HYPERBOLIC HIERARCHICAL MUSIC ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Music production commonly works with separate components such as drum parts, bass lines, stems and submixes. As generative models are developed to produce such components, their training data may also include individual stems or mixtures of stems rather than only complete recordings. This makes provenance a structured problem: can exposure to a component during adaptation be detected, and can that effect be localised to the component rather than extending to the song it belongs to? This distinction matters for auditing generative music systems, including applications in content identification and copyright analysis. Because stems, submixes, and full recordings are related by containment, provenance also has a hierarchical structure, motivating representations that can capture these relations explicitly. We first study training-data membership under controlled partial exposure. A latent-diffusion text-to-music model is fine-tuned with a low-rank adapter on selected stems and mixtures, and we compare trained mixtures, untrained mixtures sharing constituent stems with trained mixtures, and clips containing no exposed stems. This allows us to examine whether exposure can be detected at the component level rather than only at the level of the complete song. We then ask whether the structure among these musical parts is better represented in hyperbolic than Euclidean space. We compare hyperbolic and Euclidean projection heads of identical size over a frozen audio encoder, trained on known stem-containment relations and evaluated on unseen songs. To the best of our knowledge, this is one of the first studies to jointly examine component-level training exposure and hyperbolic representations of the hierarchical relation among stems, submixes, and complete recordings. Together, these experiments provide a controlled framework for studying partial training exposure in structured data, where the relevant unit of provenance may be a component, a composition of components, or a larger hierarchy containing them.

1 INTRODUCTION

Music generation can operate on individual components of a recording: latent diffusion models have been trained to generate bass stems conditioned on an accompanying mixture (Pasini et al., 2024). When a generator is fine-tuned on stems or submixes, training exposure cannot always be described at the level of a whole recording. A model may encounter a bass–drums mixture without encountering either the bass alone or the full mix. Existing work examines training exposure in audio generators (Proboszcz et al., 2025) and retrieves similar training material for generated music (Barnett et al., 2024). For creators and catalogue owners auditing the use of their recordings, a signal associated with a song therefore leaves open which of its parts entered training.

Multitrack datasets such as MoisesDB provide isolated sources for the same recording (Pereira et al., 2023), making this ambiguity concrete. Rendering subsets of a song's stems yields mixtures related by containment: a bass stem is contained in both a bass–drums and a bass–vocals mixture, while neither of those mixtures contains the other. If a generator is trained on some of these mixtures, every rendered mixture also receives an exposure label. It may be a trained mixture, an untrained mixture that shares stems with a trained one, or an untrained mixture that shares no stems with any trained mixture from the same song segment. The same audio thus carries known containment relations and controlled exposure labels.

This setting supports two related questions about exposure and containment. The first is membership specificity: whether a probe responds only to the exact mixtures used for fine-tuning, or also to un-

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.

trained mixtures built from exposed stems. Exact membership and constituent exposure are different labels, and conventional membership evaluations for audio and music generators (Proboszcz et al., 2025; Liu et al., 2026b) do not necessarily distinguish them. The second is containment representation: whether an embedding encodes which mixture contains which, rather than only how much content two mixtures share. Acoustic similarity is symmetric, whereas order embeddings (Vendrov et al., 2016) and entailment cones (Ganea et al., 2018) model direction explicitly, and hyperbolic representations have been effective on hierarchical data (Nickel & Kiela, 2017). The two questions are linked: containment describes which material candidates share, while membership measurements ask whether exposure to that material changes the model's response.

Within this shared experimental setting, we make two contributions:

- **Component-level training-exposure evaluation.** We distinguish exact-mixture membership from constituent exposure using candidate-matched denoising scores. We find loss differences consistent with constituent exposure in mixtures that were not themselves used for fine-tuning.
- **Containment representation comparison.** We compare equally sized hyperbolic and Euclidean heads over frozen audio features. In the selected configuration, hyperbolic heads have higher validation-calibrated direction accuracy, while both geometries improve mean count-matched superset retrieval over their corresponding untrained references.

An exploratory reconstruction-based evaluation finds no consistent improvement in exposure discrimination from the containment representations in the evaluated settings (§5.3). The reported experiments do not attribute individual generated outputs to training examples.

2 RELATED WORK

Membership inference in generative audio. Membership inference asks whether a candidate was included in a model's training data. Diffusion-model attacks use denoising and reconstruction behaviour to assess membership: SecMI compares errors at individual diffusion steps (Duan et al., 2023), while PIA uses proximal noise initialisation (Kong et al., 2024). Their effectiveness also depends on the evaluation: differences between member and non-member populations can provide apparent membership signals (Dubiński et al., 2024).

Proboszcz et al. (2025) report weak individual-example membership inference for AudioGen, FIGARO and AudioLDM2, but a stronger signal for TANGO. Their dataset-inference procedure combines evidence across examples to detect training-set exposure. Music-specific approaches include a white-box probe based on the perturbation budget needed to degrade reconstruction (Liu et al., 2026b) and a black-box auditor trained on candidate audio and caption-conditioned generations from shadow models (Liu et al., 2026a). The latter learns a membership classifier, so its results do not establish the performance of a fixed audio-similarity score. We distinguish exact-mixture membership from constituent exposure by comparing denoising scores for the same candidate across adapters, with the pretrained model as a reference.

Memorisation, attribution and audio similarity. MusicLM evaluates memorisation through exact and approximate matches between generated semantic-token continuations and training sequences (Agostinelli, 2023). For audio latent diffusion models, Bralios et al. (2024) compare generated audio with training examples and show that replication detection depends on the similarity measure. Such analyses examine reproduction of training material, which a loss-based membership signal does not by itself demonstrate.

Training-data attribution asks which training examples contributed to a particular output. Barnett et al. (2024) compare CLAP (Wu et al., 2023) and CLMR embeddings for retrieving similar training material for a music generator, evaluating their matches through a listening study and their sensitivity to audio transformations. Choi et al. (2026) use unlearning to estimate attribution in a music diffusion model and compare the resulting attributions with similarity-based retrieval. These approaches relate generated audio to training material, whereas our denoising probe evaluates candidate exposure during adaptation. For mixtures constructed from overlapping stem subsets, we also ask whether representations capture containment rather than only shared content.

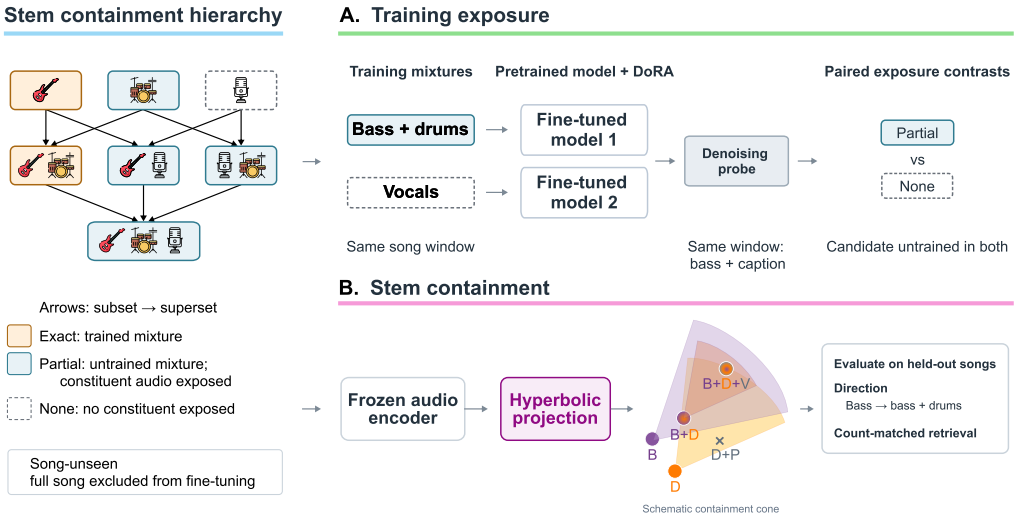

Figure 1: Overview of our methods for studying training exposure and stem containment. **Left:** containment among bass (B), drums (D) and vocals (V), with arrows from subsets to supersets. Exposure labels illustrate a training pool containing bass and bass–drums, different from the example in A. **A:** the same bass clip and caption are scored under adapters trained on bass–drums or vocals from that window, giving a partial-versus-none contrast. The candidate is untrained in both. **B:** hyperbolic and Euclidean projections of frozen audio features are evaluated on held-out songs using direction accuracy and count-matched superset retrieval. The cone sketch is illustrative, not a measured embedding; its colours distinguish stems rather than exposure classes.

Hyperbolic representation learning. Hyperbolic geometry permits low-distortion embeddings of trees (Sarkar, 2011). Poincaré and Lorentz embeddings use this geometry to learn continuous representations of hierarchies (Nickel & Kiela, 2017; 2018). Distance alone is symmetric, however, and does not specify the direction of a hierarchical relation. Order embeddings (Vendrov et al., 2016) and entailment cones (Ganea et al., 2018) model such direction explicitly. Entailment cones represent a directed relation through inclusion in a region rooted at each embedding; constructions exist in both hyperbolic and Euclidean space.

In vision–language learning, MERU uses hyperbolic contrastive learning and entailment constraints to represent relations between textual concepts and images (Desai et al., 2023). Compositional entailment learning extends this setting to images, object regions and their descriptions (Pal et al., 2025). HySAC combines contrastive and entailment objectives to organise safe and unsafe image–text representations, including a Euclidean cone comparison (Poppi et al., 2025).

Music and audio applications include genre taxonomies inferred from metadata (Astefanoaei & Collignon, 2020), hierarchical instrument-timbre representations (Nakashima et al., 2022), and source separation (Petermann et al., 2023). We study inclusion relations between rendered stem subsets. Because these form a multi-parent partial order, tree-embedding advantages cannot simply be assumed. Building on HySAC's contrastive and entailment-cone formulation, our containment method compares hyperbolic and Euclidean representations of this structure.

