capacity, four times the data, and gradient supervision).

Generality. We measure two teacher/student pairs, one modality, five frame-level scalar attributes, four target families on loudness and one each on the other attributes, and clip lengths of 10 s (AudioX) and 20 s (ACE). Within that range the finding is consistent. Outside it we claim nothing. In particular we do not claim it for image or video latents, for attributes that are not frame-wise functions of the decoded signal, for generators whose decoder is not shared between teacher and student, or for samplers with enough steps to re-project a perturbation, where Table 31 shows every direction source performs the same. The screen of Section 4.3 is how we suggest testing the boundary rather than extrapolating across it: it costs three backward passes per state and it is defined for any attribute and any latent generator.

Perceptual Evaluation. Every number in this paper is an attribute error against a target that signal processing attains exactly, plus a distribution guard. That is deliberate: the question is whether a guidance direction reaches a specified frame-level target, which is a measurement question and not a perceptual one, and a listening test would answer a different question with more noise. It also means we cannot claim that any of these clips sounds better, and we do not. The guard in Table 49 rules out the one confound that would invalidate the measurement, namely a direction source winning by damaging the audio more than the others: the signal-processing reference, which is by construction the intended change and nothing else, moves the guard numbers by as much as the guidance methods do. No listening test was run; a non-inferiority listening test (64 blinded decoder-gradient against mean-Jacobian pairs next to the unguided reference) is prepared and its stimulus list is released with the code.

Distribution Guard on Brightness. On the brightness panel the Fréchet distance between CLAP embeddings of the guided and unguided sets is 0.345 for every direction source, against 0.020 for the signal-processing reference that attains the target exactly. We cannot explain the gap and we do not want to explain it away, so we record what we know. It is identical to three decimals across the decoder gradient, both heads and the mean Jacobian, so it is a property of perturbing the latent rather than of which direction is used, and it therefore cancels in every comparison this paper makes. The same panel's waveform correlation to the unguided clip is 0.995, so the change is not large in the time domain. The statistic is computed over sixteen clips, where a Fréchet distance between Gaussian fits is noisy and biased upward. And the unseen-instrument brightness panel, which uses the same targets and the same rule, sits at 0.12. Taken together we read this as an artefact of the intervention that the shelf equaliser does not produce, of unknown audibility, and we list it as a limitation rather than a result.

Statistical Language. We avoid the word significant in its inferential sense. Where we say an interval excludes zero we mean the $95\,\%$ cluster bootstrap interval over instrument clusters, computed with $10{,}000$ draws and a seed fixed before the panels were run, and it is descriptive for the panels as run rather than a test against a population. The new-prompt panel has sixteen prompts in six clusters and the unseen-instrument panel has eight prompts in two, so resolution differs sharply between them. We apply no correction for multiple comparisons; the number of comparisons we report is large, and a reader who wants a family-wise guarantee should read the per-prompt tables of Appendix H.3 rather than the intervals.

H Complete Records

H.1 All Fitted Heads

Table 71 lists every fitted head with its training budget and its prediction error at the guidance step, split by whether the state comes from an instrument class the head trained on or from one of the two classes held out. The table exists to make one point checkable: these heads are good at the job they were trained for. A loudness head predicts the envelope of the states it sees to about half a decibel, and its error on unseen instrument classes is close to its error on familiar ones, so the failure documented in Section 6.2 is not a failure of fitting or of generalisation in the usual sense. It is a failure of the derivative: the same heads point 40 to 60 degrees away from the decoder gradient at exactly the states they predict best (Table 14), and that is the quantity guidance consumes.

The inventory also shows how much compute the head route costs. An AudioX loudness regression fit is seven to eight minutes and a VJP-matched fit is 107 minutes, the difference being the double backward; the equal-cost regression fits are matched to the VJP fits in wall time at 195 minutes because extra data is cheaper per update than gradient labels. The ACE fits are slower throughout because the clips are longer. The mean Jacobian is a quarter of an hour of decoder backward passes and no training at all.

Table 71: Every fitted head (80 fits). Source: which generator's stored states the head was trained on. Budget: parameters, AdamW updates and training wall time in minutes on one A6000. Prediction error is the head's own attribute prediction at $t{=}0.75$ states, in dB for loudness, brightness and high-band loudness and in $20\log_{10}$Hz for the centroid, split by whether the state comes from a training instrument class or from one of the two classes held out of head training. Names beginning `jac` are VJP-matched fits and names beginning `xreg` are the extra-data regression fits of Table 46. A dash in the parameter column means the fit record does not store the count; those fits use the same architecture as the others. Prediction accuracy is not direction accuracy: Table 14 shows the same heads pointing 40 to 60 degrees away from the decoder gradient on the states they predict best.

