DEMAND GEOMETRY AND SPARSE INTERFACES FOR CONTINUOUS MUSIC TOKENS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous music tokenizers support reconstruction and downstream modeling, but different tasks may benefit from different low-dimensional views of the same tokens. For frozen tokenizers, we study tradeoffs among active input width, shared capacity, and demand-specific utility. A minimax certificate quantifies unavoidable compromise and rules out exact common optimality for the tested reconstruction and prediction operators. Using target-canonical normalization, we define demand relations invariant to invertible linear input and target transformations under fixed whitening supports and ridge strengths. For each tokenizer, *Atom Atlas* assigns each known demand K of M shared orthonormal atoms. Within a fixed union subspace, chosen dense-view projectors admit exact realization by shared atom subsets if and only if they commute pairwise. Support-aware certificates bound operator utility loss from approximate realization for fitted operators. Demand relations recur across three tokenizers on Moises and two on Slakh with a different teacher. Across six primary MUSDB/Moises settings at $M/K = 4$, Atlas retains 84–97% of the dense-over-shared reduction in worst-demand operator regret. With matched nonlinear predictors on Moises, Atlas improves future-token prediction and prediction of frequency-band teacher embeddings over the shared view. On the SAME-L band task, native Atlas retains 85.3% of the improvement from the shared view to native dense, supporting prediction directly from stored atom coordinates.

1 INTRODUCTION

An audio tokenizer is both a compression device and a downstream interface. Discrete codecs define this interface through code rate and stream layout (Défossez et al., 2023; Kumar et al., 2023; Copet et al., 2023). For continuous tokens, we ask which directions in token space a fixed-width interface should expose to each downstream demand. Using the same prefix of a reconstruction-trained coordinate order gives every demand the same view, which need not suit a finite causal model or a particular source, frequency region, or time scale.

Locodec (Luo et al., 2026) represents speech by low-rate spherical vectors with progressively decodable coordinates. Our stereo-music adaptation tests whether one order serves demands ranging from local fidelity to source, frequency, and long-range structure.

For a frozen tokenizer, one shared K-dimensional view may compromise utility across known demands. Different views within a shared M-dimensional span can relieve that compromise while keeping each consumer's input width at K. Atom Atlas implements these views as subsets of one dictionary. We analyze the resulting access loss and separately compare demand relations across tokenizers (Figure 1).

This separation leads to three contributions:

- We quantify shared-subspace compromise at fixed rank. A minimax certificate bounds unavoidable regret and rules out exact common optimality for the tested reconstruction and prediction operators.
- We define target-canonical demand geometry whose relations are invariant to invertible linear changes of input and target coordinates, enabling comparison without aligning tokenizer axes. Using this geometry, we find recurring relations among matched demands across Locodec, SAME-L (Parker et al., 2026), and Music2Latent (Pasini et al., 2024).

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

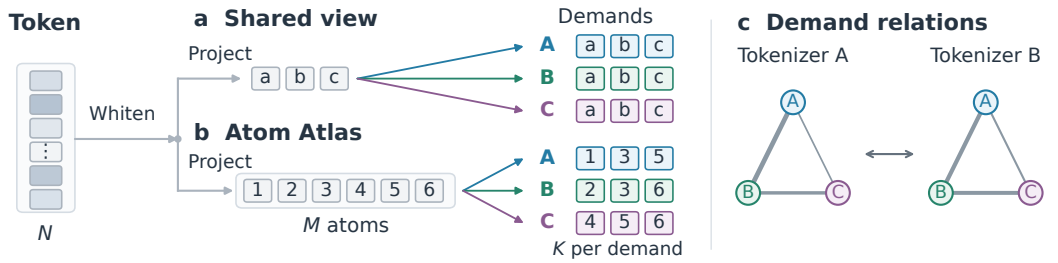

Figure 1: Compact demand-specific views of frozen music tokens. With common whitening, (a) demands share one K-coordinate view; (b) each demand accesses K of M locally fitted atoms. (c) Target-canonical relations across tokenizers without axis alignment. Schematic: $K = 3$, $M = 6$.

- We propose Atom Atlas for demand-specific K-of-M access to shared orthonormal atoms. We characterize exact realization of dense views within a fixed union and give support-aware loss bounds for fitted operators. Matched nonlinear consumers show gains over a shared view on construction-related demands.

Together, these results characterize the tradeoff between active input width, shared capacity, and sparse access loss for frozen music representations. In the primary experiments, Atom Atlas retains most of the improvement of demand-specific dense views over a shared view while exposing only K coordinates to each consumer.

2 Related Work

Audio tokenization. Audio tokenizer evaluations distinguish reconstruction fidelity from downstream utility (Mousavi et al., 2025; Wang et al., 2025). Discrete codecs supply quantized streams (Défossez et al., 2023; Kumar et al., 2023; Ji et al., 2025) to generative models (Borsos et al., 2023; Copet et al., 2023). Continuous autoencoders learn decodable latents (Pasini et al., 2024; Parker et al., 2026), and continuous audio language models generate latent sequences (Yang et al., 2025; Rouard et al., 2026). Access also depends on structure introduced during tokenization: source, time–frequency, and semantic/reconstruction organization shape token interfaces (Lin et al., 2025; Cheng et al., 2026; Yang et al., 2026), while frequency-aware autoencoders structure spectral processing (Wang et al., 2026). F3 adapts frozen autoencoder latents through a learned representation encoder (Zhou et al., 2026). On frozen representations, we study the utility retained by demand-specific K-coordinate views of one shared M-atom orthonormal bank.

Ordered and task-conditioned representations. Compact inputs can use predictive subspaces, as in reduced-rank regression (Izenman, 1975), or nested prefixes across input widths (Rippel et al., 2014; Kusupati et al., 2022; Luo et al., 2026). A shared prefix exposes the same coordinates to every demand at a given width. Matryoshka sparse autoencoders learn nested dictionaries of increasing size (Bussmann et al., 2025). Sparse coding and multitask learning provide reusable atoms and task-specific predictors (Olshausen & Field, 1996; Argyriou et al., 2008; Lachapelle et al., 2023). In sparse multitask learning, Maurer et al. (2013) bound the generalization error of predictors learned jointly with a shared dictionary. Atlas fits its orthonormal atom bank and demand-specific supports to estimated operators at fixed active width K, then freezes the interface before training consumers. With a linear consumer, Atlas has the prediction form of a sparse dictionary model; the demand-specific coordinate restriction also applies to nonlinear consumers. Our analysis separates active width from shared capacity and, for fitted operators, bounds realization loss relative to demand-specific dense views within the same union (Appendix M).

Representation comparison and shared subspaces. Representation comparison uses canonical correlations for paired views (Hotelling, 1936; Raghu et al., 2017) or relations among samples: representational similarity analysis and centered kernel alignment compare the latter without matching individual coordinates (Kriegeskorte et al., 2008; Kornblith et al., 2019). For task comparison, Task2Vec constructs embeddings from the Fisher information of probe network parameters (Achille

et al., 2019). To analyze shared structure across operators, common principal components and joint or block diagonalization seek common bases or invariant subspaces (Flury, 1984; Cardoso & Souloumiac, 1996; Maehara & Murota, 2011). Music information retrieval studies complement structural comparison with evaluations of downstream utility (Pasini et al., 2024; Yuan et al., 2023). We compare tokenizers through relation matrices of matched demand operators, computed separately within each tokenizer. Target-canonical normalization makes these relations invariant to invertible linear input and target coordinate changes under the whitening and ridge conditions of Proposition 1. Within each tokenizer, Atlas uses approximate joint diagonalization to construct its atom bank in a fixed union. We characterize when dense demand views in this union admit exact realization by coordinate subsets. Matched consumers evaluate the utility retained by coordinate access relative to demand-specific dense views at fixed input width.

3 Demand Geometry and Sparse Realization

We analyze how shared capacity supports demand-specific views at fixed consumer input width.

3.1 Demands and compatibility

For a frozen tokenizer, we use centered tokens $h \in \mathbb{R}^N$ and globally whitened coordinates $x = \Sigma_{hh}^{-1/2} h$ on a fixed support with $\Sigma_{hh} \succ 0$. A demand $d = (c, q, b, \tau, K)$ specifies a consumer c, optional source q, spectral region b, physical horizon τ, and active width K. We compare a finite demand family at a common K, with nonzero positive-semidefinite (PSD) utility operators A_d in these coordinates. Let $\mathcal{P}_K^n = \{P \in \mathbb{R}^{n\times n} : P = P^\top = P^2, \mathrm{tr}(P) = K\}$ denote rank-K orthogonal projectors. By the Ky Fan principle (Fan, 1949), a leading spectral projector $\Pi_d = \mathrm{Top}_K(A_d)$ maximizes $\mathrm{tr}(A_d P)$. For $P \in \mathcal{P}_K^N$, normalized regret measures operator utility lost relative to this demand-specific optimum:

$$\delta_d(P) = \frac{\mathrm{tr}[A_d(\Pi_d - P)]}{\mathrm{tr}(A_d)}. \tag{1}$$

A common rank-K view is optimal for every demand exactly when $\Xi_K = \min_{P\in\mathcal{P}_K^N} \max_d \delta_d(P) = 0$. We bracket Ξ_K with a spectral simplex lower bound and a feasible-projector upper bound (Appendix A); a positive lower bound certifies unavoidable compromise for these operators. Across widths, an exactly compatible progressive order also requires nested common optima.

To relieve this compromise at fixed K, demand-specific views can share a larger span. Their union rank m measures shared capacity; under budget $K \le M \le N$, the best worst-demand regret is

$$m(\{P_d\}) = \dim\left(\mathrm{span}\bigcup_d \mathrm{range}(P_d)\right), \qquad \Xi_{\mathrm{branch}}(K, M) = \min_{\substack{P_d\in\mathcal{P}_K^N \\ m(\{P_d\})\le M}} \max_d \delta_d(P_d). \tag{2}$$

$M = K$ forces one common subspace; $M > K$ allows different views at the same active width. Atlas exposes this larger span through demand-specific K-coordinate subsets of one bank.

3.2 Target-canonical demand geometry

To compare demand relations across tokenizers, we remove invertible linear input and target coordinate choices. Let $y_d \in \mathbb{R}^{p_d}$ be centered. Demand-specific sampling need not give $\Sigma_{xx,d} = I$ under this global whitening. For $\Sigma_{yy,d} \succ 0$ and scalar $\rho_d > 0$, define the target-whitened ridge loading using symmetric inverse square roots (Hotelling, 1936; Hoerl & Kennard, 1970)

$$R_d^\star = (\Sigma_{xx,d} + \rho_d I)^{-1}\Sigma_{xy,d}\Sigma_{yy,d}^{-1/2}, \qquad M_d^\star = R_d^\star R_d^{\star\top}. \tag{3}$$

Proposition 1 (bi-invariance). With fixed ρ_d and full-rank supports, recompute global input whitening after any invertible linear tokenizer-wide input transformation and target whitening after an independent invertible linear transformation of every y_d. Then one orthogonal Q shared by all demands of that tokenizer satisfies $M_d^{\star\prime} = Q^\top M_d^\star Q$ for every d.

Common orthogonal conjugation preserves weighted similarity of full operators and overlap of their leading subspaces:

$$\mathcal{K}_{de} = \frac{\mathrm{tr}(M_d^\star M_e^\star)}{\|M_d^\star\|_F \|M_e^\star\|_F}, \qquad \mathcal{O}_{de}^{(r)} = \frac{\mathrm{tr}(\Pi_{d,r}\Pi_{e,r})}{r}. \tag{4}$$

Here $\Pi_{d,r} = \mathrm{Top}_r(M_d^\star)$ for $1 \le r < N$; $\mathcal{K}$ requires nonzero operators, and $\mathcal{O}$ requires $\lambda_r(M_d^\star) > \lambda_{r+1}(M_d^\star)$ for unique leading subspaces. These matrices compare matched demands without aligning tokenizer axes. Invariance requires invertible linear changes and fixed whitening supports; Appendix C gives further invariant statistics. Local Atlas construction uses operators in each tokenizer's own whitened coordinates.

3.3 Music demands as positive-semidefinite operators

For local Atlas construction, we express musical demands through the utility operators A_d in Eq. 1. Predictive loadings, source-induced changes, and local loss sensitivity give the following PSD forms:

$$A_d \in \left\{ R_d R_d^\top,\ \mathbb{E}[\Delta_q \Delta_q^\top],\ \mathrm{GN}_{\mathrm{local}}(\mathcal{L}_{\mathrm{rel},\tau}),\ \mathbb{E}[\widehat{J}_{\mathrm{dec},t}^\top W_{q,b} \widehat{J}_{\mathrm{dec},t}] \right\}. \tag{5}$$

For prediction, ordinary-target ridge gives $R_d = (\Sigma_{xx,d} + \rho_d I)^{-1} \Sigma_{xy,d}$ (Hoerl & Kennard, 1970). Its score $\mathrm{tr}(R_d R_d^\top P) = \|P R_d\|_F^2$ measures retained loading energy. Target whitening in Eq. 3 serves cross-tokenizer relation comparison. For source interventions, the whitened displacement Δ_q measures sensitivity to controlled replacement of source q; its second moment scores retained change through $\mathrm{tr}(P\mathbb{E}[\Delta_q \Delta_q^\top]) = \mathbb{E}\|P\Delta_q\|_2^2$.

For relations and reconstruction, we score views by local loss sensitivity. The relation operator $\mathrm{GN}_{\mathrm{local}}$ sums per-token diagonal blocks of the generalized Gauss–Newton (GN) matrix (Martens, 2020). For decoder loss, the Jacobian $\widehat{J}_{\mathrm{dec},t}$ pulls the PSD probe metric $W_{q,b}$ into whitened token coordinates, with tangent restriction when applicable (Arvanitidis et al., 2022; Nazari et al., 2023). Both curvatures omit cross-time blocks (Appendix B). These operators define preferred views without assuming disentangled latent axes. We next use them to select a shared union and construct the Atlas dictionary and demand-specific coordinate supports.

3.4 The Atom Atlas

To implement K-of-M access with $1 \le K < M \le N$, we use a common union $U \in \mathbb{R}^{N \times M}$, $U^\top U = I$, and rotation $D \in \mathbb{O}(M)$. For $S \subseteq \{1, \dots, M\}$, let $E_S = \mathrm{diag}(\mathbf{1}_S)$. The dictionary is $\Phi = UD$; each consumer reads K coordinates through

$$V_d = \Phi_{:,S_d}, \quad P_d^{\mathrm{A}} = V_d V_d^\top, \quad z_d = V_d^\top x \in \mathbb{R}^K, \quad |S_d| = K. \tag{6}$$

Within a fixed U, the optimal regrets compare arbitrary dense views with subsets of one bank:

$$\Xi_{\mathrm{dense}}(U) = \max_d \min_{Z_d^\top Z_d = I_K} \delta_d(U Z_d Z_d^\top U^\top), \tag{7}$$

$$\Xi_{\mathrm{atom}}(U) = \min_{D \in \mathbb{O}(M)} \max_d \min_{|S_d| = K} \delta_d(U D E_{S_d} D^\top U^\top). \tag{8}$$

Dense choices $Z_d \in \mathbb{R}^{M \times K}$ separate across demands; Atlas restricts them to columns of one UD. Thus $\Gamma_{\mathrm{atom}} = \Xi_{\mathrm{atom}}(U) - \Xi_{\mathrm{dense}}(U) \ge 0$ measures the optimal cost of this access restriction.

We retain the candidate union U with lowest training worst-demand dense regret (Appendix Algorithm 1). Within this frozen union, define

$$C_d = \frac{U^\top A_d U}{\|U^\top A_d U\|_F} \tag{9}$$

for nonzero projections, and set $C_d = 0$ otherwise. This rescaling preserves leading projectors and support rankings. Fix a leading rank-K projector $\bar{\Pi}_d$ of each C_d. Exact realization expresses these dense optima as coordinate subsets of one orthonormal basis.

Proposition 2 (exact realization and obstruction). By simultaneous orthogonal diagonalization (Horn & Johnson, 2012), one orthogonal D and sets S_d satisfy $\bar{\Pi}_d = D E_{S_d} D^\top$ for every demand iff $[\bar{\Pi}_d, \bar{\Pi}_e] = 0$ for every pair. Then $\mathrm{tr}(\bar{\Pi}_d \bar{\Pi}_e) = |S_d \cap S_e|$. For approximate realization by any Atlas coordinate projectors $\bar{P}_d^{\mathrm{A}} = D E_{S_d} D^\top$ and $\bar{P}_e^{\mathrm{A}} = D E_{S_e} D^\top$, define $\Delta_d(P) = \mathrm{tr}[C_d(\bar{\Pi}_d - P)]$ and $\gamma = \min\{\lambda_K(C_d) - \lambda_{K+1}(C_d), \lambda_K(C_e) - \lambda_{K+1}(C_e)\}$. If $\gamma > 0$, then

$$\begin{aligned} \max\{\|\bar{\Pi}_d - \bar{P}_d^{\mathrm{A}}\|_F, \|\bar{\Pi}_e - \bar{P}_e^{\mathrm{A}}\|_F\} &\ge \tfrac{1}{4}\|[\bar{\Pi}_d, \bar{\Pi}_e]\|_F, \\ \max\{\Delta_d(\bar{P}_d^{\mathrm{A}}), \Delta_e(\bar{P}_e^{\mathrm{A}})\} &\ge \frac{\gamma}{32}\|[\bar{\Pi}_d, \bar{\Pi}_e]\|_F^2. \end{aligned} \tag{10}$$

Exact realization encodes subspace overlap as atom reuse. Added regret over the within-union dense optimum is $\|U^\top A_d U\|_F \Delta_d(P)/\operatorname{tr}(A_d)$. Noncommutation need not give $\Gamma_{\text{atom}} > 0$: affected demands can remain below the worst dense regret (Appendix D).

Approximate joint diagonalization (AJD) constructs candidate dictionaries by reducing the C_d's off-diagonal energy, a surrogate for coordinate-selection loss (Cardoso & Souloumiac, 1996). For fixed D, the K largest diagonal entries of $D^\top C_d D$ maximize demand d's utility. Training worst-demand atom regret selects the fit; global optimality is not guaranteed.

For a fitted support, set $P_d = DE_{S_d}D^\top$ and $Q_d = I - P_d$ in union coordinates. Splitting $C_d = C_{d,0} + X_d$, with $C_{d,0} = P_d C_d P_d + Q_d C_d Q_d$, separates within-block utility from coupling to excluded coordinates. Write $u_K(C) = \sum_{j=1}^K \lambda_j(C)$ and $E_d = \bar{\Pi}_d - P_d$ for the dense-projector error.

Proposition 3 (support-aware certificates). For every frozen rank-K Atlas projector,

$$\Delta_d(P_d) \le \underbrace{u_K(C_{d,0}) - \operatorname{tr}(P_d C_d)}_{\text{support choice}} + \underbrace{u_K(X_d)}_{\text{cross-support leakage}}, \tag{11}$$

$$\frac{\operatorname{tr}(\bar{\Pi}_d \bar{\Pi}_e) - \operatorname{tr}(P_d P_e)}{K} = \frac{\operatorname{tr}(E_d P_e) + \operatorname{tr}(P_d E_e) + \operatorname{tr}(E_d E_e)}{K}. \tag{12}$$

Ky Fan subadditivity (Fan, 1949) gives Eq. 11. The first term measures the cost of the selected block after removing coupling; the second bounds coupling by $u_K(X_d) = \|Q_d C_d P_d\|_*$. Unlike the full AJD residual, this leakage term excludes off-diagonal energy within the selected block. With a positive block eigengap $\eta_d = \lambda_K(C_{d,0}) - \lambda_{K+1}(C_{d,0}) > 0$, a spectral variational inequality (Davis & Kahan, 1970) can sharpen this second term:

$$u_K(C_d) - \operatorname{tr}(P_d C_d) \le u_K(C_{d,0}) - \operatorname{tr}(P_d C_d) + \min\left\{ u_K(X_d), \frac{\|X_d\|_F^2}{2\eta_d} \right\}. \tag{13}$$

Alongside the utility bounds, Eq. 12 quantifies how coordinate selection changes pairwise subspace overlaps within the fixed union. The Atlas overlap $\operatorname{tr}(P_d P_e)/K = |S_d \cap S_e|/K$ is the fraction of shared atoms. The three numerator terms can cancel; summing their absolute values and dividing by K bounds the absolute overlap error. These a posteriori bounds concern fitted operators. Appendix D gives proofs and a perturbation analysis conditional on supplied operator-estimation error bounds. Experiments with matched nonlinear consumers evaluate the predictive utility retained by these compact views.

4 Experimental Design

We test the utility gains from a larger union at fixed input width and their retention by Atlas. Cross-tokenizer comparisons examine demand relations (Table 1).