3 Method

We separately evaluate training exposure and stem containment on the same rendered subsets (Figure 1). Exposure is measured through candidate-matched denoising scores across adapters; containment is measured through direction accuracy and count-matched superset retrieval. An exploratory crossover compares raw CLAP with the learned heads using reconstruction gain: the decrease in source–generation embedding distance from the pretrained to the fine-tuned generator (App. E).

3.1 TRAINING EXPOSURE

Candidates and component exposure. A candidate c is a rendered subset of stems from one window of a recording. Relative to an adapter's training pool M_{train}, we use four exposure labels:

- **Exact:** $c \in M_{\text{train}}$.
- **Partial:** $c \notin M_{\text{train}}$, but at least one of its stems appears in a trained mixture of the same window.
- **None:** The candidate's song is represented in the adapter's training set, but none of the candidate's stems appears in a trained mixture of the same window.
- **Song-unseen:** The candidate's song has no mixtures in the adapter's training set.

Only exact candidates are members. Classifying a partial candidate as a member is therefore a false positive, even when its score reflects constituent exposure.

Training pools M_{train} are constructed per song (§4.1): one stem is withheld entirely from a tenth of the training songs, and a balanced quarter of the remaining rendered nodes is held out at each mixture size, leaving every size eligible for training. Most partial candidates contain only exposed stems; they include untrained full mixes.

The fine-tuning intervention. Let G_0 be a pretrained text-to-music generator and D the fine-tuning set. We obtain G_D by training a low-rank adapter while keeping the autoencoder and text encoder frozen. Both models are evaluated on the same candidate audio and caption, with matched noise draws for the denoising probe and generation settings for the generation-based scores (App. A).

We study membership in D, rather than membership in the pretrained model's unknown training data. For each candidate x, we evaluate the same score under G_0 and G_D, and report $\Delta s(x) = s_{G_D}(x) - s_{G_0}(x)$. Scores are oriented such that larger values indicate greater familiarity; for the denoising probe, $s_G(x) = -\ell_G(x)$, with the candidate caption included in the loss conditioning. The difference measures the change in the candidate's score following fine-tuning. It adjusts for differences already present in pretrained scores, but does not eliminate possible effects of pretraining overlap. A positive difference alone does not establish membership: fine-tuning can also change scores for unexposed candidates.

Denoising score. The primary exposure probe uses negative rectified-flow velocity MSE on the candidate latent z, conditioned on its caption. The noised input is $z_t = (1-t)z + t\epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and $t = 1$ denotes pure noise; the prediction target is $\epsilon - z$. Squared errors are averaged over latent channels and unpadded frames, without target-magnitude normalisation and with conditioning dropout disabled. We evaluate at fixed timesteps $t \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, averaging over timesteps and two fixed noise draws per timestep. The draws are identical across models. We subtract the pretrained score from the adapted score as defined above. This probe requires no generation; access requirements are given in Appendix B.

AUC and paired share. AUC is the probability that a randomly drawn positive score exceeds a randomly drawn negative, counting ties as half; it pools scores without using the candidate pairing. For a paired contrast between exposure roles A and B, define $\delta(x) = \Delta s_A(x) - \Delta s_B(x)$, averaging scores from multiple adapters with the same role for each candidate before differencing. Paired share is the fraction of candidates with $\delta(x) > 0$. Appendix B gives tie and baseline conventions.

Paired exposure contrasts. We compare each candidate under two adapters with different exposure labels: exact versus song-unseen tests example membership under a change of song-entry and generally artist exposure; exact versus none compares exact training with exposure to other material from the same song; partial versus none tests constituent exposure among non-members. Exact versus partial tests the increment associated with exact-mixture exposure. An increment beyond *full* constituent exposure requires restricting partial candidates to those whose every stem was exposed. Interpretation therefore depends on exposed-stem fraction.

Pairing holds the clip and its level fixed. Unpaired class comparisons do not, since renders are peak-limited rather than loudness-normalised (Table 17). The two-halves comparison and aggregation over artists' trained songs provide additional readouts (App. B).

3.2 CONTAINMENT REPRESENTATION

For stem families S in a recording window, the *stem-containment hierarchy* consists of non-empty subsets of S, ordered by inclusion. Each node represents audio rendered from one subset. A positive directed pair satisfies $u \subsetneq v$ within the same window, and a chain is a sequence of nested subsets directed from less to more content. Four families give 15 non-empty nodes and 24 maximal chains. The nodes form the non-empty part of the Boolean subset lattice; the empty node is not rendered.

We model this relation using entailment cones (Ganea et al., 2018), following the contrastive and cone construction from HySAC (Poppi et al., 2025). Equally sized linear audio and text projections act on frozen CLAP features. Hyperbolic features are mapped to the Lorentz model; Euclidean features use the identity map. Both heads use the same rendered chains and objective terms.

Writing u and v also for their embeddings, let $\theta_u(O, v)$ be the exterior angle between the outward radial direction at u from origin O and the geodesic from u to v. Let $a(u)$ be the implemented cone half-aperture. The angular region used for the containment test is

$$\mathcal{C}_s(u) = \{v \neq u : \theta_u(O, v) \leq s\, a(u)\}.$$

For a known relation $u \subsetneq v$, training encourages $v \in \mathcal{C}_1(u)$: the superset lies in the subset's cone. Reversing the pair tests a different cone, making this relation directed in both geometries. Appendix A specifies the exterior angles and half-aperture implementation.

Training combines contrastive losses between the caption and the smallest subset and between adjacent audio nodes, cone penalties on these same links, and a ranking hinge anchored on the full mix. App. A specifies the exact objective, including overlapping cone penalties and loss normalisation. A *sibling* hinge additionally compares a true superset v with a same-window node q satisfying $|q| = |v|$ and $u \not\subseteq q$:

$$\mathcal{L}_{\text{sib}}(u, v, q) = \max\{0,\, m + \theta_u(O, v) - \theta_u(O, q)\}.$$

This explicitly penalises ranking a count-matched non-superset ahead of the true superset. Its weight is 30 and its margin is $m = 0.3$; the other loss weights and the selection procedure are given in Appendix A.

Direction accuracy requires forward inclusion and reverse exclusion. Define $C_s(u, v) = \mathbf{1}[v \in \mathcal{C}_s(u)]$. The score is the mean of $C_s(u, v)[1 - C_s(v, u)]$ over positive pairs. Both-inside and neither-inside cases score zero; the statistic has no specified 0.5 chance baseline. Violation rate counts failed forward tests. After training, a global aperture scale s is selected on validation data with embeddings held fixed (§4.3). This calibrated angular test assesses pairwise direction; it does not establish a transitive containment relation over the embedding space.

For count-matched superset retrieval, the query is a subset and candidates are its true superset and same-window, same-size nodes not containing it. Ranking uses exterior angle at the query, with smaller angles first. Ties are misses and full-mix supersets are skipped. Retrieval does not depend on aperture scale. Sibling and cross-song distances provide supporting diagnostics (App. C).

4 EXPERIMENTAL SETUP

4.1 DATA, EXPOSURE AND PARTITIONS

MoisesDB-4 and MoisesDB-11 (Pereira et al., 2023) render the same 240 multitrack recordings under four-family and native-family taxonomies; the analysed manifests contain 47 distinct artist labels. Both use 710 windows; they contain 10,382 and 20,876 rendered nodes, respectively. Billboard-4 (Burgoyne et al., 2011) contains 1,380 recording entries with 518 artist labels and 1,087 distinct title–year pairs; release years parsed from filenames span 1957–1991. These recordings are separated into vocals, drums, bass and other accompaniment using Demucs, a neural music source-separation model (Rouard et al., 2023), giving 4,140 windows and 58,928 nodes. Four-family windows are rendered exhaustively except nodes with RMS below 0.0025; 3,504 Billboard windows contain all 15 nodes. The native taxonomy uses sampled chains. The full mix is the stem sum. Renders are peak-limited to 0.95 without loudness normalisation.

Artist metadata define five folds, stratified by genre on MoisesDB and decade on Billboard-4. For the heads, each run uses one test fold, the next validation fold, and the remaining three for training.

Three MoisesDB and 30 Billboard windows with fewer than three families are excluded from head training and evaluation. Adapter i trains on the three training folds; both its validation and test folds are song-unseen, so each song is excluded from two adapters. Adapters are not selected by validation: exports use fixed training fractions. Exposure readouts compare the same candidate across adapters under which it has different roles. Labels come from the actual training pools per track and window; exact clips must be in the pool, and partial/none clips must not.

These partitions do not completely isolate artists or recordings. Four MoisesDB artists have spelling variants spanning folds (five tracks). Billboard repeats are usually grouped, but artist-name variants can place copies across folds. Labels are per entry, so audio absent under one ID may be trained under another. Song-exposure contrasts generally also change artist exposure, and baseline subtraction does not correct these split-leakage issues. Appendix A gives corpus, caption and sampling details.

4.2 Models and measurements

We adapt Stable Audio 3 Small (base) (Evans et al., 2026), a rectified-flow text-to-music generator, with weight-decomposed low-rank adaptation (DoRA) (Liu et al., 2024), while freezing its autoencoder and text encoder. The two settings are rank 16 at learning rate 10^{-4} and rank 64 at 10^{-3}, both trained for 20 epochs. They change rank and learning rate jointly. Checkpoints at 25, 50 and 100 % of steps measure training duration within each run; equal epochs imply different update counts across corpora. A 256-clip, 200-epoch positive control checks probe sensitivity under strong exposure. To characterise how adaptation changes generation, we examine Fréchet audio distance (FAD) and output diversity across conditioning strengths on fold 0 (App. B). For each strength, the implemented rule requires the FAD reduction from pretrained to adapted generations towards the adaptation corpus to exceed both the reduction towards the comparison corpus and a reference quantity: the mean of signed split-half differences over eight splits. The post hoc flag (†) marks adapters that fail this rule at every evaluated strength, including text-only. Because the signed differences can cancel, the reference quantity does not estimate variability. The flag is descriptive; it neither establishes generation quality nor determines inclusion in the membership analysis.