fit	attribute	source	seed	params	updates	min	own	held out
ace/fits/seed29101/jac1_mixed_direct	ACE loudness	mixed	29101	n/r	10000	343	0.56	0.52
ace/fits/seed29101/jac1_student_direct	ACE loudness	student	29101	n/r	10000	343	0.66	0.63
ace/fits/seed29101/jac1_teacher_direct	ACE loudness	teacher	29101	n/r	10000	342	0.73	0.65
ace/fits/seed29101/mixed_direct	ACE loudness	mixed	29101	n/r	10000	32	0.47	1.50
ace/fits/seed29101/student_direct	ACE loudness	student	29101	n/r	10000	32	2.01	2.40
ace/fits/seed29101/teacher_direct	ACE loudness	teacher	29101	n/r	10000	33	1.70	3.04
ace/fits/seed29101/xreg_student_extra_u107000_direct	ACE loudness	student_extra	29101	n/r	107000	382	0.70	1.65
ace/fits/seed29103/mixed_direct	ACE loudness	mixed	29103	n/r	10000	21	0.48	1.53
ace/fits/seed29103/student_direct	ACE loudness	student	29103	n/r	10000	26	2.15	2.65
ace/fits/seed29103/teacher_direct	ACE loudness	teacher	29103	n/r	10000	27	1.72	3.06
ace/fits/seed29107/mixed_direct	ACE loudness	mixed	29107	n/r	10000	21	0.45	1.57
ace/fits/seed29107/student_direct	ACE loudness	student	29107	n/r	10000	27	2.06	2.59
ace/fits/seed29107/teacher_direct	ACE loudness	teacher	29107	n/r	10000	26	1.83	3.23
centroid-head/fits/seed29101/jac1_mixed_direct	centroid	mixed	29101	n/r	10000	71	0.62	1.18
centroid-head/fits/seed29101/jac1_student_direct	centroid	student	29101	n/r	10000	84	0.80	1.24
centroid-head/fits/seed29101/jac1_teacher_direct	centroid	teacher	29101	n/r	10000	45	1.17	1.49
centroid-head/fits/seed29101/mixed_direct	centroid	mixed	29101	6921216	10000	12	0.72	2.04
centroid-head/fits/seed29101/student_direct	centroid	student	29101	6921216	10000	13	1.55	2.86
centroid-head/fits/seed29101/teacher_direct	centroid	teacher	29101	6921216	10000	14	1.73	3.13
centroid-head/fits/seed29101/xreg_student_extra_u100000_direct	centroid	student_extra	29101	6921216	100000	136	0.63	1.51
centroid-head/fits/seed29101/xreg_student_extra_u10000_direct	centroid	student_extra	29101	6921216	10000	11	1.07	1.99
clap-head/fits/seed29101/mixed_direct	clap-head	mixed	29101	6921216	10000	9	1.14	8.53
clap-head/fits/seed29101/student_direct	clap-head	student	29101	6921216	10000	10	4.66	10.09
enriched-head/fits/seed29101/clean_both_direct	loudness	clean_both	29101	6921216	10000	23	2.15	3.11
enriched-head/fits/seed29101/clean_student_direct	loudness	clean_student	29101	6921216	10000	23	3.64	4.80
enriched-head/fits/seed29101/clean_teacher_direct	loudness	clean_teacher	29101	6921216	10000	23	2.96	3.92
enriched-head/fits/seed29101/jac0.1_student_direct	loudness	student	29101	n/r	10000	132	0.93	1.56
enriched-head/fits/seed29101/jac10_student_direct	loudness	student	29101	n/r	10000	138	0.94	1.19
enriched-head/fits/seed29101/jac1_mixed_direct	loudness	mixed	29101	n/r	10000	107	0.67	1.00
enriched-head/fits/seed29101/jac1_student_direct	loudness	student	29101	n/r	10000	107	0.85	1.13
enriched-head/fits/seed29101/jac1_teacher_direct	loudness	teacher	29101	n/r	10000	107	0.80	0.99
enriched-head/fits/seed29101/jac1d1_student_direct	loudness	student	29101	n/r	10000	38	0.89	1.27
enriched-head/fits/seed29101/mixed_direct	loudness	mixed	29101	6921216	10000	22	0.85	1.70
enriched-head/fits/seed29101/mixed_direct_d512x12	loudness	mixed	29101	41254912	10000	36	0.80	2.47
enriched-head/fits/seed29101/mixed_direct_d512x8	loudness	mixed	29101	27604992	10000	17	0.76	2.26
enriched-head/fits/seed29101/student_direct	loudness	student	29101	6921216	10000	8	n/r	3.47
enriched-head/fits/seed29101/student_direct_frac0.125	loudness	student	29101	6921216	10000	19	5.79	6.40
enriched-head/fits/seed29101/student_direct_frac0.25	loudness	student	29101	6921216	10000	19	4.35	5.41
enriched-head/fits/seed29101/student_direct_frac0.5	loudness	student	29101	6921216	10000	19	2.82	4.34
enriched-head/fits/seed29101/teacher_direct	loudness	teacher	29101	6921216	10000	7	n/r	2.93
enriched-head/fits/seed29101/xreg_mixed_extra_u10000_direct	loudness	mixed_extra	29101	6921216	10000	6	1.01	1.36
enriched-head/fits/seed29101/xreg_mixed_extra_u175000_direct	loudness	mixed_extra	29101	6921216	175000	195	0.31	1.06
enriched-head/fits/seed29101/xreg_student_extra_u10000_direct	loudness	student_extra	29101	6921216	10000	6	1.35	1.69
enriched-head/fits/seed29101/xreg_student_extra_u175000_direct	loudness	student_extra	29101	6921216	175000	136	0.68	1.14
enriched-head/fits/seed29103/jac1_student_direct	loudness	student	29103	n/r	10000	139	0.82	1.06
enriched-head/fits/seed29103/jac1_teacher_direct	loudness	teacher	29103	n/r	10000	138	0.75	0.90
enriched-head/fits/seed29103/student_direct	loudness	student	29103	6921216	10000	15	1.82	3.29

continued

fit	attribute	source	seed	params	updates	min	own	held out
enriched-head/fits/seed29103/teacher_direct	loudness	teacher	29103	6921216	10000	20	2.15	3.42
enriched-head/fits/seed29107/jac1_student_direct	loudness	student	29107	n/r	10000	117	0.83	1.16
enriched-head/fits/seed29107/jac1_teacher_direct	loudness	teacher	29107	n/r	10000	117	0.73	0.88
enriched-head/fits/seed29107/student_direct	loudness	student	29107	6921216	10000	15	1.84	3.50
enriched-head/fits/seed29107/teacher_direct	loudness	teacher	29107	6921216	10000	20	1.92	3.27
flux-head/fits/seed29101/jac1_mixed_direct	onset strength	mixed	29101	n/r	10000	80	0.68	0.58
flux-head/fits/seed29101/jac1_student_direct	onset strength	student	29101	n/r	10000	93	0.76	0.67
flux-head/fits/seed29101/mixed_direct	onset strength	mixed	29101	6921216	10000	14	0.47	0.84
flux-head/fits/seed29101/student_direct	onset strength	student	29101	6921216	10000	16	1.55	1.96
flux-head/fits/seed29101/teacher_direct	onset strength	teacher	29101	6921216	10000	14	1.08	1.42
flux-head/fits/seed29101/xreg_student_extra_u175000_direct	onset strength	student_extra	29101	6921216	175000	365	0.34	0.58
highband-head/fits/seed29101/jac1_student_direct	high-band loudness	student	29101	n/r	10000	65	1.38	1.40
highband-head/fits/seed29101/mixed_direct	high-band loudness	mixed	29101	6921216	10000	11	1.18	2.35
highband-head/fits/seed29101/student_direct	high-band loudness	student	29101	6921216	10000	7	2.08	2.91
highband-head/fits/seed29101/teacher_direct	high-band loudness	teacher	29101	6921216	10000	7	2.91	5.06
highband-head/fits/seed29101/xreg_student_extra_u175000_direct	high-band loudness	student_extra	29101	6921216	175000	290	1.02	1.61
spectral-head/fits/seed29101/jac1_mixed_direct	brightness	mixed	29101	n/r	10000	125	0.99	1.74
spectral-head/fits/seed29101/jac1_student_direct	brightness	student	29101	n/r	10000	124	1.27	1.82
spectral-head/fits/seed29101/jac1_teacher_direct	brightness	teacher	29101	n/r	10000	125	1.29	1.91
spectral-head/fits/seed29101/mixed_direct	brightness	mixed	29101	6921216	10000	7	1.06	3.63
spectral-head/fits/seed29101/student_direct	brightness	student	29101	6921216	10000	15	2.62	4.45
spectral-head/fits/seed29101/teacher_direct	brightness	teacher	29101	6921216	10000	15	2.43	4.58
spectral-head/fits/seed29101/xreg_student_extra_u10000_direct	brightness	student_extra	29101	6921216	10000	6	2.04	3.25
spectral-head/fits/seed29101/xreg_student_extra_u180000_direct	brightness	student_extra	29101	6921216	180000	144	1.28	2.58
spectral-head/fits/seed29103/student_direct	brightness	student	29103	6921216	10000	15	2.67	4.77
spectral-head/fits/seed29103/teacher_direct	brightness	teacher	29103	6921216	10000	15	2.33	4.51
spectral-head/fits/seed29107/student_direct	brightness	student	29107	6921216	10000	15	2.69	4.68
spectral-head/fits/seed29107/teacher_direct	brightness	teacher	29107	6921216	10000	15	2.51	4.88
ace/spectral/fits/seed29101/jac1_student_direct	ACE loudness	student	29101	n/r	10000	370	3.79	4.06
ace/spectral/fits/seed29101/mixed_direct	ACE loudness	mixed	29101	n/r	10000	9	0.75	3.94
ace/spectral/fits/seed29101/student_direct	ACE loudness	student	29101	n/r	10000	9	3.24	5.38
ace/spectral/fits/seed29101/teacher_direct	ACE loudness	teacher	29101	n/r	10000	10	2.78	6.04
ace/spectral/fits/seed29101/xreg_student_extra_u107000_direct	ACE loudness	student_extra	29101	n/r	107000	175	1.19	3.32

H.2 All Panels

Table 72 indexes every stored control panel, including the ones the paper does not quote. We list them all for two reasons. First, the strength calibration is visible: the panels whose tags carry a numeric suffix are the development-prompt runs at 1, 2, 3 and $5\,\%$ from which one strength per attribute was chosen with the decoder gradient, and the reader can check that no other direction source was given its own strength. Second, the single-step panels that preceded the calibrated multi-step protocol are still stored, so the effect of the protocol change is auditable rather than asserted; Appendix B.1 records when the change was made and why.

