Measured Influence: Pre-registered, Power-gated Validation and Calibration of Per-output Training-data Attribution for Consented Music Adapters

Chris Green
Khaos Foundation
London, United Kingdom

9 October 2026
Working paper, version 3

Abstract

When a generative music model is adapted to one artist's consented catalogue, the rights in each output can be computed from how much each input track influenced it and who holds the rights in each input track. The second is a record; the first must be measured, per output, with a stated error. We report a pre-registered, power-gated validation of such a measurement on low-rank adapters (LoKr, rank 32) to a 2.4 B-parameter flow-matching audio model (ACE-Step 1.5), against influence by subset retraining as ground truth: 220 subset adapters on a twelve-input-track catalogue and 165 on a second. Every confirmatory test had its bars committed before scoring, and every per-output test had to demonstrate the ground truth's own resolving power before a verdict was read from it. On ordering, the D-TRAK kernel with velocity-norm features reaches a linear datamodeling score of 0.70 (0.67 to 0.74) over twenty fixed half-size subsets, above cosine (0.46) and a size-only prediction (0.38) on 92% of outputs. On per-output magnitude, the quantity a payment requires, we define the share of a unit payment that lands on the same tracks under the method's split and the ground truth's (one minus their total variation distance), set a bar of 0.70, and find 0.802 (0.797 to 0.807) with calibration slope 1.00 against 0.695 for a split by training-data share. On the second catalogue the first reading drew no verdict because one of 165 adapters had diverged in training and the instrument failed its power gate; under a run-health rule committed before re-scoring, it passed at 0.749 on tracks and 0.916 on the right writers under invented co-writer splits. A calibrated split, the kernel's share times one fixed factor per track fitted once on the retraining record, then passed its own pre-registered tests on fresh outputs at new seeds: 0.879 on the first catalogue, 0.874 on the second, and 0.900 on the production path. Four further tests are reported with their outcomes: dose response, whose expected direction retraining showed to be wrong; delivery invariance, where shares move under 0.05 but the first-named track changes on about one output in ten; known answers, failed three times; and abstention, failed on an embedding rule and passed on a loss-margin rule. An independent unlearning-based method failed by construction of its registered rule. A float16 fault that had reversed the sign of a baseline is disclosed. Results are for two catalogues by one composer; nothing here is evidence of generality.

1 Introduction

A generative music model adapted to an artist's own recordings raises a question with no analogue in the discriminative setting: who holds the rights in what it produces? The premise of this paper is that the base model is a canvas that holds no rights, and that the rights in each output belong to the input tracks of the consented catalogue in proportion to their influence on that output. Each input track's rights are already

recorded, with writers, publishers and splits, so if the influence percentages are correct the output's rights follow by arithmetic. The whole scheme rests on one measurement: for an individual output, how much did each input track influence it?

Training-data attribution offers candidate answers. Influence functions [13], gradient tracing [20], datamodels [11] and TRAK [19] were built for classifiers; D-TRAK [27] carried the TRAK kernel to diffusion models and found, against theoretical expectation, that the gradient of the squared prediction norm is a far better attribution feature than the gradient of the loss. Recent work applies these methods to music [3, 26]. But a score that ranks training examples correctly is not one whose magnitudes can divide money. A royalty needs each share to be the right size, per output, to a tolerance stated in advance, and the standard evaluation of attribution methods, the linear datamodeling score, measures rank agreement only.

This paper reports a protocol of ten tests on that question, nine of them scored against bars committed before scoring and one a recorded listening judgement, run on two catalogues, and the results of each, including the tests that failed. Its contributions are these.

1. **A retraining ground truth whose power is measured before it is used.** We train 220 subset adapters on one catalogue and 165 on a second, show on the first fifteen that subset membership explains 91% of held-out loss variance, and require each per-output test to pass a power gate on the ground truth itself before any verdict is read. Two under-powered designs are reported as such, and on the second catalogue one diverged adapter was found to have poisoned the instrument, which produced a run-health rule committed before re-scoring (Sections 4, 10).
2. **A payment-grade metric and pre-registered passes on it.** The share of a unit payment landing on the same tracks under the method and the ground truth, with a bar of 0.70: 0.802 on the first catalogue, 0.750 on the production path, 0.749 on the second catalogue, each beating a training-share rule by a paired margin (Sections 8, 10).
3. **A calibrated split.** One fixed factor per track, fitted once on the retraining record, raises the figure to 0.879, 0.874 and 0.900 on fresh outputs at new seeds, and on the second catalogue moves the correct first-named track from one output in eight to five in eight (Section 11).
4. **Published failures and surprises.** Known answers failed three times; delivery invariance failed on the first-named track; a dose-response test showed that the expected direction was wrong, because retraining itself shows that repeating a track makes the adapter's outputs slightly worse; an unlearning-based method failed by construction (Sections 12, 9).

Section 15 sets out what is not shown: two catalogues by one composer, a ground truth defined by retraining loss rather than any listener's judgement, co-writer splits that were invented for the test, and a run-health rule whose first independent test is the next catalogue.

2 Related work

Attribution by retraining. Datamodels [11] regress a model output on training-subset membership across many retrainings and are the ground truth against which cheaper methods are scored; Feldman and Zhang [5] and Data Shapley [7] likewise rest on retraining. Their cost confines them to small settings, which is why they serve here as the instrument of proof rather than the product.

Gradient-feature methods. Influence functions [13] approximate upweighting a training example through the inverse Hessian; TracIn [20] traces gradient agreement across checkpoints. TRAK [19] projects per-example gradients to low dimension, whitens them by the projected Gram matrix, and introduces the linear

datamodeling score (LDS): the Spearman correlation [22], over random training subsets, between the model output under retraining on each subset and the summed attribution scores of that subset. TRAK's own results show a single model gives a weak score and that ensembling over independently trained models improves it substantially. Bae et al. [1] and Mlodozeniec et al. [17] extend influence estimation to unrolled training and to diffusion models; dattri [4] packages several methods.

Attribution on diffusion models. Georgiev et al. [6] applied TRAK to diffusion at the level of the noisy trajectory. Zheng et al. [27] (D-TRAK) evaluated attribution on DDPMs and LoRA-finetuned Stable Diffusion by LDS and counterfactual retraining and reported that the gradient of the squared prediction norm ($\mathcal{L}_{\text{Square}}$), through the same kernel, raised LDS by a factor of two to five at ten timesteps over the loss gradient, while gradient cosine without the kernel was near chance at ten timesteps on every benchmark: on generation sets their D-TRAK reaches 0.11 to 0.24 against 0.01 to 0.05 for cosine, each with 64 fixed half-size subsets and three seeds per subset. Xie et al. [24] showed raw gradient norms depend on the timestep; Kwon et al. [14] gave a closed form for LoRA [10] that does not transfer as written to the Kronecker factorisation used here [25]. Wang et al. [23] evaluate attribution for text-to-image models against customisation-based ground truth.