Table 1: Experimental tests of demand geometry, sparse realization, and consumer utility. Appendix Table 8 defines the readouts and their aggregation.

Research question	Evaluation design	Readout
Shared-view compromise (§5.1)	Common rank-K view versus demand-wise optima	Minimax bounds on Ξ_K
Demand relations (§5.2)	Matched relation matrices across tokenizers; factor-preserving permutations	Cross-tokenizer relation correlations
Sparse realization (§5.3)	Single–dense at fixed K; dense–Atlas within the same union	Observed closure and certified lower bounds
Consumer utility (§5.4)	Matched nonlinear consumers; address-selection controls and compact/full baselines	Prediction error, retained improvement, macro AUROC

Post-hoc comparisons freeze music-adapted Locodec (768-D, 8.52 Hz) and released SAME-L/Music2Latent checkpoints. MUSDB18-HQ supports method development; MoisesDB v0.1 supports interface construction and prediction evaluation (Rafii et al., 2019; Pereira et al., 2023). Within each construction setting, views share training-fitted whitening; validation selects ridge strengths.

Shared-subspace compromise. To assess view choice, we compare Locodec's core, trained jointly with the tokenizer, with a raw prefix at matched width; post-hoc controls match projection updates.

We then ask whether even the best shared view must compromise, bracketing Ξ_K for reconstruction and three output-side prediction operators. A separate input frontier measures predictive utility under a reconstruction-regret constraint (Appendix G).

Cross-tokenizer recurrence of demand relations. We correlate off-diagonal relations (Eq. 4) across tokenizers for 32 Moises guitar/piano–band–horizon demands reserved from Atlas construction, matching physical horizons. Factor-preserving permutations assess agreement against shuffled demand correspondences. Slakh2100-redux with MERT targets extends this comparison to another corpus and teacher (Manilow et al., 2019; Li et al., 2024). Native future-token controls remove the shared teacher; paired Slakh refits assess estimation stability (Appendix F.2).

Sparse realization and certificate tightness. At fixed K, Single–dense tests the utility of demand-specific views in a larger union; dense–Atlas measures the loss from restricting these views to shared-atom subsets. We fit training operators (Eq. 5) for 25 construction demands, then select unions, dictionaries, and supports locally (Appendix Table 7 and Algorithm 1). For Locodec $K = 32$ and SAME-L $K \in \{10, 26\}$, Single uses $M = K$; dense, Atlas, and controls share a separately fitted, frozen U at $M = 4K$ ($M = 3K$ sensitivity). To assess dictionary orientation, raw and 64 Haar controls select K coordinates by diagonal utility in the unrotated union basis and random rotations.

Let S, D, and A denote the worst-demand regrets of Single, dense, and Atlas, respectively (Eq. 1). For $S > D$, closure $(S - A)/(S - D)$ measures the Single-to-dense improvement retained by Atlas. Comparing support and dense-projector overlaps tests preservation of subspace relations. Construction audits compare Proposition 3's utility and overlap bounds with realized losses and errors; evaluation audits compare closure lower bounds with observed closure (Appendix E). Evaluation uses test-estimated source, relation, and decoder operators; readouts lacking independent test operators reuse frozen training fits (Appendix F).

Nonlinear prediction and downstream utility. To test predictive utility, we match architectures, optimizers, and three seeds across consumers fitted on fixed Single, dense, and Atlas views within each primary task. Procrustes aligns dense/Atlas bases to Single; inputs are radius-normalized and validation selects checkpoints. Endpoint Transformers predict future coordinates in their input view (about .47 seconds ahead); lifted outputs are scored against common native targets within each tokenizer. We apply the frozen Moises interfaces and consumers to 800 FMA-small test tracks (Defferrard et al., 2017). To hold training targets fixed across views, a temporal convolutional network predicts PupuJEPA embeddings (Gu et al., 2026) for four frequency bands. The native ablation tests stored Atlas coordinates without alignment. Prediction retention uses the same ratio for cosine errors. Moises validation requires dense and then Atlas to beat Single and persistence (endpoint) or the training-target centroid (band), with at least 50% retention by Atlas before testing (Appendix F).

To test address adaptation, we select 16 of 64 atoms from a frozen Moises Music2Latent dictionary on MTG-Jamendo's official training partition (Bogdanov et al., 2019). Genre and Mood/theme each use 20 selected labels and a separate consumer. Reselected shared addresses are compared with source-construction addresses; comparing task-specific with shared and swapped addresses tests task matching. With addresses fixed, we fit fresh consumers for comparison with full-token and compact baselines, including dense projections learned jointly with each consumer. Macro AUROC is averaged over development folds and seeds (Appendix P). The Slakh/MERT comparison also tests prediction with separately fitted interfaces and consumers (Appendix L). Public controls for shared projectors and native addresses appear in Appendices J, N, and O.

Unless noted, reported consumer 95% confidence intervals (CIs) use 10,000 paired song bootstraps (artists for OpenMIC; Humphrey et al., 2018), conditional on fitted models and unadjusted for adaptive development. Moises refits, native ablations, and later OpenMIC baselines reuse test data.

5 Results

5.1 Shared-subspace compromise

At fixed 32-D width, Locodec's jointly trained core has higher primary MIR scores than a raw prefix in 28/30 comparisons (five tasks, three tokenizer seeds with and without progressive-dimension dropout).

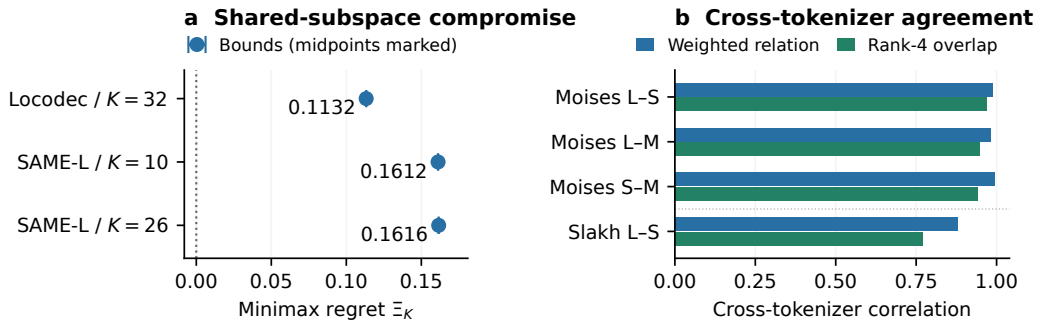

Figure 2: Shared-subspace compromise and cross-tokenizer demand relations. (a) Intervals bound Ξ_K for fitted MUSDB18-HQ operators; points and labels mark midpoints. SAME-L ranks 10 and 26 approximately match Locodec ($K = 32$) in latent fraction and coordinate rate, respectively. (b) Point estimates of correlations between off-diagonal target-canonical relations: Pearson for 32 Moises demands; Spearman for 48 Slakh/MERT demands. Appendix Tables 14 and 18 give values, and Appendix F.2 describes factor-preserving permutation tests. L/S/M: Locodec/SAME-L/Music2Latent.

The core has lower causal flow loss in all six core–prefix comparisons with matched predictor-training budgets (Appendix Table 12). In a separate comparison with matched predictors, the core also has lower causal flow loss than PCA and random projections (Appendix Table 13a). For the frozen tokenizer, post-hoc views with matched projection updates have similar STFT reconstruction losses, while reconstruction and causal flow losses rank these views differently (Appendix Table 13b).

The minimax certificate quantifies compromise even for the best shared view. For the fitted reconstruction and three output-side prediction operators on MUSDB18-HQ, it gives $.11321 \leq \Xi_{32} \leq .11327$ for Locodec (Figure 2a). Every common rank-32 view thus incurs positive regret for at least one demand relative to that demand's rank-32 optimum. Positive lower bounds for SAME-L under both latent-fraction and coordinate-rate matching extend this finding across tokenizers.

To assess prediction from compact inputs, we select a view for each horizon along a separate input frontier, subject to a 5% training-operator budget on local reconstruction regret. Held-out prediction improves over the reconstruction-optimal input view at the same rank in 9/10 tokenizer–rank–horizon conditions (Appendix G). These gains and the operator compromise motivate demand-specific views within a larger shared union at fixed consumer input width (Eq. 2).

5.2 Cross-tokenizer recurrence of demand relations

Using the invariant relations of Proposition 1, we find agreement across tokenizers without axis alignment. On 32 Moises guitar/piano demands reserved from Atlas construction, Locodec, SAME-L, and Music2Latent have pairwise Pearson correlations of .981–.995 between off-diagonal target-canonical weighted relations. Agreement extends to rank-4 projector relations (.942–.968), which compare leading subspaces without eigenvalue weighting (Figure 2b). Factor-preserving permutation tests support agreement on both measures in all three pairs. These demands' relations to the 64-demand development grid recur across tokenizers without target whitening (Appendix Table 18).

On the 64-demand grid, separate controls for mean or isotropic operator components and for target means and sample counts retain weighted correlations above .95. Omitting individual source, band, or horizon levels also preserves agreement (Appendix F.2). To examine recurrence without a shared teacher, we use each tokenizer's own future tokens as targets. Across five matched physical horizons, Locodec–SAME-L weighted relations have Pearson correlations of .863 on MUSDB and .977 on Moises (exact horizon-permutation $p = .0083$ each).

On Slakh with MERT targets, cross-tokenizer agreement recurs across 48 source–layer–horizon demands. Canonical weighted and rank-4 projector Spearman correlations for Locodec–SAME-L are .878 and .770, supported by factor-preserving permutation tests. With whitening, ridge, and operators re-estimated, both prespecified correlation thresholds are met in all 200 paired song-bootstrap refits and six 50%/75% training-song subsets across three sampling seeds (Appendix Table 23).

Within each tokenizer, demands favor distinct leading subspaces. Pairwise loading gaps are mean normalized utility differences between separate and shared rank-4 optima for truncated operators (Appendix F, Eq. 39). On the 32 Moises demands, medians across tokenizers span .332–.400 for canonical rank-4 chance-adjusted overlap and .185–.218 for the loading gap. We next test utility retention on the separate 25-demand construction family, fitting each Atlas to its tokenizer's operators.

5.3 Sparse realization and certificate tightness

At fixed active width, fitted Atlas views retain 83.5–96.5% of the single-to-dense reduction in worst-demand operator regret across six primary MUSDB/Moises cells at $M/K = 4$, with lower regret than coordinate selection in the unrotated union basis and all 64 Haar controls per cell (Appendix Table 15). Positive closure persists on Moises across 18 construction refits (six 50%/75% song subsets across three tokenizer/rank settings) and three independently trained Locodec models (Appendix Tables 16–17). We compare AJD with direct dictionary refinement using the hard-top-K worst-demand regret objective in the same union at $M/K = 4$. Evaluation regret is lower for AJD in all three Moises cells and for direct refinement in all three MUSDB cells (Appendix Table 6).

Support overlap measures coordinate reuse (Proposition 2); in the fitted Atlases, it closely tracks the ordering of dense-view overlaps. On Moises, Spearman correlations are .950–.984 across three tokenizer/rank settings at $M/K \in \{3, 4\}$, exceeding those of all 64 Haar controls per cell. At $M/K = 4$, the Locodec $K = 32$ and SAME-L $K = 10$ supports collectively use 118/128 and 40/40 atoms, respectively: compact inputs draw on a larger shared bank. Representing native views by atom indices saves 50.8%/71.4% in interface storage relative to factorized dense for Locodec $K = 32$/SAME-L $K = 26$, using 25 demands, float32 matrices, and uint16 indices (Appendix F.3, Table 10).

Proposition 3's within-union utility bounds quantify how much improvement can be certified, yielding worst-demand regret upper bounds and closure lower bounds (Appendix Eq. 26). On construction operators in 12 MUSDB/Moises cells at $M/K \in \{3, 4\}$, support-aware bounds fall below trivial caps for all 300 utility losses and 3,591/3,600 pairwise overlap errors. Under frozen evaluation, they certify positive worst-demand closure in 10/12 cells, with a median certified-to-observed closure ratio of 54.1% across all 12 (Appendix E). We next test the fitted interfaces with matched nonlinear consumers.

5.4 Nonlinear prediction and downstream utility

On Moises, Atlas improves future-token and frequency-band embedding prediction over Single with matched nonlinear consumers. In future-token prediction, Atlas retains 86.8% and 100.0% of the dense-over-single improvement for Locodec and SAME-L, respectively. Atlas has lower cosine error than both single and persistence in every model seed (Appendix Table 19). For SAME-L, this benefit persists with the Moises interfaces and consumers frozen: on 800 FMA-small test tracks, Atlas retains 96.4% of the dense-over-single improvement and has lower error than persistence (Appendix K).

With PupuJEPA band targets fixed across views, the native ablation removes Procrustes alignment to test prediction directly from stored atom coordinates. On the same Moises split, native SAME-L Atlas retains 85.3% of the single-to-native-dense improvement. Its cosine-error difference from Single is $-.01159$ (paired 95% CI $[-.01339, -.00987]$; Figure 3). All three seed means and all four band means favor Atlas over Single, supporting direct use of stored coordinates for this task.

Reselecting shared addresses also improves downstream tagging within a fixed atom bank. Using only MTG-Jamendo's official training partition, we select a 16-atom subset shared by Genre and Mood/theme from a frozen 64-atom Music2Latent dictionary fitted on Moises, then train separate task consumers. This shared subset improves mean development macro AUROC over the source-construction support by $+.01519$ and $+.01155$, respectively, and yields higher mean scores than task-specific addresses for both tasks (Table 2; Appendix P.1). Comparisons with full/PCA/joint dense inputs vary by task and fold (Appendix P.2).

On Slakh/MERT, demand relations recur across Locodec and SAME-L; with separately fitted interfaces and consumers, paired 95% CIs for Atlas–Single prediction-error differences include zero for both (Appendix L). For the primary construction-related prediction tasks, compact coordinate access retains most of the measured dense-view benefit from a larger shared union.

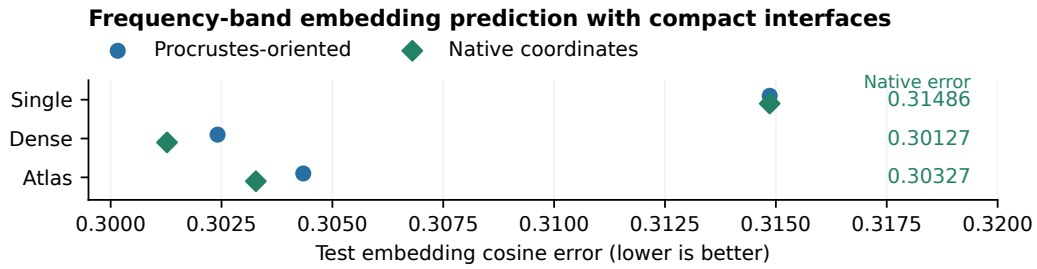

Figure 3: Native and Procrustes-oriented SAME-L views for PupuJEPA frequency-band embedding prediction on MoisesDB ($K = 26$; dense/Atlas union $M = 104$). Points show song–band mean errors over three seeds; Single is shared across conditions. Native Atlas retains 85.3% of the single-to-native-dense improvement. Its error difference from Single has a paired song-bootstrap 95% CI $[-.01339, -.00987]$. The centroid control (.41327) is off scale. Exact values: Appendix Table 20.

Table 2: Native-address adaptation on MTG-Jamendo with matched final consumer training. Macro AUROC over 20 selected labels per task, averaged over two outer development folds and two fresh seeds (higher is better; numerical column maxima in bold). Address reselection uses inner artist folds; shared subsets feed separate task consumers.

Interface	Input dim.	Genre	Mood/theme
Source construction	16	.75470	.67069
Optimized shared	16	**.76989**	.68224
Task-specific (own)	16	.76789	.68069
Other task's subset (swapped)	16	.76211	.68509
Full	64	.76184	**.68510**
PCA	16	.76572	.66959
Joint dense	16	.76354	.67994
Random native	16	.75518	.67469

6 Limitations and Conclusion

Shared capacity can relieve the compromise of a common compact view without increasing each consumer's input width. Our framework quantifies this compromise and the utility loss from restricting demand-specific dense views to K-of-M atom subsets. Within a fixed union, the chosen dense-view projectors admit exact realization by subsets of a common orthonormal basis if and only if they commute pairwise. Support-aware certificates bound the loss in operator utility from approximate realization.

Demand relations recur across the three evaluated tokenizers without axis alignment. At fixed active width, Atlas retains 84–97% of the dense-over-single reduction in worst-demand operator regret across six primary cells at $M/K = 4$. With matched nonlinear consumers, Atlas improves future-token and frequency-band embedding prediction over the shared view on Moises. For SAME-L, native Atlas retains 85.3% of the improvement from single to native dense on the band task, supporting prediction directly from stored atom coordinates.

The certificates concern fitted operators, and invariance of target-canonical relations assumes invertible linear coordinate changes under the stated whitening and ridge conditions. In the Jamendo development setting, reselecting shared addresses within a frozen dictionary improves mean macro AUROC over source-construction supports. Testing interface transfer across tokenizers and assessing end-to-end deployment efficiency are directions for future work. The primary results support sharing capacity across known demands through compact coordinate access that retains much of the measured dense-view benefit.

ETHICS STATEMENT

The approximately 20,000-hour source collection is an in-house research corpus whose use was authorized under the company's internal data-governance and access policies; access was restricted to the controlled training environment. Experiments use a fixed 20,214-track (1,080.6-hour) subset, including 18,201 training tracks. Public evaluations use MUSDB18-HQ, MoisesDB v0.1, Slakh2100-redux, FMA, EmoMusic, GiantSteps, GTZAN, MagnaTagATune, NSynth, VocalSet, MTG-Jamendo, and the FMA-derived OpenMIC-2018 annotations under their research terms. We redistribute no audio, stems, token arrays, captions, lyrics, or private metadata. Private-corpus manifests contain only hashed identifiers and non-content fields; the public replay inventory retains public MoisesDB identifiers and crop offsets needed for replay. The OpenMIC prediction extension retains public sample keys and integer artist groups, but no labels or artist identities. Source-separation corpora and synthesized Slakh audio do not represent all musical traditions, performance practices, or production styles, and frozen targets can inherit their own biases. The study deploys no generator and claims no perceptual improvement. Any model release requires rights review, privacy and memorization screening, and broader evaluation.

REPRODUCIBILITY STATEMENT

The supplement specifies preprocessing, data roles, estimators, budgets, and proofs; the artifact supplies locks, atom sets, seed-level outcomes, gate results, and verifiers. For the public band-embedding experiment it also bundles the exact public inventory, compact SAME-L interface state, encoding/evaluation code, and all nine validation-selected TCN checkpoints. These enable consumer replay from upstream MoisesDB audio and SAME-L/PupuJEPA weights. A song-indexed draw fixes SAME-L bottleneck noise; a fresh raw-audio audit preserves every gate and reproduces the three method means within 2.4×10^{-6}. This is conclusion-level numerical reproduction, not bitwise recovery of the original cached draw.

A separate hash-locked full-refit path starts from the same public waveforms and released checkpoints, rebuilds all 25 operators and the Atlas, retrains all nine consumers, and passes both validation gates and the held-out test. A locked follow-up removes the Procrustes transforms, records the ordered Atlas indices, retrains the six native dense/Atlas consumers, and again passes both gates and test. This is end-to-end computational reproduction, not an independent corpus/task replication. The released nine checkpoints belong to the earlier frozen-interface run, not the full-refit or native runs. The latter are supported by result-level records, not model replay. Most other experiments remain result-level reaggregations, and exact Locodec retraining requires authorized in-house data.

For the original OpenMIC comparison and retrospective seven-arm joint-classifier audit, the artifact supplies logits/probabilities, public sample keys, pseudonymous artist groups, split roles, construction/checkpoint/calibration hashes, and test-access locks. A CPU-only script uses upstream labels to recompute seed/ensemble AUROC/AP, original calibrated risks, and the paired artist-bootstrap intervals. This is prediction-level reproduction, not checkpoint or training replay; OpenMIC checkpoints, labels, and token caches are not bundled.

Sanitized GTZAN, NSynth, and VocalSet follow-ups remain result-level records. Jamendo development bundles retain arm/seed/fold aggregates and native supports, including reused references and policy aliases. They cover dictionary and fresh-consumer controls, address/history crossover, cross-artist selection, and matched-baseline evaluations; CPU-only verifiers reaggregate the reported comparisons. These are scalar-result records, not candidate-search replay. They contain no labels, predictions, tokens, or checkpoints and provide no new official-test confirmation.

No public repository or model-host identity is disclosed during double-blind review. Subject to upstream licenses and rights review, author-written scripts and permitted checkpoints will be released publicly after acceptance. Appendix Q gives the submission-time availability matrix.