Every panel in the index was produced by the same script with different flags, writes one result record per case, and asserts three invariants before it reports anything: that a zero-strength hook reproduces the unguided output bit-exactly, that repeating a guided run reproduces its output bit-exactly, and that the gradient at the first intervention is computed from the same clean estimate the unguided run produced. A panel that fails any of the three does not write a result file.

Table 72: Index of all 230 stored control panels. Each row is one complete run of the probe over a prompt panel: the attribute, the prompt panel, the target family, the relative step size, the intervention times, the number of cases, and the feature functions compared. Tags beginning `ace` are the second model family. Panels whose tag ends in `cal` use the calibrated multi-step probe; panels with a numeric suffix are strength-calibration runs on the development prompts. The paper quotes the calibrated panels and reports the others here for completeness.

tag	attribute	panel	target	step	times	cases	feature functions
ace-student-rms-dev	rms	dev	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu), reg. (both)
ace-student-rms-dev-cal	rms	dev	gain	0.03	0.875 to 0.125 (7)	8	decoder, reg. (tea), reg. (stu), reg. (both)
ace-student-rms-dev-cal0.01	rms	dev	gain	0.01	0.875 to 0.125 (7)	8	decoder
ace-student-rms-dev-cal0.02	rms	dev	gain	0.02	0.875 to 0.125 (7)	8	decoder
ace-student-rms-dev-cal0.03	rms	dev	gain	0.03	0.875 to 0.125 (7)	8	decoder
ace-student-rms-dev-cal0.05	rms	dev	gain	0.05	0.875 to 0.125 (7)	8	decoder
ace-student-rms-dev-calinits	rms	dev	gain	0.03	0.875 to 0.125 (7)	8	decoder, teacher:29103, student:29103, teacher:29107, student:29107
ace-student-rms-dev-calmixed	rms	dev	gain	0.03	0.875 to 0.125 (7)	8	decoder, reg. (tea), mixed:29103, mixed:29107
ace-student-rms-dev-jaccal	rms	dev	gain	0.03	0.875 to 0.125 (7)	8	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea)
ace-student-rms-heldout	rms	heldout	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu), reg. (both)
ace-student-rms-heldout-cal	rms	heldout	gain	0.03	0.875 to 0.125 (7)	8	decoder, reg. (tea), reg. (stu), reg. (both)
ace-student-rms-heldout-calinits	rms	heldout	gain	0.03	0.875 to 0.125 (7)	8	decoder, teacher:29103, student:29103, teacher:29107, student:29107
ace-student-rms-heldout-calmixed	rms	heldout	gain	0.03	0.875 to 0.125 (7)	8	decoder, reg. (tea), mixed:29103, mixed:29107
ace-student-rms-heldout-fgfill	rms	heldout	gain	0.03	0.875 to 0.125 (7)	8	fixed dir.
ace-student-rms-heldout-fixedjac48cal	rms	heldout	gain	0.03	0.875 to 0.125 (7)	8	decoder, $\bar{J}$, VJP (stu), reg. (both)
ace-student-rms-heldout-jaccal	rms	heldout	gain	0.03	0.875 to 0.125 (7)	8	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea)
ace-student-rms-heldout-xrfill	rms	heldout	gain	0.03	0.875 to 0.125 (7)	8	xreg student extra u107000
ace-student-rms-rep	rms	rep	gain	0.05	0.75	24	decoder, reg. (tea), reg. (stu), reg. (both)
ace-student-rms-rep-cal	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, reg. (tea), reg. (stu), reg. (both)
ace-student-rms-rep-calinits	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, teacher:29103, student:29103, teacher:29107, student:29107
ace-student-rms-rep-calmixed	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, reg. (tea), mixed:29103, mixed:29107
ace-student-rms-rep-crossattr	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, $\bar{J}$, $\bar{J}$: brightness operator
ace-student-rms-rep-fgfill	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	fixed dir.
ace-student-rms-rep-fixedjac48cal	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, $\bar{J}$, VJP (stu)
ace-student-rms-rep-fixedjaccal	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, $\bar{J}$, VJP (stu), reg. (both)
ace-student-rms-rep-jaccal	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea)
ace-student-rms-rep-pertime	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, $\bar{J}$, $\bar{J}$: t875, $\bar{J}$: t250, $\bar{J}$: t125
ace-student-rms-rep-xrfill	rms	rep	gain	0.03	0.875 to 0.125 (7)	24	xreg student extra u107000
ace-student-spectral-dev-cal0.01	rms	dev	gain	0.01	0.875 to 0.125 (7)	8	decoder
ace-student-spectral-dev-cal0.02	rms	dev	gain	0.02	0.875 to 0.125 (7)	8	decoder
ace-student-spectral-dev-cal0.03	rms	dev	gain	0.03	0.875 to 0.125 (7)	8	decoder
ace-student-spectral-dev-cal0.05	rms	dev	gain	0.05	0.875 to 0.125 (7)	8	decoder
ace-student-spectral-heldout-fgfill	spectral	heldout	gain	0.03	0.875 to 0.125 (7)	8	fixed dir.
ace-student-spectral-heldout-fjcal	spectral	heldout	gain	0.03	0.875 to 0.125 (7)	8	decoder, $\bar{J}$, reg. (stu), reg. (both), reg. (tea)
ace-student-spectral-heldout-vjpfill	spectral	heldout	gain	0.03	0.875 to 0.125 (7)	8	VJP (stu)
ace-student-spectral-heldout-xrfill	spectral	heldout	gain	0.03	0.875 to 0.125 (7)	8	xreg student extra u107000
ace-student-spectral-rep-crossattr	spectral	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, $\bar{J}$, $\bar{J}$: loudness operator
ace-student-spectral-rep-fgfill	spectral	rep	gain	0.03	0.875 to 0.125 (7)	24	fixed dir.
ace-student-spectral-rep-fjcal	spectral	rep	gain	0.03	0.875 to 0.125 (7)	24	decoder, $\bar{J}$, reg. (stu), reg. (both), reg. (tea)
ace-student-spectral-rep-vjpfill	spectral	rep	gain	0.03	0.875 to 0.125 (7)	24	VJP (stu)
ace-student-spectral-rep-xrfill	spectral	rep	gain	0.03	0.875 to 0.125 (7)	24	xreg student extra u107000
student-centroid-dev-cal0.01	centroid	dev	gain	0.01	0.75,0.5,0.25	8	decoder
student-centroid-dev-cal0.02	centroid	dev	gain	0.02	0.75,0.5,0.25	8	decoder
student-centroid-dev-cal0.03	centroid	dev	gain	0.03	0.75,0.5,0.25	8	decoder
student-centroid-dev-cal0.05	centroid	dev	gain	0.05	0.75,0.5,0.25	8	decoder
student-centroid-heldout-fgfill	centroid	heldout	gain	0.03	0.75,0.5,0.25	8	fixed dir.