Attribution for music. Choi et al. [3] attribute outputs of a music generative model by unlearning at scale. Serrà et al. [21] (MUCS) attribute by mirrored unlearning with a noise-consistent skew score, on image datasets and full models, reporting an AUC of 0.90 against D-TRAK's 0.67 on CIFAR-10; we read MUCS against our instrument in Section 9. Morreale et al. [18] argue for attribution designed into the inference path. Closest to this paper is Zhang et al. [26], who attribute Stable Audio Open outputs to 22 catalogues with EKFAC and D-TRAK and derive each creator's payment form from the attribution's informativeness; their D-TRAK reaches a calibrated signal-to-noise above one for one catalogue of 22, with a ground truth retrained once per leave-out and no measured noise floor. We adopt their informativeness measure and differ in measuring the ground truth's noise before reading any method against it.

3 Setting and claim

Model and adapters. The base model is ACE-Step 1.5 (turbo variant) [8, 9], a decoder-only flow-matching model [15, 16] over audio latents, at a pinned engine revision. Adapters are LoKr [25] on 264 attention and MLP projections at rank 32, $5{,}128{,}704$ trainable parameters (0.21%). The base model is frozen; every gradient in this paper is with respect to adapter parameters only.

Catalogues. The first catalogue (*The Leap*) is fifteen tracks by the author; three are held out of every training run as query tracks and the remaining twelve are the input tracks, sliced into sections and 25 s overlapping crops (hop 10 s) with their stems, 396 training pairs. One input track supplies about 35% of the pairs and two supply about 7% each. The second catalogue (*HiJinGo*) is fourteen short cues by the same composer, eleven of which vary across the subset plan; six of the eleven have five pairs or fewer. For the second catalogue the rights file was given invented splits between three writers so that the writer-level arithmetic could be tested.

Rights arithmetic. For an output x the scheme computes an influence share $s_t(x) \geq 0$ per input track t, $\sum_t s_t(x) = 1$, reads each track's rights record (writer w holds $r_{t,w}$ of track t), and assigns writer w the share $\sum_t s_t(x)\, r_{t,w}$ of the output; publishers likewise (Appendix C). Two policies are stated as policies: the base model holds no rights, which is why shares sum to one; and a track the method scores as making the

output less likely receives no share. A co-written track divides between its writers by its record, never by measurement.

Claim under test. That an individual output's input-track shares are measured correctly, with a known error, cheaply enough to sit on its provenance record. We do not claim anything about how a co-written track divides, about aesthetic quality, or about catalogues not measured.

4 Ground truth: influence by subset retraining

We take the datamodel view [11, 19]: the influence of input track t on a quantity $f(x)$ is what happens to $f(x)$ when t is present or absent from training, estimated by retraining on many random subsets and regressing.

Design, first catalogue. A subset adapter θ_S is trained on the pairs of $S \subseteq \mathcal{T}$, $|\mathcal{T}| = 12$. The initial plan drew fifteen Bernoulli-half subsets (3 to 8 tracks, 56 to 330 pairs; varying sizes keep the centred design full rank at twelve tracks) plus a reference adapter on all twelve. All runs: 20 epochs, rank 32, held-out pairs capped at 63, ten fixed timesteps $t \in \{0.05, \ldots, 0.95\}$, per-pair seeded noise. After the first per-output test showed fifteen adapters resolve the catalogue's pooled ordering but not one output's deviation from it, the plan was extended at its own seed, keeping the first fifteen: to 70 subsets (sizes 2 to 10), then to 220 with 150 further draws banded to four to nine tracks, because a two- or three-track adapter's loss sits about 0.1 above the rest, 25 times seed noise, and the linear model assumes each track adds a constant. Power was computed against the effect the test must resolve (0.0032 in loss against a coefficient standard error of 0.0113 in the half-subset regime): 70 runs $1.6\times$, 100 runs $2.0\times$, 150 runs $2.5\times$, 250 runs $3.3\times$.

Two instruments. The first 70 subsets were trained on an Apple M4 Max (MPS); all 220 and the reference on a CUDA workstation (RTX PRO 6000, bfloat16). On the 70 both trained, every held-out loss is lower on CUDA by a mean of 0.036 (s.d. 0.020), ten times the per-track effect, so the two sets are separate instruments and are never pooled. Section 8.4 measures what the offset can and cannot do.

Measurement. For a query q and run r, $L(q, r)$ is the trainer's own bucketed velocity mean-squared error at the final epoch. Ridge regression ($\lambda = 1$) of $L(q, \cdot)$ on the $0/1$ membership vector gives $\vartheta(q, t)$, the change in q's loss attributable to t's presence, negative when t helps. Over the first fifteen adapters, mean r^2 across the three queries is 0.906; design rank 12 of 12. Membership explains 91% of the variance in a held-out track's loss, so the adapters learned track-specific structure and the regression is well posed. For a generated output, the same fit of the output's loss on membership gives its ground-truth shares: $-\vartheta$ over the helpful tracks, normalised.

Power before verdict. Each per-output test carries a power gate on the ground truth: the mean per-output signal-to-noise, defined as the spread of an output's fitted coefficients across tracks over their standard error, must be at least 2.0, with at least 60 fits stable under a bootstrap over runs. When the gate is not met the test returns "instrument insufficient" rather than a verdict. Section 8.1 reports the two designs that produced this rule and Section 10 the catalogue on which the gate did its work.

5 Methods under test

D-TRAK kernel. For training pair n under adapter θ, the feature is the adapter gradient of a scalar target, averaged over the ten timesteps with one noise draw per pair, projected to $k = 8192$ dimensions by a seeded

random projection [2, 12]: $\phi_n = P^\top \nabla_\theta g(z_n; \theta)$. Following Zheng et al. [27] the target is the squared norm of the predicted velocity (their $\mathcal{L}_{\text{Square}}$; we say *velocity-norm*); the loss-gradient target is compared in Appendix A. An output is sliced as training audio is, captioned with the caption it was rendered under (trigger phrase and prompt), and its query vector is the mean of its crops' features under the same adapter. With $\Phi \in \mathbb{R}^{N \times k}$ the training feature matrix,

$$\tau(x)_n = \left[\phi(x)^\top (\Phi^\top \Phi + \lambda I)^{-1} \Phi^\top\right]_n, \tag{1}$$

with $\lambda = 10\times$ the mean squared row norm, fixed before any generated output was scored. A track's score is the mean of $\tau(x)_n$ over its pairs; shares are the positive track scores normalised. One index is built per adapter, from a single model: no ensembling.

Training-share rule. Each track's share of the training pairs. It is the cheapest split available and the one any per-output method must beat.

Calibrated split. The kernel's errors are partly a fixed habit per track: across the outputs already scored against the ground truth, some tracks are over-named every time and others under-named. One factor per track, c_t, the ground truth's mean share over the kernel's mean share across the outputs of the fit record, corrects it: the calibrated share is $\hat{s}_t(x)\, c_t$, renormalised. Nothing is fitted per output and the factors are a function of the fit record alone, so a fresh output cannot move them (Section 11).

Controls. Cosine similarity of the same features, L2-normalised and aggregated per track by the mean, which is the "Gradient (cosine)" baseline of Zheng et al. [27]; and a size-only null that scores every pair 1.

6 Evaluation protocol

Ordering. The LDS [19]: for output x and subsets S_m, the Spearman correlation between $-L(x; \theta_{S_m})$ and $\sum_{n \in S_m} \tau(x)_n$. The confirmatory test uses twenty adapters each trained on a random half of the pairs, so subset size is fixed and no correction is needed. The exploratory results on the varying-size subsets, which required a size correction disclosed as post hoc, are in Appendix A.