The frozen encoder is `laion/clap-htsat-unfused` (Wu et al., 2023), with 512-dimensional features before final ℓ_2 normalisation. The final heads use two 512×512 modality projections, depth-three audio chains and the sibling term of §3.2. They are trained with AdamW and evaluated at the checkpoint with lowest validation loss; the final run settings are given in Appendix A. Repaired Audio Flamingo 2 captions (Ghosh et al., 2025) condition the generator; the heads use templated stem-name captions. Denoising probes evaluate the first 10 s of each encoded node and the first 2,500 latents in directory order from each fold's training and held-out pools. Timesteps and noise averaging follow §3.1. The crossover evaluates both adapter settings and both stem-only head configurations, with and without the sibling term, on fold 0. All representations use identical candidates and generated audio. Generation uses 10 s outputs, 50 sampling steps and guidance 4.0; audio-conditioning strength is $\lambda = 1 - \sigma$, where σ is the noise fraction in latent initialisation. Source references represent the first 10 s used for conditioning, and generated clips are encoded at their native duration without padding (App. E). Appendix A gives further pipeline settings.

4.3 Calibration, selection and uncertainty

For each trained head, aperture scale is selected from $\{0.5, 0.75, 1, 1.25, 1.5, 2\}$ to maximise validation direction accuracy, with ties resolved toward the trained scale; untrained references use $s = 1$. We report direction at both calibrated and trained scales so that the effect of this decision boundary is visible separately from retrieval.

Base-configuration selection used fold-0 test containment margins and, for a later rejection decision, five-fold test results. The initial tuning was performed only for the hyperbolic head; the Euclidean head inherited its cone and objective settings without separate tuning. The sibling weight and margin were subsequently selected on fold-0 validation data for the hyperbolic head and applied to both geometries. The final comparison therefore depends on prior test-set inspection, which excluding fold 0 does not fully address. Appendix A documents the selection history and final settings.

Geometry tables report five-fold means and t-based 95 % intervals where available, with one seed (0) per fold for the final sibling configuration. The three-seed checks use fold 0 of all three corpora

Table 1: Paired denoising contrasts, five folds pooled: AUC / percent positive within-candidate differences. E = exact, P = partial, N = none, U = song-unseen. Partial includes incomplete constituent exposure. Adapters: R16 = rank 16, learning rate 10^{-4}; R64 = rank 64, learning rate 10^{-3} (rank and learning rate vary together). †Fails the implemented FAD rule at every evaluated conditioning strength on fold 0 (post hoc; App. B).

corpus	adapter	E–U	E–N	P–N	E–P
MoisesDB-11	R16	0.83 / 96	0.58 / 80	0.57 / 78	0.56 / 70
	R64	0.95 / 96	0.70 / 82	0.68 / 81	0.60 / 71
	n	13,072	147	80	5,871
MoisesDB-4	R16	0.73 / 94	0.58 / 77	0.55 / 74	0.53 / 67
	R64	0.93 / 96	0.70 / 84	0.66 / 82	0.59 / 69
	n	9,827	189	84	6,298
Billboard-4	R16	0.60 / 91	0.54 / 73	0.54 / 74	0.52 / 64
	R64†	0.70 / 85	0.57 / 66	0.56 / 63	0.54 / 59
	n	20,573	470	242	14,463

and the earlier configuration without the sibling term. Overlapping training folds and selection limit an independent-replicate interpretation. Membership AUC pools candidate scores; paired share uses within-candidate differences. Nodes share stem audio and windows share songs, so clip-level resampling understates uncertainty; the denoising membership tables therefore report no intervals. Crossover intervals are reported separately in Appendix E.

5 Results

5.1 Training exposure

Fine-tuning changes generation similarity for both exposed and unexposed candidates. Fold-0 diagnostics show movement toward the training corpus at different conditioning strengths, while the rank-64 Billboard run shows generation degradation (Appendix B). These descriptive diagnostics motivate model-relative scores but do not themselves establish membership.

Exact-versus-song-unseen AUC reaches 0.83 and 0.73 at rank 16 on MoisesDB-11 and MoisesDB-4, respectively, rising to 0.95 and 0.93 under the rank-64, higher-learning-rate setting (Table 1). Billboard-4 reaches 0.60 at rank 16. Paired shares are higher than pooled AUC because the latter also reflects variation across candidates. This contrast does not isolate song-specific from artist-associated effects. Checkpoint AUCs increase numerically with training in every setting (Table 6); the final Billboard rank-64 checkpoint shows generation degradation, and the implemented FAD rule does not establish comparable generation quality across checkpoints.

The song-unseen comparison leaves open how specific the signal is to exposed material within a song. Exact-versus-none comparisons address this by holding the clip fixed while both fine-tuned models have seen its song. On MoisesDB the AUC is 0.58 at rank 16 and 0.70 at rank 64, with 77–84 % positive paired differences. Billboard rank 16 gives 0.54 and 73 %. The contrast is estimated from a much smaller candidate set than the song-level comparison, largely single stems from withheld-stem tracks. Pairing addresses candidate composition, but the adapters still differ in their whole training sets.

Among non-members, partial-versus-none AUC is 0.55–0.68 with paired shares of 74–82 % on MoisesDB, and 0.54 with 74 % on Billboard at rank 16. These paired contrasts provide evidence of sensitivity to constituent exposure in mixtures that were not themselves used for training. The small none groups and dependencies between candidates limit the precision of this result.

Given this constituent-exposure signal, exact-versus-partial comparisons ask what exact training adds. The increment over observed partial exposure is smaller: AUC is 0.52–0.60 and paired share 64–71 % across the MoisesDB settings and Billboard rank 16. This comparison includes candidates with only some stems exposed; an increment beyond *full* constituent exposure requires the all-stems-exposed restriction. None-versus-song-unseen AUC is 0.529–0.635 over the same settings (Table 3,

Table 2: Containment on songs excluded from head training, with the sibling term (weight 30, margin 0.3): five-fold means and 95 % t-intervals where available. Direction is evaluated at the trained scale ($s = 1$) and a validation-selected aperture scale, with embeddings fixed. Untrained references use $s = 1$ only. Superset retrieval ranks exterior angles and is independent of aperture scale. The base-configuration selection used test results (§4.3); the sibling setting used fold-0 validation data.

corpus	representation	direction ($s = 1$)	direction (calibrated)	superset first
MoisesDB-11	hyperbolic	0.18 ± 0.04	0.66 ± 0.05	0.71 ± 0.06
	Euclidean	0.22 ± 0.04	0.63 ± 0.03	0.69 ± 0.06
	untrained, Lorentz ($s = 1$)	0.02	—	0.70 ±0.05
	untrained, Euclidean ($s = 1$)	0.04	—	0.61 ±0.05
MoisesDB-4	hyperbolic	0.22 ± 0.07	0.76 ± 0.07	0.90 ± 0.03
	Euclidean	0.24 ± 0.06	0.71 ± 0.06	0.90 ± 0.03
	untrained, Lorentz ($s = 1$)	0.02	—	0.86 ±0.01
	untrained, Euclidean ($s = 1$)	0.03	—	0.77 ±0.03
Billboard-4	hyperbolic	0.32 ± 0.02	0.85 ± 0.02	0.91 ± 0.01
	Euclidean	0.33 ± 0.02	0.78 ± 0.01	0.90 ± 0.01
	untrained, Lorentz ($s = 1$)	0.03	—	0.89 ±0.00
	untrained, Euclidean ($s = 1$)	0.04	—	0.81 ±0.01

App. B), consistent with sensitivity to other material from the same song or artist. The degraded Billboard rank-64 run is reported separately in that table. Entry duplicates further limit localisation on Billboard-4 and unpaired class AUCs fail to order exposure classes (Table 4, App. B).

The Billboard two-halves swap provides a complementary paired comparison (App. B). At rank 16, baseline adjustment changes pooled AUC from 0.50 to 0.62 while the paired share remains 87 %. Aggregating the scores over each artist's trained songs gives AUC 0.75. This measures pooling of evidence over trained works, not generalisation to untrained songs by a seen artist.

5.2 Containment representation

With the sibling term, hyperbolic heads have higher validation-calibrated direction accuracy in all fifteen corpus–fold comparisons (Table 2). The paired gaps are 0.04 ± 0.02, 0.05 ± 0.02 and 0.07 ± 0.01 on MoisesDB-11, MoisesDB-4 and Billboard-4, respectively. At the trained aperture scale, Euclidean direction accuracy is instead higher by 0.04, 0.02 and 0.01. The direction comparison therefore depends on aperture calibration.

Superset retrieval does not depend on this aperture decision. The hyperbolic head ranks the true superset first in 0.71, 0.90 and 0.91 of cases, compared with 0.69, 0.90 and 0.90 for Euclidean. Both trained heads improve mean retrieval over their corresponding untrained references. The untrained Lorentz reference already outperforms the untrained Euclidean reference, so the geometry difference is not wholly attributable to head training. The earlier configuration without the sibling term reduced retrieval in both geometries; its results and fold-exclusion check are retained in Appendix C.

Supporting loudness controls, transfer tests and ablations use the earlier configuration without the sibling term (App. C). They do not establish robustness of the final configuration or exclude acoustic level cues. Exploratory joint-training containment results are reported in Appendix D.

5.3 Containment for exposure scoring

At $\lambda = 0.4$ on fold 0, reconstruction-gain AUCs for exact versus song-unseen candidates range from 0.43 to 0.55 across the three corpora, two adapter settings, raw CLAP and both head configurations (App. E). Hyperbolic and Euclidean AUCs differ by at most 0.02 within each setting, and no representation consistently performs best. These results provide no clear evidence that containment representations improve reconstruction-based exposure discrimination under the evaluated conditions. Using radial and angular readouts does not change this conclusion. Angular reconstruction-gain AUC exceeds full-distance AUC in several settings and reaches its highest reported value, 0.58, for the hyperbolic headwithout the sibling term on rank-64 MoisesDB-11 (Table 21).