continued

tag	attribute	panel	target	step	times	cases	feature functions
student-centroid-heldout-jaccal	centroid	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, VJP (stu), jac1 mixed
student-centroid-heldout-regcal	centroid	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), reg. (stu)
student-centroid-heldout-xregcal	centroid	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, xreg student extra u10000, xreg student extra u100000
student-centroid-rep-ce02	centroid	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both)
student-centroid-rep-ce05	centroid	rep	gain	0.05	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both)
student-centroid-rep-crossattr	centroid	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: loudness operator, $\bar{J}$: brightness operator, $\bar{J}$: onset-strength operator
student-centroid-rep-fgfill	centroid	rep	gain	0.03	0.75,0.5,0.25	32	fixed dir.
student-centroid-rep-jaccal	centroid	rep	gain	0.03	0.75,0.5,0.25	32	decoder, VJP (stu), jac1 mixed, VJP (tea)
student-centroid-rep-opstruct	centroid	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: separable (one channel direction), $\bar{J}$: 8-bit quantised, $\bar{J}$: rank 4, $\bar{J}$: frame-local band only
student-centroid-rep-regcal	centroid	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both), reg. (stu), reg. (tea)
student-centroid-rep-xregcal	centroid	rep	gain	0.03	0.75,0.5,0.25	32	decoder, xreg student extra u10000, xreg student extra u100000
student-clap-dev-cal0.01	clap	dev	gain	0.01	0.75,0.5,0.25	8	decoder
student-clap-dev-cal0.02	clap	dev	gain	0.02	0.75,0.5,0.25	8	decoder
student-clap-dev-cal0.03	clap	dev	gain	0.03	0.75,0.5,0.25	8	decoder
student-clap-dev-cal0.05	clap	dev	gain	0.05	0.75,0.5,0.25	8	decoder
student-clap-heldout-fj1cal	clap	heldout	gain	0.01	0.75,0.5,0.25	8	decoder, $\bar{J}$, $\bar{J}$: random matrix, matched row norms, reg. (stu), reg. (both)
student-clap-heldout-fjcal	clap	heldout	gain	0.02	0.75,0.5,0.25	8	decoder, $\bar{J}$, $\bar{J}$: random matrix, matched row norms, reg. (stu), reg. (both)
student-clap-rep-fj1cal	clap	rep	gain	0.01	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: random matrix, matched row norms, reg. (stu), reg. (both)
student-clap-rep-fjcal	clap	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: random matrix, matched row norms, reg. (stu), reg. (both)
student-flux-dev-cal0.01	flux	dev	gain	0.01	0.75,0.5,0.25	8	decoder
student-flux-dev-cal0.02	flux	dev	gain	0.02	0.75,0.5,0.25	8	decoder
student-flux-dev-cal0.03	flux	dev	gain	0.03	0.75,0.5,0.25	8	decoder
student-flux-dev-cal0.05	flux	dev	gain	0.05	0.75,0.5,0.25	8	decoder
student-flux-heldout-fgfill	flux	heldout	gain	0.02	0.75,0.5,0.25	8	fixed dir.
student-flux-heldout-fjcal	flux	heldout	gain	0.02	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (stu), reg. (both), reg. (tea)
student-flux-heldout-vjpfill	flux	heldout	gain	0.02	0.75,0.5,0.25	8	VJP (stu)
student-flux-heldout-xrfill	flux	heldout	gain	0.02	0.75,0.5,0.25	8	xreg student extra u175000
student-flux-rep-crossattr	flux	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: loudness operator, $\bar{J}$: brightness operator, $\bar{J}$: centroid operator
student-flux-rep-fgfill	flux	rep	gain	0.02	0.75,0.5,0.25	32	fixed dir.
student-flux-rep-fjcal	flux	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (stu), reg. (both), reg. (tea)
student-flux-rep-vjpfill	flux	rep	gain	0.02	0.75,0.5,0.25	32	VJP (stu)
student-flux-rep-xrfill	flux	rep	gain	0.02	0.75,0.5,0.25	32	xreg student extra u175000
student-highband-dev-cal0.01	highband	dev	gain	0.01	0.75,0.5,0.25	8	decoder
student-highband-dev-cal0.02	highband	dev	gain	0.02	0.75,0.5,0.25	8	decoder
student-highband-dev-cal0.03	highband	dev	gain	0.03	0.75,0.5,0.25	8	decoder
student-highband-dev-cal0.05	highband	dev	gain	0.05	0.75,0.5,0.25	8	decoder
student-highband-heldout-fgfill	highband	heldout	gain	0.02	0.75,0.5,0.25	8	fixed dir.
student-highband-heldout-fjcal	highband	heldout	gain	0.02	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (stu), reg. (both), reg. (tea)
student-highband-heldout-vjpfill	highband	heldout	gain	0.02	0.75,0.5,0.25	8	VJP (stu)
student-highband-heldout-xrfill	highband	heldout	gain	0.02	0.75,0.5,0.25	8	xreg student extra u175000
student-highband-rep-fgfill	highband	rep	gain	0.02	0.75,0.5,0.25	32	fixed dir.
student-highband-rep-fjcal	highband	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (stu), reg. (both), reg. (tea)
student-highband-rep-vjpfill	highband	rep	gain	0.02	0.75,0.5,0.25	32	VJP (stu)
student-highband-rep-xrfill	highband	rep	gain	0.02	0.75,0.5,0.25	32	xreg student extra u175000
student-joint-heldout-jointcal	joint	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, heads
student-joint-rep-jointcal	joint	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, heads
student-rms-dev	rms	dev	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107, student frac0.125, student frac0.25, student frac0.5
student-rms-dev-cal	rms	dev	gain	0.03	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107
student-rms-dev-cal0.01	rms	dev	gain	0.01	0.75,0.5,0.25	8	decoder
student-rms-dev-cal0.02	rms	dev	gain	0.02	0.75,0.5,0.25	8	decoder
student-rms-dev-cal0.03	rms	dev	gain	0.03	0.75,0.5,0.25	8	decoder
student-rms-dev-dpscal	rms	dev	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-dev-dpsx0.5	rms	dev	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-dev-dpsx1	rms	dev	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-dev-dpsx2	rms	dev	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-dev-e2ecal0.01	rms	dev	gain	0.01	0.75,0.5,0.25	8	end-to-end
student-rms-dev-e2ecal0.02	rms	dev	gain	0.02	0.75,0.5,0.25	8	end-to-end
student-rms-dev-e2ecal0.03	rms	dev	gain	0.03	0.75,0.5,0.25	8	end-to-end
student-rms-dev-e2ecal0.05	rms	dev	gain	0.05	0.75,0.5,0.25	8	end-to-end
student-rms-dev-e2esmoke	rms	dev	gain	0.03	0.75,0.5,0.25	1	decoder, $\bar{J}$, end-to-end
student-rms-dev-multi	rms	dev	gain	0.05	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both)
student-rms-dev-perstr0.01	rms	dev	gain	0.01	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)