Magnitude. For a unit payment split by the method as $\hat{s}(x)$ and by the ground truth as $s(x)$,

$$M(x) = 1 - \tfrac{1}{2} \sum_t \left|\hat{s}_t(x) - s_t(x)\right|, \tag{2}$$

the share of the payment that lands on the same tracks under both splits: one minus the total variation distance. The same quantity over writers, after the rights arithmetic, is money on the right writers. Beside it: the slope of s on $\hat{s}$ (calibration), mean absolute error per track, Spearman rank agreement, and top-track agreement. Informativeness in the sense of Zhang et al. [26] is the slope of the ground truth's per-output deviation from the pooled split on the method's predicted deviation.

Pre-registration. Every confirmatory test has a committed record with content hash, written before scoring, fixing the ground truth, the outputs, the metric, the bar and the reading rule; result files carry the hash. Bars are read on the lower bound of a 20,000-draw bootstrap resampling prompts with their seeds together. Analyses not pre-registered are labelled exploratory. The 0.70 bar for M was set, before any adapter was scored, at the level a royalty collection needs.

Table 1: The protocol's ten tests at sign-off, 9 October 2026. "Lower bound" means the pass is read on the bootstrap lower bound.

	Question	Bar	Standing
T1	Does it rank input tracks as retraining does?	LDS ≥ 0.25; above cosine and size on 80% of outputs	Passed: 0.70; 92%
T2	Is each share the right size?	$M \geq 0.70$ (lower bound); slope in $[0.5, 1.5]$; beats flat split	Passed on both catalogues and the served path; calibrated split passed too
T3	Does removing the top tracks hurt most?	True on 80% of outputs	No verdict: gap under seed noise
T4	Dose: does a track's share rise when it is repeated in training?	Rises at each step, in the method and in the measured effect (first records); direction only (third)	Passed on direction, third record; first two records stand failed
T5	Known answers: is a training section's own track named first?	First on 90%	Failed, three readings
T6	Do shares survive MP3, level and trim?	Shares within 0.10; same top track on 90%	Failed on top track under MP3 and trim; shares within 0.05
T7	Does it refuse outputs it should not attribute?	Controls refused on 90%	Failed on an embedding rule; passed on a loss-margin rule
T8	Does it hold on other catalogues?	Each passes on its own	One second catalogue passed
T9	Same output, same answer?	Identical on one device; within 2 points across; an independent repeat	Passed on one device; across devices met narrowly once; no independent repeat yet
T10	Where method and resemblance disagree, which does the artist believe?	Recorded, not scored	10 of 10, labelled

The ten tests. Table 1 lists the protocol's tests, their bars and their standing at sign-off. The rest of the paper takes them in the order the evidence was built.

7 Ordering (T1)

The confirmatory ranking test, run on 30 September 2026 against twenty adapters each trained on a random half of the pairs, gave mean LDS $+0.702$ ($+0.667$ to $+0.735$) against a bar of 0.25, with the kernel above both cosine ($+0.460$) and a size-only prediction ($+0.380$) on 92% of outputs (86% to 98%) against a bar of 80%. The exploratory path that preceded it, on the varying-size calibration subsets with a size correction, gave 0.36 over 48 outputs (20 of them in the pre-registered confirmation, 28 in a fresh sample) and 0.43 over an 80-output stratified set across five prompt kinds, with cosine at chance and the loss-gradient target failing its own pre-registered comparison (Appendix A). For scale, Zheng et al. [27] report generation-set LDS of 0.11 to 0.24 for D-TRAK on image benchmarks at ten timesteps; we do not claim the comparison is exact.

Table 2: Per-output payment test on the first catalogue, 80 outputs. Primary set: 213 CUDA-trained subset adapters, sizes 4 to 9. Intervals are 95% bootstrap over prompts with seeds together. Bars were fixed before scoring.

Reading	Research index	Served path (9 tracks)	Bar
Power gate: signal-to-noise	6.42	6.09	≥ 2.0
Power gate: stable fits	80 of 80	80 of 80	≥ 60
M, money on the right tracks	0.802 (0.797–0.807)	0.750 (0.744–0.756)	≥ 0.70 lower bound
Slope of ground-truth shares on method's	1.00	0.94	0.5 to 1.5
Paired gain over training-share rule	+0.107 (+0.093–+0.122)	+0.099 (+0.083–+0.116)	above 0; ≥ 0.02
Training-share rule alone	0.695 (0.680–0.710)	0.651 (0.634–0.667)	reported
Mean absolute error per track (points)	3.3 (3.2–3.4)	5.6 (5.4–5.7)	≤ 10, reported
Rank agreement	0.64 (0.62–0.66)	0.45 (0.41–0.48)	≥ 0.7: not met
Informativeness (deviations)	0.55	0.95	reported

8 Magnitude on the first catalogue (T2)

8.1 Two under-powered designs

Leave-one-track-out (withdrawn as uninformative). Twelve adapters each retrained without one input track, plus the full adapter on three seeds as a noise floor, read against 80 outputs with effects counted above twice the seed noise. The pre-registered reading at first reported a failure (mean error 11.8 points, slope 1.80). Re-analysis showed the instrument could not measure the quantity: seed noise on one adapter's loss is 0.0042 against a per-track effect of 0.0032 (signal-to-noise 0.77), 12% of the measured spread is real, and on 53 of 80 outputs nothing survives the gate. Two runs of that ground truth would agree at 0.11 on order; a perfect method would score 0.49 against it and D-TRAK scored 0.47. The gate was manufacturing concentration: with signal-to-noise under one, whichever track's noise spike clears it takes everything. The result file and record stand; the withdrawal is not a pass. The author's listening prompted the check: the shares did not match what the outputs sounded like.

Two ground truths (inconclusive as pre-registered). Leave-one-track-out against the fifteen-adapter regression agreed at 0.21 against a floor of 0.50; the regression was a thirteen-parameter fit on fifteen points (mean leverage 0.87). Pooled over 80 outputs the catalogue's per-track ordering was reliable (split-half +0.97), but per-output resolution was absent. The extension to 220 adapters followed.

8.2 The regression test (passed, confirmatory)

Fixed before scoring: the CUDA-trained adapters alone, subsets of four to nine tracks as the primary set (213 runs; at least 100 required) and every size as a sensitivity set (220); one ridge fit per output; the power gate; the same 80 outputs as the exploratory set (forty prompts over five kinds, two seeds); and the comparison against the training-share rule, paired per output, margin 0.02.

The gate was met and the test passed (Table 2, first column). The D-TRAK split places 80.2% of a unit payment where the retraining ground truth places it, with slope 1.00; the training-share rule reaches 69.5% and D-TRAK adds 10.7 points on top. The sensitivity set agrees (0.800, slope 0.96, gain +0.108, gate 5.57). Per-output fits explain the loss well (median $r^2 = 0.82$, lowest 0.72). Rank agreement of 0.64 misses its reported target: the method gets the money about right and orders the smaller shares poorly. Informativeness in the sense of Zhang et al. [26] is 0.55 (correlation 0.56, signal-to-noise 1.2); at their risk aversion of 2 this implies a benchmark royalty rate of 0.27, a hybrid contract rather than a pure royalty. On shares rather than deviations the slope is 1.00 and the correlation 0.86.