6 DISCUSSION AND LIMITATIONS

Training exposure. The constituent-exposure result changes how a membership score should be read. A candidate can respond to adaptation without being a training example because its stems appeared inside other mixtures. Such a response is informative about constituent exposure, but classifying that candidate as an exact member would conflate two different properties. Exact-versus-partial separation suggests an additional increment over the observed partial exposure; because that class includes incomplete exposure, it does not yet isolate an increment beyond full constituent exposure. The distinction is therefore between measured forms of exposure, not between established memorisation mechanisms. Loss separation also does not demonstrate copying.

Containment representation. With the sibling term, both geometries improve mean retrieval over their corresponding untrained references, while hyperbolic heads retain a smaller advantage in validation-calibrated direction accuracy. This advantage is conditional on aperture calibration and the selected objective; both geometries use directed cone tests. Configuration selection used test results and did not tune the Euclidean head independently, limiting the generality of the comparison. The earlier loudness and count controls qualify radius interpretation but neither exclude acoustic shortcuts nor establish robustness of the final sibling configuration.

Containment for exposure scoring. Containment describes which stems a mixture contains; exposure depends on which mixtures were used to train an adapter. The crossover provides no clear evidence that the learned containment representations improve reconstruction-based exposure discrimination. Limited to the evaluated fold, candidates, and generation settings, this finding does not establish that exposure information is absent from audio representations or specific to denoising loss.

Scope and access. The exposure evaluation covers one generator family adapted under two joint rank/learning-rate settings; it does not evaluate full fine-tuning or vary rank and learning rate independently. Membership labels refer only to adaptation data, and unknown pretraining overlap may affect the measured responses. Artist grouping couples song and artist exposure, while spelling variants and repeated chart entries limit partition isolation. The containment evaluation uses two taxonomies of the same small multitrack collection and one larger collection of separated stems. Cross-corpus transfer does not fully isolate separation artifacts, and containment defined by rendered stem inclusion captures only one aspect of musical hierarchy.

Uncertainty estimates are limited by dependencies between observations: geometry results use five overlapping cross-validation runs, while membership candidates share audio, songs and adapters. All exposure probes require candidate audio and a caption; model-relative scores additionally require the pretrained reference. These access requirements and evaluation limitations constrain the study's implications for practical auditing: the experiments do not validate a deployment-ready auditor.

7 CONCLUSION

In this paper, we studied training exposure and stem containment in music through controlled adaptation and hyperbolic representation learning. We performed three sets of experiments. First we evaluated loss-based exposure signals, finding that untrained mixtures are affected when their constituent stems have been exposed during fine-tuning. This emphasizes the importance of distinguishing exact membership from constituent exposure. Second we compared hyperbolic and Euclidean representations of stem containment. With the sibling term and the reported configuration-selection procedure, hyperbolic heads achieve higher validation-calibrated direction accuracy than Euclidean heads in every evaluated corpus and fold. Both geometries improve mean count-matched retrieval over their corresponding untrained references. Third, we explored whether these containment representations improve reconstruction-based exposure discrimination, finding no clear improvement in the evaluated settings. In short an untrained mixture can show a training-exposure signal when it shares stems with trained material, but in our experiments, better representations of which mixtures contain which stems do not consistently improve exposure detection. Code and configurations are available at `withheld.for.doubleblind.review`

AI USE STATEMENT

The core research ideas—including the training-exposure and containment-representation investigations, the use of stem-containment hierarchies, and the construction of a larger dataset through source separation—originated with the authors. In this work, we used generative AI tools to assist with implementing and debugging experiments, obtain feedback on experimental design, suggest additional diagnostic tests, and support the interpretation and reporting of results. Claude Code assisted with experimental code and the implementation of dataset stratification.

Additionally, we used generative AI tools to suggest potentially missing literature, co-draft portions of the manuscript, and provide feedback on clarity, structure and flow. Figure 1 was created by the authors. Audio Flamingo 2 was used to generate audio captions, as described in the experimental setup.

The authors repeatedly reviewed and revised the manuscript, assessed AI suggestions concerning the implementation, and checked unfamiliar suggested references for their relevance and support for the associated claims. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

Our methods and experimental procedures are described in Sections 3 and 4. Additional data-processing details, operational settings, and configuration-selection procedures are provided in Appendix A. Upon publication, we will release the implementation through a public GitHub repository, including preprocessing and data-construction scripts, specifications for the separate software environments required by different model components, and text manifests documenting the data partitions and rendered stem combinations. To respect the interests of the recording rights holders and dataset contributors, our release will exclude source recordings, derived audio, and fine-tuned model weights; it will instead provide code and metadata for reconstructing the experiments. Reproducing the experiments will therefore require independently obtaining the source recordings under the applicable access terms and rerunning the training procedures.

REFERENCES

Andrea Agostinelli. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Maria Astefanoaei and Nicolas Collignon. Hyperbolic embeddings for music taxonomy. In *Proceedings of the 1st Workshop on NLP for Music and Audio (NLP4MusA)*, pp. 38–42, 2020.

Julia Barnett, Hugo Flores Garcia, and Bryan Pardo. Exploring musical roots: Applying audio embeddings to empower influence attribution for a generative music model. *arXiv preprint arXiv:2401.14542*, 2024.

Dimitrios Bralios, Gordon Wichern, François G Germain, Zexu Pan, Sameer Khurana, Chiori Hori, and Jonathan Le Roux. Generation or replication: Auscultating audio latent diffusion models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1156–1160. IEEE, 2024.

John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga. An expert ground truth set for audio chord recognition and music analysis. In *Proc. ISMIR*, pp. 633–638, 2011.

Woosung Choi, Junghyun Koo, Kin Wai Cheuk, Joan Serrà, Marco 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. *Advances in Neural Information Processing Systems*, 38, 2026.

Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pp. 7694–7731. PMLR, 2023.

Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *Proceedings of the 40th International Conference*

on Machine Learning (ICML), volume 202 of *Proceedings of Machine Learning Research*, pp. 8717–8730, 2023. URL https://proceedings.mlr.press/v202/duan23b.html.

Jan Dubiński, Antoni Kowalczuk, Stanisław Pawlak, Przemysław Rokita, Tomasz Trzciński, and Paweł Morawiecki. Towards more realistic membership inference attacks on large diffusion models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. doi: 10.1109/WACV57701.2024.00479. URL https://arxiv.org/abs/2306.12983. arXiv:2306.12983.

Zach Evans, Julian D. Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3, 2026.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*. PMLR, 2018.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Proc. ICML*, volume 267 of *PMLR*, pp. 19358–19405, 2025.

Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.18355. arXiv:2305.18355, 2023.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proc. ICML*, volume 235 of *PMLR*, pp. 32100–32121, 2024.

Yi Chen Liu, Jiawei Yu, Kexin Cao, Syed Irfan Ali Meerza, Trishika Movva, and Jian Liu. Auditing training data in generative music models via black-box membership inference, 2026a. URL https://arxiv.org/abs/2605.29202. ArtSec 2026 workshop; submitted 28 May 2026.

Yuxuan Liu, Peihong Zhang, Rui Sang, Zhixin Li, Yizhou Tan, Yiqiang Cai, and Shengchen Li. Membership inference attack against music diffusion models via generative manifold perturbation, 2026b. URL https://arxiv.org/abs/2602.01645. Submitted 2 February 2026.

Futa Nakashima, Tomohiko Nakamura, Norihiro Takamune, Satoru Fukayama, and Hiroshi Saruwatari. Hyperbolic timbre embedding for musical instrument sound synthesis based on variational autoencoders. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 735–742. IEEE, 2022.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *NeurIPS*, 2017.

Maximillian Nickel and Douwe Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *ICML*. PMLR, 2018.

Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Marco Pasini, Maarten Grachten, and Stefan Lattner. Bass accompaniment generation via latent diffusion. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. URL https://arxiv.org/abs/2402.01412.

Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl. MoisesDB: A dataset for source separation beyond 4-stems. In *Proc. ISMIR*, pp. 619–626, 2023.

Darius Petermann, Gordon Wichern, Aswin Subramanian, and Jonathan Le Roux. Hyperbolic audio source separation. In *ICASSP*. IEEE, 2023.

Tobia Poppi, Tejaswi Kasarla, Pascal Mettes, Lorenzo Baraldi, and Rita Cucchiara. Hyperbolic safety-aware vision-language models. In *CVPR*, 2025.

Jakub Proboszcz, Paweł Kochański, Karol Korszun, Donato Crisostomi, Giorgio Strano, Emanuele Rodolà, Kamil Deja, and Jan Dubiński. Membership and dataset inference attacks on large audio generative models, 2025. URL https://arxiv.org/abs/2512.09654. NeurIPS 2025 workshops; v1 December 2025; v2 February 2026.

Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *Proc. ICASSP*, 2023. doi: 10.1109/ICASSP49357.2023.10096956.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *GD*. Springer, 2011.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. Order-embeddings of images and language. In *International Conference on Learning Representations (ICLR)*, 2016. URL https://arxiv.org/abs/1511.06361.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proc. ICASSP*, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

APPENDIX

A ADDITIONAL IMPLEMENTATION DETAILS

Dataset definitions, exposure pools, model settings and evaluation criteria are given in §3–4. This section records additional implementation choices. Slash-separated corpus values below use MoisesDB-11/MoisesDB-4/Billboard-4 order.

Sampling and captions. MoisesDB recordings supply three fixed 30 s windows where available: 231 supply three, eight supply two, and one supplies one. Its folds contain 83/41/40/38/38 tracks; Billboard folds contain 276 entries each and use `htdemucs_ft` separation. Native-family rendering adds the nodes of eight sampled nested removal chains per window; head training draws up to eight chains per window once with a fixed seed. Generator prompts combine genre, the repaired AF2 caption and a known-instrumentation sentence. Billboard uses a constant genre placeholder; plural instrument mentions may survive repair, and identity cues were not checked. Exposure readouts reject labels inconsistent with actual per-window training pools. Split-isolation limitations are stated in §4.1.