ARTIFICIAL INTELLIGENCE USE STATEMENT

Generative artificial intelligence tools assisted literature discovery, hypothesis refinement, mathematical formulation, experimental design, code implementation and debugging, result organization and

interpretation, and language editing. The authors specified the objectives, inspected primary sources, ran the experiments, verified code, formulas, values, citations, and claims, and made all inclusion and interpretation decisions. Artificial intelligence tools did not create evaluation labels and do not replace author responsibility.

REFERENCES

Alessandro Achille, Michael Lam, Rahul Tewari, Avinash Ravichandran, Subhransu Maji, Charless C. Fowlkes, Stefano Soatto, and Pietro Perona. Task2Vec: Task embedding for meta-learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6430–6439, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Achille_Task2Vec_Task_Embedding_for_Meta-Learning_ICCV_2019_paper.html.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008. doi: 10.1007/s10994-007-5040-8.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: A meta-algorithm and applications. *Theory of Computing*, 8(6):121–164, 2012. doi: 10.4086/toc.2012.v008a006.

Georgios Arvanitidis, Miguel González-Duque, Alison Pouplin, Dimitrios Kalatzis, and Søren Hauberg. Pulling back information geometry. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pp. 4872–4894, 2022.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop at ICML*, 2019. URL https://repositori.upf.edu/handle/10230/42015.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. arXiv:2209.03143.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning multi-level features with matryoshka sparse autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 6077–6101, 2025.

Jean-François Cardoso and Antoine Souloumiac. Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 17(1):161–164, 1996. doi: 10.1137/S0895479893259546.

Yuqing Cheng, Xingyu Ma, Guochen Yu, and Xiaotao Gu. Modeling music as a time-frequency image: A 2D tokenizer for music generation. *arXiv preprint arXiv:2605.15831*, 2026.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/94b472a1842cd7c56dcb125fb2765fbd-Abstract-Conference.html.

Chandler Davis and William M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970. doi: 10.1137/0707001.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *Proceedings of the 18th International Society for Music Information Retrieval Conference*, pp. 316–323, 2017. arXiv:1612.01840.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research*, 2023. arXiv:2210.13438.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with WaveNet autoencoders. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 1068–1077, 2017. URL https://proceedings.mlr.press/v70/engel17a.html.

Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949. doi: 10.1073/pnas.35.11.652.

Bernhard N. Flury. Common principal components in k groups. *Journal of the American Statistical Association*, 79(388):892–898, 1984. doi: 10.1080/01621459.1984.10477108.

Yicheng Gu, Junan Zhang, Jerry Li, Zhizheng Wu, and Lauri Juvela. Frequency-aware self-supervised music representation learning. *arXiv preprint arXiv:2606.25713v2*, 2026. URL https://arxiv.org/abs/2606.25713v2. PupuJEPA version.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012.

Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. doi: 10.2307/2333955.

Eric J. Humphrey, Simon Durand, and Brian McFee. OpenMIC-2018: An open dataset for multiple instrument recognition. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018.

Michael F. Hutchinson. A stochastic estimator of the trace of the influence matrix for laplacian smoothing splines. *Communications in Statistics—Simulation and Computation*, 18(3):1059–1076, 1989. doi: 10.1080/03610918908812806.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975. doi: 10.1016/0047-259X(75)90042-1.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, Ziang Zhang, Xiaoda Yang, Rongjie Huang, Yidi Jiang, Qian Chen, Siqi Zheng, and Zhou Zhao. WavTokenizer: An efficient acoustic discrete codec tokenizer for audio language modeling. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/ea1f5f0878d43ff4fb8bf64ef4a2326c-Abstract-Conference.html.

James E. Kelley, Jr. The cutting-plane method for solving convex programs. *Journal of the Society for Industrial and Applied Mathematics*, 8(4):703–712, 1960. doi: 10.1137/0108053.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 3519–3529, 2019.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008. URL https://www.frontiersin.org/journals/systems-neuroscience/articles/10.3389/neuro.06.004.2008/full.

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

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Sébastien Lachapelle, Tristan Deleu, Divyat Mahajan, Ioannis Mitliagkas, Yoshua Bengio, Simon Lacoste-Julien, and Quentin Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 18171–18206, 2023.

Jianze Li, Konstantin Usevich, and Pierre Comon. Globally convergent jacobi-type algorithms for simultaneous orthogonal symmetric tensor diagonalization. *SIAM Journal on Matrix Analysis and Applications*, 39(1):1–22, 2018. doi: 10.1137/17M1116295.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/33dffa2e3d2ab74a783d1a8c292f66d9-Abstract-Conference.html.

Rui Lin, Zhiyue Wu, Jiahe Lei, Kangdi Wang, Weixiong Chen, Junyu Dai, and Tao Jiang. DuoTok: Source-aware dual-track music tokenization for vocal-accompaniment generation. *arXiv preprint arXiv:2511.20224v3*, 2025. URL https://arxiv.org/abs/2511.20224v3. Version revised September 2026.

Yi Luo, Rongzhi Gu, and Jixun Yao. Stable autoregressive speech generation with low-frame-rate high-dimensional continuous tokens. *arXiv preprint arXiv:2607.29363*, 2026.

Takanori Maehara and Kazuo Murota. Algorithm for error-controlled simultaneous block-diagonalization of matrices. *SIAM Journal on Matrix Analysis and Applications*, 32(2):605–620, 2011. doi: 10.1137/090779966.

Ethan Manilow, Gordon Wichern, Prem Seetharaman, and Jonathan Le Roux. Cutting music source separation some Slakh: A dataset to study the impact of training data quality and quantity. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. IEEE, 2019.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020. URL https://www.jmlr.org/papers/v21/17-678.html.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. Sparse coding for multitask and transfer learning. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28, pp. 343–351, 2013.

Pooneh Mousavi, Gallil Maimon, Adel Moumen, Darius Petermann, Jiatong Shi, Haibin Wu, Haici Yang, Anastasia Kuznetsova, Artem Ploujnikov, Ricard Marxer, Bhuvana Ramabhadran, Benjamin Elizalde, Loren Lugosch, Jinyu Li, Cem Subakan, Phil Woodland, Minje Kim, Hung-yi Lee, Shinji Watanabe, Yossi Adi, and Mirco Ravanelli. Discrete audio tokens: More than a survey! *Transactions on Machine Learning Research*, 2025.

Philipp Nazari, Sebastian Damrich, and Fred A. Hamprecht. Geometric autoencoders—what you see is what you decode. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, 2023.

Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996. doi: 10.1038/381607a0.

Michael L. Overton and Robert S. Womersley. On the sum of the largest eigenvalues of a symmetric matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):41–45, 1992. doi: 10.1137/0613006.

Julian D. Parker, Zach Evans, CJ Carr, Zachary Zukowski, Josiah Taylor, Matthew Rice, and Jordi Pons. SAME: A semantically-aligned music autoencoder. *arXiv preprint arXiv:2605.18613*, 2026.

Marco Pasini, Stefan Lattner, and George Fazekas. Music2Latent: Consistency autoencoders for latent audio compression. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2024. arXiv:2408.06500.

Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl. MoisesDB: A dataset for source separation beyond 4-stems. In *Proceedings of the 24th International Society for Music Information Retrieval Conference*, 2023.

Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner. MUSDB18-HQ: An uncompressed version of MUSDB18, 2019. URL https://zenodo.org/records/3338373.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Oren Rippel, Michael Gelbart, and Ryan P. Adams. Learning ordered representations with nested dropout. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 1746–1754, 2014.

Simon Rouard, Manu Orsini, Axel Roebel, Neil Zeghidour, and Alexandre Défossez. Continuous audio language models. In *International Conference on Learning Representations*, 2026.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958. doi: 10.2140/pjm.1958.8.171.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560. URL https://ieeexplore.ieee.org/document/1021072/.

Kangdi Wang, Yusheng Dai, and Jin Xu. Fourier is frontier: Frequency-aware autoencoding for high-fidelity music reconstruction. *arXiv preprint arXiv:2608.19843*, 2026.

Lu Wang, Hao Chen, Siyu Wu, Zhiyue Wu, Hao Zhou, Chengfeng Zhang, Ting Wang, and Haodi Zhang. AudioCodecBench: A comprehensive benchmark for audio codec evaluation. *arXiv preprint arXiv:2509.02349*, 2025.

Julia Wilkins, Prem Seetharaman, Alison Wahl, and Bryan Pardo. VocalSet: A singing voice dataset. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, 2018. URL https://archives.ismir.net/ismir2018/paper/000114.pdf.

Dongchao Yang, Yuanyuan Wang, Dading Chong, Songxiang Liu, Xixin Wu, and Helen M. Meng. UniAudio 2.0: A unified audio language model with text-aligned factorized audio tokenization. *arXiv preprint arXiv:2602.04683*, 2026.

Shu-Wen Yang, Byeonggeun Kim, Kuan-Po Huang, Qingming Tang, Huy Phan, Bo-Ru Lu, Harshavardhan Sundar, Shalini Ghosh, Hung-Yi Lee, Chieh-Chi Kao, and Chao Wang. Generative audio language modeling with continuous-valued tokens and masked next-token prediction. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 70793–70812, 2025.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Roger Dannenberg, Wenhu Chen, Gus Xia, Wei Xue, Si Liu, Shi Wang, Ruibo Liu, Yike Guo, and Jie Fu. MARBLE: Music audio representation benchmark for universal evaluation. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7cbeec46f979618beafb4f46d8f39f36-Abstract-Datasets_and_Benchmarks.html.

Dinghao Zhou, Xingchen Song, Di Wu, Pengyu Cheng, Shengfan Shen, and Sixiang Lv. F3-Tokenizer: Taming audio autoencoder latents for understanding and generation. *arXiv preprint arXiv:2606.06357*, 2026.

A Notation and Compatibility Certificates

This appendix follows the main argument: assumptions and proofs establish the geometric statements, numerical comparisons assess certificate tightness and sparse construction, and consumer studies test the scope of interface utility. Protocol details and internal experiment identifiers link these comparisons to their frozen designs and result records. Utility operators are real symmetric; $\|\cdot\|_2$, $\|\cdot\|_F$, and $\|\cdot\|_*$ denote spectral, Frobenius, and nuclear norms, respectively. Ranks are integers in $[1, n]$; statements using λ_{K+1} require $K < n$. Covariances use centered variables with finite second moments.

For a symmetric matrix $H \in \mathbb{R}^{n\times n}$, define its rank-$K$ Ky Fan functional $u_K(H) = \sum_{j=1}^{K} \lambda_j^{\downarrow}(H) = \max_{P\in\mathcal{P}_K^n} \operatorname{tr}(PH)$ (Fan, 1949). Let $\mathcal{F}_K^n = \{X : 0 \preceq X \preceq I, \operatorname{tr}(X) = K\}$ be the Fantope, the convex hull of $\mathcal{P}_K^n$ (Overton & Womersley, 1992). For normalized consumer operators $B_c = A_c/\operatorname{tr}(A_c)$, let $u_{c,K} = u_K(B_c)$. For any probability vector α over consumers and any feasible rank-K projector $\widehat{P}$,

$$\sum_c \alpha_c u_{c,K} - \sum_{j=1}^{K} \lambda_j^{\downarrow}\left(\sum_c \alpha_c B_c\right) \le \Xi_K \le \max_c\{u_{c,K} - \operatorname{tr}(\widehat{P}B_c)\}. \tag{14}$$

The maximum over consumers dominates every convex combination. Minimizing the resulting weighted objective selects the top-K eigenspace of the operator mixture, giving the lower bound; evaluating any feasible projector gives the upper bound. More precisely, write the finite maximum as a maximization over the simplex. Applying the minimax theorem to the compact convex Fantope and simplex (Sion, 1958) gives

$$\begin{aligned} \Xi_K^{\mathrm{cvx}} &= \min_{X\in\mathcal{F}_K^n} \max_c\{u_{c,K} - \operatorname{tr}(XB_c)\} \\ &= \max_{\alpha\in\Delta_C} g(\alpha) \le \Xi_K, \\ g(\alpha) &= \sum_c \alpha_c u_{c,K} - u_K\left(\sum_c \alpha_c B_c\right). \end{aligned} \tag{15}$$

Thus an upper bound on the dual optimum is not an upper bound on the original rank-projector problem; that endpoint must come from a feasible $\widehat{P}$.

For $P_\alpha = \operatorname{Top}_K(\sum_c \alpha_c B_c)$, $s_c(\alpha) = u_{c,K} - \operatorname{tr}(P_\alpha B_c) \in [0, 1]$ is a supergradient of g. Indeed, the Ky Fan variational formula gives, for every $\beta \in \Delta_C$, $g(\beta) \le \beta^\top u - \operatorname{tr}[P_\alpha \sum_c \beta_c B_c] = g(\alpha) + s(\alpha)^\top(\beta - \alpha)$. A uniform-start exponentiated update $\alpha_{t+1,c} \propto \alpha_{t,c}\exp(\eta s_c(\alpha_t))$ therefore obeys (Arora et al., 2012)

$$\Xi_K^{\mathrm{cvx}} - \max_{t\le T} g(\alpha_t) \le \frac{\log C}{\eta T} + \frac{\eta}{8}. \tag{16}$$

For the prespecified $C = 4$, $T = 120$, and $\eta \in \{.1, .3, 1\}$, the best a-priori bound is .0760. Equation 16 concerns the best dual iterate; the frozen implementation retains the weights paired with the best feasible projector. Each retained α and $\widehat{P}$ supplies a valid endpoint in Eq. 14, and the following audit evaluates their tightness.

We therefore audit the frozen operators independently with Kelley supporting-plane cuts (Kelley, 1960). Concavity gives $g(\beta) \le g(\alpha_i) + s(\alpha_i)^\top(\beta - \alpha_i)$. With $b_i = g(\alpha_i) - s(\alpha_i)^\top\alpha_i$ and rows $S_i = s(\alpha_i)^\top$, finite linear-programming duality yields

$$\max_{\beta\in\Delta_C} \min_i\{b_i + S_i\beta\} = \min_{\lambda\in\Delta_m} \{b^\top\lambda + \max_c(S^\top\lambda)_c\}. \tag{17}$$

The left side globally upper-bounds $\max_\alpha g(\alpha)$, while the best evaluated g is a lower bound; an explicitly feasible λ evaluates the right side without relying on premature convergence of the master solver. Together these quantities bracket Ξ_K^{cvx} to at most 9.7×10^{-8} in all reported cells. Spectral-oracle projectors then give the tightened intervals in Table 14. These certificates use float64 numerical evaluation.

The compatibility bounds apply to the frozen local operators on future-token coordinates. Let native token $\widetilde{h}_t$ have mean μ_h and covariance $C_h = \Sigma_{hh} \succ 0$. As in Section 3, write the centered token as

$h_t = \widetilde{h}_t - \mu_h$ and its whitened coordinate as $x_t = C_h^{-1/2} h_t$. For decoder D_{dec}, let $J_t = \partial D_{\mathrm{dec}} / \partial \widetilde{h}_t$ and let $T_{\widetilde{h}_t}$ be the native spherical-tangent projector (or identity for a nonspherical tokenizer). With positive-semidefinite waveform metric W, the prespecified per-token reconstruction pullback and output-side ridge-prediction operator are (Hoerl & Kennard, 1970; Arvanitidis et al., 2022):

$$A_{\mathrm{rec}} = \mathbb{E}_{t,\widetilde{h}}[C_h^{1/2} T_{\widetilde{h}_t} J_t^\top W J_t T_{\widetilde{h}_t} C_h^{1/2}], \tag{18}$$

$$A^{\mathrm{out}}_{\mathrm{pred},\ell} = C_{yx,\ell}(C_{xx,\ell} + \rho I)^{-1} C_{xy,\ell}. \tag{19}$$

Thus A_{rec} averages diagonal token blocks of the decoder pullback; it is not the full window-level matrix, whose off-diagonal blocks couple distinct times. Equivalently, $\widehat{J}_t = J_t T_{\widetilde{h}_t} C_h^{1/2}$ and $A_{\mathrm{rec}} = \mathbb{E}[\widehat{J}_t^\top W \widehat{J}_t]$. In implementation, Rademacher-derived source/band probes v define $W = \mathbb{E}[vv^\top]/L$, where L is the waveform dimension (Hutchinson, 1989). The native gradient $g_t = \nabla_{\widetilde{h}_t}(v^\top D_{\mathrm{dec}}/\sqrt{L})$ is first tangent-projected and then mapped to whitened coordinates as $g_{t,\mathrm{w}} = C_h^{1/2} T_{\widetilde{h}_t} g_t$. Its outer product has expectation $\mathbb{E}_v[g_{t,\mathrm{w}} g_{t,\mathrm{w}}^\top] = \widehat{J}_t^\top W \widehat{J}_t$, yielding the displayed reconstruction block.

For the output-side prediction operator, a linear map $B^\top x$ with regularized squared loss gives a separate identity: minimizing $\mathbb{E}\|y - B^\top x\|_2^2 + \rho\|B\|_F^2$ subtracts $\mathrm{tr}(A^{\mathrm{out}}_{\mathrm{pred},\ell})$ from $\mathrm{tr}(C_{yy,\ell})$. Writing $H = C_{xx,\ell} + \rho I$ and $B_\star = H^{-1} C_{xy,\ell}$, completing the square gives

$$\mathbb{E}\|y - B^\top x\|_2^2 + \rho\|B\|_F^2 = \mathrm{tr}(C_{yy,\ell}) - \mathrm{tr}(C_{yx,\ell} H^{-1} C_{xy,\ell}) + \mathrm{tr}[(B - B_\star)^\top H (B - B_\star)].$$

This identity concerns the regularized optimum, not its unpenalized test loss. A deployed view instead restricts current-token inputs. Its loading operator is $M^{\mathrm{in}}_{\mathrm{pred},\ell} = R_\ell R_\ell^\top$, where $R_\ell = (C_{xx,\ell} + \rho I)^{-1} C_{xy,\ell}$ with $\rho > 0$. The practical input frontier therefore addresses a different optimization problem from the output certificate. With B_{rec} and $B^{\mathrm{in}}_{\mathrm{pred},\ell}$ trace-normalized, candidate views are

$$P_{K,\ell}(\alpha) = \mathrm{Top}_K[(1-\alpha) B_{\mathrm{rec}} + \alpha B^{\mathrm{in}}_{\mathrm{pred},\ell}] \tag{20}$$

for $\alpha \in [0, 1]$. We select a view using training operators under a 5% reconstruction-regret constraint, then evaluate its projected inputs on held-out songs with the frozen ridge predictor.

B Prespecified Operator Estimators and Coordinate Maps

With whitened input x, construction uses ordinary-target ridge $R_d = (\Sigma_{xx,d} + \rho_d I)^{-1} \Sigma_{xy,d}$, preserving the target metric. For an orthogonal projector P, the score $\mathrm{tr}(R_d R_d^\top P) = \|P R_d\|_F^2$ measures retained loading energy, distinct from the risk of a refitted consumer. Canonical comparison instead right-multiplies R_d by the exact target whitener to obtain $R_d^\star$ in Eq. 3, removing dependence on invertible linear target coordinates.

The two native-to-whitened congruences point in different directions because a source difference is a vector whereas a loss gradient is a covector. With $h = \widetilde{h} - \mu_h$ and $x = C_h^{-1/2} h$,

$$A_{\Delta,x} = C_h^{-1/2} A_{\Delta,\widetilde{h}} C_h^{-1/2}, \qquad A_{g,x} = C_h^{1/2} A_{g,\widetilde{h}} C_h^{1/2}. \tag{21}$$

The first identity follows from $\Delta x = C_h^{-1/2} \Delta \widetilde{h}$; the second follows from $\nabla_x \mathcal{L} = C_h^{1/2} \nabla_{\widetilde{h}} \mathcal{L}$. It explains the inverse-covariance congruence for source interventions and covariance congruence for GN/decoder pullbacks. All reported canonical input maps had an inactive numerical eigenvalue floor; every reported target covariance was full rank and no target eigenvalue was clipped.

The long-range operator measures local sensitivity of the relation loss. For nonzero anchor a and candidates c_j, temperature $T > 0$, logits $s_j = \cos(a, c_j)/T$, and $p = \mathrm{softmax}(s)$, define $g_{0j} = \nabla_a s_j$, $g_{jj} = \nabla_{c_j} s_j$, and $\bar{g}_0 = \sum_j p_j g_{0j}$. The softmax cross-entropy output curvature is $\mathrm{diag}(p) - pp^\top$. Its anchor block and candidate diagonal blocks therefore sum to

$$\mathrm{GN}_{\mathrm{local}} = \mathbb{E}\left[\sum_j p_j (g_{0j} - \bar{g}_0)(g_{0j} - \bar{g}_0)^\top + \sum_j p_j (1 - p_j) g_{jj} g_{jj}^\top\right] \succeq 0. \tag{22}$$

This is the exact diagonal-block sum of the generalized GN matrix (Martens, 2020), not the full temporal curvature: anchor–candidate and candidate–candidate cross-time blocks are deliberately excluded. For cross-entropy, this curvature depends on p rather than the teacher target; the teacher still defines the relation loss and its held-out actual risk.