The training-share rule lands 69.5% because influence and data share are correlated, not identical: the track that is 35% of the training data receives 18 to 28% of the kernel's aggregate share and two tracks at 7% each receive 12 to 18%. A rule paying by data share would systematically overpay the longest recordings.

8.3 The served path as its own method (passed, confirmatory)

The test above read a research index on the calibration reference adapter. What the product serves is a different artefact by a different path: a 50-epoch run with epoch 8 selected by the held-out picker, an index built on those weights at publication, and the serving worker's attribution call. Pre-registered on 27 September before any output was rendered, this test rendered the same 80 captions and seeds through that adapter on the workstation, attributed each through the exact serving call, and read the splits against the same regression. The served run held out three of the regression's twelve input tracks and trained on the plan's three query tracks, so the record fixed the comparison to the nine tracks both splits can name, renormalised, with the mass outside reported: the served path put 0.31 of its mass on the query tracks and the regression 0.33 on the served run's held-out three. It passed (Table 2, second column), lower on money than the research index and worse on rank: on the nine shared tracks its largest share goes to one of two tracks on all 80 outputs, which between them hold 45% of the shared pairs. Its informativeness is higher (0.95, correlation 0.68): the served adapter's splits move more with the output and the regression follows them. Outputs were rendered and scored on one device, so no device caveat applies. One reporting defect is disclosed: the result file records the rule's mass outside the shared tracks as zero because the reader restricted the rule before measuring it; the figure is 0.30 and the comparison was unaffected.

8.4 The device offset, and a second instrument

The research-index outputs were rendered on MPS and scored under CUDA-trained adapters. On the 70 subsets both machines trained, over the same 80 outputs (exploratory), MPS losses sit 0.016 above CUDA on average; the offset decomposes into a per-adapter part (s.d. 0.0044) and a per-output part (0.0092), both of which fall into the fits' intercepts, and an interaction (0.0053) that alone could tilt a coefficient. Against the within-output spread the regression fits on (0.0248) the interaction is 0.22, and it is the size by which two trainings of one subset on one device differ (seed noise 0.0042 each, 0.0059 on their difference). The two devices order the 70 adapters alike for each output (rank agreement 0.94, lowest 0.82) and the outputs alike under each adapter (correlation 0.99). The device is a caveat, not a confound. A second instrument, never pooled with the first, agrees: the Studio's 70 adapters (63 in the four-to-nine band, under the record's floor of 100, so exploratory) pass the gate on their own (signal-to-noise 3.7, 80 of 80 stable) and read D-TRAK at 0.808 (0.799 to 0.816), slope 0.96, 0.110 above the rule's 0.698, rank agreement 0.67, informativeness 0.66.

8.5 A second confirmation on fresh outputs

A variant split, 0.7 of the kernel's share plus 0.3 of the training share (the *anchored* split), was run through the payment test on 80 fresh outputs on 5 October: 0.842 (0.836 to 0.847), slope 0.99, against the rule's 0.686. The anchor neither cost nor bought anything measurable ($+0.005$, -0.001 to $+0.011$, against the kernel alone at 0.837 on the same outputs), so the run stands as a second confirmation of the kernel's payment result on outputs it had never seen, with rank agreement (0.715) above its 0.7 bar for the first time. The anchored split itself was withdrawn the same day, having failed the dose-response record (Section 12); the kernel alone was retained.

Table 3: The second catalogue, 70 outputs, primary set. First reading: 161 runs, gate not met, every number exploratory by its record. Second reading: 160 runs under the run-health rule, confirmatory by its record.

Reading	First (161 runs)	Second (160 runs)	Bar
Power gate: signal-to-noise	1.10	5.79	≥ 2.0
Power gate: stable fits	70 of 70	70 of 70	≥ 60
M, right tracks	0.656 (0.643–0.668)	0.749 (0.733–0.763)	≥ 0.70 lower bound
Slope	0.40	0.51	0.5 to 1.5
Paired difference from training-share rule	-0.010 (-0.019–-0.001)	$+0.049$ ($+0.037$–$+0.060$)	above 0; ≥ 0.02
Training-share rule alone	0.666	0.700	reported
Mean absolute error per track (points)	6.3	4.6	≤ 10, reported
Rank agreement	0.44	0.69 (0.65–0.73)	≥ 0.7: not met
M, right writers	0.802 (0.796–0.810)	0.916 (0.907–0.925)	≥ 0.70 lower bound
Training-share rule, right writers	0.828	0.890	reported
Paired gain over rule, writers	-0.026 (-0.033–-0.019)	$+0.026$ ($+0.018$–$+0.034$)	above 0; ≥ 0.02

9 MUCS against the same instrument (failed, confirmatory)

A method that fails independently of D-TRAK was read against the same 80 outputs, the same 213-adapter regression, the same gate and bar, pre-registered on 25 September. Following Serrà et al. [21]: copy the adapter the output was rendered from, push the copy away from that output by bounded mirrored gradient ascent within the adapter alone ($\lambda = 0.2$, stop at 0.95 of the zero-velocity loss, Adam at 10^{-4}, at most 200 steps, minibatches of 8 training pairs), then score every training pair by its normalised loss skew between copy and original under the same fixed noise at the same ten timesteps; a track's score is the mean over its pairs, shares by the kernel's rule. Every output was unlearned as intended (16 to 35 ascent steps, mean 26). To our knowledge this is the first reading of MUCS on audio.

The result is a failure by construction of the registered rule: every training pair's loss *fell* under the unlearned copy on every output, so every track score was negative, the positive-normalised shares were all zero, and M reads 0.500, the value an empty split takes. The mechanism is plain in hindsight: the mirrored objective descends on the training pairs while it ascends on the output, so the copy fits the training set better than the original and the skew is negative everywhere; the method would need the relative shift between pairs, not the raw one. Two exploratory re-readings outside the record, centring each output's track scores on their mean (0.421) and reversing the sign (0.537), both land under the flat split. On this adapter and this regression MUCS as built here carries no per-output information; D-TRAK's 0.802 stands alone. We do not read this as evidence about MUCS on the full models and image data its authors report on.

10 A second catalogue, and a poisoned instrument (T8)

The payment test was pre-registered unchanged on the second catalogue, with a second record for the writer-level reading under the invented three-writer splits, both committed before any adapter of the plan existed. The instrument was 165 subset adapters over the eleven varying tracks, 161 inside a 30 to 70% pair-share band as the primary set, and 70 outputs fixed by rule (35 prompts, two seeds).

First reading: instrument insufficient. The gate was not met (signal-to-noise 1.10), so by its record nothing about the method was settled (Table 3, first column). Read as exploratory, the numbers did not favour the method: it sat no nearer the retraining split than the training-share rule at track level and a little further at writer level, and its mean writer shares ($54/25/21$) stayed close to the rule's ($52/29/20$) while retraining's ($35/30/36$) were flattened toward equal, as a noisy fit flattens them. Whether more adapters would help

Table 4: The calibrated split on fresh outputs at new seeds (confirmatory). The second catalogue's first reading used 149 runs by a runner default; the reading on the 160 runs its record named follows the semicolon. The served path is the first catalogue's production adapter on the nine shared tracks.