Generator implementation. The identifier is `stable-audio-3-small-music-base`. Training uses pre-encoded 30 s latents, caption embeddings computed at each step, bf16 base weights, batch size 8, and AdamW with weight decay 0.01, betas $(0.9, 0.95)$ and constant learning rate. Ranks 16/64 train 10.7/41.6 M parameters; rank 16 covers 193 layers. Fold-0 full runs use 18,030/8,935/62,883 updates. Inference merges adapters at strength 1.

Head parameterisation. For stem-containment training and evaluation, CLAP encodes the central 10 s of each window. Crossover scoring applies these heads to source embeddings from the first 10 s, matching the generation-conditioning segment (App. E). Projection matrices are identity-initialised. Modality scales are learned and clamped at 1; recorded initial values for the earlier heads were 0.22–0.23 for audio and 0.25–0.26 for text. The learned logit scale starts at $1/0.07$ and is clamped to $[1/0.07, 100]$. The Lorentz head uses the origin exponential map with negative curvature $-c$, where $c > 0$ is initialised at 1 and learned in $[0.1, 10]$. The fitted range 0.83–0.89 describes the earlier heads without the sibling term. The Euclidean head's curvature parameter is fixed and unused.

In the formulas below, x and y denote Euclidean coordinates for the Euclidean head and spatial coordinates for Lorentz points; $\|\cdot\|$ is the Euclidean norm of those coordinates. With $\epsilon = 10^{-5}$ and $r_{\min} = 1$, the implemented half-aperture is

$$a(x) = \arcsin\left(\min\left\{1 - \epsilon, \frac{2r_{\min}}{\sqrt{c}\,\|x\|}\right\}\right).$$

For the Euclidean head, $c = 1$ is used in this aperture expression. Its exterior angle is

$$\theta_x(O, y) = \arccos \frac{\langle x, y - x \rangle}{\|x\| \, \|y - x\|}.$$

For Lorentz points, the time coordinates are $x_0 = \sqrt{1/c + \|x\|^2}$ and $y_0 = \sqrt{1/c + \|y\|^2}$, and the exterior angle is

$$\theta_x(O, y) = \arccos \frac{y_0 + \kappa x_0}{\|x\| \sqrt{\kappa^2 - 1}}, \qquad \kappa = c(\langle x, y \rangle - x_0 y_0).$$

The displayed expressions give the ideal geometric quantities for nonzero apex embeddings and distinct point pairs. In implementation, $\epsilon = 10^{-5}$ is added to each denominator in the aperture and exterior-angle formulas, $\kappa^2 - 1$ is floored at ϵ before taking its square root, and inverse-trigonometric arguments are clipped to $[-1 + \epsilon, 1 - \epsilon]$. At a zero apex or for coincident embeddings, the stabilised exterior-angle expression returns a numerical convention rather than a geometric angle. These stabilised quantities define the containment test in §3.2.

At the aperture clamp, $a(x) = \arcsin(1 - \epsilon) \approx 89.74°$. Scale $s = 1.25$ gives a boundary angle of approximately $112.18°$ at the clamp. This scale was selected on validation data for the earlier hyperbolic heads and for both final geometries in all five MoisesDB-11 folds. The scaled boundary exceeds the $\pi/2$ bound for the transitive angular-cone construction of Ganea et al. (2018). Moreover, clamping extends the aperture formula to radii outside its unclamped domain, and this extension need not preserve transitivity even at $s = 1$. We therefore interpret both trained-scale and calibrated cone tests as pairwise angular diagnostics, without claiming a transitive containment relation.

Configuration selection. We developed the head configuration through four successive rounds of configuration search, referred to below as stages. These stages describe the development history, not successive training phases of a single model. We report this history to distinguish the final configuration from the earlier recipes used in supporting experiments and to identify where test results informed configuration choices.

The test-informed selection procedure and final architecture are disclosed in §4.3. Stages one and two tuned the hyperbolic head; the Euclidean head inherits its settings. Stage one swept minimum radius $\{0.1, 0.3, 1\}$ ($\{0.3, 1\}$ on Billboard) and aperture $\{1, 2\}$; stage two compared nine objective arms. Stage three varied scale, widths 128/256/1024, depth four and direction convention. Stage three included a weight-2 chain term averaging the consecutive-link cone penalties; this term does not directly supervise non-adjacent pairs. The rejected fourth stage added transitive chain entailment, scale initialisation 0.44 and a raised clamp; that clamp never became active. The final heads use the configuration retained after stage three, augmented with the sibling hinge; they do not use the rejected stage-four modifications. The chain-term ablation uses that recipe, while the width sweep uses its own base recipe. Calibration ties are resolved toward the trained scale; five-fold intervals use $t_4 = 2.776$.

Final head training. The final sibling configuration comprises 30 runs: three corpora, two geometries and five folds, with seed 0. Heads are trained with AdamW using $\beta = (0.9, 0.98)$, weight decay 0.2, learning rate 10^{-4} and batch size 64 for 50 epochs with mixed precision. There is no learning-rate schedule, warm-up or gradient clipping; the early-stopping patience is 999, so it does not shorten these runs. Evaluation uses the checkpoint with lowest validation loss. The optimiser updates the parameterisation with ordinary AdamW, rather than a Riemannian optimiser.

Each training item comprises the templated caption t of the smallest subset and a chain $u_1 \subsetneq u_2 \subsetneq u_3$, where u_3 is the full mix. Define $\mathcal{P} = \{(t, u_1), (u_1, u_2), (u_2, u_3)\}$. For each link, $\mathcal{L}_{\text{con}}$ is symmetric in-batch cross-entropy with diagonal targets and logits $-\tau d$, where τ is the learned logit scale and d is geodesic distance in the corresponding geometry. The cone penalty is

$$h(x, y) = \max\{0, \theta_x(O, y) - a(x)\}.$$

The implemented objective is

$$\mathcal{L} = \frac{1}{39} \left[\sum_{(x,y) \in \mathcal{P}} \mathcal{L}_{\text{con}}(x, y) + \frac{5}{3} \sum_{(x,y) \in \mathcal{P}} h(x, y) + \mathcal{L}_{\text{rank}} + 30 \mathcal{L}_{\text{sib}} \right],$$

where the contrastive, cone and ranking terms are averaged over the batch, and the per-chain ranking penalty is

$$\max\{0, 0.1 + d(u_2, u_3) - d(u_1, u_3)\}.$$

The coefficient $5/3$ combines a unit-weight penalty on each link with a weight-2 term averaging the same three penalties. The latter adds no new pairwise constraints; neither cone term directly supervises non-adjacent pairs. The sibling hinge defined in §3.2 uses one randomly sampled count-matched non-superset for (u_1, u_2), with margin 0.3, averaged over chains that have such a candidate. The link into the full mix has no sibling. The denominator 39 is the sum of the original configured weights; all other terms have weight zero.

Minimum radius and the training aperture scale are both 1.0. Each modality scale α is automatically initialised to the reciprocal mean norm of its projected features, estimated from 256 randomly sampled training chains using full-mix audio and the corresponding captions. The projections have width 512. Training uses depth-three chains, up to eight per window, and 10 s clips. The encoder remains frozen. For MoisesDB-11, validation selects aperture scale 1.25 for both final geometries in all five folds.

Sibling-term selection. Weight and margin were selected on fold 0 using the hyperbolic head, evaluated on the validation split with aperture calibration on the training split. The search used weights $\{1, 3, 10, 30\}$ and margins $\{0.1, 0.3, 0.5\}$, but not every pair. The prespecified rule required learned curvature magnitude of at least 0.7 and direction accuracy no more than 0.02 below the head without the term on every corpus; among eligible settings, it selected the largest mean gain in true-superset-first retrieval. Weight 30 and margin 0.3 passed, with MoisesDB-11 direction accuracy 0.694 against a floor of 0.684 and a mean retrieval gain of 0.11. Weight 30 and margin 0.5 failed because curvature magnitude fell to 0.48 on Billboard-4. Both final geometries use the selected setting. This search does not remove the earlier test-based choices described in §4.3.

A separate within-song contrastive term with stem-overlap (Jaccard) soft targets was also evaluated and rejected. With distance-based similarity, curvature magnitude fell from 0.85 to 0.35, 99–100 % of cones saturated, and direction accuracy fell by 0.19. With angle-based similarity, no tested setting satisfied the direction criterion, and 17 of 36 settings triggered the training collapse check.

B SUPPLEMENTARY EXPOSURE EVIDENCE

Probe implementation and access. Denoising requires candidate encoding and access to the model's loss. Generation-based scores require sampling access and candidate embeddings. Model-relative scores require the pretrained reference; paired contrasts also require a comparison adapter.

Aggregation and counts. For the metrics in Section 3.1, paired-share ties count as failures. Under a null distribution symmetric about zero, $\Pr[\delta > 0] = (1 - \Pr[\delta = 0])/2$, which equals 0.5 when there are no ties. Pretrained subtraction cancels in paired differences but can change pooled AUC. Table 1 gives the primary paired candidate counts; Table 3 adds the none-versus-song-unseen contrast. Unpaired song-unseen counts are 31,410/20,166/43,022 clip-by-adapter observations, in MoisesDB-11/MoisesDB-4/Billboard-4 order. Tables 4 and 5 give class and timestep breakdowns.

None versus song-unseen. Table 3 reports the additional candidate-matched contrast using corrected training-pool labels. AUC pools the scores for the two roles; paired share is the percentage of candidates with $\Delta s_{\text{none}} - \Delta s_{\text{song-unseen}} > 0$, with ties counted as failures. The two roles change exposure to other material from the same song and generally its artist, so the contrast does not isolate those effects.

Training duration. Table 6 gives the checkpoint breakdown. Numerical AUC increases in every row, while rank and learning rate vary together. For text-only generation, the implemented FAD rule labelled the Billboard rank-64 checkpoint at 25 % as passing and those at 50 % and 100 % as failing (Table 8). These are recorded rule outcomes, not validated quality categories: the signed split-half reference does not measure variability. The final checkpoint also shows generation degradation, so the AUC trajectory does not establish improved membership detection under comparable generation quality.