C Proofs for Target-Canonical Demand Geometry

Proof of Proposition 1. Let an unwhitened input have global covariance $S_x \succ 0$ and whitened coordinate $x = S_x^{-1/2}(\widetilde{x} - \mu_x)$. After $\widetilde{x}' = G\widetilde{x}$ with invertible G, recomputed whitening gives

$$x' = (GS_xG^\top)^{-1/2}GS_x^{1/2}x = Q_x^\top x,$$

where $Q_x^\top = (GS_xG^\top)^{-1/2}GS_x^{1/2}$ is orthogonal and is common to all demands. For a centered target $\widetilde{y}_d$ with covariance $S_{y,d} \succ 0$, define its whitened version $z_d = S_{y,d}^{-1/2}\widetilde{y}_d$. The independently transformed target $\widetilde{y}'_d = H_d\widetilde{y}_d$ similarly satisfies $z'_d = O_d^\top z_d$ for an orthogonal O_d. Consequently, if $\widetilde{\Sigma}_{xy,d} = \mathrm{Cov}(x, z_d) = \Sigma_{xy,d}\Sigma_{yy,d}^{-1/2}$, then $\Sigma_{x'x',d} = Q_x^\top\Sigma_{xx,d}Q_x$ and $\widetilde{\Sigma}_{x'z',d} = Q_x^\top\widetilde{\Sigma}_{xy,d}O_d$. Since Eq. 3 is equivalently $R_d^\star = (\Sigma_{xx,d} + \rho_d I)^{-1}\widetilde{\Sigma}_{xy,d}$, scalar ridge is equivariant:

$$R_d^{\star\prime} = Q_x^\top R_d^\star O_d, \qquad M_d^{\star\prime} = Q_x^\top M_d^\star Q_x.$$

The same Q_x applies to every demand within a tokenizer, while O_d cancels from the loading operator. Cyclicity of trace and unitary invariance of Frobenius norm prove invariance of the relation statistics. If ridge is chosen from a fixed grid by whitened validation risk, every candidate risk is unchanged, so deterministic tie-breaking preserves the selected value. The claim keeps each ρ_d fixed, or uses this invariant selection rule; an arbitrary coordinate-dependent ridge rule or active eigenvalue clipping does not satisfy its assumptions. □

Let $M_{d,r}^\star = \Pi_{d,r}M_d^\star\Pi_{d,r}$ denote the leading rank-r spectral truncation, assumed nonzero and separated by an eigengap at r. The complete relation family additionally includes

$$\mathcal{N}_{de}^{(r)} = \frac{\|M_{d,r}^\star M_{e,r}^\star - M_{e,r}^\star M_{d,r}^\star\|_F}{\|M_{d,r}^\star\|_F\|M_{e,r}^\star\|_F}, \tag{23}$$

$$\mathcal{G}_{de}^{(r)} = 1 - \max_{P\in\mathcal{P}_r^N}\frac{1}{2}\left[\frac{\mathrm{tr}(PM_{d,r}^\star)}{\mathrm{tr}(M_{d,r}^\star)} + \frac{\mathrm{tr}(PM_{e,r}^\star)}{\mathrm{tr}(M_{e,r}^\star)}\right]. \tag{24}$$

$\mathcal{N} = 0$ is insufficient for an exactly compatible progressive order: diagonal operators may commute while assigning their largest eigenvalues to different axes.

D Proofs for Sparse Realization

Proof of Proposition 2. If the projectors share a diagonalizing orthonormal basis, they commute. Conversely, real symmetric pairwise-commuting projectors admit such a basis, with K ones and $M - K$ zeros per projector. Multiplying their diagonal indicator matrices gives the overlap identity.

Every pair of Atlas coordinate projectors commutes. Decompose

$$[\bar{\Pi}_d, \bar{\Pi}_e] = [\bar{\Pi}_d - \bar{P}_d^{\mathrm{A}}, \bar{\Pi}_e] + [\bar{P}_d^{\mathrm{A}}, \bar{\Pi}_e - \bar{P}_e^{\mathrm{A}}].$$

For a projector P, $\|[E, P]\|_F \le 2\|E\|_F$. Therefore $\|[\bar{\Pi}_d, \bar{\Pi}_e]\|_F \le 2(\|\bar{\Pi}_d - \bar{P}_d^{\mathrm{A}}\|_F + \|\bar{\Pi}_e - \bar{P}_e^{\mathrm{A}}\|_F)$, proving Eq. 10. If C_d has leading rank-K eigengap γ_d, the spectral variational inequality gives

$$\mathrm{tr}[C_d(\bar{\Pi}_d - \bar{P}_d^{\mathrm{A}})] \ge \frac{\gamma_d}{2}\|\bar{\Pi}_d - \bar{P}_d^{\mathrm{A}}\|_F^2.$$

Indeed, for descending eigenvectors v_j of C_d, let $a_j = v_j^\top\bar{P}_d^{\mathrm{A}}v_j$. The mass $t = \sum_{j\le K}(1 - a_j) = \sum_{j>K}a_j$ loses at least $\gamma_d t$, while $\|\bar{\Pi}_d - \bar{P}_d^{\mathrm{A}}\|_F^2 = 2t$. Combining this with the commutator bound and $\gamma = \min(\gamma_d, \gamma_e)$ gives the factor $\gamma/32$. Positive eigengaps make the specified projectors unique; with ties, another leading-projector choice may admit exact realization. The per-demand obstruction does not imply $\Gamma_{\mathrm{atom}} > 0$: its loss may fit within another demand's worst-regret slack. □

Proof of Proposition 3. Let $P = P_d, Q = I - P, C = C_d, C_0 = PCP + QCQ$, and $X = C - C_0$. Ky Fan subadditivity (Fan, 1949) gives

$$u_K(C) = u_K(C_0 + X) \le u_K(C_0) + u_K(X).$$

Because $\mathrm{tr}(PC) = \mathrm{tr}(PC_0)$, subtracting this quantity proves Eq. 11. Choose an orthonormal basis $T = [T_P, T_Q]$ for $\mathrm{range}(P) \oplus \mathrm{range}(Q)$ and set $B = T_Q^\top C T_P \in \mathbb{R}^{(M-K)\times K}$. Then

$$T^\top X T = \begin{bmatrix} 0 & B^\top \\ B & 0 \end{bmatrix}, \qquad QCP = T_Q B T_P^\top.$$

Let $s = \mathrm{rank}(B) \le \min(K, M-K)$. The eigenvalues of X are s pairs $\pm\sigma_j(B)$ and $M - 2s \ge K - s$ zeros. Hence $u_K(X) = \sum_{j=1}^{s} \sigma_j(B) = \|QCP\|_*$ for every $1 \le K < M$, since B and QCP share nonzero singular values. Finally, $\bar{\Pi}_d = P_d + E_d$ and $\bar{\Pi}_e = P_e + E_e$ imply

$$\mathrm{tr}(\bar{\Pi}_d \bar{\Pi}_e - P_d P_e) = \mathrm{tr}(E_d P_e) + \mathrm{tr}(P_d E_e) + \mathrm{tr}(E_d E_e).$$

Division by K proves Eq. 12; the triangle inequality gives the stated certificate. □

Spectral-margin refinement and closure certificate. Assume $\eta = \lambda_K(C_0) - \lambda_{K+1}(C_0) > 0$ and let Q_0 and $Q_\star$ be the leading rank-K projectors of C_0 and $C = C_0 + X$, respectively. The eigengap makes Q_0 unique; because C_0 is block diagonal in the P/Q split, Q_0 is block diagonal as well and $\mathrm{tr}(Q_0 X) = 0$. The variational eigengap inequality and Cauchy–Schwarz give, with $t = \|Q_\star - Q_0\|_F$,

$$\begin{aligned} u_K(C) - u_K(C_0) &= \mathrm{tr}[(Q_\star - Q_0)C_0] + \mathrm{tr}[(Q_\star - Q_0)X] \\ &\le -\frac{\eta}{2} t^2 + t\|X\|_F \le \frac{\|X\|_F^2}{2\eta}. \end{aligned} \tag{25}$$

Taking the minimum with the independent Ky Fan bound $u_K(C) - u_K(C_0) \le u_K(X)$ proves Eq. 13. This is a global finite-dimensional inequality for the frozen projector; no iterative eigenspace approximation is used.

The same result bounds the paper's reported worst-demand gap closure. Set $B_d = A_d / \mathrm{tr}(A_d)$ and $H_d = U^\top B_d U$ for the frozen union. Let $r_d^U = u_K(B_d) - u_K(H_d)$ and let b_d be the right-hand side of Eq. 13 evaluated at H_d. Then $r_d^{\mathrm{A}} = r_d^U + u_K(H_d) - \mathrm{tr}(P_d H_d) \le r_d^U + b_d$. Unlike the AJD-normalized $C_d = H_d / \|H_d\|_F$, H_d retains the regret scale: a bound computed on C_d must be multiplied by $\|H_d\|_F$. If $H_d = 0$, its within-union loss and b_d are zero; if its block eigengap vanishes, use the first-order Ky Fan bound. Consequently, with $U_{\mathrm{A}} = \max_d(r_d^U + b_d)$ and frozen worst regrets $R_{\mathrm{single}} > R_{\mathrm{dense}}$,

$$\kappa = \frac{R_{\mathrm{single}} - R_{\mathrm{Atlas}}}{R_{\mathrm{single}} - R_{\mathrm{dense}}} \ge \frac{R_{\mathrm{single}} - U_{\mathrm{A}}}{R_{\mathrm{single}} - R_{\mathrm{dense}}}. \tag{26}$$

On held-out operators, r_d^U is the oracle optimum only inside the already frozen union; it selects no held-out basis or support. R_{dense} remains the regret of the frozen train-fitted dense view, so the denominator is the same comparator used in the reported closure.

Residual/eigengap comparison. For comparison with the support-aware result, define $\mathcal{P}_D(C) = D\,\mathrm{diag}(\mathrm{diag}(D^\top C D))D^\top$, $\epsilon_d = \|C_d - \mathcal{P}_D(C_d)\|_F$, and let $P_d = D E_{S_d} D^\top$ select the top-K diagonal coordinates of $\mathcal{P}_D(C_d)$. If $\gamma_d = \lambda_K(C_d) - \lambda_{K+1}(C_d) > 0$, the conventional norm chain gives

$$\Delta_d(P_d) \le \sqrt{K}\epsilon_d, \qquad \|\bar{\Pi}_d - P_d\|_F^2 \le \frac{2\sqrt{K}\epsilon_d}{\gamma_d}, \tag{27}$$

$$\left| \frac{\mathrm{tr}(\bar{\Pi}_d \bar{\Pi}_e)}{K} - \frac{|S_d \cap S_e|}{K} \right| \le \frac{\|\bar{\Pi}_d - P_d\|_F + \|\bar{\Pi}_e - P_e\|_F}{\sqrt{K}}. \tag{28}$$

These inequalities combine Ky Fan variation, a spectral eigengap argument related to Davis–Kahan (Davis & Kahan, 1970), and Cauchy–Schwarz. To verify them, write $\widehat{C}_d = \mathcal{P}_D(C_d)$ and $R_d = C_d - \widehat{C}_d$. Because P_d is diagonal in the D basis while R_d has zero diagonal, $\mathrm{tr}(P_d R_d) = 0$. Moreover, P_d is a top-K projector of $\widehat{C}_d$. Hence

$$\Delta_d(P_d) = u_K(C_d) - \mathrm{tr}(P_d C_d) = u_K(C_d) - u_K(\widehat{C}_d) \tag{29}$$

$$\le \max_{P \in \mathcal{P}_K^M} \mathrm{tr}(P R_d) \le \sqrt{K}\|R_d\|_F. \tag{30}$$

Combining this with the eigengap inequality in the preceding proof gives the projector bound. Finally,

$$|\operatorname{tr}(\bar{\Pi}_d\bar{\Pi}_e) - \operatorname{tr}(P_dP_e)| \le |\operatorname{tr}[(\bar{\Pi}_d - P_d)\bar{\Pi}_e]| + |\operatorname{tr}[P_d(\bar{\Pi}_e - P_e)]| \tag{31}$$

$$\le \sqrt{K}(\|\bar{\Pi}_d - P_d\|_F + \|\bar{\Pi}_e - P_e\|_F). \tag{32}$$

Since $\operatorname{tr}(P_dP_e) = |S_d \cap S_e|$, division by K completes the comparison. □

OPERATOR-ESTIMATION PROPAGATION TO INTERFACES

The preceding certificates condition on fitted operators. The following deterministic result propagates supplied operator-estimation radii to demand relations, support choices, and utility closure. It combines classical normalization, Ky Fan, and Davis–Kahan inequalities to quantify sensitivity of these interface properties.

Proposition 4 (operator perturbation propagation). Let $M_d^\star \neq 0$ and $\widehat{M}_d^\star \neq 0$ be reference and estimated canonical operators. Write $q_d = \|\widehat{M}_d^\star - M_d^\star\|_F / \|M_d^\star\|_F$. Then their weighted relation entries obey

$$|\widehat{\mathcal{K}}_{de} - \mathcal{K}_{de}| \le 2(q_d + q_e). \tag{33}$$

For each $j \in \{d, e\}$, suppose $\epsilon_j = \|\widehat{M}_j^\star - M_j^\star\|_2 \le \gamma_j/2$, where $\gamma_j = \lambda_r(M_j^\star) - \lambda_{r+1}(M_j^\star) > 0$. Then

$$\|\widehat{\Pi}_{d,r} - \Pi_{d,r}\|_F \le \frac{2\sqrt{2r}\,\epsilon_d}{\gamma_d}, \qquad |\widehat{\mathcal{O}}_{de}^{(r)} - \mathcal{O}_{de}^{(r)}| \le 2\sqrt{2}\left(\frac{\epsilon_d}{\gamma_d} + \frac{\epsilon_e}{\gamma_e}\right). \tag{34}$$

For support recovery inside one union, let unit columns $d_j, \widehat{d}_j$ of two dictionaries be matched by one signed permutation and set $a = \max_j \|\widehat{d}_j - d_j\|_2$. If $\|\widehat{C}_d - C_d\|_2 \le \varepsilon_d$, every diagonal utility changes by at most $\zeta_d = \varepsilon_d + 2\|C_d\|_2 a$. Hence the selected K-atom support is unchanged whenever the reference Kth-to-$(K+1)$th diagonal-score margin is greater than $2\zeta_d$.

Finally, for trace-normalized operators $B_d, \widehat{B}_d$ and any fixed rank-K projectors $P_{v,d}$,

$$|\widehat{R}_v - R_v| \le e, \qquad e = 2K \max_d \|\widehat{B}_d - B_d\|_2, \quad R_v = \max_d\{u_K(B_d) - \operatorname{tr}(P_{v,d}B_d)\}. \tag{35}$$

More generally let valid method-specific radii be e_v. If $\widehat{R}_{\text{single}} - \widehat{R}_{\text{Atlas}} \ge e_{\text{single}} + e_{\text{Atlas}}$ and $R_{\text{single}} > R_{\text{dense}}$, then

$$\frac{R_{\text{single}} - R_{\text{Atlas}}}{R_{\text{single}} - R_{\text{dense}}} \ge \frac{(\widehat{R}_{\text{single}} - \widehat{R}_{\text{Atlas}}) - (e_{\text{single}} + e_{\text{Atlas}})}{(\widehat{R}_{\text{single}} - \widehat{R}_{\text{dense}}) + (e_{\text{single}} + e_{\text{dense}})}. \tag{36}$$

To prove Eq. 33, vectorize each operator and use $\|\widehat{m}/\|\widehat{m}\| - m/\|m\|\|_2 \le 2\|\widehat{m} - m\|_2/\|m\|_2$, followed by Cauchy–Schwarz. For the projector constant, write $M = M_d^\star$, $E = \widehat{M}_d^\star - M$, $P = \Pi_{d,r}$, $\widehat{P} = \widehat{\Pi}_{d,r}$, and $t = \|\widehat{P} - P\|_F$. Optimality of $\widehat{P}$ for $M + E$ gives

$$\frac{\gamma_d}{2}t^2 \le \operatorname{tr}[M(P - \widehat{P})] \le \operatorname{tr}[E(\widehat{P} - P)] \le \epsilon_d\|\widehat{P} - P\|_* \le \epsilon_d\sqrt{2r}\,t.$$

This proves the first part of Eq. 34, including an estimated eigenspace tie at $\epsilon_d = \gamma_d/2$; expanding one projector product and using $\|P\|_F = \sqrt{r}$ gives the second. For a matched atom,

$$|\widehat{d}_j^\top \widehat{C}_d\widehat{d}_j - d_j^\top C_d d_j| \le \|\widehat{C}_d - C_d\|_2 + 2\|C_d\|_2\|\widehat{d}_j - d_j\|_2,$$

so a $2\zeta_d$ boundary margin prevents a rank crossing. Ky Fan Lipschitz continuity and $|\operatorname{tr}[P(\widehat{B}-B)]| \le K\|\widehat{B} - B\|_2$ prove Eq. 35; applying the radii separately to the numerator and denominator proves Eq. 36: its numerator is a nonnegative lower bound and its denominator is a positive upper bound. The observable condition $\widehat{R}_{\text{single}} - \widehat{R}_{\text{dense}} > e_{\text{single}} + e_{\text{dense}}$ suffices for the required true positive gap. Without the nonnegative-numerator condition, increasing the denominator need not give a lower bound. □

The result is conditional on supplied estimation radii and, for exact support recovery, a matched-dictionary radius; it does not assert that nonconvex AJD returns the same basin. The paired-song and data-fraction refits below measure empirical stability; their bootstrap quantiles are not certified operator-norm radii.

E CERTIFICATE TIGHTNESS AND DICTIONARY OPTIMIZATION

We audit Proposition 3 on every frozen MUSDB18-HQ and MoisesDB Atlas cell: two tokenizers, both SAME-L active ranks, and $M/K \in \{3, 4\}$. The spectral trivial cap for utility is $u_K(C_d) - \sum_{j=M-K+1}^{M} \lambda_j(C_d)$; the normalized relation cap is one. Table 3 shows that the support-aware utility certificate is nonvacuous for all 300 utility instances across 12 cells and the pair-specific certificate for 3,591/3,600 within-cell pairs. The latter is the triangle bound on the three signed terms in Eq. 12. The last column records negative correlations between the linear and quadratic residual terms. Their partial cancellation explains the gap between the absolute-term certificate and the observed relation error.

Table 3: Support-aware certificates on the construction operators. This audit is post hoc. "Utility ratio" is the median certificate/realized-loss ratio; "rel. bound" gives median/maximum; counts are nonvacuous cases. RMSE summarizes realized normalized overlap errors; ρ is the Pearson correlation between the summed linear and quadratic terms. Residual ratios use the full-residual bound in Eq. 27.

Corpus	Tokenizer / K/M	Utility ratio support / residual	Utility	Rel. bound	Relation / RMSE	Linear–quadratic ρ
MUSDB18-HQ	Locodec / 32/96	2.81/11.54	25/25	.125/.966	300/300 / .066	$-.845$
MUSDB18-HQ	Locodec / 32/128	2.77/10.88	25/25	.098/.992	300/300 / .061	$-.908$
MUSDB18-HQ	SAME-L / 10/30	3.22/11.92	25/25	.083/1.072	294/300 / .064	$-.818$
MUSDB18-HQ	SAME-L / 10/40	3.13/9.10	25/25	.083/1.189	297/300 / .066	$-.855$
MUSDB18-HQ	SAME-L / 26/78	3.67/16.07	25/25	.065/.631	300/300 / .043	$-.850$
MUSDB18-HQ	SAME-L / 26/104	3.65/14.47	25/25	.069/.936	300/300 / .054	$-.865$
MoisesDB	Locodec / 32/96	2.90/9.47	25/25	.094/.775	300/300 / .050	$-.864$
MoisesDB	Locodec / 32/128	2.64/8.53	25/25	.096/.887	300/300 / .060	$-.845$
MoisesDB	SAME-L / 10/30	3.38/15.49	25/25	.072/.648	300/300 / .056	$-.702$
MoisesDB	SAME-L / 10/40	3.49/10.99	25/25	.081/.709	300/300 / .067	$-.760$
MoisesDB	SAME-L / 26/78	3.86/15.13	25/25	.063/.555	300/300 / .043	$-.827$
MoisesDB	SAME-L / 26/104	3.75/13.17	25/25	.061/.806	300/300 / .050	$-.836$

The frozen all-cell criterion requires 12/12 positive closure lower endpoints. Evaluating Eq. 26 yields positive lower bounds in 10/12 cells under the operator evaluation convention in Table 15. Across all 12, the median certified-to-observed closure ratio is $.541$. Table 4 reports the primary $M/K = 4$ cells, with positive lower endpoints in five of six. The two Moises Locodec endpoints are $-.066$ and $-.065$ at ratios three and four, respectively; the corresponding observed closures are positive. Across the 600 construction/evaluation demand cases, the spectral-margin term is selected in three, with the first-order support-aware term selected in the remaining cases.