Reading	First catalogue (80)	Second catalogue (70)	Served path (80)	Bar
M, calibrated	0.879 (0.873–0.885)	0.857 (0.835–0.875); 0.874 (0.852–0.893) on 160 runs	0.900 (0.893–0.906)	$\geq$ 0.70 lb
Kernel alone, same outputs	0.811	0.714; 0.731 on 160 runs	0.753	reported
Training-share rule, same outputs	0.734	0.681		reported
Gain over kernel, paired	+0.068 (+0.061–+0.075)	+0.143 (+0.136–+0.151); unchanged on 160	+0.147 (+0.138–+0.156)	$\geq$ 0.03 lb
Slope	0.883	0.832	0.945	0.5 to 1.5
Ground truth's first track named	80 of 80 (every method)	43 of 70 vs 9 uncalibrated; 44 of 70 vs 8 on 160 runs		reported
Mean absolute error per track		2.6 vs 5.2		reported
Power gate	SNR 7.4; 80 of 80	SNR 5.2; 70 of 70		met

was then measured rather than guessed: the ground truth re-fitted on random subsets of its own runs. On the first catalogue the gate ratio grows with run count as a power law of exponent 0.70, reaching the gate at about 36 runs. On the second it did not grow (1.2, 1.9, 2.8, 1.6, 1.1 at 20, 40, 80, 120 and 161 runs, spreads ± 1.4). That shape was traced to one adapter whose training diverged (loss $0.82 \rightarrow 1.47$ at epoch 6, final 1.54 against a median of 0.72): without it the gate ratio is 5.8, and no other run moves it by more than 0.04. The instrument was not weak; it was poisoned by one failed training that the record had no rule to exclude. The first result stands as scored.

Second reading under a run-health rule (passed, confirmatory by its record). A rule was pre-registered before re-scoring, bars unchanged: a run whose final training loss exceeds twice the median of the plan's finished runs is a failed training and is left out. It excluded exactly one run of 165. The gate was met and the test passed (Table 3, second column): 0.749 on the right tracks against the rule's 0.700, 0.916 on the right writers against 0.890, slope 0.51 at the edge of its range. With the failed training out, retraining no longer flattens toward equal (mean writer shares $50/21/29$) and the method sits nearer it than the rule on two of three writers. The two readings are read together. The first stands because its record had no rule for a failed training. The second is confirmatory by its record, but the rule was written after the diverged run had been found by looking, so it excludes exactly the run it was written for: it is not an independent confirmation of the rule, whose own test is the next catalogue, where it goes into the record before any adapter trains. On what the test can say, the second catalogue agrees with the first: the method beats the training-share rule, by five pence in the pound at track level and three at writer level, and the writers' money lands where the invented splits say it should. On what it cannot say nothing has changed: short cues, invented splits, the calibration recipe, and rank agreement under its bar on both catalogues.

11 The calibrated split

The per-track factors c_t of Section 5 were fitted once on each catalogue's payment-test record and tested, as pre-registered, on outputs nobody had scored, rendered at new seeds, against the full instrument.

All three passed (Table 4). On the second catalogue the misdirected money halves and the correct first-named track goes from about one output in eight to five in eight, which is the headline for an artist who

asks whether the top track holds. On the first catalogue every method names the same first track on every output, so that catalogue cannot test it. On the second catalogue the calibrated record named the 160 primary runs of the health record while the runner read 149 under its default band; re-read from the same cached losses under the named list it gives 0.874 with the gain unchanged, and both readings stand with the record carrying the amendment. The served path's reading was separate because the served adapter is a different artefact from the research index the factors were fitted on: 80 fresh outputs through the product's own call, 0.900 on the nine shared tracks against 0.753 for the served kernel alone; fourteen outputs the abstention line of the time refused were counted and not scored.

What a calibration costs. It exists only where subset adapters exist: about 220 for a fifteen-track catalogue, roughly two days of the workstation to train and seven hours to score. An exploratory desk reading on the stored records (25 random draws of N adapters, the ground truth re-fitted and the factors applied to the fresh outputs) puts most of the gain at 60 adapters (0.853 and 0.847 on the two catalogues; every draw at 50 or more beat the kernel alone) and shows a bad draw under 50 can be worse than no calibration. Whether a record built on 60 passes its own power gate is untested. A new artist therefore receives the kernel alone, with its measured error stated, until that work is done for them.

12 Further tests: outcomes as recorded

Dose response (T4). The first record asked whether a track's share rises when its pairs are repeated two and three times in training, and failed in the opposite direction: the adapter fitted the repeated track better at each step (its own sections' loss $0.722 \to 0.678 \to 0.624$) while the kernel's share for it fell ($0.092 \to 0.046 \to 0.012$) and its mean score per index row fell to about zero. A track the adapter has learned well has small gradients, and a gradient kernel credits it less. Normalising every index row to unit length did not change the direction ($0.083 \to 0.032 \to 0.006$): the learned track's gradients lose their direction, not only their length; averaging each pair's score over the epoch-10 and final checkpoints softened the fall without reversing it ($0.096 \to 0.065 \to 0.025$). A second record then asked the question against retraining: twelve adapters with the track absent, once, twice and three times show that having the track helps the outputs (0.0036 in loss, noise 0.0015) and that each repeat makes them slightly *worse* (0.0024 a step, noise 0.0008). The first record's expectation was wrong for this catalogue. A third record, registered for the kernel alone on a second track (the largest, 137 of 396 pairs) before any of its adapters trained, passed on direction: retraining again shows the track helping (0.0090, seven times its noise) and each repeat hurting (0.0065 and 0.0068), and the kernel's share falls at both steps ($25.1\% \to 21.8\% \to 19.1\%$). At three copies the outputs are slightly worse than with the track absent and the kernel still gives it 19%: the direction of the method's response to repetition is right, its size is not established, and a gradient kernel taken after training under-credits material the adapter has learned well. The first two records stand failed.

Known answers (T5): failed, three readings. Given a training section as if it were an output, does the method name that section's own track first? On the first catalogue the own track ranked first on 38 of 56 sections (lower bound 19% against 90%) and in the top three on 55; the 137-pair track took first place in 16 of the 18 misses, and the own track's share exceeded its share of the training pairs on 55 of 56. A size-aware reading, pre-registered, divided every share by its track's training share and gave 37 of 56 (lower bound 54%): size does not explain the misses. A third readout, each track read by the mean of its three strongest pair scores, asks which track a piece of audio is from rather than how much each track shaped it; chosen on the first catalogue's sections (55 of 56) and so tested on the second where no readout had been scored, it gave 36 of 39 (lower bound 81% against 90%). Under the meaning of credit, retraining names a section's own track first and the kernel often does not; the calibration corrects a fixed habit per track, not this.

Delivery (T6): shares survive, the first name does not. On 80 fresh outputs no share moved more than 0.042 under MP3 at 128 kbit/s, ± 6 dB or a 10% trim (bar 0.10), but the top track was kept on 92.5% under MP3 and 93.8% under the trim, whose lower bounds (86.2%, 88.8%) fall under 90%; the level changes cleared it. A pre-registered replication on 240 fresh outputs failed the same way (no share moved more than 0.047; MP3 lower bound 85.4%, trim 89.2%). On 120 further files, with no change to the file at all, a second noise draw in the scoring changed the kernel's first name on 8 of 120: the first name is unstable under the scoring's own noise rather than under delivery, on about one output in ten. The limit is stated on the output's method line.