continued

tag	attribute	panel	target	step	times	cases	feature functions
student-rms-dev-perstr0.02	rms	dev	gain	0.02	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-dev-perstr0.03	rms	dev	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-dev-perstr0.05	rms	dev	gain	0.05	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-heldout	rms	heldout	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107, student frac0.125, student frac0.25, student frac0.5
student-rms-heldout-cal	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107
student-rms-heldout-cap12	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), mixed d512x12
student-rms-heldout-cap8	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), mixed d512x8
student-rms-heldout-clean	rms	heldout	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu), reg. (both), clean both, clean teacher, clean student
student-rms-heldout-cleancal	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both), clean both, clean teacher, clean student
student-rms-heldout-e2e	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, end-to-end, VJP (stu)
student-rms-heldout-fixedheldoutcal	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, fixed dir., $\bar{J}$, VJP (stu), reg. (both)
student-rms-heldout-fjheadcal	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, head+ $\bar{J}$, reg. (both), VJP (stu)
student-rms-heldout-fjvjac1_student	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, head+ $\bar{J}$, VJP (stu)
student-rms-heldout-fjvxreg_student_extra_u175000	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, head+ $\bar{J}$, xreg student extra u175000
student-rms-heldout-jacabl	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, reg. (tea), jac1 student:29103, jac1 teacher:29103, jac1 student:29107, jac1 teacher:29107, jac0.1 student, jac10 student, jac1d1 student
student-rms-heldout-jaccal	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea)
student-rms-heldout-multi	rms	heldout	gain	0.05	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both)
student-rms-heldout-random	rms	heldout	random	0.03	0.75,0.5,0.25	8	decoder, fixed dir., $\bar{J}$, reg. (stu), VJP (stu)
student-rms-heldout-step	rms	heldout	step	0.03	0.75,0.5,0.25	8	decoder, fixed dir., $\bar{J}$, reg. (stu), VJP (stu)
student-rms-heldout-tremolo	rms	heldout	tremolo	0.03	0.75,0.5,0.25	8	decoder, fixed dir., $\bar{J}$, reg. (stu), VJP (stu)
student-rms-heldout-xregcal	rms	heldout	gain	0.03	0.75,0.5,0.25	8	decoder, xreg mixed extra u10000, xreg mixed extra u175000, xreg student extra u10000, xreg student extra u175000, VJP (stu), reg. (both)
student-rms-rep	rms	rep	gain	0.05	0.75	32	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107, student frac0.125, student frac0.25, student frac0.5
student-rms-rep-cal	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107
student-rms-rep-cap12	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both), mixed d512x12
student-rms-rep-cap8	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both), mixed d512x8
student-rms-rep-cheapdec	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, cheap bf16, cheap half, cheap quarter, $\bar{J}$, fixedjactime
student-rms-rep-clean	rms	rep	gain	0.05	0.75	32	decoder, reg. (tea), reg. (stu), reg. (both), clean both, clean teacher, clean student
student-rms-rep-cleancal	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu), reg. (both), clean both, clean teacher, clean student
student-rms-rep-crossattr	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: brightness operator, $\bar{J}$: centroid operator, $\bar{J}$: onset-strength operator
student-rms-rep-dps	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-dps2	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-e2e	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, end-to-end, VJP (stu)
student-rms-rep-e2efade	rms	rep	fade	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, end-to-end, VJP (stu)
student-rms-rep-early	rms	rep	gain	0.03	0.75,0.5	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-eta0.02	rms	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, VJP (stu)
student-rms-rep-eta0.05	rms	rep	gain	0.05	0.75,0.5,0.25	32	decoder, $\bar{J}$, VJP (stu)
student-rms-rep-fade	rms	rep	fade	0.05	0.75	32	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea), jac1d1 student
student-rms-rep-fadecal	rms	rep	fade	0.03	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea), jac1d1 student
student-rms-rep-fixedfadecal	rms	rep	fade	0.03	0.75,0.5,0.25	32	decoder, fixed dir., VJP (stu), reg. (both)
student-rms-rep-fixedgaincal	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, fixed dir., VJP (stu), reg. (both)
student-rms-rep-fixedjacfadecal	rms	rep	fade	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, VJP (stu)
student-rms-rep-fixedjacgaincal	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, VJP (stu)
student-rms-rep-fixedjactransfercal	rms	rep	transfer	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, VJP (stu)
student-rms-rep-fixedtransfercal	rms	rep	transfer	0.03	0.75,0.5,0.25	32	decoder, fixed dir., VJP (stu), reg. (both)
student-rms-rep-fjheadgaincal	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, head+ $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-fjheadtransfercal	rms	rep	transfer	0.03	0.75,0.5,0.25	32	decoder, head+ $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-fjvjac1_student	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, head+ $\bar{J}$, VJP (stu)
student-rms-rep-fjvxreg_student_extra_u175000	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, head+ $\bar{J}$, xreg student extra u175000

continued

tag	attribute	panel	target	step	times	cases	feature functions
student-rms-rep-jac	rms	rep	gain	0.05	0.75	32	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea)
student-rms-rep-jacabl	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, reg. (tea), jac1 student:29103, jac1 teacher:29103, jac1 student:29107, jac1 teacher:29107, jac0.1 student, jac10 student, jac1d1 student
student-rms-rep-jaccal	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea)
student-rms-rep-jitter	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, decoder jitter, decoder jitter8, $\bar{J}$
student-rms-rep-late	rms	rep	gain	0.03	0.5,0.25	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-mag12	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-mag3	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-matrixsource	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: linear head trained by regression, $\bar{J}$: brightness operator
student-rms-rep-multi	rms	rep	gain	0.05	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu), reg. (both)
student-rms-rep-opcontrol	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: random matrix, matched row norms, $\bar{J}$: built at an unseen instrument, $\bar{J}$: built at a real recording, $\bar{J}$: single state 0, $\bar{J}$: single state 1, $\bar{J}$: single state 2, $\bar{J}$: single state 3
student-rms-rep-opcontrol2	rms	rep	gain	0.03	0.75,0.5,0.25	32	$\bar{J}$: single state 4, $\bar{J}$: single state 5, $\bar{J}$: single state 6, $\bar{J}$: single state 7
student-rms-rep-opreal	rms	rep	gain	0.03	0.75,0.5,0.25	32	$\bar{J}$: rank 1, $\bar{J}$: rank 2, $\bar{J}$: rank 3, $\bar{J}$: rank 4, $\bar{J}$: rank 5, $\bar{J}$: rank 6, $\bar{J}$: rank 7
student-rms-rep-opstructA	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: rank 1, $\bar{J}$: rank 2, $\bar{J}$: rank 4, $\bar{J}$: frame-local band only, $\bar{J}$: separable (one channel direction), $\bar{J}$: 8-bit quantised
student-rms-rep-opstructB	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: row norms equalised, $\bar{J}$: signs only, $\bar{J}$: rows permuted, $\bar{J}$: channels permuted, $\bar{J}$: latent time reversed
student-rms-rep-opteacher	rms	rep	gain	0.03	0.75,0.5,0.25	32	$\bar{J}$: single teacher state 0, $\bar{J}$: single teacher state 1, $\bar{J}$: single unseen-instrument state 1, $\bar{J}$: single unseen-instrument state 2, $\bar{J}$: single unseen-instrument state 3
student-rms-rep-opvariant	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: fp16 storage, $\bar{J}$: rank 8, $\bar{J}$: rank 32, $\bar{J}$: rank 128
student-rms-rep-permtremolo	rms	rep	tremolo	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: rows permuted, $\bar{J}$: channels permuted, $\bar{J}$: latent time reversed, $\bar{J}$: row norms equalised, $\bar{J}$: separable (one channel direction)
student-rms-rep-poolAB	rms	rep	gain	0.03	0.75,0.5,0.25	32	$\bar{J}$: pool A (disjoint states), $\bar{J}$: pool B (disjoint states)
student-rms-rep-poolsize	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: 8-state pool, $\bar{J}$: 4-state pool, $\bar{J}$: 2-state pool
student-rms-rep-precond	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, decoder clip, decoder lowpass, decoder whiten, decoder sign, $\bar{J}$
student-rms-rep-random	rms	rep	random	0.03	0.75,0.5,0.25	32	decoder, fixed dir., $\bar{J}$, reg. (stu), VJP (stu)
student-rms-rep-saturation	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: 128 states
student-rms-rep-singlejacgaincal	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, VJP (stu)
student-rms-rep-step	rms	rep	step	0.03	0.75,0.5,0.25	32	decoder, fixed dir., $\bar{J}$, reg. (stu), VJP (stu)
student-rms-rep-steps1	rms	rep	gain	0.03	0.75	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-steps2	rms	rep	gain	0.03	0.75,0.5	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-rms-rep-transfer	rms	rep	transfer	0.05	0.75	32	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea), jac1d1 student
student-rms-rep-transfercal	rms	rep	transfer	0.03	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea), jac1d1 student
student-rms-rep-tremolo	rms	rep	tremolo	0.03	0.75,0.5,0.25	32	decoder, fixed dir., $\bar{J}$, reg. (stu), VJP (stu)
student-rms-rep-xregcal	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, xreg mixed extra u10000, xreg mixed extra u175000, xreg student extra u10000, xreg student extra u175000, VJP (stu), reg. (both)
student-rms-rep-xregfadecal	rms	rep	fade	0.03	0.75,0.5,0.25	32	decoder, xreg student extra u10000, xreg student extra u175000, VJP (stu), reg. (both)
student-rms-rep-xregtransfercal	rms	rep	transfer	0.03	0.75,0.5,0.25	32	decoder, xreg student extra u10000, xreg student extra u175000, VJP (stu), reg. (both)
student-rms-rep48-cal	rms	rep48	gain	0.03	0.75,0.5,0.25	96	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-spectral-dev	spectral	dev	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107
student-spectral-dev-cal	spectral	dev	gain	0.02	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107
student-spectral-dev-cal0.01	spectral	dev	gain	0.01	0.75,0.5,0.25	8	decoder
student-spectral-dev-cal0.02	spectral	dev	gain	0.02	0.75,0.5,0.25	8	decoder
student-spectral-dev-cal0.03	spectral	dev	gain	0.03	0.75,0.5,0.25	8	decoder
student-spectral-dev-multi	spectral	dev	gain	0.05	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both)
student-spectral-dev-perstr0.01	spectral	dev	gain	0.01	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-spectral-dev-perstr0.02	spectral	dev	gain	0.02	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-spectral-dev-perstr0.03	spectral	dev	gain	0.03	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-spectral-dev-perstr0.05	spectral	dev	gain	0.05	0.75,0.5,0.25	8	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-spectral-heldout	spectral	heldout	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107