Table 4: Worst-demand closure certificates on the evaluation operators at the primary $M/K = 4$ cells. Closure values are fractions. "Observed/lower" compares realized Atlas closure with the lower endpoint from Eq. 26; "Utility ratio" is the median per-demand certificate/realized within-union loss. See Table 15 for the operator evaluation convention.

Corpus	Tokenizer / K/M	Utility ratio	Closure observed/lower
MUSDB18-HQ	Locodec / 32/128	2.66	.899/.258
MUSDB18-HQ	SAME-L / 10/40	2.53	.963/.616
MUSDB18-HQ	SAME-L / 26/104	3.65	.933/.303
MoisesDB	Locodec / 32/128	2.32	.835/$-.065$
MoisesDB	SAME-L / 10/40	2.78	.932/.553
MoisesDB	SAME-L / 26/104	3.65	.965/.455

For comparison, Table 5 evaluates the conventional residual/eigengap chain on the exact frozen ratio-four MoisesDB operators, unions, rotations, and supports. Its projector bound is nonvacuous only below $2K$, and its normalized relation bound only below one. In each cell, the counts below these caps are 0/25 demands and 0/300 pairs. The utility bound's median bound-to-actual ratios are $8.5\times$, $11.0\times$, and $13.2\times$. These measurements quantify the difference in tightness between the conventional norm chain and the support-aware certificates.

Table 5: Residual/eigengap bounds on MoisesDB construction operators. The minimum bound on squared projector error is compared with its cap $2K$; "Nonvac." counts bounds below their trivial caps.

Tokenizer / K/M	median ϵ/γ	min squared bound / $2K$	Nonvac. demands	min chained relation	Nonvac. pairs / actual RMSE
Locodec / 32/128	388.81	737.77 / 64	0/25	10.79	0/300 / .0598
SAME-L / 10/40	74.10	52.21 / 20	0/25	5.30	0/300 / .0667
SAME-L / 26/104	171.43	298.28 / 52	0/25	7.09	0/300 / .0501

Optimization and certificate evaluation. The numerical components have different logical roles. Each leading eigenspace is a global solution of its linear spectral subproblem, and the simplex dual is globally bracketed above. In contrast, the rank-M union search and the direct Stiefel baseline are nonconvex. Their returned unions and rank-K views are feasible, so their measured regrets are valid upper endpoints conditional on the frozen operators. These endpoints evaluate the fitted views; tighter feasible solutions can lower them.

AJD is nonconvex as well. Convergence results for appropriate Jacobi variants concern convergence of the iterate sequence to a stationary point (Li et al., 2018). Several rank-104 starts and every rank-128 start reach our frozen 100-sweep cap. This does not enter Proposition 2, and both Proposition 3 and the residual/eigengap comparison hold for any returned orthogonal D when its realized projector or residual is used. A better AJD solution could tighten either certificate but cannot invalidate the displayed inequalities; the commutator lower bound is independent of AJD altogether.

A continuation with up to 1,000 additional sweeps and the same tolerance stops after 321 sweeps for Locodec 32/128 and after one sweep for each SAME-L fit. Locodec off-diagonal energy decreases by 0.008%, and median ϵ/γ changes from 388.81 to 388.37. Locodec's minimum bounds on squared projector error and normalized relation error are 736.32 and 10.78, against caps 64 and one. Counts below the caps remain 0/25 demands and 0/300 pairs in each of the three fits.

We also optimize the exact fixed-union, hard-top-K worst-demand objective directly on the Stiefel manifold. Five starts comprise identity, three Haar rotations, and—for the separately named refined variant—AJD. The scratch baseline excludes AJD. Direct optimization lowers training regret in every cell, providing an objective-matched comparison with AJD. Under frozen operator evaluation, direct refinement has lower regret than AJD in all three MUSDB cells, while AJD has lower regret in all three Moises cells (Table 6). The comparison evaluates the returned fits within the same union and separates training-objective improvement from evaluation regret.

Table 6: AJD versus direct optimization of the sparse-interface objective. This comparison is post hoc. Entries are training / evaluation worst-demand operator regret under the identical union and hard top-K interface, using Table 15's evaluation convention.

Corpus	Tokenizer / K/M	AJD	Direct scratch	Direct refined
MUSDB18-HQ	Locodec / 32/128	.0841/.1331	.0544/.1226	.0523/.1218
MUSDB18-HQ	SAME-L / 10/40	.0794/.0947	.0493/.0942	.0493/.0942
MUSDB18-HQ	SAME-L / 26/104	.0582/.0758	.0352/.0843	.0312/.0735
MoisesDB	Locodec / 32/128	.1051/.2446	.0622/.2534	.0559/.2499
MoisesDB	SAME-L / 10/40	.0895/.1031	.0564/.1064	.0548/.1053
MoisesDB	SAME-L / 26/104	.0623/.0735	.0355/.0816	.0294/.0791

F Data, Demands, and Evaluation Protocols

F.1 Frozen construction algorithm and reported quantities

For the primary MUSDB18-HQ and MoisesDB constructions, let $B_d = A_d/\operatorname{tr}(A_d)$ and $u_{d,K} = u_K(B_d)$. The rank-M union objective is

$$F_{K,M}(U) = \max_d \{u_{d,K} - u_K(U^\top B_d U)\}, \qquad U^\top U = I_M. \tag{37}$$

For a fixed U, demand d's dense view is $U \operatorname{Top}_K(U^\top B_d U) U^\top$. Thus $M = K$ forces the same projector $UU^\top$ for every demand and defines the reported *single* consensus view; $M > K$ defines the

dense union comparator. In fact, $\Xi_{\mathrm{branch}}(K, M) = \min_{U^\top U = I_M} F_{K,M}(U)$: any feasible family in Eq. 2 fits inside an M-dimensional extension of its union, where its worst regret is at least $F_{K,M}(U)$. Conversely, each U admits the demand-wise leading projectors above, which attain $F_{K,M}(U)$ and form a feasible family. These two directions prove the equivalence; they do not guarantee that the finite search finds the minimum. The $M = K$ and $M > K$ searches are separate: the single view is not constrained to lie inside the larger fitted union. Dense, raw, Atlas, and Haar views share that larger union; single shares their whitening, training operators, and active rank, but not their union rank or stored basis.

Algorithm 1 (frozen union and Atom Atlas fit).

1. Compute every training B_d and its leading rank-K basis. Form initial feasible unions from the leading rank-M left singular space of their concatenation and four deterministic Gaussian mixtures of that concatenation.

2. For step sizes .15 and .5, start demand weights once uniformly and twice from seeded Dirichlet draws. For 80 fixed iterations set the columns of U to leading M eigenvectors of $\sum_d w_d B_d$, evaluate $r_d = u_{d,K} - u_K(U^\top B_d U)$, retain the feasible U with smallest $\max_d r_d$, and update $w_d \leftarrow w_d \exp\{\eta(r_d - \max_e r_e)\}$ followed by normalization. The primary public refit uses seed $20260904 + 1009K + M$; the Gaussian candidates use its fixed offset 7919. The stopping rule is the 80-iteration budget, not a convergence declaration.

3. Set each dense basis to the leading rank-K eigenspace inside the retained U. For Atlas, Frobenius-normalize $U^\top B_d U$, run real orthogonal joint diagonalization from the identity and Haar seeds 20261001–20261003 (tolerance 10^{-8}, at most 100 sweeps), and retain the start minimizing training worst-demand atom regret; off-diagonal energy and start order break ties. Demand d addresses the K largest diagonal utilities. Sixty-four seed-20260904 Haar dictionaries are controls.

Every retained union and dictionary is feasible, so its regret is a valid upper endpoint. The union and approximate joint diagonalization (AJD) problems are nonconvex, and no global optimum is claimed; the globally audited simplex certificate is used only for the separate $M = K$ incompatibility test in Appendix A. The independent Slakh boundary retained its separately frozen mean-operator single view and span-of-dense union rather than retrofitting this search after outcomes were known.

For two demands, define directional and reciprocal regret by

$$r^{(K)}_{d\leftarrow e} = \delta_d(\Pi_{e,K}), \qquad r^{(K)}_{d\leftrightarrow e} = \tfrac{1}{2}(r^{(K)}_{d\leftarrow e} + r^{(K)}_{e\leftarrow d}). \tag{38}$$

To measure estimation noise, each consumer operator is independently refit from the deterministic even and odd training-shard aggregates and their reciprocal regret is recorded as $f^{(K)}_d$. The floor for a pair is $(f^{(K)}_d + f^{(K)}_e)/2$; a horizon aggregate averages these pair floors in the same way as its observed regrets. This gives .134265 for the Locodec row in Table 14. A distinct construction tolerance used elsewhere is $\max\{.05, q_{.95}(\{f^{(K)}_d\}_d)\}$; the two quantities are not interchanged.

For $1 \le r < N$ and $\mathcal{O}^{(r)}_{de}$ in Eq. 4, the reported chance-adjusted overlap and best-shared loading gap are

$$\mathcal{O}^{(r)}_{de,\mathrm{adj}} = \frac{\mathcal{O}^{(r)}_{de} - r/N}{1 - r/N}, \qquad \mathcal{G}^{(r)}_{de} = 1 - \max_{P\in\mathcal{P}^N_r} \frac{1}{2}\left\{\frac{\mathrm{tr}(PM^\star_{d,r})}{\mathrm{tr}(M^\star_{d,r})} + \frac{\mathrm{tr}(PM^\star_{e,r})}{\mathrm{tr}(M^\star_{e,r})}\right\}. \tag{39}$$

The subtraction r/N removes the expected overlap of two independent Haar-random rank-r subspaces. $\mathcal{G}$ is the mean normalized utility lost by the best one rank-r subspace shared by the pair; it is not the gap of an arbitrary prefix.

F.2 Demand inventories and relation controls

Table 7 gives the complete 25-demand construction inventory.

Table 7: The 25-demand inventory for primary Atlas construction. SAME-L offsets are the nearest-clock matches to the Locodec offsets in the first row.

Family	Count	Demand identifiers or physical bins
Native causal	5	Locodec 1/4/16/64/128; SAME-L 1/5/20/81/162 tokens
Time–frequency target	4	PupuJEPA bands 0/1/2/3
Long-range relation	4	4–8, 8–16, 16–32, 32–64 s
Source intervention	4	vocals, drums, bass, other
Decoder frequency	4	full, low, mid, high
Decoder source mask	4	vocals, drums, bass, other

The separate 32-demand canonical confirmation is the full Cartesian product

$$\{\text{guitar}, \text{piano}\} \times \{0, 1, 2, 3\}_{\text{band}} \times \{.5, 2, 8, 32\} \text{ s}, \tag{40}$$

using the nearest token offset in each tokenizer. The guitar- and piano-specific demand identities are reserved from Atlas construction; the mixture audio and song split are shared. This tests reserved demand definitions on the same corpus.

The Moises canonical gate is conjunctive for every tokenizer pair: Pearson correlation of off-diagonal weighted relations must be at least .75, that of rank-4 projector relations at least .65, and both one-sided exact p-values at most .01. Its null permutes source, band, and horizon labels independently ($2!4!4! = 1{,}152$ assignments). The raw base–reserved comparison links these 32 demands to the earlier 64-demand grid of vocals/drums/bass/other with the same bands and horizons. Its null shares band/horizon permutations across blocks and permutes old/new source labels separately ($4!2!4!4! = 27{,}648$). The Slakh canonical gate uses the same $.75/.65$ thresholds under its factor-preserving null.

Post-hoc controls on the 64-demand development grid use ordinary-target (raw) operators $M_d^{\text{raw}} = R_d R_d^\top$. The *mean-residual* control computes Frobenius cosines after subtracting each tokenizer's mean operator, retaining cross-tokenizer correlations of .988–.990. The *isotropic-spectrum* control subtracts $\text{tr}(M_d^{\text{raw}})I/N$ demand by demand, retaining .972–.977. The target/count control correlates off-diagonal relation vectors linearly residualized against aligned pairwise target-mean relations and effective-example-count covariates, retaining .952–.965. Leave-one-demand and leave-one-level controls recompute the statistic after deleting the named unit. These checks do not remove shared target semantics.

The 200-draw Slakh refit uses seed 20260905 and resamples train and validation songs with replacement, stratified and paired by song index across tokenizers. Every draw re-estimates token and target whitening, validation-selected ridge, all raw/canonical operators, and their relation matrices; test songs never enter. A draw passes only when all canonical $.75/.65$ thresholds pass. The nine data-fraction cells cross 50/75/100% of training songs with SHA-256 order seeds 20260951/20260952/20260953 and refit whitening, ridges, operators, dictionary, and supports. The three 100% cells coincide by construction and serve as identity checks.

Table 8 links the readouts in Table 1 to their definitions, aggregation rules, and evaluation data. Algorithm 1 specifies the primary interface construction.

F.3 Interface storage and projection cost

For n_D known demands, the storage and leading multiply–add costs per already-whitened token (excluding whitening, tokenizer, and consumer) are given in Table 9.

With float32 matrices, uint16 indices, $n_D = 25$, and factorized dense storage, Table 10 gives the concrete ledger. The direct Atlas path multiplies only the selected $N \times K$ dictionary columns; caching all M coordinates instead costs NM MACs. For one fixed demand, dense can also precompose UZ_d and use NK MACs. Storing all such maps requires $n_D NK$ floats instead of the shared factorization. If union coordinates are cached across demands, factorized dense pays NM once and MK per demand. These alternatives trade stored matrices for projection work. With float32 matrices and uint16 indices, native Atlas uses less storage than precomposed dense when $4NM + 2n_D K < 4n_D NK$, equivalently $n_D > 2NM/[K(2N-1)]$. For the reported $M = 4K$ settings, five or

Table 8: Definitions, aggregation, and evaluation roles of the main experimental readouts. Rows follow the research questions in Table 1. Operator closure and prediction retention use $(S - A)/(S - D)$ for Single, dense, and Atlas scores S, D, A, with $S > D$. Table 4 gives closure fractions; percentage entries multiply this ratio by 100.

Readout	Definition and aggregation	Fitting and evaluation roles	Results
Minimax interval	Bounds on Ξ_K, the minimum worst-demand normalized regret (Eq. 14).	Frozen reconstruction/prediction operators; spectral lower bound and feasible rank-K projector upper bound audit the same operators.	Table 14
Relation correlations	Cross-tokenizer correlations of off-diagonal $\mathcal{K}$ and $\mathcal{O}$ (Eq. 4): Pearson on Moises, Spearman on Slakh. Refits report medians and 2.5/97.5 percentiles.	Whitening and operators are fitted on training data; ridge is selected on validation. Factor-preserving permutations test correspondence. Paired refits resample training and validation songs.	Tables 18, 23
Operator regret and closure	$\max_d \delta_d(P_d)$ (Eq. 1); closure uses each method's maximum. The certified lower bound follows Eq. 26.	Views fitted on training data. Evaluation uses test source/relation/decoder operators and frozen training readout operators.	Tables 4, 15
Held-out refit excess	Eq. 43: sum loss excesses and normalizers over songs, then maximize their ratio across demands.	Frozen interfaces; held-out song losses. Paired song bootstraps repeat the full aggregation.	Table 15
Prediction error and retention	Endpoint: song-mean cosine error against native future tokens (three model and five inference seeds). Bands: song–band mean cosine error against teacher embeddings (three model seeds). Retention uses these means.	Consumers train on fixed views; validation selects checkpoints before test. Moises refits reuse the same split; FMA applies frozen interfaces and consumers.	Tables 19–20
Jamendo macro AUROC	Per-task mean over 20 selected labels; scalar average over two outer development folds and two fresh consumer seeds.	Frozen Moises dictionary; addresses selected on inner artist folds, then fresh consumers evaluated on outer folds of reused official-training data.	Tables 2, 27–28
OpenMIC AUROC/AP	Mean over 20 instruments from three-seed prediction ensembles; paired artist-bootstrap AUROC CIs.	Consumers fit role 3 and checkpoint on role 4; role 5 and official test evaluate frozen models. The later baseline audit reuses the inspected test.	Tables 21–22, 25

Table 9: Interface storage and leading per-token projection work for known demands, excluding whitening, tokenizer, and consumer cost. The orientation controls add a frozen $K \times K$ transform; the native public ablation removes it.

Interface	Shared floats	Per-demand	Active / work
Single	NK	none	K / NK
Dense, factorized	NM	MK floats	K / $NM + MK$
Dense, precomposed	0	NK floats	K / NK
Atom Atlas	NM	K indices	K / NK or NM cached
Oriented Atlas	NM	K indices+K^2 floats	K / $NK + K^2$

Table 10: Interface-only storage and per-token projection work. Entries give KiB / multiply–adds (MACs). Storage assumes 25 demands, float32 matrices, and uint16 indices; MACs are per token and per demand. Parenthesized percentages are storage savings relative to factorized dense storage.

Tokenizer (N, K, M)	Factorized dense	Native Atlas	Oriented Atlas
Locodec $(768, 32, 128)$	784.0 KiB / 102,400	385.6 KiB / 24,576 (50.8%)	485.6 KiB / 25,600 (38.1%)
SAME-L $(256, 26, 104)$	368.1 KiB / 29,328	105.3 KiB / 6,656 (71.4%)	171.3 KiB / 7,332 (53.5%)

more demands meet this condition; the ledger uses 25. A $K \times K$ orientation can be folded into the consumer's first affine layer. A shared consumer facing different demand-specific rotations generally needs an explicit transform or a conditional first layer. The ledger covers interface storage and projection; whitening, tokenization, and the downstream network lie outside these counts.

F.4 CORPUS ROLES AND CONSUMER PROTOCOLS

Table 11 summarizes the core corpora. Later target-task and Jamendo data roles are detailed in Appendices O and P. The initial Moises protocol was fixed before encoding; later refits reuse the same split.

Table 11: Core corpora for geometry, Atlas, and the original OpenMIC comparison. Song or excerpt counts are train/validation/test after the stated eligibility and deterministic filtering rules.

Corpus	Songs / excerpts	Target / teacher	Role
MUSDB18-HQ	80/20/50	future token, PupuJEPA, decoder	method development
MoisesDB v0.1	188/25/24	future token, PupuJEPA, decoder	construction and two nonlinear tasks
MoisesDB six-stem	same split	reserved guitar/piano	frozen held-out geometry
Slakh2100-redux	40/10/50	MERT	independent integrated test
Slakh disjoint	320/80/80	MERT / future token	development gate; test unopened
FMA-small	0/0/800	future token	frozen zero-shot consumer transfer
OpenMIC-2018	14,915/–/5,085 excerpts	human instrument labels	original gated comparison; later retrospective audit

The 25-demand construction inventory contains five causal horizons, four time–frequency target bands, four source interventions, four long-range relation bins, and frequency/source decoder pullbacks. All representations are centered and whitened using training data. Regression ridge strengths are selected on validation data. The long-range relation operator is Eq. 22; it is an additive per-token diagnostic and does not include cross-time curvature. The Moises construction fixes both ratios $M/K \in \{3, 4\}$, identity plus three Haar AJD starts, tolerance 10^{-8}, 100 Jacobi sweeps, and 64 Haar dictionaries. Training worst-demand regret selects the AJD start. The subsequent operator audit uses test accumulators for source, relation, and decoder demands, and reuses frozen training operators for readout demands lacking independent test operators. Refit risks are evaluated separately on held-out songs.

The nonlinear physical-horizon experiment uses the same 188/25/24 Moises split. For native tokens $\widetilde{h}_t$, training mean μ_h, and training whitener W, method v with basis B_v receives $z_t^v = \mathrm{norm}_{\sqrt{K}}(B_v^\top W(\widetilde{h}_t - \mu_h))$. Here dense and Atlas bases are orthogonally Procrustes-oriented to the corresponding single basis before projection; the orientation does not change their projectors. A causal history ending at z_t^v predicts $z_{t+\ell}^v$. If s_a, s_b are two points ($a \le b$) on the same spherical interpolation from random noise to that target, the endpoint-consistency loss is

$$\mathcal{L}_{\mathrm{ec}} = 1 - \cos\big(f_\theta(s_a, z_{\le t}^v, a), z_{t+\ell}^v\big) + 1 - \cos\big(f_\theta(s_a, z_{\le t}^v, a), \mathrm{sg}\, f_\theta(s_b, z_{\le t}^v, b)\big)\,, \tag{41}$$

where sg stops gradients, $a = 0$ with probability .25, and otherwise $a = \max(0, b - .25)$ for uniform $b \in [0, 1]$. Ten percent of examples condition only on the local path; the rest use the full available causal history, with five-degree spherical history noise during training.