Targeted removal (T3): no verdict. Removing each output's top-ranked track raised its loss by $+0.0079$ on average against $+0.0096$ for random sets of the same pair count; the gap (-0.0017) is under twice the measured seed noise (0.0021), which the record fixed in advance as the limit of what could be claimed. Across the 18 removal adapters the share the method gave the removed tracks ranks the measured loss rise at $+0.62$ per output. Resolving top against random needs several seeds per removal.

Abstention (T7): failed, then passed. A closeness rule at the artist's own 5th percentile of cosine to the catalogue's CLAP centroid refused the base model's outputs on 47 of 80 and another adapter's on 52 of 80 (lower bounds 44% and 53% against 90%) while keeping 76 of 80 of the artist's own: one number on one embedding cannot separate them. A second rule, pre-registered the next day, accepts an output only when the adapter explains it better than the base model by more than a fixed margin and explains it well in absolute terms, two loss thresholds set on the adapter's own earlier outputs. On 120 fresh outputs of each kind it accepted 116 of the adapter's own (lower bound 93.3% against 85%), refused 119 of the base model's (lower bound 97.5%) and 115 of the other adapter's (90.8%); the outputs it accepts still pass T1 (LDS 0.701). The thresholds are per adapter and per device, the rule needs the caption an output was rendered under, and its behaviour on re-encoded files is unmeasured. A line drawn on 48 outputs refused one in six of the adapter's fresh work; redrawn on 128 of its own outputs, false refusals fell to one in sixteen (120 of 128 accepted) and the base-model control still cleared its bar.

Reproducibility (T9). Re-scoring 80 outputs from their audio on the same machine returned the same top track on every output and moved no share by more than 0.00018; the record had defined identical as within 0.0001 and three outputs exceeded it, so the letter of the record failed and stands. Re-recorded at the precision the output card shows (0.1 of a point) and re-scored afresh, it passed, 80 of 80 with a worst change of 0.02 of a point: the MPS forward pass is not bit-repeatable; percentages are repeatable to two hundredths of a point. Across devices (Section 13) the bar was met narrowly once.

The artist's ears (T10). On the ten outputs where measured influence and a resemblance estimate disagree most, seven with different first-named tracks, the author believed measured influence on all ten. The splits were labelled, so this is a recorded judgement and not a blind test.

13 Conditions of a served attribution (exploratory)

Putting the served path on three machines produced four findings. One had its bar fixed in advance; the rest were measured in the course of the work and have no result files.

An index built on another device type. An adapter trained and indexed on CUDA is served on Apple silicon, so the output's gradient is taken on MPS and read against rows computed on CUDA. The bar, fixed

before measurement: the cross path moves on average at most 0.05 of the money against the other device's own attribution and names the same top track on at least 90% of outputs. On 20 outputs over the five prompt kinds, against an MPS index of the same weights: 0.044 moves (agreement 0.956, lowest 0.948) and the top track agrees on 95%. Met, narrowly. Swapping the index alone moves 0.020; the two devices' own paths differ by 0.040. One reading on one catalogue; a second is owed before the policy is leaned on. The same adapter's abstention thresholds differ by device (gain 0.5607/loss 0.9989 on CUDA, 0.5222/0.9951 on Apple silicon), which is why the reader checks the device.

The caption must be the one the output was rendered with. Every adapter is trained with a trigger phrase ahead of the caption. A scorer that used the prompt as typed, without the trigger, moved 12 to 14% of the money on three 60-second outputs of a 77-track catalogue since withdrawn (agreement 0.882, 0.863, 0.866), more than the method's gain over the flat split; one output of a three-track catalogue kept 0.997. Four outputs; the reading cannot be repeated. Every test in this paper scored its outputs with the trigger in the caption.

An index scores only outputs its own weights rendered. An output rendered by an earlier adapter of the same artist has no rows of its own in a later index, and scoring it returns percentages that look like any others. The software now compares the hash of the weights on the output's record with the index's and refuses on a mismatch. The rule was written after nine outputs had been scored against the wrong index; those documents were withdrawn and how far off they were was not measured.

Memory. One plain backward pass in bfloat16 takes 12.4 GB at a 25-second window and 24.1 GB at a 120-second output, about 0.12 GB per second of audio. Gradient checkpointing over the decoder's 24 layers, with the decoder's cache disabled so the recomputation matches the first pass, removes the dependence on length: the projected feature and the shares are identical to the last bit on three windows, and the scorer peaks at 8.1 GB against 25.7 GB, about 40% slower. On a 32 GB laptop (Apple M2 Max) four outputs scored in 4.7 to 5.1 minutes for 60 seconds of audio and 8.7 minutes for 120. Checkpointing on CUDA has not been run on hardware.

14 Cost

The product runs no retraining. One index per adapter is built once (81 minutes for 396 pairs on the M4 Max) and each output is attributed in about 40 seconds on the machines that proved it, or five to nine minutes on a 32 GB laptop. Retraining is how a method version is proven: one subset adapter takes about 16 minutes on the workstation, a 221-adapter ground truth (220 subsets and the reference) about 2.5 days at two workers, scoring 80 outputs under 220 adapters about 6.5 hours, and one MUCS attribution about 28 minutes. These are recorded timings, not result files.

15 Limitations

- **Two catalogues, one composer.** Neither catalogue has real co-writers. The split only moves money between different rights holders, and that has been tested once, on invented splits. A catalogue with real co-written splits by another composer is the next proof, and with it a larger catalogue that can be read at 15, 25 and 40 tracks.
- **Ground truth is retraining loss.** "Correct" throughout means the split a retraining ground truth assigns to an output's velocity-matching loss, which is the standard datamodel criterion and not a listener's judgement. The one listening record (T10) was labelled.

- **The run-health rule.** It was written after the diverged run had been found by looking. Its first independent test is the next catalogue.
- **Known answers fail.** Under the meaning of credit, retraining names a training section's own track first and the kernel often does not; no readout tried has cleared the bar, and the calibration does not address it.
- **The first-named track is unstable** on about one output in ten under the scoring's own noise, though shares move under 0.05.
- **Repetition.** A gradient kernel taken after training under-credits material the adapter has learned well; the direction of its response to repetition matches retraining, its size does not.
- **Partial informativeness and rank.** 0.55 beyond the pooled ordering on the research index; rank agreement under 0.7 on both catalogues except on one fresh-output run.
- **Served path on nine tracks; second catalogue's served path unrun.** The calibrated split is served for the second catalogue on the strength of the first's served reading.
- **Single-model index.** No ensembling; λ chosen on three held-out tracks and consistently beaten on outputs by a smaller value not pre-registered; ten fixed timesteps unmeasured; per-timestep renormalisation [24] untested.
- **Independent replication.** Every result file is published with its record's hash, but no one other than the author has yet re-run them.
- **Eighteen runs without a dataset hash**, completed from their own files after a watermark-band failure; their trained subsets were verified against the plan.