continued

tag	attribute	panel	target	step	times	cases	feature functions
student-spectral-heldout-cal	spectral	heldout	gain	0.02	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107
student-spectral-heldout-fgfill	spectral	heldout	gain	0.02	0.75,0.5,0.25	8	fixed dir.
student-spectral-heldout-fixedjaccal	spectral	heldout	gain	0.02	0.75,0.5,0.25	8	decoder, $\bar{J}$, VJP (stu), reg. (both)
student-spectral-heldout-jaccal	spectral	heldout	gain	0.02	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea)
student-spectral-heldout-xregcal	spectral	heldout	gain	0.02	0.75,0.5,0.25	8	decoder, xreg student extra u10000, xreg student extra u180000, VJP (stu), reg. (both)
student-spectral-rep	spectral	rep	gain	0.05	0.75	32	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107
student-spectral-rep-cal	spectral	rep	gain	0.02	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu), reg. (both), teacher:29103, student:29103, teacher:29107, student:29107
student-spectral-rep-crossattr	spectral	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: loudness operator, $\bar{J}$: centroid operator, $\bar{J}$: onset-strength operator
student-spectral-rep-fgfill	spectral	rep	gain	0.02	0.75,0.5,0.25	32	fixed dir.
student-spectral-rep-fixedjaccal	spectral	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, VJP (stu), reg. (both)
student-spectral-rep-jac	spectral	rep	gain	0.05	0.75	32	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea)
student-spectral-rep-jaccal	spectral	rep	gain	0.02	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu), reg. (both), VJP (stu), jac1 mixed, VJP (tea)
student-spectral-rep-multi	spectral	rep	gain	0.05	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu), reg. (both)
student-spectral-rep-opstruct	spectral	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, $\bar{J}$: separable (one channel direction), $\bar{J}$: 8-bit quantised, $\bar{J}$: rank 4, $\bar{J}$: frame-local band only
student-spectral-rep-percal	spectral	rep	gain	0.02	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
student-spectral-rep-sp01	spectral	rep	gain	0.01	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both)
student-spectral-rep-sp03	spectral	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both)
student-spectral-rep-xregcal	spectral	rep	gain	0.02	0.75,0.5,0.25	32	decoder, xreg student extra u10000, xreg student extra u180000, VJP (stu), reg. (both)
student-spectral-rep48-cal	spectral	rep48	gain	0.02	0.75,0.5,0.25	96	decoder, $\bar{J}$, reg. (both), VJP (stu)
teacher-rms-dev-multi	rms	dev	gain	0.05	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu)
teacher-rms-dev-single	rms	dev	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu)
teacher-rms-heldout-multi	rms	heldout	gain	0.05	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu)
teacher-rms-heldout-single	rms	heldout	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu)
teacher-rms-rep-fjteachergen	rms	rep	gain	0.03	0.75,0.5,0.25	32	decoder, $\bar{J}$, reg. (both), VJP (stu)
teacher-rms-rep-multi	rms	rep	gain	0.05	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu)
teacher-spectral-dev-multi	spectral	dev	gain	0.05	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu)
teacher-spectral-dev-single	spectral	dev	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu)
teacher-spectral-heldout-multi	spectral	heldout	gain	0.05	0.75,0.5,0.25	8	decoder, reg. (tea), reg. (stu)
teacher-spectral-heldout-single	spectral	heldout	gain	0.05	0.75	8	decoder, reg. (tea), reg. (stu)
teacher-spectral-rep-multi	spectral	rep	gain	0.05	0.75,0.5,0.25	32	decoder, reg. (tea), reg. (stu)

H.3 Per-Prompt Results

The paired statistics of Tables 1, 11 and 12 are computed from the per-prompt numbers in this section, so these tables are what a reader needs in order to recompute any interval in the paper by hand. Each entry averages the two target directions and, on the new-prompt panel, the two generation seeds of that prompt; prompts are then weighted equally. Reading down a column shows how much of a panel's mean is carried by individual prompts, which is the question a mean and an interval cannot answer.

Two patterns are worth pointing out. On the loudness panels the mean Jacobian is ahead of the decoder gradient on most prompts rather than on a few by a lot, which is why the cluster bootstrap interval excludes zero at a mean difference of only 0.39 dB. On the unseen-instrument panels the spread across prompts is larger than the difference between direction sources, which is the honest reason those comparisons are ties: with eight prompts and two instrument clusters there is not enough resolution to separate methods that differ by a tenth of a decibel.