Table 3: None-versus-song-unseen denoising contrast, five folds pooled, using corrected training-pool labels and scores averaged across adapters per clip. n is the number of matched candidates; paired share counts positive within-candidate score differences. R16 and R64 use learning rates 10^{-4} and 10^{-3}, respectively, so rank and learning rate vary jointly. No uncertainty intervals are reported. †Fails the implemented FAD rule at every evaluated conditioning strength on fold 0 (post hoc; App. B); these values are retained but excluded from the primary interpretation.

corpus	adapter	n	AUC	paired share (%)
MoisesDB-11	R16	171	0.571	69.6
	R64	171	0.604	68.4
MoisesDB-4	R16	202	0.590	74.3
	R64	202	0.635	71.8
Billboard-4	R16	509	0.529	64.8
	R64†	509	0.542	62.1

Table 4: Unpaired AUC against song-unseen candidates, grouped by candidate exposure role (§3.1); n counts clip-by-adapter observations. Classes are not compared directly. Partial counts with every stem exposed are 13,821/7,356/18,330. Unlike the paired main-table readout, adapters are not averaged per clip. †Fails the implemented FAD criteria at every evaluated conditioning strength on fold 0 (post hoc; App. B); these values are retained but excluded from the primary interpretation.

			AUC ↑ vs. unseen song	
corpus	candidate exposure	n	16, 10^{-4}	64, 10^{-3}
MoisesDB-11	exact	30,858	0.83	0.95
	partial	16,078	0.80	0.91
	none	179	0.53	0.78
MoisesDB-4	exact	21,198	0.72	0.92
	partial	8,846	0.70	0.88
	none	205	0.53	0.71
Billboard-4	exact	41,951	0.60	0.69†
	partial	22,045	0.60	0.67†
	none	534	0.40	0.56†

Two-halves swap. The fixed era-stratified Billboard training split is divided by artist metadata into 489/488 songs, 215 artists per half; era distributions differ by total variation 0.003. Each clip is a training example under its own half's adapter and an unseen song-entry and artist label under the other; $\delta(x) = s_{\text{own}}(x) - s_{\text{other}}(x)$, which subtracting the pretrained reference leaves unchanged. Metadata-based isolation has the limitations described in §4.1. Table 7 retains the rank-16 comparisons. Without the pretrained reference, pooled AUC is 0.50 at every timestep; paired share is unchanged at 87 %. The artist-aggregate interval [0.71,0.78] uses 1,000 resamples of 430 artists. No held-out songs by seen artists are evaluated. Rank-64 half-adapter results are not reported here because the corresponding FAD assessment was not performed; this absence does not establish that those adapters failed an assessment.

Generation diagnostics. Each fold-0 generation grid uses 1,000 sources (250 per render level), text-only and at conditioning strengths $0.1, 0.4, 0.6, 0.8, 0.9$, with a shared seed across strengths and models. Readouts retain one source per song-entry by path order (240 MoisesDB; 769 Billboard), selecting mix-minus-bass for 236 MoisesDB songs.

For each conditioning strength, the implemented rule requires the CLAP-feature FAD reduction toward training to exceed both a split-half reference quantity and the reduction toward the other corpus. The reference quantity is the mean of signed split-half FAD differences over eight splits. The signs can cancel, so this is not a variability estimate. Source-distance and pairwise-cosine diagnostics do not enter the rule. No adapter-level aggregation rule was specified before analysis. For reporting, † flags an adapter that fails at every evaluated conditioning strength, including text-only; this conven-

Table 5: Five-fold pooled denoising AUC for exact–song-unseen and exact–none contrasts, by timestep. The averaged column applies AUC to the timestep-averaged score. †Fails the implemented FAD rule at every evaluated conditioning strength on fold 0 (post hoc; App. B); these values are retained but excluded from the primary interpretation.

		exact–song-unseen AUC ↑, per timestep t					
corpus	rank, lr	0.1	0.3	0.5	0.7	0.9	averaged
MoisesDB-11	16, 10^{-4}	0.73	0.86	0.87	0.83	0.64	0.83
	64, 10^{-3}	0.77	0.92	0.97	0.96	0.68	0.95
MoisesDB-4	16, 10^{-4}	0.69	0.80	0.79	0.71	0.57	0.73
	64, 10^{-3}	0.72	0.91	0.97	0.92	0.60	0.93
Billboard-4	16, 10^{-4}	0.57	0.66	0.67	0.63	0.53	0.60
	64, $10^{-3\dagger}$	0.55	0.66	0.76	0.75	0.55	0.70
		exact–none AUC ↑, per timestep t					
MoisesDB-11	16, 10^{-4}	0.56	0.62	0.63	0.58	0.53	0.58
	64, 10^{-3}	0.58	0.69	0.77	0.70	0.54	0.70
MoisesDB-4	16, 10^{-4}	0.58	0.64	0.62	0.56	0.52	0.58
	64, 10^{-3}	0.57	0.73	0.80	0.71	0.53	0.70
Billboard-4	16, 10^{-4}	0.53	0.56	0.56	0.55	0.51	0.54
	64, $10^{-3\dagger}$	0.51	0.55	0.60	0.60	0.53	0.57

Table 6: Exact–song-unseen AUC at fixed fractions of training, five folds pooled. Rank and learning rate vary together. †Fails the implemented FAD rule at every evaluated conditioning strength on fold 0 (post hoc; App. B). Table 8 reports text-only rule outcomes, which do not certify comparable quality across checkpoints.

corpus	rank, lr	25 %	50 %	100 %
MoisesDB-11	16, 10^{-4}	0.68	0.74	0.83
	64, 10^{-3}	0.85	0.91	0.95
MoisesDB-4	16, 10^{-4}	0.62	0.67	0.73
	64, 10^{-3}	0.70	0.75	0.93
Billboard-4	16, 10^{-4}	0.54	0.57	0.60
	64, 10^{-3}	0.62	0.65	$0.70^{\dagger}$

tion is post hoc. Only the final Billboard rank-64 adapter is flagged among the six primary settings. Recorded outcomes are descriptive and do not establish validated quality categories.

Under the implemented rule, rank 16 passes at text-only and 0.1 on MoisesDB-11 and Billboard. MoisesDB-4 rank 16 fails at text-only and passes only at strength 0.4, so it is not flagged by the post hoc all-strengths rule. Rank 64 passes at text-only and 0.1 on both MoisesDB renders but fails throughout the Billboard grid. Table 8 records duration and seed checks; Table 9 reports the strong-exposure positive control. The final Billboard rank-64 run also shows increased FAD toward training and increased pairwise generation cosine.

C Supporting containment evaluations

Table 2 reports the final sibling configuration. The supporting experiments in this section predate that configuration; their results do not establish its seed, dimension or transfer robustness. Captions identify the relevant earlier recipe and its departures from the primary comparison.

Earlier five-fold comparison. Table 10 retains the earlier results without the sibling term, including the retrieval decreases. The unchanged untrained references are given in Table 2. These heads use the stage-three base configuration; the final comparison adds the sibling hinge. The configuration-selection limitations in §4.3 apply to both comparisons.

Table 7: Billboard two-halves swap: pooled training clips scored under their own and other adapter. AUC pools scores; paired share uses strictly positive within-clip differences. *Aggregation over 430 artists' trained works, not unseen-song generalisation.

		AUC ↑, per timestep t						$\Pr[\delta > 0]$ ↑
level	rank, lr	0.1	0.3	0.5	0.7	0.9	averaged	averaged
song	16, 10^{-4}	0.61	0.73	0.73	0.65	0.54	0.62	87 %
artist aggregate*	16, 10^{-4}	aggregate, no per-t split					0.75	99.8 %

Table 8: Fold-0 text-only diagnostics by training duration. FAD columns give reductions toward training and other-corpus pools. Cosine is mean pairwise generation cosine, with pretrained values 0.62/0.58/0.50 in corpus order; larger values mean more similar embeddings. The last column reports the implemented rule for text-only generation; it is not the post hoc adapter-level flag across strengths. Its signed split-half reference is not a variability band (App. B). ‡Separate 12,577-step runs: seed 42 unless marked seed 1. Equal epochs imply different update counts across corpora.

corpus	rank, lr	steps	ΔFAD tr.	ΔFAD ood	cosine	rule (text-only)
MoisesDB-11	16, 10^{-4}	4,508	+0.008	−0.168	0.49	pass
		9,015	+0.261	+0.176	0.58	pass
		18,030	+0.283	+0.234	0.57	pass
MoisesDB-4	16, 10^{-4}	2,234	+0.125	+0.179	0.66	fail
		4,468	+0.108	+0.123	0.66	fail
		8,935	−0.009	−0.074	0.67	fail
Billboard-4	16, 10^{-4}	15,721	+0.150	+0.060	0.44	pass
		31,442	+0.162	+0.092	0.43	pass
		62,883	+0.127	+0.060	0.41	pass
	64, 10^{-3}	12,577‡	−0.294	−0.191	0.52	fail
		12,577‡, seed 1	+0.081	+0.068	0.43	pass
		15,721	+0.091	+0.036	0.52	pass
		31,442	−0.006	−0.103	0.50	fail
		62,883	−0.908	−1.003	0.74	fail

For the earlier configuration without the sibling term, excluding fold 0 assesses sensitivity to the direct configuration-selection fold. Paired gaps are 0.18 ± 0.09, 0.22 ± 0.11 and 0.23 ± 0.03 in direction, and 0.07 ± 0.01, 0.08 ± 0.02 and 0.06 ± 0.01 in retrieval (95 % t_3 intervals, corpus order as in the table). All fifteen per-fold gaps are positive on both metrics. Selection qualifications remain those of §4.3. Direction gaps exceed observed three-seed ranges (Table 11).