Each view has the same width-256, four-layer, eight-head, 5.87-million-parameter consumer. AdamW uses learning rate 3×10^{-4}, betas $(.9, .95)$, weight decay .05, batch size 8, bfloat16, gradient-norm cap 1, and 8,000 updates. For each view and model seed, the checkpoint is selected only by that run's prespecified validation loss. Locodec uses $(K, M) = (32, 128)$ and lag four; SAME-L uses $(26, 104)$ and lag five. Active coordinate rates are 272.7 and 279.9 coordinates/s. Both consumers receive up to about 60 seconds of causal history, so these four/five-token lags do not restrict the available context to roughly half a second. At evaluation, predicted, projected-target, and persistence coordinates are lifted by B_v and the shared coloring map, then compared with the same native target $\widetilde{h}_{t+\ell}$; hence native cosine errors are comparable within a tokenizer despite view-specific coordinates. Thus this experiment evaluates the compact input and output views jointly. The radius normalization is consumer preprocessing; the operator certificates audit the linear projectors defined in Section 3.

Stage A opens single and dense validation only. It requires the upper paired 95% confidence bound for dense minus persistence and dense minus single to be below zero, mean dense minus single at most $-.005$, and at least two favorable seed directions. Stage B opens Atlas validation only after Stage A passes and applies the analogous skill and $-.005$ improvement conditions, at least two favorable seed directions, and at least 50% dense-gap closure. Test view manifests, target positions, checkpoint hashes, and inference seeds are locked after both stages pass.

The public band-embedding consumer uses the same song split but a distinct target and architecture. For each frozen band demand $b \in \{0, 1, 2, 3\}$, every method provides a 26-D radius-$\sqrt{26}$ projection

of the SAME-L token. In the orientation control, dense and Atlas bases are orthogonally aligned to the corresponding single basis to control arbitrary within-subspace orientation; this convention does not guarantee equal optimization difficulty. The target is the aligned, unit-normalized 1,024-D frequency-partition embedding from the frozen PupuJEPA-v2 large checkpoint. We cite the PupuJEPA version of Gu et al. (2026); the artifact pins the teacher weight digest. A noncausal TCN maps input width 26 to width 192, adds a four-way band embedding, applies four pre-normalized residual convolutions with dilations 1, 2, 4, and 8, and projects to 1,024 outputs. It has 648,640 trainable parameters and a receptive field of 31 SAME-L tokens (2.88 s). For a uniformly sampled band b and 96-token crop Ω, every view minimizes the same objective

$$\mathcal{L}_{\mathrm{spec}}(\theta; v) = \mathbb{E}_{i,b,\Omega} \frac{1}{|\Omega|} \sum_{t \in \Omega} \left[1 - \cos\big(g_\theta(z^v_{i,b,\Omega}, b)_t, y_{i,b,t}\big)\right], \tag{42}$$

where $y_{i,b,t}$ is the common unit PupuJEPA band target. The matched runs use three seeds, batch size 8, 96-token windows, dropout .1, AdamW at 3×10^{-4} with weight decay .01, 200 warmup updates followed by cosine decay, gradient-norm cap 1, and 4,000 updates. Each run contributes the checkpoint selected by its own validation loss.

The no-input baseline is the unit train-target centroid for each band, frozen before validation. Stage A requires dense to beat that centroid and single with upper paired-song confidence endpoints below zero, a dense-minus-single mean at most $-.005$, favorable directions in at least two of three seeds and three of four bands, and finite predictions. Stage B applies the analogous rules to Atlas and requires at least 50% dense-gap closure. Only both passes unlock a single evaluation of the 24-song test split. Three fixed windows per song begin at relative positions .1, .45, and .8; songs receive equal weight in 10,000 paired bootstrap draws.

The end-to-end public refit keeps this consumer and decision rule but rebuilds its entire upstream interface. It starts from the native MoisesDB four-stem waveforms and a deterministic 70-second crop for each of the same 188/25/24 songs. Because the released SAME-L bottleneck samples during evaluation, each mixture draw is fixed by a hash of a base seed and public song identifier; paired source interventions additionally share the same random draw. Training audio re-estimates the token mean, covariance, whitening map, and all 25 causal, spectral, long-range, source, and decoder-pullback operators. For fitted readouts, validation selects only regression ridge strengths. A rank-26 single view and rank-104 union are then refit, followed by identity plus three fixed-Haar approximate joint diagonalization starts, demand-specific 26-atom supports, and 64 Haar controls. Finally, all nine TCNs are retrained. No operator, union, rotation, support, or consumer checkpoint from the earlier run enters this path. Test accumulators select no construction or checkpoint, and test consumer evaluation occurs only after both unchanged validation gates pass. This computational reproduction uses the same split, teacher, tokenizer, ontology, and task as the original experiment.

A separately locked native-coordinate ablation reuses this reconstructed operator state but applies no Procrustes or other 26×26 transform. Dense emits its stored leading basis inside the union. Atlas emits the stored dictionary columns in each demand's recorded atom order, and the view manifest records those indices. The single view is exactly unchanged (its self-alignment differs from identity by at most 1.1×10^{-15}), so its checkpoints and outcomes are reused; all six native dense/Atlas consumers are retrained with the identical architecture, optimizer, seeds, and 4,000-update budget. Stage A opens only native dense validation, Stage B only after Stage A passes, and native test views are materialized only after both passes. The oriented outcomes were known when this ablation was locked. The comparison tests the coordinate mechanism on the same songs and prediction task.

The FMA transfer freezes the exact selected physical-horizon checkpoints, single/dense/Atlas interfaces, physical horizon, five endpoints, inference seeds, and decision rule before producing an FMA token or prediction. It uses all 800 tracks in the official balanced FMA-small test partition and fits nothing on FMA. Exact hashes and offset-aligned Chromaprint find zero matches against the fixed 18,201-track Locodec training subset or any of the 240 native mixtures in the MoisesDB inventory; fixed excerpt controls recover all 20/20 planted matches. The five history endpoints are 5, 10, 15, 20, and 25 seconds, with the same four/five-token prediction lags as the Moises experiment.

F.5 Public OpenMIC task-directed projector protocol

OpenMIC-2018 comprises 20,000 distinct ten-second FMA excerpts with sparse human judgments over 20 instruments. We retain its official split-01 boundary: all M84 development uses the 14,915

official-training excerpts, and the 5,085 official-test excerpts remain unavailable until a final protocol is locked. The released Music2Latent encoder maps 44.1-kHz stereo audio to 107 native 128-D tokens per excerpt. Whitening is fitted only on the projector-training role. Because the audio comes from FMA, this changes the public task and label source without supplying a fully new audio lineage. The upstream official split contains one artist identifier on both sides, so we call it the official split rather than an artist-disjoint test boundary. The original direct risk-matched dense/Atlas M84 branch stopped at its construction gate before final-consumer or official-test access. The separate E1 study evaluates a learned shared rank-8 projector.

Official-training artists are deterministically assigned to six disjoint roles with 3,732/1,071/1,256/5,942/1,121/1,793 excerpts: projector training, projector checkpointing, a method-held-out projector gate, final-consumer training (role 3), final-consumer checkpointing (role 4), and final validation (role 5). Role 2 was unused in fitting or selecting any E1 projector or consumer evaluated there, but original-M84 models had previously been assessed on it; it is therefore method-held-out, not globally untouched. The selected rank-8 shared basis is learned while the scout consumer is frozen during each projector phase: for each of eight seeds and both directions between two artist-disjoint halves of the first role, a width-128 scout TCN is warmed for 250 updates with a seeded Haar basis, then the consumer is frozen and only the basis is updated for 500 steps. A detached Procrustes factor fixes the within-subspace gauge. We average each learned projector's displacement from its initialization across directions and seeds; the leading eight eigenvectors form the strict consensus. Directional and odd/even consensus overlap, nontrivial movement, and a positive eighth eigenvalue gate construction before fresh consumers screen the basis on the first two roles and then once on the method-held-out third role.

The final comparison freezes that single consensus, the strongest prior staged learned projector for each paired seed, and three exact seed-matched Haar projectors. New consumers train only on role 3 and checkpoint only on role 4; role 5 is opened once after the checkpoint screen passes. Each consumer is a six-block width-128 causal TCN with kernel three, dilations $1, 2, 4, 8, 16, 32$, dropout .1, and a demand-specific 8×128 input map. Its 127-token field covers the complete excerpt. Equal-demand positive/negative pairs minimize pairwise logistic ranking risk for 1,000 AdamW updates; the checkpoint rule maximizes macro AUROC, then minimizes role-4 calibrated risk, then takes the earlier update. Projectors never update with these consumers.

Before semantic test access, a task lock hashes the protocol, code, upstream sources, bases, all nine selected checkpoints, role-4 temperatures, and prior gates. Access then creates an immutable manifest, label archive, and access lock; one encoding pass is followed by a verified cache lock; and a final evaluation lock hashes those objects before any prediction. Eleven unchanged conditions require absolute skill, at least 18/20 instrument AUCs above .5, a .005 advantage over both staged and Haar, positive paired artist-bootstrap lower endpoints, at least two of three favorable seed directions, and no worse fixed-temperature risk. Passing on both final-development roles authorizes one official-test encoding and evaluation. All intervals use 10,000 paired artist-bootstrap draws; test labels fit or select nothing and cannot authorize an adaptive rerun. The strict consensus is one shared projector, not an Atom Atlas: no union dictionary, demand-specific atom support, or sparse-address claim is attached to it.

For the core studies above, evaluation units are songs (artists for OpenMIC). Paired 95% intervals resample songs/artists 10,000 times while retaining aligned demands, views, and model outputs. Permutation tests preserve the factor structure named by each null. Model-seed results are listed directly because three seeds do not support a strong population-level variance estimate.

G Shared-Subspace Compatibility and Predictive Utility

Learned compact views across tokenizer seeds. The initial comparison asks whether a learned 32-D view provides useful access beyond the first 32 native coordinates. The core uses Locodec's row-orthogonal projection and a separately learned column-orthogonal reconstruction lift (Luo et al., 2026), trained jointly with the tokenizer using reconstruction, adversarial, feature-matching, and commitment losses with token-noise corruption. The primary Single and Atlas views are fitted after freezing the tokenizer. The six core/prefix comparisons use three tokenizer seeds with progressive-dimension dropout (PDD) and three with PDD disabled (NoPDD), each trained for 250,000 updates on 18,201 in-house tracks (972.581 hours). The fixed split contains another 999 validation tracks (53.866

hours) and 1,014 test tracks (54.176 hours). Public-task labels are used only for the subsequent frozen-view probes.

Table 12 gives every seed-level difference underlying the 28/30 MIR and six causal-loss comparisons in Section 5. The five MIR tasks follow the MARBLE protocol (Yuan et al., 2023): task-specific temporal pooling, or time-preserving interpolation for beat tracking, feeds a one-hidden-layer probe with 512 ReLU units and dropout .2. Probe seed 1234 and validation-fold selection are fixed across views. The causal predictor uses a six-layer, width-256 backbone with 256-token training sequences, batch size 64, and a 5,000-update budget at predictor seed 1234. Evaluation uses 512-token sequences. Both compact predictor views are normalized to radius $\sqrt{32}$. Its teacher-forced flow loss is the cosine error between predicted and target spherical tangent directions plus .1 times endpoint cosine error. The fixed 70-second cache has 17,933/988/997 source-disjoint train/validation/test sequences. Table 13 separately reports the 20,000-update view and projection-budget controls with three predictor seeds.

Table 12: Learned 32-D core versus raw 32-D prefix across tokenizer seeds. Entries are core minus prefix. MIR columns use EmoMusic mean R^2 over arousal and valence (EMO), GiantSteps weighted key score (GS), GTZAN beat F1 (Beat), GTZAN genre accuracy (Genre), and MagnaTagATune macro average precision (MTT); larger is better. Causal flow loss is evaluated on the held-out cache; smaller is better. Bold marks the direction favoring the core. Counts summarize benchmark point-estimate directions across the six tokenizer trainings.

Tokenizer / seed	Δ EMO	Δ GS	Δ Beat	Δ Genre	Δ MTT	Δ causal loss
PDD/1234	**+.0387**	**+.1358**	**+.0405**	**+.0207**	**+.0206**	**-.0210**
PDD/2345	**+.0084**	**+.1546**	**+.0109**	**+.0103**	**+.0162**	**-.0250**
PDD/3456	**+.0172**	**+.1919**	**+.0058**	**+.0034**	**+.0170**	**-.0356**
NoPDD/1234	**+.0536**	**+.1488**	**+.1189**	-.0897	**+.0117**	**-.3799**
NoPDD/2345	**+.0582**	**+.1722**	**+.1253**	**+.0276**	**+.0230**	**-.3591**
NoPDD/3456	**+.0248**	**+.1733**	**+.1126**	-.0138	**+.0143**	**-.3768**

Table 13: Coordinate-order controls for reconstruction and causal prediction. Block (a) compares the jointly trained core with fixed views; block (b) matches 2,000 post-hoc projection updates. STFT loss is validation normalized-magnitude plus log-magnitude loss across resolutions (mean ± sample SD over three projection seeds). Teacher-forced flow loss uses 20,000-update predictors, three predictor seeds, and five evaluation seeds on representation seed 1234; SDs summarize the respective seed means. Lower is better; bold marks each within-block objective minimum.

View	STFT loss	Flow loss	Predictor-seed SD	Evaluation-seed SD
(a) Joint-tokenizer view comparison				
Joint learned core	—	**.429106**	.001819	.000106
PCA	—	.485917	.002529	.000145
Raw prefix	—	.824727	.000376	.000324
Random projection	—	.826864	.000356	.000282
(b) Projection-budget-matched post-hoc comparison				
Independent	$17.493 \pm .222$	**.471630**	.002552	.000230
Shared with corruption	$17.575 \pm .111$	.524407	.002148	.000228
Shared without corruption	$\mathbf{17.474 \pm .167}$	.545170	.001454	.000243

Table 14: Shared-subspace compatibility across tokenizers and active ranks. Reciprocal regret and the same-demand split-estimation floor average the prespecified horizons; CIs use paired song bootstraps. The minimax interval uses the globally audited Fantope lower endpoint and the best feasible rank-projector upper endpoint; it certifies the frozen estimated operators, not their population values. Minimax endpoints are rounded outward.

Tokenizer / rank scheme	Reciprocal regret [95% CI]	Split floor	Minimax interval
Locodec / 32	.2964 [.2886,.3045]	.1343	[.113217,.113269]
SAME-L / fraction 10	.3282 [.3068,.3512]	.1405	[.161127,.161206]
SAME-L / fraction 11	.3573 [.3376,.3788]	.1446	[.169883,.170175]
SAME-L / rate 25	.5069 [.4792,.5366]	.1588	[.165561,.165566]
SAME-L / rate 26	.5108 [.4821,.5411]	.1584	[.161569,.161571]

Under the 5% training-operator reconstruction budget, all ten practical frontier cells select an interior mixture. Nine have a strictly positive held-out prediction gain while retaining an upper reconstruction-regret bound below .10; every view also beats persistence in its own coordinates. For Locodec at 15 seconds, the prediction-gain estimate is $-.00289\,[-.00524, -.00067]$. These input-view comparisons complement the output-side compatibility certificate by measuring the prediction–reconstruction tradeoff.

H Atlas Utility and Demand Geometry

Table 15 reports two evaluations of frozen interfaces. The operator block uses each method's maximum normalized regret across the 25 demands. Operator closure applies $(S - A)/(S - D)$ to those three maxima; the maximizing demand can differ by method. The held-out refit-excess block uses songwise task losses. For a frozen interface $V = \{V_d\}$, its risk is

$$\mathcal{R}_{\mathrm{actual}}(V) = \max_d \frac{\sum_{i\in\mathcal{T}}[\ell_{i,d}(V_d) - \ell_{i,d}(V_d^{\mathrm{own}})]}{\sum_{i\in\mathcal{T}} z_{i,d}}, \tag{43}$$

where $\mathcal{T}$ is the held-out song set, V_d^{own} is demand d's rank-matched reference view, and $z_{i,d} \geq 0$ is the prespecified demand-specific normalizer (target energy, relation-to-uniform divergence, or full tangent utility), with $\sum_{i\in\mathcal{T}} z_{i,d} > 0$. Loss-oriented forms of source and decoder utilities are used inside ℓ. Thus the maximum is taken only after songwise numerators and normalizers have been aggregated; song bootstraps repeat this entire operation. The two blocks therefore use their respective operator and task-loss scores. Nonlinear prediction retention applies the same improvement ratio to the cosine errors in Appendix I.

In each primary cell of Table 15, Atlas has lower worst-demand operator regret than coordinate selection in the unrotated union basis and in all 64 Haar rotations.

Table 15: Operator regret and held-out refit excess of shared, dense, and Atlas views at $M/K = 4$. Operator regret uses Eq. 1, with test accumulators for source, relation, and decoder demands and frozen training operators for fitted readout demands. Held-out refit excess uses Eq. 43. Dense views are fitted within the union on training operators and then frozen. Bold marks the minimum within each three-method block; operator closure is $100(S - A)/(S - D)$ for single (S), Atlas (A), and dense (D) worst-demand regrets in the operator block.

		Operator regret				Held-out refit excess		
Corpus	Tokenizer / K	Single	Dense	Atlas	Closure	Single	Dense	Atlas
MUSDB18-HQ	Locodec / 32	.2239	**.1229**	.1331	89.9%	.1720	**.0426**	.0529
MUSDB18-HQ	SAME-L / 10	.3299	**.0858**	.0947	96.3%	.2647	**.0456**	.0545
MUSDB18-HQ	SAME-L / 26	.2219	**.0653**	.0758	93.3%	.1961	**.0054**	.0313
MoisesDB	Locodec / 32	.3111	**.2314**	.2446	83.5%	.1840	**.0302**	.0475
MoisesDB	SAME-L / 10	.3154	**.0876**	.1031	93.2%	.2454	**.0437**	.0598
MoisesDB	SAME-L / 26	.2782	**.0661**	.0735	96.5%	.2394	**.0070**	.0713

On Moises, Atlas-minus-single paired actual-risk CIs are $[-.1516, -.1178]$, $[-.2361, -.1684]$, and $[-.2309, -.1266]$ for Locodec rank 32 and SAME-L ranks 10/26. In the held-out refit-excess block of Table 15, Atlas-minus-dense differences quantify the change in this risk when dense views are restricted to shared atom subsets. Ratio-three operator closures are 79.6%, 99.4%, and 95.4%, retaining the main direction.

Table 16 reports 18 construction refits: six shared song subsets across three tokenizer/rank settings. Each subset contains 94 or 141 of 188 training songs, selected by sorting identifiers under SHA-256 seeds 4101/4202/4303. The full-data whitening map, ridge choices, decoder probes, validation set, and held-out set remain fixed, so this isolates construction sensitivity. Every row beats its refitted single view. The per-cell gates require operator median/q10 at least $.90/.75$, Atlas below single in at least five of six rows with median closure at least 50%, and active/atom medians at least $.50/.60$; all pass.

Table 16: Atlas utility and geometry under construction-data subsampling on MoisesDB. Each row rebuilds operators, union, dictionary, and supports from the indicated subset. Geometry is compared with the full-data fit: "Op." is demandwise operator-cosine median/10th percentile; "active" is median rank-K projector overlap; "atom" is median absolute matched-atom cosine. Operator closure uses Table 15's evaluation convention and is untrimmed.

Tokenizer / K	Train	Seed	Op. med./q10	Active	Atom	Closure
Locodec / 32	50%	4101	.984/.943	.859	.896	89.4%
Locodec / 32	75%	4101	.994/.976	.919	.956	90.0%
Locodec / 32	50%	4202	.982/.950	.867	.906	89.8%
Locodec / 32	75%	4202	.995/.985	.924	.963	85.6%
Locodec / 32	50%	4303	.983/.952	.882	.911	86.8%
Locodec / 32	75%	4303	.994/.979	.922	.964	82.5%
SAME-L / 10	50%	4101	.993/.981	.847	.857	96.6%
SAME-L / 10	75%	4101	.997/.994	.883	.931	98.8%
SAME-L / 10	50%	4202	.994/.981	.864	.911	91.6%
SAME-L / 10	75%	4202	.998/.996	.930	.960	98.3%
SAME-L / 10	50%	4303	.994/.986	.852	.846	93.1%
SAME-L / 10	75%	4303	.998/.996	.910	.967	96.1%
SAME-L / 26	50%	4101	.993/.981	.942	.897	95.2%
SAME-L / 26	75%	4101	.997/.994	.958	.956	98.2%
SAME-L / 26	50%	4202	.994/.981	.934	.913	97.5%
SAME-L / 26	75%	4202	.998/.996	.952	.948	96.9%
SAME-L / 26	50%	4303	.994/.986	.926	.917	96.4%
SAME-L / 26	75%	4303	.998/.996	.953	.938	98.2%

Table 17: Atlas utility across independently trained Locodec tokenizers. Every seed re-encodes the common Moises crops and independently refits all 25 operators and the $K/M = 32/128$ Atlas. Entries are worst-demand operator regret and closure, with the evaluation convention and definition in Table 15.