16 Position on use

Until 5 October 2026 no number in this paper was used in any compensation decision. On that date the author, as the artist whose catalogues these are, decided that the served split may be paid on at a stated standard: it must be clearly better than no split, most of the time, and must state how wrong it can be, with its failed tests shown beside its passes. Nothing yet pays per track. The conditions under which it would are these: every paid split shows its method, version and error per track; an adapter has its abstention check before money is taken on its splits, and a refused piece is paid nothing by this method; when the method changes, splits are recomputed from that date; a new artist is paid on the kernel alone, with its error stated, until a calibration has been run for them. The limit is stated with it: the split only moves money where tracks have different rights holders, and that has been tested once, on invented splits. The defensible sentences are the measured ones: on these two catalogues, in pre-registered tests against a retraining ground truth, the kernel's per-output split put 80 pence in the pound on the right tracks against 70 for a flat split, and the production path 75 against 65 on the nine tracks both could name; the calibrated split put 88, 87 and 90.

17 Conclusion

We asked whether a single-model D-TRAK attribution, computed in under a minute without retraining, measures an individual generated track's dependence on the input tracks of a consented catalogue well enough to divide a payment, and ran ten pre-registered tests to find out. Against a retraining ground truth whose resolving power was measured first, it orders tracks reliably (LDS 0.70), puts 80 pence in the pound where retraining says they belong on the first catalogue and 75 on the second, and with one fixed factor per track rises to 87 to 90. It also fails in ways that are now known: it does not name a training section's own track first,

its first-named track is unstable on one output in ten, and it under-credits material the adapter has learned well. One instrument was found poisoned by a single diverged training and the rule that excludes such runs awaits its first independent test. The experiment that changes what can be claimed is a catalogue with real co-writers by another composer, run under rules that are now all in the record before any adapter trains.

Acknowledgements. Both catalogues are the author's own. The research was carried out independently.

References

[1] Juhan Bae, Wu Lin, Jonathan Lorraine, and Roger Grosse. Training data attribution via approximate unrolling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2405.12186.

[2] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. *Theoretical Computer Science*, 312(1):3–15, 2004.

[3] Woosung Choi, Junghyun Koo, Kin Wai Cheuk, Joan Serrà, Marco A. Martínez-Ramírez, Yukara Ikemiya, Naoki Murata, Yuhta Takida, Wei-Hsiang Liao, and Yuki Mitsufuji. Large-scale training data attribution for music generative models via unlearning. In *NeurIPS 2025 Creative AI Track*, 2025. arXiv:2506.18312.

[4] Junwei Deng, Ting-Wei Li, Shiyuan Zhang, Shixuan Liu, Yijun Pan, Hao Huang, Xinhe Wang, Pingbang Hu, Xingjian Zhang, and Jiaqi W. Ma. dattri: A library for efficient data attribution. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2410.04555.

[5] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2008.03703.

[6] Kristian Georgiev, Joshua Vendrow, Hadi Salman, Sung Min Park, and Aleksander Madry. The journey, not the destination: How data guides diffusion models. In *ICML 2023 Workshop on Challenges in Deployable Generative AI*, 2023. arXiv:2312.06205.

[7] Amirata Ghorbani and James Zou. Data Shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, 2019. arXiv:1904.02868.

[8] Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo. ACE-Step: A step towards music generation foundation model. *arXiv preprint arXiv:2506.00045*, 2025.

[9] 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.

[10] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2106.09685.

[11] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Predicting predictions from training data. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022. arXiv:2202.00622.

[12] William B. Johnson and Joram Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. In *Conference in Modern Analysis and Probability*, volume 26 of *Contemporary Mathematics*. American Mathematical Society, 1984.

[13] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017. arXiv:1703.04730.

[14] Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs and diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2310.00902.

[15] 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. arXiv:2210.02747.

[16] 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)*, 2023. arXiv:2209.03003.

[17] Bruno Mlodozeniec, Runa Eschenhagen, Juhan Bae, Alexander Immer, David Krueger, and Richard Turner. Influence functions for scalable data attribution in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.13850.

[18] Fabio Morreale, Wiebke Hutiri, Joan Serrà, Alice Xiang, and Yuki Mitsufuji. Attribution-by-design: Ensuring inference-time provenance in generative music systems. *arXiv preprint arXiv:2510.08062*, 2025.

[19] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. TRAK: Attributing model behavior at scale. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023. arXiv:2303.14186.

[20] Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracing gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. arXiv:2002.08484.

[21] Joan Serrà, Dipam Goswami, Fabio Morreale, Wei-Hsiang Liao, and Yuki Mitsufuji. Training data attribution in diffusion models via mirrored unlearning and noise-consistent skew. *arXiv preprint arXiv:2605.17938*, 2026.

[22] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[23] Sheng-Yu Wang, Alexei A. Efros, Jun-Yan Zhu, and Richard Zhang. Evaluating data attribution for text-to-image models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. arXiv:2306.09345.

[24] Tong Xie, Haoyu Li, Andrew Bai, and Cho-Jui Hsieh. Data attribution for diffusion models: Timestep-induced bias in influence estimation. *Transactions on Machine Learning Research*, 2024. arXiv:2401.09031.

[25] Shih-Ying Yeh, Yu-Guan Hsieh, Zhidong Gao, Bernard B. W. Yang, Giyeong Oh, and Yanmin Gong. Navigating text-to-image customization: From LyCORIS fine-tuning to model evaluation. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.14859.

[26] Luyang Zhang, Xirui Jiang, Junwei Deng, Beibei Li, Jiaqi W. Ma, and Chris Donahue. What's a credit worth? a market framework for attribution-aware compensation in generative music. *arXiv preprint arXiv:2607.00641*, 2026.

[27] Xiaosen Zheng, Tianyu Pang, Chao Du, Jing Jiang, and Min Lin. Intriguing properties of data attribution on diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2311.00500.

A Ordering on the varying-size subsets (exploratory development path)

Before the fixed half-size subsets of Section 7 existed, ordering was read on the fifteen calibration subsets, which vary in size (56 to 330 pairs). Because held-out loss falls with training-set size, raw LDS on them measures size alone: the size-only null scores $+0.944$, identical to cosine. LDS was therefore reported with subset size partialled out (rank residuals), a correction made after observing the confound and disclosed as such; under it the null scores exactly zero. On the three held-out query tracks, which served to choose λ, the velocity-norm kernel gave $+0.574$ at $\lambda = 10\times$ with cosine and null at 0.000; the loss-gradient kernel gave $+0.709$ at its best λ but its own cosine $+0.600$, so on loss features the kernel added little and on velocity-norm features it added everything. Three queries cannot separate 0.57 from 0.71.