Additional diagnostics. For the earlier heads without the sibling term, at trained apertures, hyperbolic cones miss 66–81 % of true supersets and admit 3–7 % of count-matched negatives; Euclidean rates are 46–63 % and 15–18 %. Validation selected scale 1.25 for every earlier hyperbolic head. Neither geometry has mutually contained positive pairs at the trained scale. A subset is closer to its superset than to a same-size non-superset in 0.88–0.97 of tests (0.85–0.95 untrained), and closer to its own song's mix than another song's in 0.89–0.97 (0.83–0.85 untrained). An earlier sibling-hinge trial with weight 1 and margin 0.1 changed direction by +0.001 to +0.004 and retrieval by +0.005 to +0.014 (paired retrieval intervals exclude zero), with retrieval still below the untrained reference. Those trial results are distinct from the final weight-30, margin-0.3 configuration in Table 2.

C.1 Stem count and loudness controls

Lattice distance counts the symmetric difference of stem sets. Partial correlations adjust distance associations for the sum and absolute difference of stem counts, and radius–count associations for loudness. Loudness uses ungated K-weighted mono energy, not full gated integrated loudness. Ta-

Table 9: Positive control: adapters trained for 200 epochs on 256 Billboard clips, evaluated against 256 held-out clips. Per-timestep AUC tests probe sensitivity under strong exposure; it does not demonstrate extraction.

	AUC ↑, seen against unseen clips, per timestep				
adapter, 256 clips, 200 epochs	0.1	0.3	0.5	0.7	0.9
rank 16, lr 10^{-4}	0.885	0.956	0.944	0.905	0.734
rank 64, lr 10^{-3}	0.867	0.986	0.990	0.989	0.888

Table 10: Earlier heads without the sibling term: five-fold means and reported 95 % t-intervals on songs excluded from head training. Column definitions match Table 2. These measurements describe the earlier configuration, not reruns of the final sibling configuration. The unchanged untrained references are reported in the main table.

corpus	representation	direction ($s = 1$)	direction (calibrated)	superset first
MoisesDB-11	hyperbolic	0.19 ± 0.07	0.67 ± 0.05	0.62 ± 0.04
	Euclidean	0.37 ± 0.08	0.48 ± 0.03	0.55 ± 0.04
MoisesDB-4	hyperbolic	0.25 ± 0.10	0.74 ± 0.05	0.79 ± 0.03
	Euclidean	0.46 ± 0.09	0.52 ± 0.08	0.71 ± 0.02
Billboard-4	hyperbolic	0.34 ± 0.03	0.82 ± 0.01	0.80 ± 0.01
	Euclidean	0.54 ± 0.03	0.60 ± 0.03	0.73 ± 0.01

bles 17 and 18 use one earlier fold-0 hyperbolic head without the sibling term and provide no fold or seed uncertainty. Adjusted correlations are associations, not isolated cue contributions.

D Exploratory joint supervision

Containment chains and conditioning-strength sequences are sampled separately and are never interleaved. A conditioning-strength sequence varies audio-conditioning strength for a fixed full-mix source. Full-budget joint training uses an intermediate strength of 0.6; the realised fractions of conditioning sequences are 0.19–0.25 on MoisesDB and 0.044 on Billboard. Hyperbolic direction accuracy is 0.64/0.70/0.51 (stems only: 0.67/0.74/0.82); the joint Euclidean Billboard head reaches 0.55. At a fixed 2,700-chain budget, increasing the fraction of conditioning sequences from zero to 25 % reduces hyperbolic direction by 0.031 ± 0.017, 0.046 ± 0.012 and 0.037 ± 0.010 (95 % paired intervals). These designs do not isolate budget from allocation effects.

Table 19 reports containment measurements under the separate budget of 800 training chains. In this experiment, conditioning sequences pair source embeddings from the central 10 s with generations conditioned on the first 10 s. This affects both joint-head training and the conditioning-order evaluation. The containment measurements therefore characterise the implemented procedure, including its observed decreases, but do not establish the effect of correctly aligned joint supervision.

The conditioning-order statistic measures the share of the three ordered pairs among text-only generation and conditioning strengths 0.1 and 0.8 for which distance to the source decreases by the readout margin. Its evaluation uses a fixed set of 105 full-mix source chains across five folds, so its fold intervals reflect training variability only. These conditioning-order measurements are omitted. The retained containment results do not establish successful learning of both orders, improved exposure discrimination or the full product of stem subsets and conditioning levels.

E Generation-based scores and crossover evaluation

Closeness gain from text. Generate from the candidate caption and measure the negative distance between the output and candidate in a frozen audio embedding. Generation uses only text, but calculating the score requires the candidate audio or its cached embedding. Gain is the adapted-model score minus the pretrained-model score.

Table 11: Three training seeds on fold 0 for the earlier configuration without the sibling term. Columns s_0, s_1 and s_2 identify training seeds. Direction and retrieval are reported separately for each seed; these measurements do not establish seed robustness of the final sibling configuration.

	direction acc.			true superset first		
corpus	s_0	s_1	s_2	s_0	s_1	s_2
MoisesDB-11, hyp	0.713	0.705	0.727	0.599	0.604	0.602
MoisesDB-11, euc	0.463	0.468	0.487	0.544	0.546	0.544
MoisesDB-4, hyp	0.778	0.792	0.794	0.800	0.803	0.797
MoisesDB-4, euc	0.574	0.542	0.547	0.740	0.746	0.740
Billboard-4, hyp	0.818	0.821	0.832	0.806	0.807	0.809
Billboard-4, euc	0.612	0.594	0.606	0.733	0.735	0.738

Table 12: Reversed-convention training from the earlier experiments without the sibling term, five folds, evaluated with the primary-direction tests. Low direction accuracy and inverted radius ordering indicate direction specificity, not failure to learn the reversed order.

corpus	projection	direction acc.	true superset first	radius grows with stems
MoisesDB-11	hyperbolic	0.08 $\pm$0.03	0.47 $\pm$0.04	0.25
	Euclidean	0.06 $\pm$0.04	0.39 $\pm$0.04	0.25
MoisesDB-4	hyperbolic	0.06 $\pm$0.03	0.57 $\pm$0.06	0.18
	Euclidean	0.06 $\pm$0.04	0.49 $\pm$0.06	0.21
Billboard-4	hyperbolic	0.03 $\pm$0.00	0.57 $\pm$0.01	0.13
	Euclidean	0.05 $\pm$0.01	0.49 $\pm$0.00	0.22

Reconstruction gain from audio. Initialise from $z_{\text{init}} = (1-\sigma)z_x + \sigma\epsilon$ with caption and seed fixed, and measure negative source–output embedding distance; conditioning strength is $\lambda = 1-\sigma$. Reconstruction gain is the reduction in this distance from the pretrained to the adapted generator. Source references encode the first 10 s supplied to the generator. Generated clips are encoded at their native 10 s duration as a single window without padding.

Crossover protocol. The crossover tests whether learned containment representations improve exposure discrimination. We compare reconstruction gains using raw CLAP features and hyperbolic and Euclidean heads on identical candidates and generated audio. Evaluation uses fold 0 of all three corpora, both adapter settings, and both stem-only head configurations: the earlier configuration without the sibling term and the final configuration with sibling weight 30 and margin 0.3. The heads are applied without retraining. Generation uses audio initialisation, 10 s outputs, 50 sampling steps and guidance 4.0.

Exposure discrimination. Table 20 reports exact-versus-song-unseen AUCs at $\lambda = 0.4$ for the heads without the sibling term. Across raw CLAP and both geometries, AUCs range from 0.43 to 0.55. The sibling-head AUCs also range from 0.43 to 0.54; raw CLAP results are identical across the two head configurations because they do not use a projection head. Hyperbolic and Euclidean AUCs differ by at most 0.02 within each setting, and no representation consistently performs best. These results provide no clear evidence that containment training improves reconstruction-based exposure discrimination in the evaluated settings.

The exact-versus-song-unseen contrast changes song exposure and generally artist exposure, with the partition limitations described in §4.1. It does not isolate constituent exposure. The generation sample contains only 1/0/2 none candidates in MoisesDB-11/MoisesDB-4/Billboard-4 order, precluding a useful component-exposure comparison. Evaluation on one fold also limits conclusions about variation across data partitions.

Distance decomposition. The trained heads place stem subsets at radii that increase with stem count, so the full distance mixes a hierarchy coordinate into the exposure score. We therefore recompute reconstruction gain with two components of the full distance, using the same candidates and generations. Let $r(x) = \|\log_0 x\|$ be the norm of the tangent vector at the origin. The radial

Table 13: Width sweep, fold 0, validation-calibrated direction accuracy. The sweep uses unit-weight entailment under its own base recipe and predates the final sibling configuration. Dashes mark epoch-0 runs stopped without checkpoints for failing the minimum spread criterion (reported spread 0.029–0.050; threshold 0.05); these occur only on MoisesDB. Billboard favours Euclidean geometry at width eight, so the sweep establishes no general low-dimensional hyperbolic advantage.

corpus	projection	8	16	32	64
MoisesDB-11	hyperbolic	—	—	0.672	0.681
	Euclidean	—	—	—	0.500
MoisesDB-4	hyperbolic	—	—	0.749	0.770
	Euclidean	—	—	0.571	0.571
Billboard-4	hyperbolic	0.763	0.699	0.774	0.795
	Euclidean	0.791	0.672	0.639	0.576

Table 14: Chain-term ablation, five folds, stage-four recipe (stage-four scale initialisation 0.44; its raised clamp is inactive). The weight-2 chain term is removed and adjacent-node contrastive training added; adjacent cones and ranking remain. Full-objective direction baselines are 0.66/0.74/0.83 (hyperbolic) and 0.44/0.48/0.55 (Euclidean), in corpus order. This is not a single-factor ablation of the final sibling configuration.

corpus	projection	direction acc.	true superset first	radius grows with stems
MoisesDB-11	hyperbolic	0.38 $\pm$0.10	0.65 $\pm$0.03	0.73
	Euclidean	0.34 $\pm$0.06	0.57 $\pm$0.04	0.71
MoisesDB-4	hyperbolic	0.51 $\pm$0.15	0.83 $\pm$0.02	0.78
	Euclidean	0.38 $\pm$0.08	0.74 $\pm$0.03	0.76
Billboard-4	hyperbolic	0.68 $\pm$0.04	0.83 $\pm$0.01	0.86
	Euclidean	0.48 $\pm$0.01	0.77 $\pm$0.01	0.86

score is $|r(u) - r(v)|$ and the angular score is the angle between $\log_0 u$ and $\log_0 v$. Both readouts use tangent coordinates at the origin, with $\log_0$ equal to the identity in Euclidean space. For a raw feature f and a fixed positive scale α, the identity lift and the exact Lorentz exponential lift both recover αf in these coordinates. The angular readout is unchanged by α, whereas radial differences and their reconstruction gains are multiplied by α. Consequently, both reconstruction-gain AUCs are invariant to this common positive scaling and agree across the two lifts, so one raw AUC is reported for each readout.