Tokenizer seed	Single	Dense	Atlas	Closure
1234	.3111	.2314	.2446	83.5%
2345	.3263	.2373	.2529	82.4%
3456	.3037	.2327	.2489	77.3%

All three Atlas regrets in Table 17 are below single. Pairwise correlations between the three 25-demand operator-relation matrices are $.9981/.9966/.9992$ (median .9981). Atom indices are not compared across independently trained latent spaces; the cross-seed quantities are coordinate-invariant relations and operator regret under the frozen evaluation.

Table 18: Cross-tokenizer correlations of canonical and raw demand relations. Figure 2b shows canonical results. Moises rows use Pearson correlations on 32 reserved guitar/piano demands; Slakh/MERT uses Spearman correlations on 48 demands. Projector rank is 4. The last column gives raw weighted/projector correlations between 64 development and 32 reserved Moises demands.

Pair	Raw weighted	Canon. weighted	Canon. projector	Raw base–reserved
Locodec–SAME-L	.9939	.9877	.9678	.9833 / .9434
Locodec–Music2Latent	.9896	.9805	.9466	.9764 / .9405
SAME-L–Music2Latent	.9976	.9953	.9416	.9828 / .9343
Slakh: Locodec–SAME-L	.8538	.8780	.7700	—

For relations among the 32 reserved Moises demands, all raw tests and all but one canonical test are the unique maxima among 1,152 structured label assignments. The remaining canonical Locodec–Music2Latent projector test has $p = .00347$. Base–reserved placement uses 27,648 block-preserving permutations. Target-canonical rank-16 noncommutation also recurs, with correlations $.926/.918/.978$, but was secondary to the frozen held-out gate.

I NONLINEAR PREDICTION AND INTERFACE CONTROLS

Table 19: Endpoint prediction with compact interfaces on MoisesDB and FMA-small. Values average three model and five inference seeds. Confidence intervals (CIs) resample 24 Moises songs or 800 FMA tracks; lower is better and bold marks the minimum. Rows use their prespecified corpus-specific history endpoints. Values are absolute native cosine errors for within-row interface comparisons. "Retained" is the single-to-dense error reduction retained by Atlas.

Corpus	Tokenizer / horizon	Single	Dense	Atlas	Atlas−single [95% CI]	Retained
Moises	Locodec / .469 s	.71228	**.67707**	.68171	$-.03057$ $[-.03898, -.02263]$	86.8%
Moises	SAME-L / .464 s	.57869	.51888	**.51885**	$-.05983$ $[-.07089, -.04923]$	100.0%
FMA	Locodec / .469 s	.87469	**.81822**	.83533	$-.03936$ $[-.04186, -.03681]$	69.7%
FMA	SAME-L / .464 s	.59164	**.53762**	.53958	$-.05206$ $[-.05421, -.04991]$	96.4%

For Moises endpoint prediction, model seeds 1234/2345/3456 yield Atlas-minus-single means of $-.03407/-.03049/-.02715$ for Locodec and $-.06208/-.05779/-.05963$ for SAME-L. Their six seed-specific paired-song 95% intervals are respectively $[-.04391, -.02464]$, $[-.03912, -.02261]$, $[-.03576, -.01907]$, $[-.07446, -.05036]$, $[-.06963, -.04617]$, and $[-.07036, -.04877]$. All six intervals favor Atlas, with uncertainty conditional on each fitted model. In the same seed order, Atlas-minus-persistence means are $-.02881/-.02626/-.02728$ for Locodec and $-.02764/-.02485/-.02186$ for SAME-L.

We also evaluate the output view by projecting each observed future target and lifting it back through the shared coloring map. The cosine errors relative to the native future target average $.512765/.540151$ for single/Atlas on Locodec and $.307169/.389879$ on SAME-L.

Table 20: Frequency-band embedding prediction on MoisesDB with SAME-L: interface refitting and coordinate-orientation controls ($K = 26$; dense/Atlas union $M = 104$). Targets are frozen PupuJEPA embeddings. The frozen-interface run retrains only consumers; the full refit starts from public waveforms and rebuilds the 25-demand interface and all consumers. "Oriented" applies the Procrustes control; "native" emits each stored basis and Atlas atom order without a 26×26 transform. Errors are three-seed means of song–band macro cosine distances; lower is better. Confidence intervals (CIs) use paired song bootstraps. Bold marks the lowest learned-interface error; "Retained" is the single-to-dense error reduction retained by Atlas. Figure 3 shows the two full-refit test rows.

Run / split	Centroid	Single	Dense	Atlas	Atlas−single [95% CI]	Retained
Frozen / validation	.41835	.30606	**.29418**	.29570	$-.01036$ $[-.01227, -.00847]$	87.2%
Frozen / test	.41327	.31471	**.30168**	.30368	$-.01103$ $[-.01277, -.00929]$	84.7%
Full refit, oriented / validation	.41835	.30706	**.29552**	.29744	$-.00962$ $[-.01146, -.00787]$	83.4%
Full refit, oriented / test	.41327	.31486	**.30241**	.30434	$-.01052$ $[-.01228, -.00878]$	84.5%
Full refit, native / validation	.41835	.30706	**.29332**	.29538	$-.01168$ $[-.01381, -.00959]$	85.0%
Full refit, native / test	.41327	.31486	**.30127**	.30327	$-.01159$ $[-.01339, -.00987]$	85.3%

In the oriented full-refit test, Atlas minus the no-input centroid is $-.10893$ $[-.11843, -.09898]$. In the native test that contrast is $-.11000$ $[-.11960, -.09978]$, and Atlas minus single is negative for model seeds 1234/2345/3456 ($-.01239/-.01190/-.01047$) and for all four band means ($-.00677/-.01297/-.01192/-.01469$). Native minus oriented error is $-.00114$ $[-.00169, -.00063]$ for dense and $-.00107$ $[-.00180, -.00038]$ for Atlas. These comparisons support direct prediction from stored atom coordinates on this task. Both native validation stages and the unchanged test gate pass. The reconstructed operator Atlas uses 95/104 union atoms; its single/dense/Atlas worst-demand operator regrets are $.28962/.06653/.07378$, so it closes 96.8% of dense operator headroom and beats all 64 Haar controls under the sampled-dictionary Monte Carlo rank test ($p = 1/65$). For the native test, $S = .31486$, $D = .30127$, and $A = .30327$ are mean cosine errors. The 85.3% retention is $(S - A)/(S - D)$, using unrounded scores: an absolute error reduction of .01159, or 3.68% of Single's error.

Construction demand family. The primary interfaces serve all 25 construction demands, with band prediction evaluating four of them. A separate refit constructs the interfaces from the four bands alone. Both refits use the same public corpus split, target definition, TCN, and consumer

training budget, with each interface rebuilt under its recorded protocol. The four-demand refit gives validation single/dense errors of $.29348/.29209$ and a dense-minus-single difference of $-.00139$ $[-.00164, -.00115]$. The 25-demand refit gives $.30706/.29552$ and $-.01154$ $[-.01367, -.00955]$, respectively (Table 20). The four-demand protocol ends at Stage A validation under the prespecified $.005$ improvement rule.

Additional target checks. For source-activity prediction, single/dense song-macro AUROC is $.9145/.8558$, with dense-minus-single $-.0588$ $[-.0784, -.0381]$; the TCN exceeds its no-input comparator. A finite SAME-L decoder gives single/dense source-weighted complex-STFT losses of $1.0323/.9606$, relative to its full-token output. The geometric mean of paired dense/single cell-loss ratios is $.8378$ $[.8156, .8606]$, averaging cell log ratios within songs, then over songs within each source, and finally equally over sources. Two of four sources (bass and vocals) improve, against the prespecified requirement of three. Both protocols stopped after Stage A validation.

J SHARED-PROJECTOR UTILITY ON OPENMIC

The construction gate passes with directional and odd/even consensus overlaps $.362/.271$, median learned-to-initial overlaps $.973/.975$, and positive eighth consensus eigenvalue $.02099$. Before the fresh final-consumer run, the method-held-out role-2 screen also passes 11/11 conditions, with strict/staged/Haar AUROCs $.7211/.6576/.6121$; as stated above, this role was not globally untouched. Table 21 shows the subsequent fresh-consumer and official-test progression. The strict consensus remains stable from its role-4 checkpoint screen through role-5 validation to the official test; neither later role updates a projector, consumer checkpoint, ensemble, or temperature. The official test contains 5,085 excerpts from 1,709 artists. All 20 strict-consensus instrument AUCs are above $.5$, and its absolute artist-bootstrap 95% interval is $[.735835, .760322]$.

Every one of the 11 frozen conditions passes on role 4, role 5, and official test. The test result therefore confirms that the cross-fitted, task-directed shared projector improves a fresh finite nonlinear consumer over both the strongest staged learned control and data-independent rank-8 views. The result concerns shared-projection utility for a known instrument-ranking task: construction and consumer use that same task, and the known instrument selects a demand-specific input map. The evaluated interface is the learned shared projector defined in Appendix F. OpenMIC and the FMA cross-corpus check use audio from the same source collection. The matched joint-classifier audit in Appendix N separately tests stronger compact baselines.

Table 21: OpenMIC instrument recognition with shared rank-8 projectors. Nonlinear consumers are evaluated as three-seed ensembles. Role 4 selects checkpoints and temperatures; role 5 and the official test are one-shot evaluations. Test average precision (AP) uses raw ensemble probabilities, while test risk uses the frozen role-4 arm temperature. Higher AUC/AP and lower risk are better; bold marks their column optima.

Projector	Role-4 AUC	Role-5 AUC	Test AUC	Test AP	Test risk	AUC$>.5$
Strict consensus	**.751535**	**.743666**	**.748542**	**.631199**	**.480304**	20/20
Staged learned	.694245	.695531	.694668	.585561	.557265	20/20
Frozen Haar	.660925	.665873	.664739	.565080	.595680	20/20

Table 22 reports the original paired test contrasts.

Table 22: Prespecified OpenMIC official-test contrasts: strict consensus minus each rank-8 comparator. Ensemble AUC intervals use 10,000 paired artist-bootstrap draws. The last column instead contrasts individual consumer seeds 1234/2345/3456. Risk difference uses the fixed role-4 temperatures.

Comparator	ΔAUC	95% CI	Δrisk	Paired seed ΔAUCs
Staged learned	+.053874	[.040160,.067532]	$-.076962$	+.102084, +.071667, +.074100
Frozen Haar	+.083803	[.069906,.097889]	$-.115376$	+.132952, +.099707, +.105901

K Cross-Corpus Transfer to FMA

SAME-L passes every frozen transfer conjunct: its Atlas is skilled over persistence, improves on single by $-.05206$ $[-.05421, -.04991]$, and closes 96.4% of dense headroom. Locodec dense is skilled and its Atlas improves on single by $-.03936$ $[-.04186, -.03681]$. Its Atlas-minus-persistence difference is $+.00104$ $[-.00116, .00329]$; the frozen joint criterion requires the upper endpoint of this interval to be below zero. The three Atlas-minus-single seed means are $-.03964/-.04270/-.03574$ for Locodec and $-.04811/-.05379/-.05429$ for SAME-L; all six seed-specific intervals are strictly negative.

All eight prespecified genre strata (100 tracks each) retain negative Atlas-minus-single intervals for both tokenizers. Mean differences range from $-.05576$ to $-.02745$ for Locodec and $-.05925$ to $-.03804$ for SAME-L. Thus the Atlas–single comparison retains its direction within each genre stratum; the separate persistence comparison is reported above.

L Demand Relations and Consumer Utility on Slakh

Table 23: Stability of demand relations on Slakh/MERT. Entries are median Locodec–SAME-L Spearman correlations across 200 paired song-level operator refits. Intervals are 2.5/97.5 percentiles of full refits, not entrywise matrix bootstraps. Parentheses give the frozen threshold and the proportion of draws clearing it.

Relation	Raw	Target-canonical
Weighted similarity	.902 [.827,.964] $(.80; 100\%)$	.909 [.835,.954] $(.75; 100\%)$
Rank-4 projector	.839 [.722,.930] $(.75; 93.5\%)$	.866 [.749,.935] $(.65; 100\%)$
Rank-16 commutator	.813 [.657,.903] $(.65; 98.5\%)$	.466 [.202,.696]
Rank-4 compromise gap	.909 [.774,.963]	.895 [.747,.955]

In Table 23, the complete raw conjunction passes 186/200 draws, whereas the canonical conjunction passes 200/200. Across the separately refitted 50/75/100%–by–three-seed grid, the canonical conjunction passes 9/9 cells and the raw conjunction 7/9. These refits assess relation stability; the consumer comparisons below separately test predictive utility.

Table 24: Nonlinear and linear prediction of MERT targets on Slakh, measured by normalized mean-squared error (lower is better). Nonlinear values average 48 demands, 50 songs, and three Transformer seeds; linear values use target-canonical normalization. The paired Atlas-minus-single 95% intervals include zero for both tokenizers.

Tokenizer	Nonlinear single / dense / Atlas	Atlas−single [95% CI]	Linear single / dense / Atlas
Locodec	1.037473/1.038269/1.037543	$+.000069$ $[-.000119, +.000298]$	1.002252/1.004308/1.002297
SAME-L	1.037319/1.037338/1.037210	$-.000110$ $[-.000233, +.000018]$	1.001419/1.001985/1.001354

Table 24 gives the consumer boundary. All 48 linear validation fits per tokenizer select the largest prespecified ridge, and held-out normalized risks are near one. A subsequent disjoint 320/80-song development gate widens the ridge grid and uses 120-second crops. The follow-up records zero demands meeting the joint full-token-skill and nonlinear-dense-headroom criterion. Execution ends at development screening, with the 80-song test unused. The relation and consumer comparisons therefore report separate structural and predictive measurements under the changed corpus and teacher.

M Formal Comparison and Evidential Scope

Sparse predictors versus a sparse access interface. In the sparse multitask formulation of Maurer et al. (2013), a dictionary D with bounded columns and task coefficients a_d with $\|a_d\|_1 \leq \alpha$ minimize an average empirical loss of predictions $\langle Da_d, x\rangle$. The dictionary and coefficients are learned jointly; the paper supplies generalization results under its sampling and loss assumptions. By contrast, our

primary Atlas first fits demand operators on frozen tokens, then selects an orthonormal dictionary $\Phi \in \mathbb{R}^{N\times M}$ and binary selectors $s_d \in \{0,1\}^M$, $\mathbf{1}^\top s_d = K$, exposing

$$z_d = \Phi_{S_d}^\top x, \qquad P_d = \Phi \,\mathrm{diag}(s_d)\Phi^\top, \qquad \widehat{y}_d = f_{\theta_d}(z_d).$$

The access restriction precedes a separately refitted, potentially nonlinear consumer. For a scalar linear consumer, the prediction has the sparse dictionary form $\widehat{y}_d = b_d^\top \Phi_{S_d}^\top x = \langle \Phi a_d, x\rangle$ with $\mathrm{supp}(a_d) \subseteq S_d$. The interface analysis therefore focuses on the shared orthogonal coordinate bank, separate active/union ranks, and the access restriction imposed before the consumer.

Matryoshka training supervises nested prefix widths within a learned representation (Kusupati et al., 2022). A global prefix has $S_{d,K} = \{1,\dots,K\}$ for every demand, whereas Atlas permits different, nonnested supports at the same K after tokenization. Our empirical comparisons concern frozen-token shared, dense, and coordinate-subset views. Comparisons with retrained Matryoshka encoders, continuous sparse coefficients, and task-conditioned hypernetworks remain outside this evaluation.

Certificate scope. Proposition 3 gives a posteriori certificates for fitted operators and their chosen supports. Evaluating the relation certificate requires dense reference projectors. In the numerical audit, the utility bounds yield positive closure lower bounds in 10/12 evaluated cells (Appendix E). Proposition 4 propagates externally supplied operator-error radii deterministically. Generalization bounds would additionally require a sampling model and concentration estimates for temporally dependent music tokens. The commutation criterion and spectral inequalities are classical ingredients specialized to this interface.

Chronology and independence. The primary endpoint and band-embedding experiments use the same 24 Moises test songs. The later replay, full refit, and native-coordinate ablation assess computational reproducibility or coordinate access on that same sample. The original OpenMIC sequence gates test access internally, but its method development role 2 had been inspected by earlier models, and one artist spans the official train/test split. The matched OpenMIC baseline audit was designed after the original test results and is retrospective. The two GTZAN studies use new Music2Latent fits on a corpus previously evaluated with other representations. NSynth and VocalSet use the data roles specified below. Protocol locks are internal execution records. Reported bootstrap intervals condition on the fitted methods and named evaluation protocol. Where inspected outcomes informed later method choices, the reported intervals leave that adaptive search outside their uncertainty calculation.

N OpenMIC Baselines and Computational Costs

Design and budget matching. This audit was frozen after the original M84-E4 outcome and before training 21 new models. It reuses the 107-token Music2Latent cache and the six artist roles in Appendix F. All final consumers fit role 3 (5,942 excerpts), select on role 4 (1,121), and report role 5 (1,793) before the already-inspected official test (5,085). Test labels are used for scoring; the comparison is a retrospective audit of the fitted methods.

Every arm uses one six-block width-128 TCN with an unconditional input map, the same dilations, dropout .1, and one 20-logit output. Full-128 exposes all packed coordinates; strict-8 reuses exactly the M84 consensus; PCA-8 and native-prefix-8 are mapped to orthonormal views in the shared whitened space; Haar-8 retains the paired controls. For RRR-8, masked per-instrument ridge regressions fit role 0, select from $10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10$ on role 1, and supply the leading eight left singular vectors of the coefficient matrix. Learned-8 jointly optimizes an orthonormal QR-parameterized projector and the final consumer from Haar initialization. PCA uses role 0 only; the raw prefix requires no label fitting. All seven arms have seeds 1234/2345/3456, 4,000 AdamW updates, learning rate 3×10^{-4}, weight decay .01, 200-step warmup, cosine decay, gradient clipping at one, and masked class-balanced binary cross-entropy. Checkpoints are selected every 250 updates by role-4 AUROC, AP, then earlier update. Three-seed mean logits precede the sigmoid. These are matched final-consumer update budgets, not equal total method-development compute. Full-128 has a larger input layer. Architecture and loss both differ from M84's query-conditioned pairwise-ranking consumer, so cross-experiment score differences do not isolate either change.

Table 25: Compact-interface and full-token baselines on the OpenMIC official test, using three-seed ensembles. The contrast is strict-8 minus the row's arm. Intervals are paired artist bootstraps, 10,000 draws, seed 20260917. These are pointwise, retrospective intervals, not a new confirmatory gate or a familywise multiplicity correction.

Arm	AUROC	AP	AUROC contrast	95% interval
Full-128	.735415	.623000	+.006717	$[-.004930, .017411]$
Strict-8	.742133	.619549	.000000	—
PCA-8	.745070	.633674	-.002937	$[-.014727, .007939]$
Prefix-8	.674535	.574397	+.067598	$[.054492, .080236]$
RRR-8	.752767	.636153	-.010635	$[-.020044, -.001881]$
Haar-8	.666059	.566134	+.076074	$[.060454, .091421]$
Learned-8	.697669	.594154	+.044464	$[.030402, .058509]$

Table 25 compares interfaces within the same joint-classifier family. Appendix J reports the original three-arm comparison with a query-conditioned ranking consumer.

Original strict versus staged projector compute. The original three arms share fresh 1,000-update consumers and role-3/4 data, but strict consensus uses 12,000 scout updates versus 2,250 for the selected staged construction, before its historical screening cost. Strict freezes the scout and fixes its gauge; staged also updates its demand input adapter. The comparison therefore combines changes in projector construction, scout budget, and input adaptation. The learned-to-initial overlaps in Appendix J describe the individual fitted projectors.

Complete-label computation. The original query-conditioned consumer needs 20 nonlinear branches per model to score every instrument, whereas the joint classifier produces all 20 scores in one pass. The archived original evaluation scores only annotated excerpt–instrument pairs. Whitening and projection can be cached, but the query-dependent hidden activations require branch-specific computation. The complete-label evaluation cost therefore depends on the consumer architecture as well as the interface. Appendix F.3 gives the separate interface storage and projection counts.