Table 73: Per-prompt target error in dB, AudioX loudness, ± 6 dB gain, new prompts, calibrated multi-step probe. Each entry averages the two target directions and the two generation seeds of that prompt. Columns: decoder = decoder gradient; $\bar{J}$ = MaJa ($\bar{J}$); VJP (stu) = VJP-matched head (student); unguided = unguided; DSP = signal-processing reference.

prompt	decoder	$\bar{J}$	VJP (stu)	unguided	DSP
rep_i00_s00	0.97	0.97	1.14	6.00	0.00
rep_i00_s01	2.37	1.53	2.46	5.18	0.00
rep_i01_s00	1.88	1.93	1.93	6.00	0.00

continued

prompt	decoder	$\bar{J}$	VJP (stu)	unguided	DSP
rep_i01_s01	2.95	2.53	3.01	5.91	0.00
rep_i02_s00	1.82	1.33	1.78	6.00	0.00
rep_i02_s01	1.95	1.22	1.91	4.86	0.00
rep_i03_s00	1.36	1.59	1.51	6.00	0.00
rep_i03_s01	2.66	1.73	2.63	4.98	0.00
rep_i04_s00	1.38	1.12	1.38	6.00	0.00
rep_i04_s01	2.10	2.12	2.42	5.90	0.00
rep_i05_s00	1.54	1.29	1.55	6.00	0.00
rep_i05_s01	2.29	1.39	2.24	4.69	0.00
rep_i06_s00	2.79	2.53	2.90	5.98	0.00
rep_i06_s01	2.63	1.85	2.61	4.11	0.00
rep_i07_s00	1.00	0.89	1.15	6.00	0.00
rep_i07_s01	2.87	2.27	2.81	5.78	0.00
mean	**2.03**	**1.64**	**2.09**	**5.59**	**0.00**

Table 74: Per-prompt target error in dB, AudioX loudness, linear fade, new prompts. Columns: decoder = decoder gradient; $\bar{J}$ = MaJa ($\bar{J}$); VJP (stu) = VJP-matched head (student); unguided = unguided; DSP = signal-processing reference.

prompt	decoder	$\bar{J}$	VJP (stu)	unguided	DSP
rep_i00_s00	2.01	1.61	2.10	3.48	0.00
rep_i00_s01	3.07	2.53	3.32	3.16	0.00
rep_i01_s00	2.45	2.30	2.50	3.48	0.00
rep_i01_s01	3.72	3.29	3.92	3.39	0.00
rep_i02_s00	2.18	1.27	2.17	3.48	0.00
rep_i02_s01	2.91	2.36	2.65	2.99	0.00
rep_i03_s00	2.15	1.89	2.23	3.48	0.00
rep_i03_s01	3.06	2.32	3.16	2.86	0.00
rep_i04_s00	1.63	1.30	1.65	3.48	0.00
rep_i04_s01	3.39	2.60	3.47	3.48	0.00
rep_i05_s00	1.37	0.82	1.39	3.48	0.00
rep_i05_s01	3.22	2.76	2.98	2.68	0.00
rep_i06_s00	3.05	2.76	3.44	3.46	0.00
rep_i06_s01	3.00	2.73	3.02	2.26	0.00
rep_i07_s00	1.40	0.96	1.50	3.48	0.00
rep_i07_s01	3.61	3.25	3.83	3.44	0.00
mean	**2.64**	**2.17**	**2.71**	**3.26**	**0.00**

Table 75: Per-prompt target error in dB, AudioX loudness, envelope transfer, new prompts. Columns: decoder = decoder gradient; $\bar{J}$ = MaJa ($\bar{J}$); VJP (stu) = VJP-matched head (student); unguided = unguided; DSP = signal-processing reference.

prompt	decoder	$\bar{J}$	VJP (stu)	unguided	DSP
rep_i00_s00	2.43	2.40	2.45	7.04	0.00
rep_i00_s01	4.55	4.51	4.44	11.72	0.00
rep_i01_s00	2.15	2.28	2.15	8.25	0.00
rep_i01_s01	3.45	3.17	3.48	11.42	0.00
rep_i02_s00	4.63	4.19	4.59	8.39	0.00
rep_i02_s01	4.74	4.76	4.47	11.65	0.00
rep_i03_s00	2.90	3.07	2.93	7.83	0.00
rep_i03_s01	5.06	4.88	4.85	11.05	0.00
rep_i04_s00	3.37	3.39	3.39	7.75	0.00
rep_i04_s01	3.70	3.20	3.67	10.71	0.00
rep_i05_s00	4.61	4.24	4.61	8.58	0.00
rep_i05_s01	5.12	5.27	5.07	11.31	0.00
rep_i06_s00	2.90	3.20	2.92	10.38	0.00
rep_i06_s01	4.50	4.14	4.35	9.80	0.00
rep_i07_s00	2.51	2.33	2.46	6.47	0.00
rep_i07_s01	2.78	3.18	2.86	11.89	0.00
mean	**3.71**	**3.64**	**3.67**	**9.64**	**0.00**

Table 76: Per-prompt target error in dB, AudioX loudness, gain, unseen instruments. Columns: decoder = decoder gradient; fixed dir. = fixed direction; $\bar{J}$ = MaJa ($\bar{J}$); VJP (stu) = VJP-matched head (student); reg. (both) = regression head (teacher and student states); unguided = unguided; DSP = signal-processing reference.

prompt	decoder	fixed dir.	$\bar{J}$	VJP (stu)	reg. (both)	unguided	DSP
heldout_08	0.68	0.75	0.58	0.73	1.14	6.00	0.00
heldout_09	0.88	1.18	0.80	0.89	1.04	6.00	0.00
heldout_10	0.73	0.94	0.94	0.79	1.89	6.00	0.00
heldout_11	1.35	2.68	1.48	1.36	2.64	6.00	0.00
heldout_12	1.14	2.16	1.05	1.21	2.23	6.00	0.00
heldout_13	0.70	0.80	0.71	0.71	1.35	6.00	0.00
heldout_14	0.85	0.86	0.86	0.86	1.33	6.00	0.00
heldout_15	1.08	2.61	1.36	1.11	2.24	6.00	0.00
mean	**0.93**	**1.50**	**0.97**	**0.96**	**1.73**	**6.00**	**0.00**

Table 77: Per-prompt target error in dB, AudioX brightness, ± 6 dB shelf, new prompts. Columns: decoder = decoder gradient; $\bar{J}$ = MaJa ($\bar{J}$); VJP (stu) = VJP-matched head (student); reg. (both) = regression head (teacher and student states); unguided = unguided; DSP = signal-processing reference.

prompt	decoder	$\bar{J}$	VJP (stu)	reg. (both)	unguided	DSP
rep_i00_s00	1.08	1.24	1.44	2.43	4.86	0.00
rep_i01_s00	0.83	0.79	1.00	2.38	5.25	0.00
rep_i02_s00	2.60	2.23	2.99	3.10	4.65	1.96
rep_i03_s00	2.99	1.64	2.91	3.15	5.10	0.00
rep_i04_s00	1.76	1.38	1.60	2.76	4.10	0.00
rep_i04_s01	1.56	2.36	2.05	3.30	4.32	4.89
rep_i05_s00	2.73	2.11	2.66	2.86	4.86	0.70
rep_i06_s00	5.62	2.66	6.07	4.34	5.43	2.98
rep_i07_s00	2.28	2.44	2.40	2.85	4.84	0.00
mean	**2.38**	**1.87**	**2.57**	**3.02**	**4.82**	**1.17**

Table 78: Per-prompt target error in $20 \log_{10}$Hz, AudioX spectral centroid, new prompts. Columns: decoder = decoder gradient; $\bar{J}$ = MaJa ($\bar{J}$); reg. (both) = regression head (teacher and student states); reg. (stu) = regression head (student); reg. (tea) = regression head (teacher); unguided = unguided; DSP = signal-processing reference.