Table 21 reports both head configurations. No component yields consistent exposure discrimination. The angular score exceeds the full distance in the rank-64 MoisesDB-11 setting, reaching 0.58 [0.51, 0.66] for the hyperbolic head without the sibling term, and 0.57 [0.50, 0.65] with it. The corresponding raw angular score is 0.50. In the other settings, head angular AUCs lie between 0.46 and 0.55. The radial score falls below 0.5 in several head settings, most clearly for the sibling heads on the flagged rank-64 Billboard-4 adapter, where the Euclidean value is 0.45 [0.40, 0.49]. The highest radial value, 0.56 [0.48, 0.64], comes from raw CLAP on rank-16 MoisesDB-4 rather than from a trained head. With 96 cells, one fold and candidate-level intervals, we treat the angular gain in one setting as exploratory. It does not change the conclusion that containment training provides no clear improvement in reconstruction-based exposure discrimination.

Separate generation-readout limitations. The independent seen-song analysis uses a different preprocessing path that centre-pads 10 s generations to 30 s, introducing silent windows. Its affected reconstruction, decomposition and source-distance values remain omitted and do not support membership or conditioning-order conclusions. These limitations do not apply to the crossover protocol above. The separate joint-supervision experiment and its training limitations are described in Appendix D.

Table 15: Direction accuracy on training and held-out songs for the earlier heads without the sibling term, five folds, validation-calibrated apertures. Gaps describe these partitions, not the isolated effect of seeing a song; test-informed selection remains as disclosed in §4.3.

		direction acc.		
corpus	projection	seen	unseen	gap
MoisesDB-11	hyperbolic	0.714 ±0.014	0.674 ±0.049	+0.040
	Euclidean	0.507 ±0.011	0.479 ±0.034	+0.028
MoisesDB-4	hyperbolic	0.787 ±0.015	0.739 ±0.045	+0.048
	Euclidean	0.541 ±0.021	0.519 ±0.083	+0.022
Billboard-4	hyperbolic	0.828 ±0.009	0.821 ±0.011	+0.007
	Euclidean	0.603 ±0.019	0.597 ±0.031	+0.007

Table 16: Cross-corpus evaluation of the earlier heads without the sibling term, fold 0, with aperture calibrated on target validation data. MoisesDB taxonomies share recordings; transfer does not isolate the contribution of separation artifacts. Untrained references are retained for comparison.

	direction acc.		true superset first	
projection trained on	hyp	euc	hyp	euc
read on MoisesDB-4				
MoisesDB-4 (own corpus)	0.778	0.574	0.800	0.740
MoisesDB-11	0.749	0.516	0.802	0.748
Billboard-4	0.716	0.494	0.820	0.703
untrained encoder	0.017	0.027	0.855	0.757
read on MoisesDB-11				
MoisesDB-11 (own corpus)	0.713	0.463	0.599	0.544
MoisesDB-4	0.728	0.519	0.594	0.528
Billboard-4	0.625	0.470	0.619	0.518
untrained encoder	0.027	0.038	0.644	0.551
read on Billboard-4				
Billboard-4 (own corpus)	0.818	0.612	0.806	0.733
MoisesDB-11	0.704	0.441	0.816	0.751
MoisesDB-4	0.726	0.485	0.818	0.763
untrained encoder	0.024	0.037	0.889	0.810

Table 17: Fold-0 within-song size ordering for the earlier hyperbolic head without the sibling term: fraction of differently sized pairs whose loudness or embedding radius grows with stem count. This statistic alone does not establish containment learning.

corpus	loudness ordering	radius ordering
MoisesDB-11	0.88	0.84
MoisesDB-4	0.94	0.92
Billboard-4	0.91	0.93

Table 18: Fold-0 correlations for the earlier hyperbolic head without the sibling term and frozen features. Distance–set-difference correlations adjust for stem counts; radius–count correlations adjust for loudness.

	distance–set difference adjusted for counts		radius–loudness		radius–count adjusted for loudness	
corpus	head	frozen	head	frozen	head	frozen
MoisesDB-11	0.47	0.44	0.63	−0.24	0.45	−0.12
MoisesDB-4	0.54	0.50	0.57	−0.24	0.60	−0.22
Billboard-4	0.58	0.52	0.56	0.06	0.65	−0.09

Table 19: Containment readout from the hyperbolic joint-supervision check: 800 training chains, depth four and five folds. Test containment is validation-calibrated direction accuracy. Conditioning sequences constitute 0/1/0.5 of training sequences for containment-only/conditioning-only/joint training, respectively; the frozen encoder is also reported. Conditioning sequences use mismatched source segments during training and evaluation; conditioning-order results are omitted, and containment scores describe that implemented training procedure. These budgets and depths differ from the final sibling comparison.

training setting	test containment
MoisesDB-11	
frozen encoder	0.02 $\pm$0.01
containment only	0.62 $\pm$0.07
conditioning only	0.11 $\pm$0.06
joint	0.54 $\pm$0.06
MoisesDB-4	
frozen encoder	0.02 $\pm$0.01
containment only	0.72 $\pm$0.08
conditioning only	0.08 $\pm$0.05
joint	0.64 $\pm$0.05
Billboard-4	
frozen encoder	0.03 $\pm$0.00
containment only	0.81 $\pm$0.01
conditioning only	0.15 $\pm$0.03
joint	0.73 $\pm$0.01

Table 20: Exact-versus-song-unseen reconstruction-gain AUC [95 % interval] on fold 0 at $\lambda = 0.4$, using the stem-only heads without the sibling term. Source references match the 10 s conditioning segments; generated clips are encoded without padding. All representations use identical candidates and generations. Scores use the full geodesic distance. Raw CLAP is scaled by the reciprocal of its mean norm and lifted with the Lorentz origin map; the Euclidean-space raw reference differs by at most 0.01. Intervals are percentile bootstraps (1000 draws) resampling candidates independently within each class; candidates share songs, so they understate uncertainty. R16 and R64 use learning rates 10^{-4} and 10^{-3}, respectively, so rank and learning rate vary jointly. † denotes the descriptive FAD flag (App. B); the setting is retained in the comparison.

Corpus	Adapter	Hyperbolic	Euclidean	Raw CLAP
MoisesDB-11	R16	0.46 [0.38, 0.54]	0.47 [0.39, 0.55]	0.43 [0.34, 0.51]
	R64	0.54 [0.47, 0.62]	0.55 [0.48, 0.63]	0.50 [0.42, 0.58]
MoisesDB-4	R16	0.44 [0.36, 0.52]	0.45 [0.37, 0.54]	0.45 [0.37, 0.53]
	R64	0.47 [0.38, 0.55]	0.46 [0.39, 0.54]	0.45 [0.37, 0.53]
Billboard-4	R16	0.48 [0.44, 0.52]	0.49 [0.44, 0.53]	0.49 [0.44, 0.53]
	R64†	0.53 [0.49, 0.58]	0.52 [0.47, 0.56]	0.50 [0.46, 0.54]

Table 21: Distance decomposition of exact-versus-song-unseen reconstruction-gain AUC on fold 0 at $\lambda = 0.4$, using the same candidates and generations as Table 20. Full: geodesic distance. Rad.: $|r(u) - r(v)|$ with $r(x) = \|\log_0 x\|$. Ang.: angle between the tangent vectors at the origin. Raw radial and angular scores are identical for the Lorentz and Euclidean lifts. Point estimates only; candidate-bootstrap 95 % intervals, which understate uncertainty because candidates share songs, have half-widths of about 0.08 on MoisesDB and 0.045 on Billboard-4. † denotes the descriptive FAD flag (App. B).

		Hyperbolic head			Euclidean head			Raw CLAP	
Corpus	Adapter	Full	Rad.	Ang.	Full	Rad.	Ang.	Rad.	Ang.
Heads without the sibling term									
MoisesDB-11	R16	0.46	0.42	0.51	0.47	0.45	0.50	0.45	0.46
	R64	0.54	0.48	0.58	0.55	0.49	0.57	0.48	0.50
MoisesDB-4	R16	0.44	0.49	0.47	0.45	0.49	0.47	0.56	0.45
	R64	0.47	0.47	0.49	0.46	0.48	0.48	0.43	0.45
Billboard-4	R16	0.48	0.49	0.49	0.49	0.47	0.50	0.52	0.50
	R64†	0.53	0.51	0.52	0.52	0.52	0.50	0.53	0.51
Heads with sibling weight 30, margin 0.3									
MoisesDB-11	R16	0.43	0.46	0.48	0.43	0.48	0.46	0.45	0.46
	R64	0.54	0.45	0.57	0.54	0.47	0.55	0.48	0.50
MoisesDB-4	R16	0.45	0.46	0.48	0.44	0.47	0.47	0.56	0.45
	R64	0.47	0.45	0.48	0.46	0.47	0.46	0.43	0.45
Billboard-4	R16	0.48	0.49	0.50	0.49	0.48	0.51	0.52	0.50
	R64†	0.54	0.47	0.54	0.54	0.45	0.55	0.53	0.51