O Native Atlas on Tasks Excluded from Dictionary Construction

These completed follow-ups use a released Music2Latent encoder and a native rank-128 orthogonal dictionary constructed from 64 Moises source–band–horizon operators. Target training labels select supports within this source-fitted dictionary. These are supervised interface adaptations with a frozen source dictionary. All consumers are fresh, and selected Atlas columns retain global atom order without Procrustes rotation. Dictionary storage and exposed widths refer to the packed 128-coordinate implementation; mono waveforms are duplicated across two channels. The artifact retains all arm and seed outcomes. Construction exclusion concerns the fitted dictionary; overlap with the released tokenizer's full pretraining inventory remains unaudited.

GTZAN genre. We use GTZAN (Tzanetakis & Cook, 2002) with the fault-filtered MARBLE split of 442/197/290 tracks (Yuan et al., 2023). The joint-classifier study uses a train-only ridge operator to choose one support for the joint ten-class task, with widths 8/16/32, seven compact families and full-128 (66 fits). A width-128 six-block TCN trains for 2,500 updates with three paired seeds; validation selects checkpoints, and three fixed windows supply each test prediction. Native Atlas-32 reaches .493103 accuracy versus .489655 for single and full. Its paired difference from single is $+.003448$ $[-.034483, .041379]$. The dense reference is fitted using linear operators; all reported accuracies use the matched nonlinear consumers.

The class-specific study uses ten one-vs-rest operators and ten rank-eight supports. Its primary control cyclically reassigns exactly those addresses across demands; all eight arms use ten branches and three seeds. Correct assignment reaches .386207 test accuracy versus .427586 for permuted and .441379 for shared single-8. Both studies use the same previously inspected GTZAN corpus; their contrasts evaluate joint-task and class-specific address selection, respectively.

NSynth pitch and instrument family. The two tasks are 128-way MIDI pitch and 11-way instrument family (Engel et al., 2017). Pitch is the principal excluded task family; instrument family is a secondary demand semantically closer to source construction. Four-second examples produce 42 packed tokens. Five train-only ridge folds choose stable top-K scores in the frozen dictionary; the other task's address is the swapped control. Ten compact families at $K = 8, 16, 32$ plus full-128, two tasks, and three seeds give 186 fits. Consumers normalize projected tokens, mean-pool, and use a 512-unit ReLU layer with dropout .2; they train for 8,000 updates, selecting by validation accuracy, balanced accuracy, then earlier checkpoint. Ensembles average probabilities.

The splits use disjoint examples, with zero train/test instrument-instance overlap and 53 instances shared by validation and test. The 10,000 example-bootstrap intervals describe uncertainty conditional on this instance pool. Atlas-32 reaches .852295 pitch and .524414 instrument accuracy, versus .892090/.583252 full-token accuracy. Both matched-minus-swapped intervals include zero. Stable and plug-in rank-32 supports are identical and fold Jaccards are .94–1.00. The support-stability and prediction comparisons measure these two properties separately.

VocalSet technique. The MARBLE ten-technique protocol uses 15 development and five held-out singers, with three-second chunks (Yuan et al., 2023; Wilkins et al., 2018). Consumer-aligned soft gates and per-atom held-out ablations across five singer folds rank atoms; hard top-32 supports are then fixed. A matched dense projector uses the same proxy budget. No proxy consumer weights enter the 48 fresh final fits (eight arms, two demands, three seeds), each trained for 100 epochs with a 512-unit ReLU consumer and no final early stopping. The official test comprises 1,138 chunks from 410 recordings and five singers. The auxiliary demand is the dataset's binary singer-sex attribute; its use does not imply a broader demographic inference claim.

Technique accuracies are .532513 for the selected Atlas address, .414763 for the swapped address, .510545 for random atoms, and .569420 for construction-only atoms. The selected-minus-swapped interval is $[.074757, .161566]$; the selected-minus-random and selected-minus-construction intervals are $[-.015844, .058981]$ and $[-.072587, -.000836]$, respectively. On the auxiliary task, its own address gives .920035 accuracy and the technique address gives .957821. The paired bootstrap resamples the 410 recordings within the fixed set of five test singers.

Table 26 collects the primary contrasts across these tasks.

Table 26: Native Atlas on tasks excluded from dictionary construction. Target-task training labels select supports in the frozen dictionary. Contrasts are accuracy differences; A denotes the named target-specific Atlas, S swapped/permuted addresses, C construction-only atoms, R random atoms. The joint GTZAN classifier instead compares to its shared single subspace. Intervals are pointwise paired 95% bootstraps under each task's stated resampling unit.

Task	Contrast	Atlas accuracy	Difference	95% interval
GTZAN joint, $K = 32$	$A-$single	.493103	+.003448	$[-.034483, .041379]$
GTZAN class, $K = 8$	$A - S$	.386207	-.041379	$[-.093103, .006897]$
NSynth pitch, $K = 32$	$A - S$	.852295	+.001221	$[-.006104, .008545]$
NSynth family, $K = 32$	$A - S$	.524414	-.001953	$[-.011475, .007568]$
VocalSet technique	$A - S$	.532513	+.117750	$[.074757, .161566]$
VocalSet technique	$A - R$	.532513	+.021968	$[-.015844, .058981]$
VocalSet technique	$A - C$	.532513	-.036907	$[-.072587, -.000836]$
VocalSet auxiliary	$A - S$	.920035	-.037786	$[-.056886, -.019146]$

P Native-Address Selection on MTG-Jamendo Development Data

Question and data roles. These studies ask whether selecting native addresses improves downstream utility and whether task-specific addresses outperform an equally optimized shared address. We first report the cross-artist selection and matched baselines used in the main text, then summarize the earlier controls that distinguish dictionary choice, consumer co-adaptation, and address reuse.

All studies use 9,949 cached tracks from the official MTG-Jamendo training partition (Bogdanov et al., 2019), with 20 previously selected labels each for Genre and Mood/theme. These are restricted

vocabularies. Five artist-separated folds define fit and selection roles; outer folds 0 and 1 are evaluated separately. Artists do not cross roles within a run, but the runs share data and all folds have been repeatedly inspected during development. The sequence accesses neither official validation nor test; an earlier development study had already used 3,802 official-validation tracks for checkpoint selection. The reported outcomes are development-set comparisons on these reused folds; fresh seeds vary the consumer fits. Internal locks document the execution sequence.

Frozen dictionary and common consumer. Mono Music2Latent caches contain two identical 64-coordinate blocks; we verify and use one. The frozen 64×64 orthogonal dictionary is fitted to 64 Moises source–band–horizon operators using the original 188 training songs, with source-only whitening and joint diagonalization. Channels are views of the same song, not independent songs. Jamendo labels train support selectors and consumers, but do not fit this dictionary.

An address contains exactly 16 ordered dictionary columns without a within-support rotation. A shared address still feeds separate task consumers. Each is a six-block width-128 TCN with dilations 1/2/4/8/16/32 and dropout .2. Training uses track-uniform batches of 64, class-balanced binary cross-entropy (BBCE), AdamW with learning rate and weight decay .001, gradient clipping at one, and a 4,000-update cosine schedule with 100-update warmup. Evaluation averages three crop logits and reports track-level macro AUROC, average precision (AP), and BBCE. Model seeds are averaged as scalar metrics, without prediction ensembling; no labels are dropped by outcome. The following subsections specify how addresses and checkpoints are selected.

P.1 Cross-artist address selection with fresh consumers

Candidate generation and selection roles. Cross-artist selection tests whether averaging utility over more artist groups improves the stability of native-address selection. The dictionary, native 16-of-64 interface, tasks, and five artist folds remain fixed. Candidate generation does not import supports selected by earlier target-task experiments. Nine whole-coordinate proposals are generated using only fold 2: Genre, Mood, and equal-task ridge objectives, each fit to all generation artists or one of two prespecified artist-bootstrap draws. The generator uses artist-balanced, standardized track-level atom means and standard deviations, ridge penalty .1, and forward selection of whole coordinate groups. Together with source construction and four prespecified random supports, this gives 14 unique addresses. Generation and selection use the separate roles specified for this round; the shared development-data history is described in Appendix P.

For outer fold 0, candidate consumers use inner folds 1, 3, and 4; for outer fold 1, they use 0, 3, and 4. Each fit excludes its current inner and outer folds, so generation fold 2 always belongs to its fit set. Both tasks score every address with search seeds 112101/112102: 14 addresses $\times$ two tasks $\times$ two outer contexts $\times$ three inner roles $\times$ two seeds gives 336 new fits. Each receives 4,000 updates, with only the final checkpoint eligible; inner data no longer select training duration. The primary *cross-artist* policy averages scalar AUROC over the three inner roles and both seeds. The prespecified *single-inner* control uses only inner fold 4. Each task chooses its best address, while shared maximizes the equal-task mean over the same pool. Neither averaging rule ensembles predictions.

Fresh-consumer evaluation. Both policies' own, swapped, shared, and construction addresses are locked before new consumers train with seeds 112201/112202. Final fits use inner role 4, a physical width-16 input from initialization, and the fixed 4,000-update checkpoint. Weights, optimizer states, and random streams are reset; no search-consumer state is inherited. Exact task/fold/seed/support aliases yield 28 unique fits for 64 logical evaluation cells across the two policies. All final models are locked before outer scoring. The round adds 1,456,000 consumer updates, excluding dictionary construction and ridge generation. Three-inner selection costs three times the candidate fitting of the single-inner policy alone; shared aliases do not count as extra fits. This comparison holds the candidate pool and training rule fixed.

Selection criteria and observed contrasts. Reciprocal specificity requires, for each task, own-minus-swapped mean AUROC at least .002, at least 3/4 positive fold/seed cells, and both fold means positive. Shared replacement requires equal-task own-minus-shared at least .002, neither task mean negative, and positive equal-task means in both folds. These thresholds summarize descriptive

development comparisons. Each policy meets 0/2 criteria. Table 27 gives every primary-policy AUROC pair and the source-construction comparisons. Own-minus-swapped is positive in 2/4 Genre and 0/4 Mood fold/seed pairs. The single-inner policy selects one identical own/swapped/shared address within each outer context, giving exactly zero specialization contrasts; its mean AUROC is .767333 for Genre and .676545 for Mood.

Table 27: Paired macro AUROC contrasts after cross-artist address selection on MTG-Jamendo. Columns identify outer development fold and fresh-consumer seed suffix (112201/112202). Contrasts compare task-selected (own), other-task (swapped), optimized shared, and source-construction supports under matched consumers.

Task	Contrast	Fold 0/201	Fold 0/202	Fold 1/201	Fold 1/202	Mean
Genre	Own–swapped	+.012450	+.018680	-.003620	-.004378	+.005783
	Own–shared	.000000	.000000	-.003620	-.004378	-.002000
	Own–construction	+.015589	+.029335	+.004895	+.002946	+.013191
	Shared–construction	+.015589	+.029335	+.008515	+.007324	+.015191
Mood/theme	Own–swapped	-.002565	-.003627	-.009341	-.002063	-.004399
	Own–shared	-.002565	-.003627	.000000	.000000	-.001548
	Own–construction	+.016721	+.014456	-.003282	+.012115	+.010003
	Shared–construction	+.019286	+.018083	-.003282	+.012115	+.011551

Shared-address adaptation improves utility relative to source construction: all four Genre and three of four Mood comparisons favor optimized shared, with positive means in both folds. Mean AP and BBCE also improve. This comparison measures adaptation of a shared native subset within the frozen source dictionary.

P.2 Matched baselines for cross-artist address selection

Matched final training. The baseline protocol is frozen before the cross-artist outer results become available. It adds 32 fits for Full (64 coordinates), centered PCA (16), Joint dense (16), and one prespecified Random native support (16; seed 112301), with the same two tasks, outer folds, fresh seeds, final fit roles, and 4,000-update final checkpoint as the address-selection study. Consumer architecture and random-stream keys are matched; Full has a larger input layer. PCA uses only fit-frame FP64 covariance over all crops and subtracts its fitted mean at inference. Joint dense starts from one archived Haar frame and learns a QR-parameterized projector jointly with each task's consumer. It is not a learned interface shared jointly across tasks. All baseline models are locked before outer scoring.

Native arms reuse the 28 unique final fits; baseline fits are not replicated for each policy. The baselines add 128,000 consumer updates and 32,000 projector optimizer updates. Matching final updates does not match FLOPs, input-layer parameters, preprocessing, or address-search cost. The random control measures one fixed support, not uncertainty over random draws. These fixed training recipes do not measure fully tuned performance or an accuracy–cost frontier.

Variation across folds and comparators. Table 28 gives the mean scores. For optimized shared addresses, Genre AUROC differences against Full, PCA, and Joint dense are $+.017948/-.001849$, $+.009457/-.001121$, and $+.013514/-.000814$, respectively, as fold-0/fold-1 means over the two fresh consumer seeds. The corresponding Mood/theme differences are $-.002606/-.003109$, $+.017424/+.007887$, and $+.002049/+.002554$ in the same baseline order. Own-minus-dense16 AUROC is $+.004350/+.000754$ for Genre/Mood, with 2/4 positive pairs per task and opposite fold directions. Genre's own-minus-full64 and own-minus-PCA16 contrasts have positive overall means and negative fold-1 means. For Mood, the own-minus-full64 mean is negative, and own-minus-PCA16 is positive in all four pairs. Own has higher AUROC than random native16 in 3/4 pairs per task; the table reports AP and BBCE alongside AUROC. The earlier fresh-reuse controls in Appendix P.3 use different seeds and an inner-selected checkpoint, so their dense advantage is a separate within-protocol result. Complete results for both selection policies and all baseline arms remain in the `jamendo_cross_artist/` artifact.

Table 28: Cross-artist address selection and matched baselines on MTG-Jamendo. The interfaces and AUROC means match Table 2, here shown to six decimal places with AP and balanced binary cross-entropy (BBCE). Means average two outer development folds and two fresh seeds over 20 selected labels per task, without prediction ensembling. Higher AUROC/AP and lower BBCE are better; bold marks each metric's numerical optimum.

		Genre			Mood/theme		
Interface	Input dim.	AUROC	AP	BBCE	AUROC	AP	BBCE
Source construction	16	.754700	.258006	.767195	.670692	.132174	1.008327
Optimized shared	16	**.769891**	**.279973**	**.681718**	.682242	.136092	.864716
Task-specific (own)	16	.767892	.276852	.686994	.680694	.132762	.893659
Other task's subset (swapped)	16	.762109	.272150	.714438	.685093	.136497	**.858721**
Full	64	.761842	.268893	.933478	**.685100**	**.137010**	1.218066
PCA	16	.765723	.268907	.705248	.669587	.128875	.919015
Joint dense	16	.763541	.276469	.715548	.679941	.133561	.908408
Random native	16	.755180	.259264	.701159	.674693	.125097	.859221

P.3 Controls for native-address selection

Table 29 separates address–consumer co-adaptation, dictionary choice, reuse with fresh consumers, and direct support selection. Each block compares methods under its own matched training protocol.

Table 29: Controls for native-address selection on MTG-Jamendo development folds. Entries are mean macro AUROC differences within each matched comparison. Co-adaptation and dictionary comparisons use four source seeds and two folds; fresh-reuse comparisons additionally average two new consumer seeds, and utility selection uses two fresh evaluation seeds and two folds. All entries are descriptive means on the reused artist development folds.

Control	Contrast	Genre	Mood/theme
Co-adaptation	Learned shared − fixed anchor	+.005540	+.006763
	Unrestricted task-specific − learned shared	-.000990	-.003090
Dictionary	AJD − mean of three Haar draws	+.012281	+.002425
Fresh reuse	Learned shared − fixed anchor	+.001364	+.002727
Fresh baseline	Learned shared − dense16	-.003356	-.003302
Utility selection	Own − swapped	-.001172	+.007265
	Own − optimized shared	-.003607	-.005221
	Optimized shared − construction	+.007167	+.012657

Co-adaptation and dictionary choice. The continuation comparison starts all arms from the same update-2,000 consumer, optimizer, random streams, and target-supervised shared anchor. During updates 2,001–2,800, shared or task-specific straight-through gates select 16 coordinates; fit-only 200-step occupancy then fixes the support. After physical width-16 compaction, consumers continue to 4,000, with inner-fold checkpoint selection over updates 3,000:200:4,000. The table reports the shared-minus-anchor and task-specific-minus-shared AUROC contrasts. Dictionary controls repeat this procedure with three fixed Haar dictionaries and whitened identity; the Haar comparison averages the three draws. For learned shared addresses, AJD has higher mean Genre AUROC than both dictionary controls. Mood AUROC differences vary across seed/fold groups, and AJD has higher mean Mood BBCE than both controls.

Reuse after consumer resetting. The exact fixed and learned shared supports are frozen while two new consumer seeds train for 4,000 updates at physical width 16. Their scalar metrics are averaged within each of four address-source seeds and two folds. The prespecified directional criterion is 6/8 favorable pairs; the observed counts are 4/8 for Genre and 5/8 for Mood. Mean AP is lower for the learned address on both tasks. The fixed anchor in this comparison is target-supervised. Matched full64, PCA16, and jointly learned dense16 consumers use the same fresh seeds and checkpoint rule. Dense16 has higher mean AUROC and AP and lower BBCE than learned shared for both tasks;

PCA's ordering depends on the task. The matching applies to final consumer updates; address-search history, input-layer size, and preprocessing depend on the method.

Direct utility selection. The utility-selection control ranks 28/27 previously generated supports in outer contexts 0/1 using two search-consumer seeds, one inner checkpoint/selection role, and 4,000 updates per fit. Each task selects its best support; shared maximizes the equal-task mean over the same pool. Two fresh evaluation seeds fit the locked own, swapped, shared, and construction addresses. Own-minus-swapped AUROC is positive in 2/4 cells per task. Table 29 reports the mean contrasts; its dense16 comparison comes from the separate fresh-reuse protocol. Complete arm/seed/fold records and reused-row identities are provided in the accompanying artifact.

Q Artifact Availability and Reproduction

Table 30 summarizes artifact availability and reproduction scope.

Table 30: Availability and reproduction scope at submission. "Upstream" means obtainable from the cited public project, not bundled here.

Component	Availability	What can be reproduced
Frozen reports, locks, manifests, atom sets	artifact	Exact integrity checks and deterministic result reaggregation
MUSDB18-HQ, MoisesDB, Slakh, FMA/OpenMIC audio and labels	upstream	Corpus acquisition under upstream terms; audio is not redistributed
SAME-L, Music2Latent, PupuJEPA, MERT checkpoints	upstream	Upstream inference; the public experiment supplies its SAME-L/PupuJEPA encoding and replay path
Three music-Locodec checkpoints	not released	Recorded hashes and result summaries; model weights cannot be checked from this bundle
Compact public band-embedding interface and nine TCN checkpoints	artifact	Re-encode public audio and replay the second nonlinear consumer
Public full-refit/native locks and summaries	artifact	Verify gates, aggregate results, and provenance; checkpoints and caches are excluded
OpenMIC protocols, seed predictions, integer artist groups	artifact	With upstream labels, recompute metrics and paired CIs; checkpoints, labels, and caches excluded
GTZAN/NSynth/VocalSet follow-up protocols and results	artifact	Result-level verification of all arms, reported contrasts, and stated test boundaries
Jamendo development rows and native supports	artifact	Three separate verifiers reaggregate arms and paired contrasts; no official-test, prediction, or training replay
Other interface and consumer checkpoints	not released	Aggregate and seed-level result verification only
Audio and cached token arrays	not released	The public path regenerates them locally; no arrays are bundled

The `REPRODUCTION_GUIDE.md` maps claims and figures to their frozen records, verification scripts, and replay entry points. Hashes, split provenance, selected atoms, gate outcomes, and reused-row identities support integrity checks and deterministic reaggregation. Internal locks record the execution chronology.

The public band-embedding bundle supplies the fitted SAME-L single/dense/Atlas interfaces, consumer code, and nine checkpoints from the earlier frozen-interface experiment. With upstream audio and models, it regenerates inputs using song-indexed bottleneck draws and supports consumer replay. The later full-refit and native-coordinate studies rebuild interfaces or remove their rotations, but their checkpoints and intermediate caches are not bundled; their protocols and aggregates support result-level verification. These studies reuse the same corpus split.

For OpenMIC, saved predictions, public sample keys, and integer artist groups permit metric and paired-interval reproduction with upstream labels. The classifier checkpoints and token caches are excluded, and the stronger baseline audit remains retrospective. GTZAN, NSynth, VocalSet, and

the three Jamendo extensions provide scalar results, supports, and protocol records. The Jamendo records retain the controls summarized in the manuscript and support scalar-result reaggregation on the reused development folds.

No audio or token arrays are redistributed. Exact Locodec retraining still requires the authorized in-house corpus; public release of author-written code and permitted checkpoints remains subject to upstream rights.