prompt	decoder	$\bar{J}$	reg. (both)	reg. (stu)	reg. (tea)	unguided	DSP
rep_i00_s00	0.99	1.08	1.41	1.30	1.86	2.37	0.00
rep_i01_s00	1.36	1.18	1.53	1.46	1.85	2.46	0.00
rep_i02_s00	1.62	1.50	1.83	1.81	2.49	3.69	0.00
rep_i03_s00	1.21	0.96	1.52	1.43	2.23	3.51	0.00
rep_i04_s00	1.56	1.62	1.64	1.62	2.22	3.29	0.00
rep_i04_s01	1.62	1.39	1.64	1.64	2.17	2.50	0.00
rep_i05_s00	1.24	1.11	1.41	1.39	1.99	3.20	0.00
rep_i06_s00	1.33	1.34	1.87	1.77	2.33	3.66	0.00
rep_i07_s00	0.83	1.09	1.39	1.28	2.04	3.02	0.00
mean	**1.31**	**1.25**	**1.58**	**1.52**	**2.13**	**3.08**	**0.00**

Table 79: Per-prompt target error, AudioX onset strength, new prompts. Columns: decoder = decoder gradient; $\bar{J}$ = MaJa ($\bar{J}$); reg. (stu) = regression head (student); reg. (both) = regression head (teacher and student states); reg. (tea) = regression head (teacher); unguided = unguided; DSP = signal-processing reference.

prompt	decoder	$\bar{J}$	reg. (stu)	reg. (both)	reg. (tea)	unguided	DSP
rep_i00_s00	0.80	0.82	0.99	0.95	1.10	1.11	0.00
rep_i01_s00	0.64	0.60	0.74	0.70	0.84	0.85	0.00
rep_i02_s00	0.82	0.76	1.05	1.02	1.07	1.07	0.00
rep_i03_s00	1.80	1.87	2.31	2.13	2.61	2.64	0.00
rep_i04_s00	0.83	0.78	0.99	0.98	1.09	1.11	0.00
rep_i04_s01	0.53	0.53	0.68	0.70	0.80	0.81	0.00
rep_i05_s00	0.88	0.89	1.34	1.24	1.63	1.61	0.00
rep_i06_s00	0.82	0.86	1.20	1.16	1.45	1.43	0.00
rep_i07_s00	0.85	0.74	0.97	0.96	1.01	1.02	0.00
mean	**0.88**	**0.87**	**1.14**	**1.09**	**1.29**	**1.29**	**0.00**

Table 80: Per-prompt target error in dB, ACE-Step loudness, gain, new prompts. Columns: decoder = decoder gradient; $\bar{J}$ = MaJa ($\bar{J}$); VJP (stu) = VJP-matched head (student); unguided = unguided; DSP = signal-processing reference.

prompt	decoder	$\bar{J}$	VJP (stu)	unguided	DSP
acerep_i00_s00	2.74	2.95	2.81	6.00	0.00
acerep_i00_s01	2.43	2.60	2.47	6.00	0.00
acerep_i01_s00	2.46	2.65	2.52	6.00	0.00
acerep_i01_s01	2.27	2.49	2.37	6.00	0.00
acerep_i02_s00	2.77	2.70	2.78	6.00	0.00
acerep_i02_s01	2.87	2.61	2.95	6.00	0.00
acerep_i03_s00	2.65	3.03	2.74	6.00	0.00
acerep_i03_s01	2.42	2.41	2.47	6.00	0.00
acerep_i04_s00	2.33	2.48	2.37	6.00	0.00
acerep_i04_s01	2.42	2.38	2.42	6.00	0.00
acerep_i05_s00	2.59	3.00	2.68	6.00	0.00
acerep_i05_s01	2.48	2.77	2.53	6.00	0.00
mean	**2.54**	**2.67**	**2.59**	**6.00**	**0.00**

Table 81: Per-prompt target error in dB, ACE-Step brightness, shelf, new prompts. Columns: decoder = decoder gradient; $\bar{J}$ = MaJa ($\bar{J}$); reg. (stu) = regression head (student); reg. (both) = regression head (teacher and student states); reg. (tea) = regression head (teacher); unguided = unguided; DSP = signal-processing reference.

prompt	decoder	$\bar{J}$	reg. (stu)	reg. (both)	reg. (tea)	unguided	DSP
acerep_i00_s00	1.57	2.28	4.37	4.23	4.20	5.38	0.97
acerep_i00_s01	1.51	1.80	4.11	4.00	4.04	5.33	0.51
acerep_i01_s00	1.06	1.45	3.94	3.79	3.87	4.95	0.46
acerep_i01_s01	1.38	1.57	3.77	3.56	3.71	4.90	0.00
acerep_i02_s00	1.50	1.90	4.06	3.87	3.93	5.00	0.51
acerep_i02_s01	2.32	1.89	3.77	3.57	3.85	4.78	0.00
acerep_i03_s00	1.04	1.59	4.30	4.12	4.10	5.17	0.00
acerep_i03_s01	1.66	1.18	3.64	3.59	3.86	4.79	0.00
acerep_i04_s00	0.77	1.50	4.12	3.77	4.32	5.33	0.00
acerep_i04_s01	1.80	1.82	4.39	4.01	4.28	5.26	0.50
acerep_i05_s00	1.49	2.29	4.78	4.54	4.59	5.59	0.50
acerep_i05_s01	1.44	2.33	4.72	4.59	4.44	5.60	0.00
mean	**1.46**	**1.80**	**4.16**	**3.97**	**4.10**	**5.17**	**0.29**

H.4 Released Artefacts

Table 82 is an inventory of the release by kind of artefact. Alongside the files it records the git commit and, for every table and figure in this paper, the command that regenerates it from those files. Nothing in the paper is computed from anything outside that inventory.

Two properties of the release matter more than its size. The readout scripts take a panel directory and produce the paper's numbers, so a reader can change an aggregation choice, for instance weighting cases instead of prompts, and see what happens without touching a GPU. And the audit described in Appendix B.2 re-scores every panel from its saved audio with an independent implementation of the attributes, so a reader who distrusts our feature code can check every number in the paper without running either generator.

Table 82: The release, by kind of artefact: every fitted head, every precomputed Jacobian, every panel result and the diagnostic and calibration records, with counts and size on disk in mebibytes. The saved guided and unguided audio of every panel is released alongside these and is larger than all of them together. The release also records the git commit and, for each table and figure of this paper, the command that regenerates it.

artefact kind	files	MiB	example path
fitted heads	61	1824	experiments/h1-control-transfer/ace/fits/seed29101/jac1_mixed_direct.pt
precomputed Jacobians	101	1585	experiments/h1-control-transfer/ace/mean-jacobian/jattr_spectral_student.pt
panel results	179	53	experiments/h1-control-transfer/panels/ace-student-rms-dev-cal/generation-result.json
diagnostics and calibration	16	2	data/pilot_prompts.json
total	357	3464	

I AI ASSISTANCE CLAIM

- **Yes, to aid or polish writing.** An AI assistant was used to edit and tighten prose written by the authors, to check terminology and cross-references for consistency, and to generate the LaTeX for tables and figures from the stored results. The authors wrote the claims and reviewed every sentence.
- **Yes, for retrieval and discovery.** An AI assistant was used to search for and summarise related work and to check bibliographic details. Every cited paper was read and its relevance judged by the authors.
- **Yes, for research ideation or execution.** An AI coding assistant, working under the authors' direction, wrote and ran much of the experiment code (data collection, head fitting, panel runners, re-scoring, statistics and figure scripts) and proposed some of the control experiments. The study design, the hypotheses, the protocols fixed in advance and the interpretation of the results are the authors'; every panel is independently re-scored from saved audio and every number in the paper is regenerated by the released scripts.