The pre-registered confirmation on twenty generated outputs (twelve held-out captions, four out-of-catalogue, four adversarial), bar 0.25 above null and cosine, passed: kernel $+0.316$ ($+0.15$ to $+0.45$), cosine -0.137; 19% of resamples below the bar, so twenty outputs established direction and approximate size. A fresh sample of 28 outputs gave $+0.391$ ($+0.32$ to $+0.46$); pooled over 48 outputs and 16 captions, $+0.360$ ($+0.267$ to $+0.438$). The loss-gradient target, pre-registered on the same outputs with the rule that its kernel must beat its cosine, was indeterminate on the first twenty (kernel $+0.300$, cosine $+0.294$, lower bound exactly 0.250) and failed on the 28 fresh (kernel $+0.273$, cosine $+0.285$); over 48, $+0.284$ ($+0.21$ to $+0.36$) against velocity-norm's $+0.360$, paired difference $+0.076$ (-0.026 to $+0.172$). The held-out-track ordering, where the loss kernel scored highest, did not transfer to generated outputs, in the direction of Zheng et al. [27]'s finding. An exploratory stratified set of 80 outputs over five prompt kinds (held-out, in-catalogue, out-of-catalogue, adversarial, lyric-heavy) gave $+0.380$, $+0.512$, $+0.439$, $+0.456$, $+0.370$ by kind and $+0.431$ ($+0.383$ to $+0.478$) overall, cosine -0.046. In every sample $\lambda = 1\times$ scored above the pre-registered $10\times$; the pre-registered figure is the one reported.

Two seeds of one prompt draw on the catalogue differently: mean share change 0.102, worst 0.191, top track kept on 14 of 40, against 0.013 between outputs on the same seed with different prompts. Both seeds' attributions predict retraining equally well (LDS $+0.417$ and $+0.446$). A new seed is a new output; the protocol had listed a second seed among changes that leave the music the same and removed it after this measurement.

B The float16 precision fault and withdrawn results

The measurements marked float16 used a research loader that requested a compute dtype without passing the requested precision; on the engine revision in use the default on Apple silicon is float16, while the adapters had been trained in bfloat16. Renders were unaffected. Float16 overflowed on four training pairs and introduced a per-pair loss offset of 0.02 to 0.045 below the trainer's logged values. The loader was fixed and the affected measurements re-taken in bfloat16 except where marked; the re-measured evaluator reproduces the trainer's logged held-out losses exactly (maximum difference 0 over fifteen runs and three queries; coefficient rank agreement 1.000).

Withdrawn: gradient cosine ranks influence backwards. In float16, plain cosine of loss-gradient features (count-sketch to $k = 2048$, L2-normalised, mean over pairs) against the retraining coefficients gave per-query Spearman ρ of -0.31, -0.46 and -0.23 at the final checkpoint and a similar sign at epoch 10 (Table 5), stable under centring, and had been read as evidence that single-model cosine attribution was worse than uninformative. Rebuilt in bfloat16 at $k = 8192$ the same construction gives $\rho = +0.154$ (float16, same construction: -0.35) and size-controlled LDS $+0.600$ (float16: -0.016). The sign was a precision artefact. A plausible mechanism, untested, is that loss gradients are small and float16 without loss scaling loses them to underflow. The $k = 2048$ index has not been rebuilt in bfloat16; the withdrawal rests on the $k = 8192$ index measured in both precisions.

Table 5: Withdrawn float16 results: mean per-query ρ between cosine attribution and retraining coefficients, $n = 36$ (query, track) pairs, twelve post-hoc scoring variants. Pre-set support bar $\rho \geq +0.5$.

Scoring variant (float16, $k = 2048$)	epoch 10	epoch 20
plain cosine, mean over pairs	-0.31	-0.34
mean-centred rows and query	-0.24	-0.31
top-1 principal component removed	-0.39	-0.27
top-3 principal components removed	$+0.33$	$+0.30$
plain, max over pairs	$+0.26$	$+0.10$
mean-centred, top-3 pairs	$+0.31$	$+0.19$
3 PCs removed, top-3 pairs	$+0.28$	$+0.48$
3 PCs removed, max over pairs	$+0.34$	$+0.26$

The best float16 number ($+0.48$) was one of twelve post-hoc variants at one checkpoint and did not repeat at the other; it was never treated as evidence. In float16 the first held-out pre-test gave the velocity-norm kernel $+0.513$ and the loss kernel $+0.345$, the ordering D-TRAK reports on images; in bfloat16 that ordering reversed on held-out tracks and was restored on generated outputs (Appendix A). The float16 confirmation on twenty outputs gave $+0.328$ ($+0.16$ to $+0.47$) against bfloat16's $+0.316$.

Reproduction, and a single known answer. Re-scoring each subset run's own held-out pairs under its own exported adapter gives coefficient rank agreement with the trainer's logged fit of 0.958, 0.993 and 0.993 in float16 (pass at 0.95) and exactly 1.000 in bfloat16. One training track attributed with its own caption by the retraining regression (float16, not re-measured) gave top share to itself (0.38; next 0.20, 0.17), $r^2 = 0.896$, bootstrap rank agreement 0.72 against a stability bar of 0.8; a single case kept as a demonstration, superseded by T5.

C Worked example of the rights arithmetic

Suppose an output's measured shares are Track A 40%, Track B 35%, Track C 25%, and the records say Track A is Writer 1 50% and Writer 2 50%, Track B is Writer 1 100%, Track C is Writer 3 100%. The output's rights are Writer 1: $0.40 \times 0.50 + 0.35 = 55\%$; Writer 2: 20%; Writer 3: 25%. Publishers follow from their splits the same way. Only the first column is measured.

D Reproducibility

Every measured number is in a JSON result file named by checkpoint and scoring variant, committed under the research results directory of the source repository with the plan file and records beside it; every confirmatory test has a pre-registration record whose content hash and commit the result file carries, and

every result file is stamped with its record's standing. Seeds, plan hash, adapter rank, pair cap and bucket count are pinned in the plan file and each run's manifest; the projection is seeded and fingerprinted; query vectors are cached beside each index.

Records named in this paper: subset plan `ce883ca91f89b691`; payment test `6caa44b1bf9a79e1`; MUCS `7c711d86fc7359a6`; served path `d4e466198937f511`; second catalogue `c325b480525b3770` and `2b81fd6ccca75e9c`, and under the run-health rule `f519c708f26836dd` and `638fed289c2d9095`; known-answer third readout `2eb42b821cbf8d4a`.

Machines. Apple M4 Max, 64 GB, MPS, for the first 70 subset adapters, the first catalogue's output renders, and everything through the two-ground-truth test; a Threadripper workstation with an RTX PRO 6000 (96 GB, CUDA, bfloat16) for the 220-adapter ground truth, the second catalogue's 165 adapters, the payment tests' scoring, MUCS and the served path; an Apple M2 Max 32 GB laptop for the timings in Section 13. The workstation trains at 0.24 s per pair-epoch against the Studio's 2.05.

Precision. Appendix B's tables, the single known answer, the first held-out pre-test and the first confirmation table on twenty outputs were computed in float16; every other figure, and every figure in the abstract, in bfloat16. The cost figures and the memory, timing and caption measurements of Section 13 are recorded timings, not result files.

Entry points. Calibration planning, training and fitting (`gradient_calibration.py`); retraining-based attribution (`datamodel_attribution.py`); kernel pre-test, output scoring and invariance (`dtrak_pretest.py`, `dtrak_outputs.py`, `dtrak_invariance.py`); the per-output designs (`leave_one_out.py`, `ground_truth_compare.py`, `payment_test_v2.py`); MUCS (`mucs_outputs.py`); the served path (`served_outputs.py`); device comparison and cross-device agreement (`device_comparison.py`, `cross_device_agreement.py`). Code and result files are to be released with the final version of this paper.